KTH Matematik

Statistik för bioteknik sf1911
Övningsuppgifter
2017

Timo Koski

Innehåll

Förord

- Följande text är en blandning av synpunkter från webben:

  > Biotechnology can focus on a whole range of topics, from genetic modification of plants and animals to gene therapy, medicine and drug manufacturing, reproductive therapy, and even energy production (through the bioengineering of bacteria and algae). In all cases, the work is carried out by developing something and testing whether or not it has the desired performance. Determining performance requires statistical analysis of results.
  >
  > In today's world of high-throughput experiments, biotechnology deals with laboratory equipment constantly churning out mountains of data. But without an understanding of statistics and a knowledge of the techniques required to analyse, summarize and interpret these data, we are very limited in what we can learn from our observations, which will in turn inhibit our ability to move forward in our activity. Even with experiments that generate very little data, there is a need to simulate phenomena by modelling the behaviour of systems and their parameters, which again often needs to be done statistically. It is therefore imperative to understand the basics of probability, statistical distributions, descriptive statistics, and some simple parametric hypothesis tests.

- The solutions to (99.9 % of) these exercises are available in a separate document on page `https://www.math.kth.se/matstat/gru/sf1911/exercises.html`

- Further exercises (with solutions) will be added during 2nd quarter, 2017.

Kapitel 1

Data-analys Del I

1.1 Datatyper

Heltal, flyttal, textsträngar och vektorer är datorns datatyper. De statistiskt viktiga datatyperna hos ett datamaterial är (verklighetens datatyper):

Kategoridata

- Dikotoma data (binära data, två kategorier)
- Kategoridata utan ordningsstruktur (nominaldata)
- Kategoridata med ordningsstruktur (ordinaldata)

Kvantitativa data

- Kvantitativa antalsdata (diskreta data)
- Kvantitativa kontinuerliga data, indelade i icke-negativa, intervallbegränsade eller obegränsade

Man talar även om de mostvarande mätnivåerna hos ett datamaterial: nominalskala, ordinalskala, intervallskala och kvotskala. Eller om mostvarande variabler: nominal-, ordinalvariabler o.s.v.

Problem 1.1.1.

Hur kan kategoridata omvandlas till kvantitativa antalsdata?

Hur kan kontinuerliga data omvandlas till kategoridata?

Problem 1.1.2.

Mätningar av temperatur i C^o utgör data

a) i nominalskala

b) i ordinalskala

c) i intervallskala

d) i kvotskala.

Problem 1.1.3.

Betygen på olika typer av tandborstar, som har betygsatts med avseende på reningskapacitet från 1 till 5, där 5 är bäst, utgör data

a) i nominalskala

b) i ordinalskala

c) i intervallskala

d) i kvotskala.

Problem 1.1.4.

Which of the following variables is quantitative?

a) sex

b) diastolic blood pressure

c) eye color

d) diagnosis

e) height

f) genotype

g) CD4 count = the number of CD4 T lymphocytes (CD4 cells) in a sample of your blood.

Problem 1.1.5.

Which of the following variables is continuous?

a) blood glucose

b) dendritic length

c) age at last birth day

d) allele frequency

e) exact age

f) family size

Problem 1.1.6.

In one of the first truly random trials in Britain, patients with pulmonary tuberculosis received either streptomycin or no drug (Medical Research Council, 1948). Patients were classified after six months into the following: considerable improvement, moderate/slight improvement, no material change, moderate/slight deterioration, considerable deterioration, or death.

Which data type do we have with these patient classifications?

Problem 1.1.7.

Huntington disease is a dominantly transmitted neurodegenerative disorder that arises from expansion of a CAG trinucleotide repeat on chromosome 4p16.3. CAG repeat allele lengths are defined as fully penetrant at $\geq$ 40, reduced penetrance at 36-39, high normal at 27-35, and normal at $\leq$ 26.

a) Which data type/scale is CAG repeat allele length?

b) Which data type/scale is penetrance of CAG repeat allele lengths?

1.2 Statistikor: numeriska sammanfattningar av kvantitativa data

En **statistika** är ett tal som beskriver/sammanfattar en kvantitiv datamängd (data set, även kallad ett stickprov av data). Statistikans numeriska värde är känt när datamängden är analyserad.

En statistikas numeriska värde varierar från stickprov till stickprov från samma källa av mätdata. Ofta används en statistika för att göra en skattning av en okänd parameter i en statistisk modell (mer om detta senare).

Detta och senare avsnitt innehåller många exempel på statistikor, bl.a., medelvärde, median (stickprovs)varians, skevhet, variationskoeffient, korrelationskoefficient.

Medelvärdet av $x_1, x_2, \ldots, x_n$ är
$\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$.
Medelvärdet ger läget för en datamängd. I Matlab `mean(x)`.

Avvikelserna från medelvärdet är

$$x_1 - \bar{x}, x_2 - \bar{x}, \ldots, x_n - \bar{x}$$

Variansen/Stickprovsvariansen är summan av de n kvadrerade avvikelserna från medelvärdet dividerad med $n - 1$:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}.$$

Variansen ger spridningen i en datamängd kring dess medelvärde. I Matlab `std(x)`.
Variansen kan också beräknas som $s^2 = \dfrac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$ (jfr. Bilagan nedan).
Standardavvikelsen är
$s = \sqrt{\dfrac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$.
$\bar{x}$ och s har samma enhet (sort).

Median är det värde för en kvantitiv datamängd, som ordnats efter storlek, som delar materialet i två lika stora delar. Medianen är ett lägesmått. Medianen överskrides lika ofta som det underskrides av värden i den givna datamängden. Medianen kan även beräknas för data i ordinalskala.

Beräkning av median
Antag en ordnad datamängd av n mätvärden. Medianen är det mittersta värdet om n är udda. Om n är jämnt beräknas medianen som medelvärdet av de mittersta värdena. I Matlab `median(x)`.

Problem 1.2.1.

Under under 10 dagar man mätt avkastningen (enhet ton) från två kemiska processer A och B.

Följande mätvärden erhölls:

A:	2.11	0.82	7.60	0.09	8.10	8.47	4.52	8.07	4.83	6.13
B:	5.27	5.88	5.65	5.49	5.77	5.96	5.99	5.84	5.75	5.04

Man ser genast att A-data är mer utspridda än B-data. Men det gäller uttrycka denna insikt med siffervärden.

Beräkna medelvärde, standardavvikelse och median både för A och för B. Kommentera resultaten.

Problem 1.2.2.

Tjebysjovs regel (teorem)
För varje dataset och för varje tal k *(*$k \geq 1$*), ligger minst procentdelen (fraction)* $1 - \frac{1}{k^2}$ *av data inom intervallet* $[\bar{x} - ks, \bar{x} + ks]$.
Denna regel säger, till exempel, att minst 8/9 av värdena i ett dataset ligger inom $[\bar{x}-3s, \bar{x}+3s]$, och man säger att 8/9 av värdena ligger inom tre standardavvikelser från dess medelvärde.
Tjebysjovs regel utlovar alltså att om s är litet, så kommer data att ligga tätt kring dess medelvärde.

Checka Tjebysjovs regel med avseende på datamängden nedan, IQ för 112 barn, med $k = 2$ och $k = 3$:

61
72 75 76 77 79
80 80 80 80 81 82 84 85 85 85 86 86 88
90 90 90 90 91 91 92 93 93 93 94 94 96 96 96 96 97 97 98 98 98 99
100 100 100 101 101 102 102 102 102 102 103 103 103 105 106 106 106 107 107 107 107 108 108 109 109 109 109 109
110 110 110 110 111 111 112 112 113 113 113 113 114 114 114 114 114 114 117 117 118 119 119
120 121 121 121 121 121 122 123 124 124 125 126 126 129
130 130 136
142 146
150

Problem 1.2.3.

Sammanslagning av mätserier Två teknologer har gjort var sin likadan laboration. Det gällde att bestämma ljudhastigheten i järn. Den förste teknologen gjorde 10 mätningar $x_1, \ldots, x_{10}$ och den andre gjorde 5 mätningar $y_1, \ldots, y_5$. De erhöll (enhet m/s):

$\bar{x} = 5313$	$s_x = 5.2$
$\bar{y} = 5309$	$s_y = 3.0$

Vilket medelvärde och vilken standardavvikelse fås om de 15 värdena betraktas som en mätserie?

Problem 1.2.4.

För data med 800 observationer har man räknat ut medelvärdet och standardavvikelsen. Man har fått $\bar{x} = 9.496$, $s = 33.32$. Vid kontroll av datafilen visar det sig att en observation som skulle ha varit 9.56 har införts som 956.

Vilket medelvärde och vilken standardavvikelse hade man fått om den nämnda felet ej hade funnits?

Relativ frekvens
Vi har en datamängd $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ med n datapunkter i någon skala $\mathcal{D}$. x är ett värde i skalan $\mathcal{D}$. Då är $f_{\mathcal{X}}(x)$, den relativa frekvensen av x (m.a.p. $\mathcal{X}$), lika med

$$f_{\mathcal{X}}(x) = \frac{\text{antalet gånger } x \text{ förekommer i } \mathcal{X}}{n}.$$

eller

$$f_{\mathcal{X}}(x) = \frac{\text{frekvensen av } x \text{ i } \mathcal{X}}{\text{totalantalet data i } \mathcal{X}} = \frac{\text{frekvensen av } x \text{ i } \mathcal{X}}{n}.$$

Denna definition gäller för alla skalor/datatyper. Observera att $f_{\mathcal{X}}(x) = 0$ om x inte återfinns bland $\mathcal{X}$.

Medelvärde och varians m.h.a. relativ frekvens
Om vi har en kvantitativ datatyp, kan medelvärdet $\bar{x}$ beräknas som ett *viktat* medelvärde av värdena i $\mathcal{X}$, med vikterna $f_{\mathcal{X}}(x)$,

$$\frac{x_1 + x_2 + \cdots + x_n}{n} = \sum_{\text{de olika värdena } x \text{ i } \mathcal{X}} x \cdot f_{\mathcal{X}}(x)$$

P.s.s. fås att

$$\sum_{i=!}^{n} x_i^2 = x_1^2 + x_2^2 + \cdots + x_n^2 = n\dot{\sum}_{\text{de olika värdena } x \text{ i } \mathcal{X}} x^2 \cdot f_{\mathcal{X}}(x).$$

Problem 1.2.5.

$\mathcal{D}$ = de positiva heltalen. $\mathcal{X} = \{7 \quad 5 \quad 5 \quad 9 \quad 6 \quad 6 \quad 8 \quad 8 \quad 6 \quad 2 \quad 3 \quad 8\}$.

a) Beräkna de relativa frekvenserna $f_{\mathcal{X}}(x)$ för $x = 2, 3, 5, 6, 7, 8, 9$.

b) Beräkna

$$f_{\mathcal{X}}(2) + f_{\mathcal{X}}(3) + f_{\mathcal{X}}(5) + f_{\mathcal{X}}(6) + f_{\mathcal{X}}(7) + f_{\mathcal{X}}(8) + f_{\mathcal{X}}(9).$$

c) Beräkna medelvärdet

$$\sum_{\text{de olika värdena } x \text{ i } \mathcal{X}} x \cdot f_{\mathcal{X}}(x) = 2 \cdot f_{\mathcal{X}}(2) + 3 \cdot f_{\mathcal{X}}(3) + 5 \cdot f_{\mathcal{X}}(5) + 6 \cdot f_{\mathcal{X}}(6) + 7 \cdot f_{\mathcal{X}}(7) + 8 \cdot f_{\mathcal{X}}(8) + 9 \cdot f_{\mathcal{X}}(9)$$

och jämför med $\bar{x}$.

d) Beräkna s^2 m.h.a. $f_{\mathcal{X}}(x)$. *Ledning:* Vi har att

$$s^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$$

där $\sum_{i=1}^{n} x_i^2$ och $\bar{x}$ kan beräknas m.h.a. $f_{\mathcal{X}}(x)$. Kontrollräkna gärna med `mean` och `std` i Matlab.

Problem 1.2.6.

I en hamn räknade man för varje dygn under en månad ankommande lastfartyg. Man erhöll följande antal:

3	3	2	4	2	5	2	7	2	1	2	0	3	3	4
1	3	9	3	2	3	4	1	2	1	4	7	6	3	1

a) Bestäm de relativa frekvenserna $f_{\mathcal{X}}(x)$(=antalet dagar med x ankommande båtar/antalet dagar).

b) Rita ett stolpdiagram

Ett stolpdiagram visar frekvens/relativ frekvens hos olika värden i en datamängd med diskreta antalsdata, d.v.s höjden på stolparna visar hur ofta ett visst värde förekommer.

c) Beräkna medelvärde och standardavvikelse och relatera dem till stolpdiagrammet.

d) Checka Tjebysjovs teorem med $k = 2$.

Skevhet, Kurtosis
Skevheten av $x_1, x_2, \ldots, x_n$ är definierad som

$$g_1 \stackrel{\text{def}}{=} \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^{3/2}}.$$

Den mäter **hur sned, eller osymmetrisk en datamängd är**. Om $g_1 < 0$, så lutar datamängden åt vänster. Om $g_1 > 0$, så lutar datamängden åt höger. Om $g_1 = 0$, så är datamängden symmetrisk.
Kurtosis definieras av

$$g_2 \stackrel{\text{def}}{=} \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2}.$$

På engelska talar man om kurtosis som mått på flatness och peakedness (toppighet). Ibland har man kurtosis som

$$g_2 \stackrel{\text{def}}{=} \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right)^2} - 3.$$

Pearson 2 skevhet koefficient
En annan statistika att mäta skevhet är Sk_2, Pearson 2 skevhet koefficient, som definieras som

$$Sk_2 \stackrel{\text{def}}{=} 3 \cdot \frac{\overline{x} - \text{median}(x)}{s}.$$

Det kan visas att $-3 \leq Sk_2 \leq 3$. Denna statistika är oförtjänt (?) bortsedd.

Problem 1.2.7.

Man har följande data på diastoliskt blodtryck (mmHg) för nio patienter

Patient	1	2	3	4	5	6	7	8	9
Blodtrycket	96	119	119	108	126	128	110	105	94

a) **Bestäm** median och **range ≡ variationsbredd =största värdet - minsta värdet**.

b) **Bestäm** medelvärde, standardavvikelse och varians.

c) **Bestäm** skevhet g_1 och Sk_2.

d) **Hur** ändras medelvärdet, om vi lägger till 10 mmHg till var och en av mätningarna? **Hur** ändras standardavvikelsen? **Hur** ändras skevhetsmåtten? *Ledning:* se bilagan om Linjära transformationer.

d) **Hur** ändras medelvärdet, om vi tar bort värdet 128 från mätningarna? **Hur** ändras medianen om vi tar bort värdet 128 från mätningarna?

e) **Hur** ändras medelvärdet, om vi multiplicerar med -2.5 var och en av mätningarna? **Hur** ändras standardavvikelsen? **Hur** ändras skevhetsmåtten? *Ledning:* se bilagan om Linjära transformationer.

f) **Hur** tror Du att medelvärdet kommer att ändras, om vi mäter diastoliskt blodtryck av nya individer? Hur ändras standardavvikelsen, om vi mäter diastoliskt blodtryck av flera individer?

Problem 1.2.8.

We continue with the Huntington disease (HD) as introduced above. To find the prevalence (=förekomsten) of the high normal allele in the general population a group of doctors plans to use samples of primarily Parkinson disease (PD) cases. They compare the HD allele lengths in the PD sample to the non-penetrant length HD alleles in the HD sample. The result is shown below.

	Number of Alleles	Range	Mean	Median
Both Alleles for PD Sample	1276	14-34	20.0	19
Non-penetrant length Allele for HD Subjects	2065	6-35	18.7	18

They conclude:

> *The similarity further supports the assumption that the length of the HD allele is not associated with the etiology (etiologi studerar orsakssamband bakom sjukdomar) of PD.*
> Hendricks, Audrey E and Latourelle, Jeanne C and Lunetta, Kathryn L and Cupples, L Adrienne and Wheeler, Vanessa and MacDonald, Marcy E et.al.: *Estimating the probability of de novo HD cases from transmissions of expanded penetrant CAG alleles in the Huntington disease gene from male carriers of high normal alleles (27–35 CAG)*, **American Journal of Medical Genetics Part A**, 149, pp. 1375–1381 , 2009.

How do you draw this conclusion?

Problem 1.2.9.

Vid ett laboratorium har man för 50 proteiner av en viss typ (i vissa celler) noterat livslängderna (livslängd = tiden till ett proteins degradering). Följande tabell ger livslängderna i minuter ordnade i storleksordning.

87	105	108	120	142	151	155	155	161	162
169	173	174	183	185	186	186	192	193	196
199	205	207	211	215	217	217	222	224	226
227	230	231	231	237	242	244	246	251	258
263	267	278	286	294	312	338	341	362	390

a) **Beräkna** datasetets medelvärde, median standardavvikelse och skevhet g_1 och Sk_2.

b) Ta bort datapunkten 390. Vad händer med datasetets medelvärde, median standardavvikelse och skevhet g_1 och Sk_2?

b) Checka **Tjebysjovs teorem** med avseende på det här datat med $k = 3$.

Variationskoefficient, **Coefficient of variation (CV)**
Variationskoefficient är en statistika, ett mått på spridning. Om vi t.ex. har data på vikt (kg) och på längd (m), är det omöjligt att direkt jämföra spridningen i längd med spridningen i vikt. Eller, om man väger ett antal människor i Sverige och anger deras vikt i kg, och gör samma sak i USA och anger de uppmätta vikterna i uns, är det omöjligt att direkt jämföra spridningarna.
För det andra, med observationer på olika skalor ex. 1,2,3,4,5 och 1000,2000,3000,4000,5000 kommer standardavvikelserna vara olika (större vid högre skalor) även om de procentuellt sett är lika.
Variationskoefficienten för en datamängd definieras som

$$\text{CV} = \frac{s}{\bar{x}} \times 100[\%]$$

Således uttrycker CV standardavvikelsen som procentandel av medelvärdet. För mätdata med en viss enhet har $\bar{x}$ och s samma sort och därför är CV utan enhet (dimensionless). Variationskoefficienten gör alltså standardavvikelser på olika skalor och i olika enheter jämförbara. CV används för positiva data.

Problem 1.2.10.

a) Vad händer med variationskoefficienten CV om till alla observationer i en datamängd adderas ett positivt tal a (t.ex. $a = 1000$)? *Ledning:* se bilagan om linjära transformationer.

b) Vad händer med variationskoefficienten CV om alla observationer i en datamängd multipliceras med ett positivt tal b? *Ledning:* se bilagan om linjära transformationer.

Problem 1.2.11.

A study was conducted to examine effects of dose, renal function, and arthritic status on the pharmacokinetics of ketoprofen in elderly subjects. The study consisted of five non-arthritic and six arthritic subjects, each receiving 50 and 150 mg of racemic ketoprofen (a crossover study). Among the pharmacokinetic indices reported is $AUC_{0-\infty}$ (mg/L)hr for S- ketoprofen at the 50 mg dose.

Non-arthritic:	6.84	9.29	3.83	5.95	5.77	
Arthritic:	7.06	8.63	5.95	4.75	3.00	8.04

a) Compute the mean and standard deviation for the arthritic and non-arthritic groups, separately.

b) Compute the coefficient of variation for the arthritic and non-arthritic groups, separately.

c) Do these groups tend to differ in terms of the extent of absorption, as measured by AUC?

Problem 1.2.12.

A study was conducted to observe the effect of grapefruit juice on cyclosporine and prednisone metabolism in transplant patients. Among the measurements made was creatinine clearance at the beginning of the study. The values [ml/min] for the $n = 8$ patients in the study are as follow:

38 66 74 99 80 64 80 120

Find the mean creatinine clearance, the variance, standard deviation and coefficient of variation.

What can you say about standard deviation in relation to the mean?

Problem 1.2.13.

The duration of time from first exposure to HIV infection to AIDS diagnosis is called the incubation period. The incubation periods of a random sample of 7 HIV infected individuals is given below (in years):

	12.0	9.5	13.5	7.2	10.5	6.3	12.5

a) Calculate the sample mean and the sample median. Calculate the sample standard deviation.

b) If the number 6.3 above were changed to 1.5, what would happen to the sample mean, median, and standard deviation? State whether each would increase, decrease, or remain the same.

c) Suppose instead of 7 individuals, we had 14 individuals. (We add 7 additional randomly selected observations to the original 7)

	12.0	9.5	13.5	7.2	8.1	10.5	6.3	12.5	14.9	7.9	5.2	13.1	10.7	6.5

Make an educated guess of whether the sample mean and sample standard deviation for the 14 observations would increase, decrease, or remain roughly the same compared to your answers in part **a) -b)** based on only 7 observations.
Next actually calculate the sample mean and standard deviation to see if you were right. How does your calculation compare to your educated guess? Why do you think this is?

Coefficient of correlation (korrelationskoefficient)
The *covariance* between x- and y-values in $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ $(n > 1$ is

$$c_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}),$$

where $\bar{x} = \frac{1}{n-1}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n-1}\sum_{i=1}^{n} y_i$.
We standardize (to get back to the original units of measurement) with s_x and s_y,

$$s_x = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2}, s_y = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2}$$

and get the *correlation coefficient* as

Definition 1.1.

$$r \stackrel{\text{def}}{=} \frac{c_{xy}}{s_x s_y}.$$

Observera att $-1 \leq r \leq 1$. I Matlab `corrcoef(X,Y)`.

Problem 1.2.14.

You can clearly also write (why?)

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{j=1}^{n}(x_j - \bar{x})^2}\sqrt{\sum_{j=1}^{n}(y_j - \bar{y})^2}}.$$

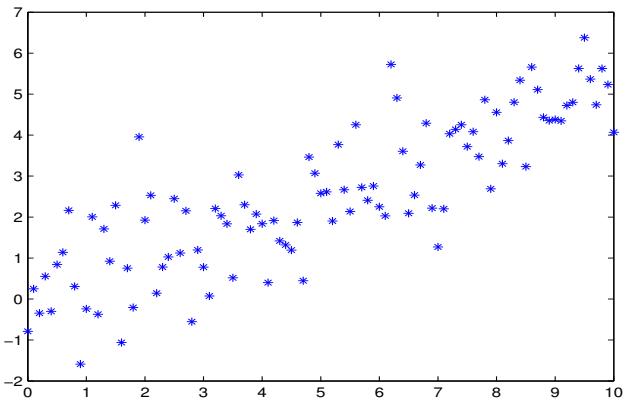
Check now that

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}\sqrt{n\sum_{j=1}^{n} y_j^2 - (\sum_{i=1}^{n} y_i)^2}},$$

which may perhaps be a clever form to use with simple electronic calculators. *Hint:* Bilaga 'Formler för beräkningar med summatecken' below.

Problem 1.2.15.

Which of the following statements is correct by inspection of the *scatterplot* for the paired data (x, y) ($*$ in



the figure) ?

a) Coefficient of correlation r is $= 0.8334$

b) Coefficient of correlation r is $= -0.7334$

c) Coefficient of correlation r is $= 1$

d) Coefficient of correlation r is $= 0$

Problem 1.2.16.

The kinetics of zidovudine in pregnant baboons was investigated in an effort to determine dosing regimens in pregnant women, with the goal to maintain AZT levels in the therapeutic range to prevent HIV infection in children. As part of the study, $n = 25$ measurements of AZT concentration (y) were made at various doses (x). The values of AZT concentration (μg/ml) and dose (mg/kg/hr) are given in the Table below.

Dose	0.67	0.86	0.96	0.6	0.94	1.12	1.4	1.17	1.35	1.76	1.92
AZT Conc.	0.169	0.178	0.206	0.391	0.387	0.333	0.349	0.437	0.428	0.597	0.587
Dose	1.43	1.77	1.55	1.51	1.82	1.91	1.89	2.5	2.5	2.02	2.12
AZT Conc.	0.653	0.66	0.688	0.704	0.746	0.797	0.875	0.549	0.666	0.759	0.806
Dose	2.5	2.5	2.5								
AZT Conc.	0.83	0.897	0.99								

a) Plot this data set in a scatterplot. Does it look like that there is a linear association between Dose x and AZT Concentration y?

b) Find the coefficient of correlation between Dose x and AZT Concentration y. How does the sign of the coefficient of correlation reflect the data in the scatterplot?

To be continued !

Problem 1.2.17.

Betrakta följande beskrivning[1] av det statistiska tänket

> **Information is gained by discarding information (discarding individual identification)**.

Diskutera poängen med detta påstående.

1.3 Kategoridata

Typvärde (på engelska **mode**) i en datamängd är det värde som förekommer flest gånger. En datamängd kan ha mer än ett typvärde, ty det kan hända att flera olika värden är lika (och mest) förekommande. Vi talar om **unimodala** (ett typvärde), **bimodala** (två typvärden) och **multimodala** (flera typvärden) data. En datamängd kan sakna typvärde, om inget värde upprepas. Typvärdet används mest i praktiken för beskrivning av kategoridata.

Problem 1.3.1.

Bestäm typvärdena för följande datamängder:

a) $\{5.40, 1.10, 0.42, 0.73, 0.48, 1.10\}$

b) $\{27, 27, 27, 55, 55, 55, 88, 88, 99\}$

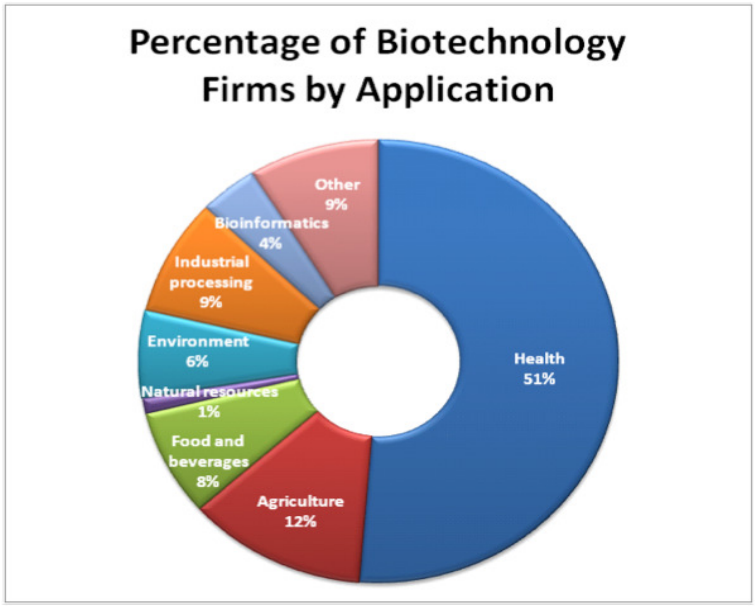
c) $\{1, 2, 3, 6, 7, 8, 9, 10\}$

Stapeldiagram (bar chart) och **cirkeldiagram (pie chart)** Ett stapeldiagram framställer kategoriska data med rektangulära staplar med längder proportionella mot kategorins frekvens i data. Diagrammet används för att jämföra olika kategorier. Ett **Paretodiagram** är ett stapeldiagram, där staplarna ges i storleksordning med den högsta stapeln längst till vänster.

[1] Stephen M. Stigler: Seven Pillars of Statistical Wisdom 2015

Problem 1.3.2.

Framställ cirkeldiagrammet i bilden som ett Paretodiagram.



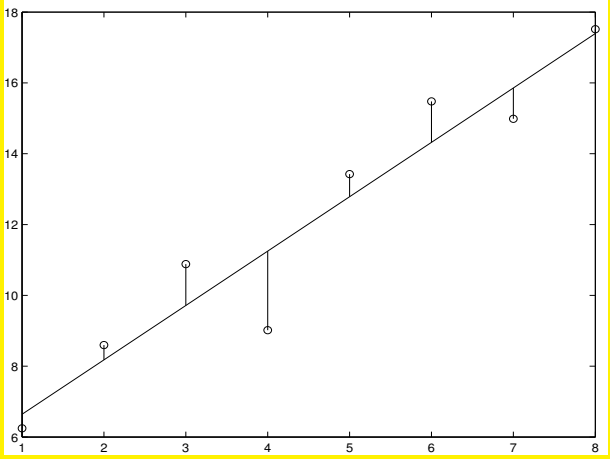
1.4 Linjär regression

$$y = \alpha + \beta x$$

kallas en *regressionslinje*.
För att bestämma α och β från data använder man minsta kvadratmetoden som innebär att man skall minimera kvadratsumman (minimal mean square regression)

$$Q(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2$$

Geometriskt innebär det att summan av de lodräta kvadratavstånden mellan observationerna y_i och regressionslinjen skall minimeras, se figuren. Genom att derivera med avseende på α och β och sätta derivatorna lika med 0, finner man att $Q(\alpha, \beta)$ minimeras av $\widehat{\beta}$ och $\widehat{\alpha}$ givna av



$$\widehat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\widehat{\alpha} = \bar{y} - \widehat{\beta}\bar{x}.$$

Man kan notera en algebraisk likhet

$$\sum_{i=1}^{n}(x_i - \bar{x})y_i = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i(y_i - \bar{y}),$$

varför uttrycket för $\widehat{\beta}$ kan skrivas/räknas på flera sätt.

Problem 1.4.1.

The kinetics of zidovudine in pregnant baboons was investigated in an effort to determine dosing regimens in pregnant women, with the goal to maintain AZT levels in the therapeutic range to prevent HIV infection in children. As part of the study, $n = 25$ measurements of AZT concentration (y) were made at various doses (x). The values of AZT concentration (μg/ml) and dose (mg/kg/hr) are given in the Table in problem **1.2.16** above. It holds for this data:

$$\sum_{i=1}^{25}(x_i - \bar{x})^2 = 6.1744, \sum_{i=1}^{25}(x_i - \bar{x})(y_i - \bar{y}) = 1.8692.$$

a) Find the least square regression equation relating y to x.

b) What is the expected AZT concentration $\widehat{y}(= \widehat{\alpha} + \widehat{\beta}x)$ for $x = 1.7$?

c) $\widehat{y}_i = \widehat{\alpha} + \widehat{\beta}x_i$. Compute the **residuals**

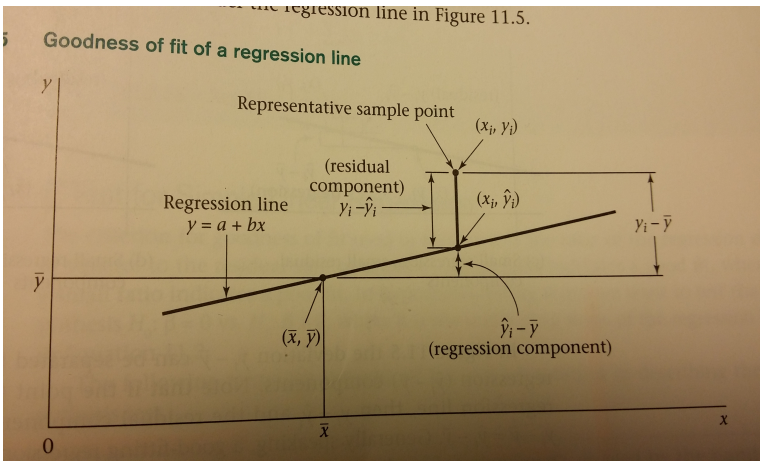
$$e_i \stackrel{\text{def}}{=} y_i - \widehat{y}_i, \quad i = 1, 2, \ldots, 25.$$

Plot the pairs (x_i, e_i) in a scatterplot. What do you find?

d) We have found in the preceding $\widehat{\alpha} = \bar{y} - \widehat{\beta}\bar{x}$. Thus $\widehat{y} = \widehat{\alpha} + \widehat{\beta}x = \bar{y} + \widehat{\beta}(x - \bar{x})$. Hence if $x = \bar{x}$, we have

$$\widehat{y} = \bar{y} + \widehat{\beta}(\bar{x} - \bar{x}) = \bar{y}.$$

This means that the estimated regression line always passes through the point $(\bar{x}, \bar{y})$.



Goodness of fit & Coefficient of determination
For any (x_i, y_i) the **regression component** of that pair about the regression line is $\widehat{y}_i - \bar{y}$.
A good-fitting regression line will have regression components large in absolute value relative to the residuals. Then we have

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 \leftrightarrow \textbf{Total Sum of Squares, Total SS}$$

$$\sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2 \leftrightarrow \textbf{ Regression Sum of Squares, Reg SS}$$

It can be shown

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2,$$

i.e.,

$$\textbf{Total SS} = \textbf{ Reg SS } + \textbf{ Residual Sum of Squares}$$

For any (x_i, y_i) the **explained deviation** of that pair about the regression line is $\widehat{y}_i - \bar{y}$.
For any (x_i, y_i) the residual or the **unexplained deviation** of that pair about the regression line is $e_i = y_i - \widehat{y}_i$.

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 \leftrightarrow \textbf{Total Variation}$$

$$\sum_{i=1}^{n} (\widehat{y}_i - \bar{y})^2 \leftrightarrow \textbf{ Explained Variation}$$

$$\sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 \leftrightarrow \textbf{ Unexplained Variation}$$

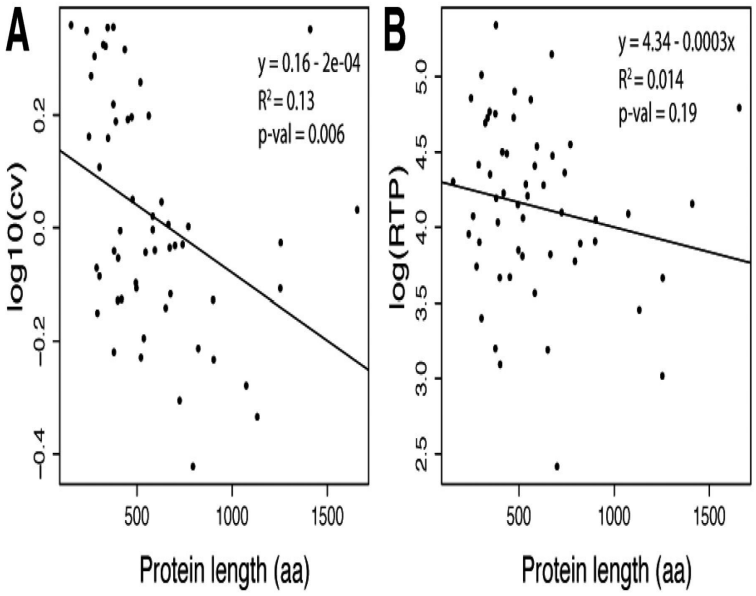
The **coefficient of determination** R^2 is the amount of variation in y that is explained by the regression line.

$$R^2 \stackrel{\text{def}}{=} \frac{\textbf{Explained Variation}}{\textbf{Total Variation}}$$

Compute R^2 for the regression with (dose,AZT) $(= (x, y))$ data. Coment your result.

Problem 1.4.2.

Let us study below a figure from:
Edfors, Fredrik and Danielsson, Frida and Hallström, Björn M and Käll, Lukas and Lundberg, Emma and Pontén, Fredrik and Forsström, Björn and Uhlén, Mathias: *Gene-specific correlation of RNA and protein levels in human cells and tissues*, **Molecular Systems Biology**, 12, 883 , 2016.



In the left hand field (**A**) we see the variation measured as coefficient of variation (CV) across samples plotted against the protein length.

In the right hand field (**B**) the protein lengths for the 55 target proteins are plotted against the RNA-to-protein (RTP) ratio.

Data information from the paper cited: **Test** based on correlation coefficient and follows a **t-distribution with length(x)-2 degrees of freedom** if the samples follow **independent normal distributions**. An asymptotic **confidence interval** (the boldfaced stuff to be explained later in the course) is given based on **Fisher´s z-transform**, this not necessarily included later in the course, but is given by as

$$z \stackrel{\text{def}}{=} \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \operatorname{arctanh}(r),$$

and derived and discussed in Vidakovic p. 577-578.

Discuss the scatterplots, the regression lines, and the values of R^2. ***p*-values will be treated later in the course**.

Problem 1.4.3.

a) Check that

$$\widehat{\beta} = \frac{s_y}{s_x} r,$$

where r is the coefficient of correlation.

b) Check now that

$$\frac{\widehat{y} - \bar{y}}{s_y} = r\frac{x - \bar{x}}{s_x}.$$

Aid: Check the formulas in **d)** of the problem **1.4.1**.

c) Hence we see that, if $-1 < r < 1$ (i.e. $|r| < 1$),

$$\frac{|\,\widehat{y} - \bar{y}\,|}{s_y} < \frac{|\,x - \bar{x}\,|}{s_x}.$$

Sir Francis Galton, 1822-1911, (statistician and proto-geneticist), bought family records that contained the heights of 205 sets of parents and their adult children. In this data the smallest parents had offspring who were slightly bigger and closer to the mean. The largest parents had offspring who were slightly

smaller and once again closer to the mean. This lead Galton to invent the word regression. **Regression was defined by Galton as the process of returning to the mean**.
How is 'returning to the mean" mathematically expressed above? What does the mathematical assumption $|r| < 1$ mean in terms of heredity?

d) Returning to the mean was in earlier times often called 'regression to mediocrity". What is the implication of this?

1.5 Bilaga: Formler för statistiska beräkningar

Denna bilaga är avsedd som ett koncist uppslagsverk för formler som används i statistiska beräkningar. Man kan och får använda dessa utan att ha läst de motsvarande bevisen.
Bevisen kommer ej på en tenta.

1.5.1 Lite exercis med summatecken

(1) $\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n.$

(2) $\sum_{i=1}^{n} a \cdot x_i = a \sum_{i=1}^{n} x_i.$
Bevis: Definitionen (1) ger $\sum_{i=1}^{n} a \cdot x_i = ax_1 + ax_2 + \ldots + ax_n = a\left(x_1 + x_2 + \ldots + x_n\right) = a \sum_{i=1}^{n} x_i.$
Exempel: $x_i = 1$, $i = 1. \ldots, n$
$\sum_{i=1}^{n} a = a + a + \ldots + a = a\left(1 + 1 + \ldots + 1\right) = a \cdot n.$

(3) $\sum_{i=1}^{n} \left(x_i + y_i\right) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i.$
Bevis: Definition (1) ger $\sum_{i=1}^{n} \left(x_i + y_i\right) = (x_1 + y_1) + (x_2 + y_2) + \ldots + (x_n + y_n) = x_1 + x_2 + \ldots + x_n + y_1 + y_2 + \ldots + y_n = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i.$

(4) $\sum_{i=1}^{n} \left(ax_i + by_i\right) = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} y_i$
Bevis: Detta fås av (3) och (2).

(5) $\sum_{i=1}^{n} \left(x_i + y_i\right)^2 = \sum_{i=1}^{n} x_i^2 + 2 \sum_{i=1}^{n} x_i y_i + \sum_{i=1}^{n} y_i^2.$
Bevis: Använd $\left(x_i + y_i\right)^2 = x_i^2 + 2x_i y_i + y_i^2$ och (4) samt (2) med $a = 2$.

Låt $\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$. Då gäller

(6) $\sum_{i=1}^{n} \left(x_i - \overline{x}\right) = 0.$
Bevis: $\sum_{i=1}^{n} \left(x_i - \overline{x}\right) = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \overline{x}$ enligt (4). Men här har vi med $a = \overline{x}$ i (2) att $\sum_{i=1}^{n} \overline{x} = \overline{x} \sum_{i=1}^{n} 1 = \overline{x} \cdot n$ enligt exemplet i (2). Men $\overline{x} \cdot n = \sum_{i=1}^{n} x_i$ och detta ger påståendet i (6).

(7) $\sum_{i=1}^{n} \left(x_i - \overline{x}\right) \cdot \left(y_i - \overline{y}\right) = \sum_{i=1}^{n} \left(x_i - \overline{x}\right) y_i = \sum_{i=1}^{n} x_i \left(y_i - \overline{y}\right).$
Bevis: $\left(x_i - \overline{x}\right) \cdot \left(y_i - \overline{y}\right) = \left(x_i - \overline{x}\right) \cdot y_i - \left(x_i - \overline{x}\right) \cdot \overline{y}.$ Då fås enligt (4) att $\sum_{i=1}^{n} \left(x_i - \overline{x}\right) \cdot \left(y_i - \overline{y}\right) = \sum_{i=1}^{n} \left(x_i - \overline{x}\right) \cdot y_i - \sum_{i=1}^{n} \left(x_i - \overline{x}\right) \cdot \overline{y}$. Men med $a = \overline{y}$ i (2) fås att $\sum_{i=1}^{n} \left(x_i - \overline{x}\right) \cdot \overline{y} = \overline{y} \cdot \sum_{i=1}^{n} \left(x_i - \overline{x}\right)$ och då ger (6) att $\sum_{i=1}^{n} \left(x_i - \overline{x}\right) \cdot \left(y_i - \overline{y}\right) = \sum_{i=1}^{n} \left(x_i - \overline{x}\right) y_i$. Analogt tar vi fram den andra likheten.

(8) $\sum_{i=1}^{n} \left(x_i - \overline{x}\right) \cdot \left(y_i - \overline{y}\right) = \sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}.$
Bevis: Utveckla t.ex. $\sum_{i=1}^{n} x_i \left(y_i - \overline{y}\right)$ i högra ledet av (7) och använd (2) och definitionen på $\overline{x}$.

(9) $\sum_{i=1}^{n}(x_i-\overline{x})^2=\sum_{i=1}^{n}x_i^2-n\overline{x}^2$.
Bevis: Ur (5) fås att $\sum_{i=1}^{n}(x_i-\overline{x})^2=\sum_{i=1}^{n}x_i^2-2\sum_{i=1}^{n}x_i\overline{x}+\sum_{i=1}^{n}\overline{x}^2$. Då ger (2) med $a=\overline{x}$ och exemplet i (2) att $\sum_{i=1}^{n}x_i^2-2\sum_{i=1}^{n}x_i\overline{x}+\sum_{i=1}^{n}\overline{x}^2=\sum_{i=1}^{n}x_i^2-2\overline{x}\sum_{i=1}^{n}x_i+n\overline{x}^2$. Definitionen på $\overline{x}$ ger $\sum_{i=1}^{n}x_i=n\overline{x}$, så att $\sum_{i=1}^{n}x_i^2-2\overline{x}\sum_{i=1}^{n}x_i+n\overline{x}^2=\sum_{i=1}^{n}x_i^2-2\overline{x}\cdot n\overline{x}+n\overline{x}^2=\sum_{i=1}^{n}x_i^2-n\overline{x}^2$.

(10) $\sum_{i=1}^{n}(x_i-a)^2=\sum_{i=1}^{n}(x_i-\overline{x})^2+n(\overline{x}-a)^2$.

Bevis: En identisk omskrivning

$$\sum_{i=1}^{n}(x_i-a)^2=\sum_{i=1}^{n}\left[(x_i-\overline{x})-(a-\overline{x})\right]^2$$

och enligt (5)

$$=\sum_{i=1}^{n}(x_i-\overline{x})^2-2\sum_{i=1}^{n}(x_i-\overline{x})\cdot(a-\overline{x})+\sum_{i=1}^{n}(a-\overline{x})^2.$$

Med (2)

$$=\sum_{i=1}^{n}(x_i-\overline{x})^2-2(a-\overline{x})\sum_{i=1}^{n}(x_i-\overline{x})+\sum_{i=1}^{n}(a-\overline{x})^2$$

och (2) och dess Exempel igen

$$=\sum_{i=1}^{n}(x_i-\overline{x})^2-2(a-\overline{x})\sum_{i=1}^{n}(x_i-\overline{x})+n(a-\overline{x})^2$$

och (6) ger $\sum_{i=1}^{n}(x_i-\overline{x})=0$ så att

$$=\sum_{i=1}^{n}(x_i-\overline{x})^2+n(a-\overline{x})^2,$$

vilket är högra ledet i (10), vilket skulle visas.

1.5.2 Derivator av summor

Låt $g(\theta)=\sum_{i=1}^{n}(x_i-\mu(\theta))^2$. Då gäller

(11) $\frac{d}{d\theta}g(\theta)=-2\mu'(\theta)\sum_{i=1}^{n}(x_i-\mu(\theta))$, där $\mu'(\theta)=\frac{d}{d\theta}\mu(\theta)$.
Bevis: $\frac{d}{d\theta}g(\theta)=\frac{d}{d\theta}\sum_{i=1}^{n}(x_i-\mu(\theta))^2=\frac{d}{d\theta}\left((x_1-\mu(\theta))^2+\ldots+(x_n-\mu(\theta))^2\right)=\frac{d}{d\theta}(x_1-\mu(\theta))^2+\ldots+\frac{d}{d\theta}(x_n-\mu(\theta))^2$, ty derivatan av en summa är summan av termernas derivator. För varje enskild term i summan gäller $\frac{d}{d\theta}(x_i-\mu(\theta))^2=2(x_i-\mu(\theta))\left(-\mu'(\theta)\right)$. Detta ger $\frac{d}{d\theta}(x_1-\mu(\theta))^2+\ldots+\frac{d}{d\theta}(x_n-\mu(\theta))^2=2(x_1-\mu(\theta))\left(-\mu'(\theta)\right)+\ldots+2(x_n-\mu(\theta))\left(-\mu'(\theta)\right)=2\left(-\mu'(\theta)\right)((x_1-\mu(\theta))+\ldots+(x_n-\mu(\theta)))=2\left(-\mu'(\theta)\right)\sum_{i=1}^{n}(x_i-\mu(\theta))$, där vi utnyttjade (2) med $a=2\left(-\mu'(\theta)\right)$.

(12) $\frac{d}{d\theta}g(\theta)=0\Rightarrow\mu(\theta)=\overline{x}$.
Bevis: Enligt (11) $\frac{d}{d\theta}g(\theta)=0\Leftrightarrow-2\mu'(\theta)\sum_{i=1}^{n}(x_i-\mu(\theta))=0\Leftrightarrow\sum_{i=1}^{n}(x_i-\mu(\theta))=0$. Då ger (3) och (2) att $\Rightarrow\sum_{i=1}^{n}x_i-\sum_{i=1}^{n}\mu(\theta)=0$ och Exemplet i (2) ger $\Rightarrow\sum_{i=1}^{n}x_i-n\mu(\theta)=0$ och detta $\Rightarrow\mu(\theta)=\frac{1}{n}\sum_{i=1}^{n}x_i$.

Problem 1.5.1.

Bestäm ett tal θ så att funktionen

$$Q(\theta) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \theta)^2$$

minimeras. Det finns två olika sätt, som bör ge samma svar:

a) Använd (11) och (12).

b) Använd (10).

Problem 1.5.2.

Visa M =median$(x_1, \ldots, x_n)$ minimerar funktionen

$$Q(M) = \frac{1}{n}\sum_{i=1}^{n}|x_i - M|.$$

Problem 1.5.3.

Valbar, kommer ej på en tenta
Visa för en minsta kvadrat linjär regression med beteckningarna ovan att

$$\textbf{Total SS} = \textbf{ Reg SS } + \textbf{ Residual Sum of Squares}$$

d.v.s. att

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} e_i^2.$$

Ledning: Kolla hur man härleder (10) ovan och använd här att

$$\widehat{y}_i = \widehat{\alpha} + \widehat{\beta}x_i,$$

där $\widehat{\alpha}$ och $\widehat{\beta}$ minimerar $Q(\alpha, \beta)$.

Problem 1.5.4.

We use $\sum_{i=1}^{n}(x_i - \bar{x})^2$ to measure dispersion.

Why is it not possible to use $\sum_{i=1}^{n}(x_i - \bar{x})$ for this purpose ?

1.5.3 Linjära transformationer

Om man adderar ett tal a till varje tal i en datamängd kommer medelvärdet att förändras med samma värde. Det är lätt att se att standardavvikelsen blir oförändrad; spridningen förändras ju inte, endast läget för datamängden.

Om varje tal multipliceras med ett tal b kommer medelvärdet att förändras på samma sätt medan standardavvikelsen blir den gamla multiplicerat med absolutvärdet på b. I formler har vi följande.

(13) Om $y_i = bx_i + a$, $i = 1, \ldots, n$, blir

$$\underline{\bar{y} = b\bar{x} + a \text{ och } s_y = \mid b \mid s_x.}$$

Bevis: Detta följer med stöd av reglerna (4) och (2) i avsnittet ovan i denna bilaga och definitionerna

$$\bar{y} = \frac{\sum_{i=1}^{n}(bx_i + a)}{n} = \frac{\sum_{i=1}^{n} bx_i + n \cdot a}{n} = \frac{b\sum_{i=1}^{n} x_i}{n} + a = b\bar{x} + a$$

och

$$s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1} = \frac{\sum_{i=1}^{n}(bx_i + a - (b\bar{x} + a))^2}{n-1} = \frac{\sum_{i=1}^{n}(bx_i - b\bar{x})^2}{n-1} = \frac{b^2\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = b^2 s_x^2.$$

Problem 1.5.5.

Låt $z_i = \frac{x_i - \bar{x}}{s_x}$, $i = 1, \ldots, n$. Dessa z_i är kända som **z-scores** eller **z-statistikor**.

Bestäm $\bar{z}$ och s_z.

Problem 1.5.6.

Låt $y_i = bx_i + a$, $i = 1, \ldots, n$.

Bestäm värdet på korrelationskoefficienten

$$r = \frac{c_{xy}}{s_x s_y}$$

för dessa x_i och y_i.

Problem 1.5.7.

Låt oss skriva

$$r_{xy} \stackrel{\text{def}}{=} \frac{c_{xy}}{s_x s_y}.$$

Vad är r_{yx}? Vad lär vi oss av svaret?

Kapitel 2

Data-analys Del II

2.1 Metoder för jämförelse av data

z-score
I statistiken definieras **z-score** eller **standardpoängen** för en datapunkt x som

$$z \stackrel{\text{def}}{=} \frac{x - \bar{x}}{s}.$$

Det finns andra betydelser av z-score utanför statistiken. z-score är nyttig för jämförelse av olika datamängder och jämförelse av en datapunkt med avseende på en mängd: z saknar sort (dimensionless quantity).
z-score kan vara negativ eller positiv. Ofta säges det att om z-score ≤ -2 eller ≥ 2, så är x en ovanlig observation.

Problem 2.1.1.

Enligt en undersökning gjord av SCB(= Statistiska centralbyrån) under 1998−2000 är medellängden på svenska kvinnor 165.5 [cm] med standardavvikelsen 6.15 [cm] och medellängden på svenska män 179.0 [cm] med standardavvikelsen 6.85 [cm]. Detta kommer från ett stickprov på svenska kvinnor och män i åldersintervallet 16-84 år.

Lotta Schelin är fotbollsspelare i svenskt damlandslag och har kroppslängd 179.0 [cm]. John Guidetti är fotbollsspelare i svenskt herrlandslag och har kroppslängd 185.0 [cm].

Beräkna z-score för dessa två kroppslängder och kommentera resultaten.

Kvantiler, percentiler och kvartiler
Vi har ett data set med n värden som i storleksordning är $x_1 \leq x_2 \leq \ldots \leq x_n$.

1. För varje tal p som är av formen $p = \frac{i-0.5}{n}$, där i är ett heltal i $\{1, \ldots, n\}$, är **p-kvantilen** lika med x_i.
2. För tal p mellan $\frac{0.5}{n}$ och $\frac{n-0.5}{n}$, som inte är lika med $\frac{i-0.5}{n}$, fås **p-kvantilen** för detta data set genom (linjär) interpolation mellan två tal av formen $\frac{i-0.5}{n}$ sådana att p ligger mellan dem.

I båda fallen betecknas **p-kvantilen** med $Q(p)$. Olika programpaket utnyttjar var sin interpolationsmetod i fall 2..
En **percentil** är det värde nedanför vilken en viss procent av observationerna i datat ligger. Så är till exempel den tjugonde percentilen P_{20} det värde som delar observationsvärden så att 20 procent av dem är mindre än P_{20} och 80 procent är större.
De percentiler som delar in materialet i fyra delar, P_{25} d.v.s. $Q(0.25)$(även betecknad med Q_1), $P_{0.50}$, d.v.s. $Q(50)$(även betecknad med Q_2) och P_{75} d.v.s. $Q(0.75)$(även betecknad som Q_3)), kallas **kvartiler**.

Problem 2.1.2.

Ten strength measurements of a paper towel were performed. The breaking strengths were reported as follows.

Test	Breaking strength (g)
1	8.577
2	9.471
3	9.011
4	7.583
5	8.572
6	10.688
7	9.614
8	9.614
9	8.527
10	9.165

a) Find **first (första) (or lower (nedre)) qvartile** $Q(0.25)(= Q_1)$.

b) Find **median** $Q(0.5)$.

c) Find **third (tredje) (or upper (övre)) qvartile** $Q(0.75)(= Q_3)$.

Interquartile range, (på sv. kvartilavstånd) (IQR) is defined as

$$IQR \stackrel{\text{def}}{=} Q(0.75) - Q(0.25) = Q_3 - Q_1.$$

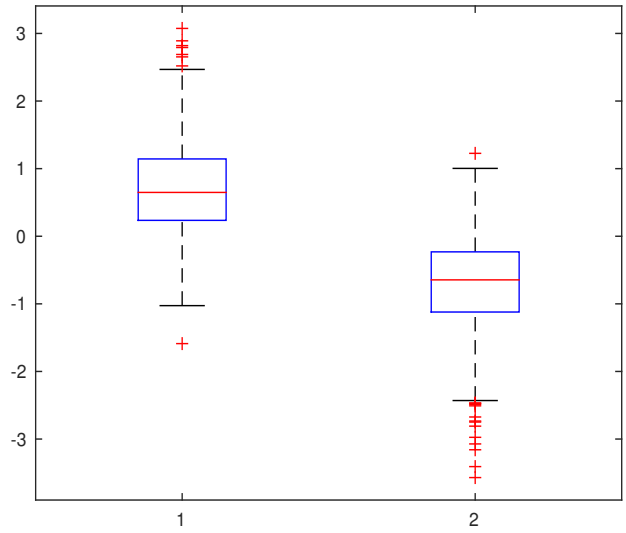
and is implemented in every boxplot, see next.

Lådagram, låddiagram eller boxplot
är ett diagram där ett statistiskt material åskådliggörs i form av en låda, som anger den mittersta hälften av datamaterialet. Diagrammet sammanfattar datamaterialet med hjälp av fem värden: medianen, $Q(0.25)$ och $Q(0.75)$ samt minimum och maximum. Eventuella extrema värden betraktas som utliggare (outliers) och markeras med egna symboler, t.ex. +.
Lådan begränsas nedåt av $Q(0.25)$ och uppåt av $Q(0.75)$. Medianen ritas med ett streck genom lådan. Värden som ligger längre ifrån boxen än 1.5 gånger kvartilavståndet IQR som utliggare och är markerade med +. Värden som ligger mer än 3 gånger kvartilavståndet IQR från lådan betraktas som avlägsna utliggare, och betecknas i Matlab med +. De strecken som går ut från boxen dras till det lägsta värdet och det högsta bland de värden som inte är utliggare.

Problem 2.1.3.

Lådagrammen (boxplottarna) i bilden har framtagits för två olika dataset. Följande statistikor har beräknats: IQR = 0.89, IQR= -0.6454, skewness= 0.3525, skewness= -0.4757 och median = -0.6454, median = 0.6482.

Vilken trippel (IQR, skewness, median) av statistikorna svarar mot boxplotten till vänster?

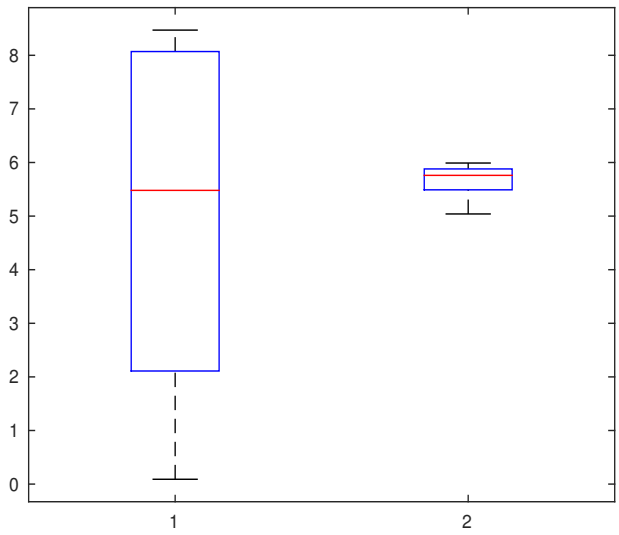


Problem 2.1.4.

Lådagrammen i bilden svarar mot data i problem 1.2.1. Data A har boxplottats till vänster.

Bestäm med ögonmått kvartilavståndet IQR för både A och B från plotten.

Jämför med analysen Du gjorde tidigare med problem 1.2.1..



Problem 2.1.5.

Vilka mått på läge och spridning skulle Du använda för en skev datamängd i ordinalskala?

Problem 2.1.6.

An **outlier** (utliggare) is an observation that lies an abnormal distance from other values in a random sample from a population.

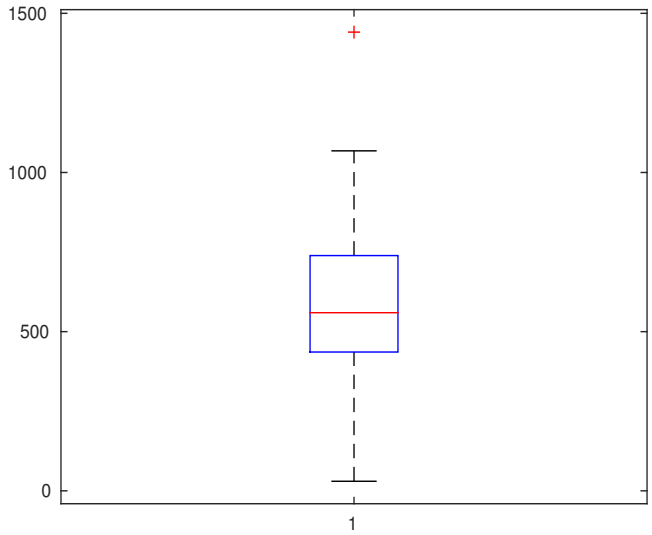
To detect abnormality we define

- **lower inner fence**: $\stackrel{\text{def}}{=} Q_1 - 1.5 \cdot IQR$,
- **upper inner fence**: $\stackrel{\text{def}}{=} Q_3 + 1.5 \cdot IQR$,
- **lower outer fence**: $\stackrel{\text{def}}{=} Q_1 - 3 \cdot IQR$,
- **upper outer fence**: $\stackrel{\text{def}}{=} Q_3 + 3 \cdot IQR$.

A data point beyond an inner fence on either side is considered a **mild outlier**. A data point beyond an outer fence is considered an **extreme outlier**.

Consider the data set x (without context) below. This data set can be loaded down from the course webpage.

30	171	184	201	212	250	265	270	272	289
305	306	322	322	336	346	351	370	390	404
409	411	436	437	439	441	444	448	451	453	470
480	482	487	494	495	499	503	514	521	522	527
548	550	559	560	570	572	574	578	585	592	592
607	616	618	621	629	637	638	640	656	668	707
709	719	737	739	752	758	766	792	792	794	802
818	830	832	843	858	860	869	918	925	953	991
			1000	1005	1068	1441			



Its boxplot is found in the figure.

a) Determine **Pearson 2 skewness coefficient**

$$Sk_2 = 3 \cdot \frac{\bar{x} - \text{median}(x)}{s}.$$

What is the conclusion? Remember that $-3 \leq Sk_2 \leq 3$.

b) Determine lower inner fence, upper inner fence, lower outer fence and upper outer fence and decide if there is an outlier and point it out. *Hint:* Look at the boxplot, too.

c) What is the z-score of your chosen outlier?

d) You have now chosen an outlier. Remove it form the data set x and compute the mean and the median of data set so trimmed. What do you find? How does skewness g_1 change in the trimmed data set? How does skewness sk_2 change in the trimmed data set? Why does this make sense?

e) Determine the kurtosis of x. Determine the kurtosis of trimmed data set in **c)**. What do you find and why should this make sense?

Problem 2.1.7.

Mer om utliggarna

Consider the following 24 determinations (replicates) of copper (μgg^{-1}) in wholemeal flour

2.9	3.1	3.4	3.4	3.70	3.7	2.8	2.5
2.4	2.4	2.7	2.2	5.28	3.37	3.03	3.03
28.95	3.77	3.4	2.2	3.5	3.6	3.7	3.7

a) Find the boxplot. What are the largest and second largest outliers?

b) Compute mean and variance. Remove the largest outlier and compute mean and variance.

c) One value, 28.95, stands out from the remainder even without any statistical computation. In this instance we may be particularly suspicious of the value, as a simple explanation suggests itself. Although recording and range errors are almost certainly the major cause of outliers, mistakes can also occur in many other parts of the analytical process and from contamination and transposition of specimens.

The almost universal practice amongst (analytical) chemists has been to regard outliers as errors, and to delete them from the set of data. In some circumstances this is plainly wrong, and in others there are much safer procedures. Why should we be interested in outliers? One good reason is to catch transcription errors while the original laboratory records are easily accessible. In such an instance we would want to check all the extreme results and remove them.

Remove even the second largest outlier and compute mean and variance. What can be concluded?

Analytical Methods Committee and others: *Robust statistics–how not to reject outliers. Part 1. Basic concepts*, **Analyst**, 114, pp. 1693–1697, 1989, Royal Society of Chemistry.

2.2 Data-analys med histogram och relativ frekvens (empirisk sannolikhet)

Histogram
Ett histogram är en grafisk framställning eller skattning av en sannolikhetskurva (eller sannolikhetstäthet, jf. senare kapitel) för en kvantitativ kontinuerlig datamängd. Konstruktion av ett histogram kallas 'data binning" och görs enligt följande (1.-3.):

1. intervallet mellan det största och minsta värdet av datapunkter uppdelas i ett antal intervall (bin), vanligtvis är dessa lika långa och icke-överlappande.
2. räkna antalet datapunkter som ligger i varje intervall.
3. rita en rektangel ovanför varje intervall. Rektangelns höjd är proportionell mot frekvensen, d.v.s. mot antalet datapunkter i intervallet. Rektangeln kan även normaliseras, så att rektangelns höjd är proportionell mot den relativa frekvensen av antalet datapunkter i intervallet.

Det finns tre avgörande **skillnader mellan histogram och stapeldiagram**: histogram används för att visa fördelningar medan stapeldiagram används för att jämföra kategorier. Med ett histogram plottas "binned' kvantitativa data, när med ett stapeldiagram plottas frekvenser för kategoriska data. Staplarna kan byta plats och ordningsföljd i ett stapeldiagram men inte i ett histogram.

Problem 2.2.1.

Bl.a. Matlab har automatiserat stegen 1.-3. ovan för konstruktion av histogram:

- **N = hist(X)** bins the elements of X into 10 equally spaced containers and returns the number of elements in each container.
- **hist(X)** bins the elements of X into 10 equally spaced containers and produces a histogram bar plot.
- **hist(Y,M),** where M is a scalar, uses M bins.

Den kvarstående frågan är tydligen att välja M, antalet intervall. Det finns ingen universell lösning. Experimentering behövs vanligtvis.

Statistikerna har utvecklat även följande strategi. Vi bestämmer först h = bredden av intervallet (=längden av bin). Sedan beräknar vi M = antalet bins eller intervall enligt formeln:

$$M = \left\lceil \frac{\max X - \min X}{h} \right\rceil .$$

där $\lceil a \rceil$ avrundar talet a uppåt till närmaste heltalet större än a, i Matlab `ceil`. Det återstår att välja h. Här har statistikerna härlett/hittat på många regler.

Betrakta datamängden X nedan:

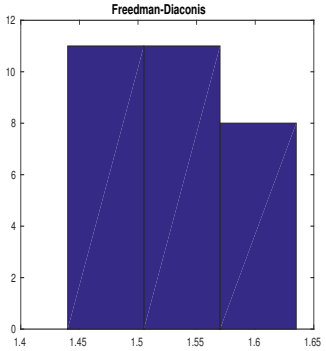
1.5450	1.4400	1.4400	1.5200	1.5800	1.5400	1.5550	1.4900
1.5600	1.4950	1.5950	1.5500	1.6050	1.5100	1.5600	1.4450
1.4400	1.5950	1.4650	1.5450	1.5950	1.6300	1.5150	1.6350
	1.6250	1.5200	1.4550	1.4500	1.4800	1.4450	

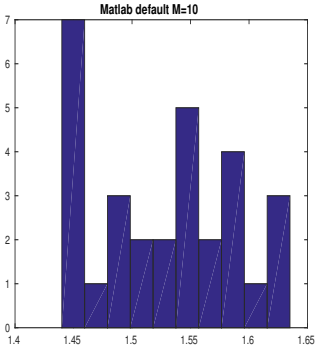
1. **Freedman−Diaconis regel för val av bredden** h (IQR= interquartile range, se tidigare)

$$h = 2\frac{\mathrm{IQR}(X)}{n^{1/3}}$$

ger $M = 3$ för X och histogrammet

```
>>   hist(X,ceil((max(X)-min(X))/(2*iqr(X)/((30^(1/3))))))
```





2. **Default i Matlab tar M=10**, `hist(X)` ger

Vilket av dessa två sätt att välja h d.v.s. antalet bin är bättre för detta data och varför? *Hjälp:* Kolla medelvärde, median och skevhet (skewness).

Problem 2.2.2.

Från problem **1.2.9.**:

87	105	108	120	142	151	155	155	161	162
169	173	174	183	185	186	186	192	193	196
199	205	207	211	215	217	217	222	224	226
227	230	231	231	237	242	244	246	251	258
263	267	278	286	294	312	338	341	362	390

Klassindela materialet med några lämpliga klassbredder och **rita upp** motsvarande histogram.

Problem 2.2.3.

Ett antal uttryck kan med fördel användas för att beskriva det allmänna utseendet eller den allmänna formen av ett histogram eller sannolikhetskurva.

Vi har fem olika histogram i figurerna (i) - (v).

Tillordna för vart och ett av histogrammen (i)-(v) den av fraserna **a)- e)** som histogrammet bäst svarar mot.

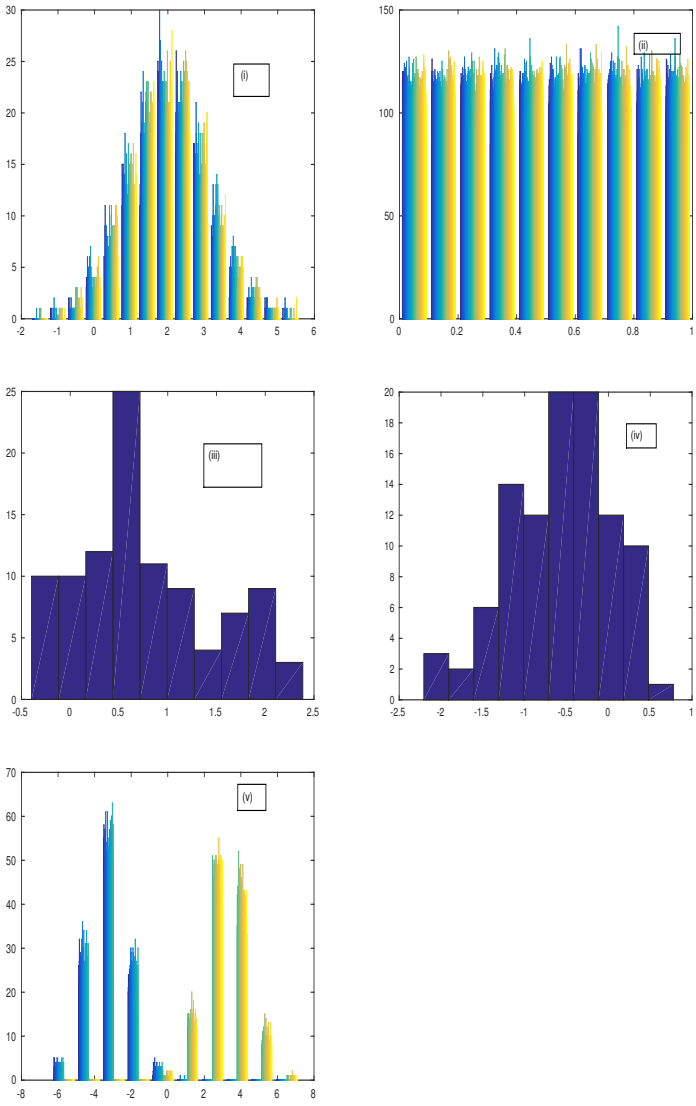
a) skev åt höger

b) skev åt vänster

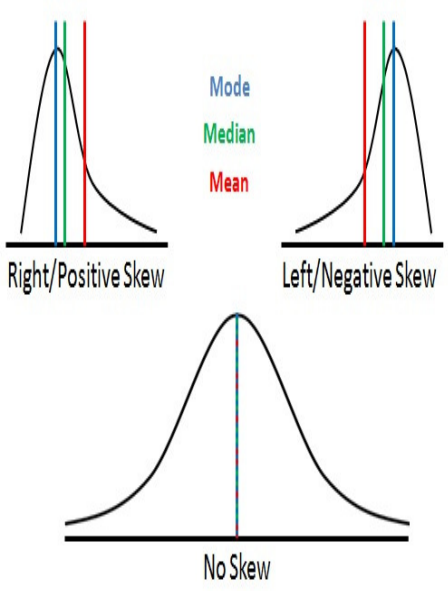
c) symmetrisk

d) bimodal

e) likformig (uniform på engelska)



Problem 2.2.4.



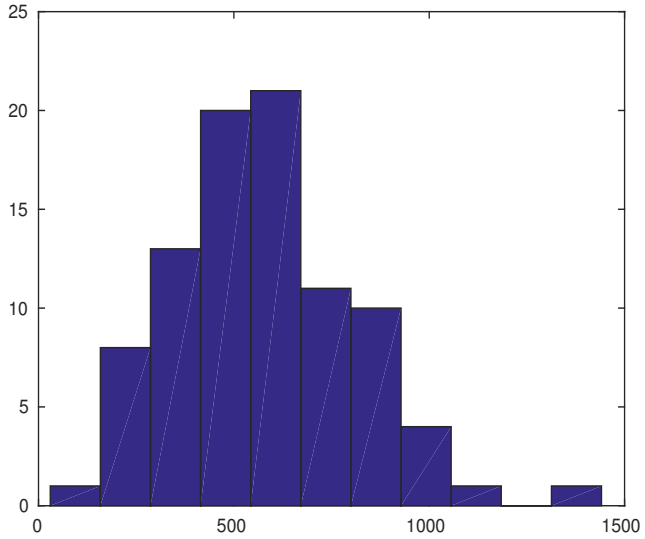
Vilken idé vill de tre bilderna förmedla?

Problem 2.2.5.

Från tidigare: Betrakta datamängden x nedan.

30	171	184	201	212	250	265	270	272	289
305	306	322	322	336	346	351	370	390	404
409	411	436	437	439	441	444	448	451	453	470
480	482	487	494	495	499	503	514	521	522	527
548	550	559	560	570	572	574	578	585	592	592
607	616	618	621	629	637	638	640	656	668	707
709	719	737	739	752	758	766	792	792	794	802
818	830	832	843	858	860	869	918	925	953	991
			1000	1005	1068	1441			

Dess histogram enligt Diaconis-Freedmans regel för bredd är givet i figuren.



a) Bestäm $\bar{x}$, median(x) och mode. Vilken ordningsföljd efter storlek har dessa sinsemellan?

b) Bestäm skevhet g_1 och Sk_2 och jämför med **a)**.

c) Hur diskuteras outliers i detta fall?

Problem 2.2.6.

The table below shows the frequencies of major causes of death in England and Wales in 1989 by sex.

	Males	Females
Circulatory system	127435	137165
Neoplasms (cancers)	75172	69948
Respiratory system	33489	33223
Digestive system	7900	10779
Injury and poisoning	11073	6427
Mental disorders	4493	9225
Other	19811	27460
Total	279373	294227

a) Calculate the relative frequency of each cause of death for both sexes combined.

b) Calculate the relative frequency of each cause of death for each sex separately. Compare the two frequency distributions and comment.

c) For each cause of death, calculate the relative frequencies of males and females. How does this compare to **b)**?

d) Histogram or bar chart for this data?

d) Can you speculate why there are almost 15000 more deaths among females than among males?

Problem 2.2.7.

Vid ett bioraffinaderi har man mätt ett antal fläckar på 50 lika stora pappersprover:

Antal fläckar	0	1	2	3	4	5
Frekvens	22	18	7	2	0	1

a) **Rita upp ett stolpdiagram** och **beräkna** medelvärde $\bar{x}$ och standardavvikelse s för observationsserien.

b) Betrakta följande funktion av de icke-negativa heltalen

$$p_x = e^{-\lambda}\frac{\lambda^x}{x!}, \quad x = 0, 1, \ldots.$$

Denna funktion kallas sannolikhetsfunktionen för **en Poissonfördelning med parametern** λ. Här är $\lambda > 0$.

Beräkna värdet på funktionen $e^{-\bar{x}}\frac{\bar{x}^k}{k!}$ för $k = 0, 1, 2, 3, 4, 5$, där $\bar{x}$ är medelvärdet från den här uppgiftens del **a)**, med de relativa frekvenserna för antalet fläckar.

I **b)** har vi gjort en s.k. **statistisk skattning** av sannolikhetsfunktionen för den stokastiska variabeln X = antalet fläckar i ett prov. Om denna skattning verkar bra, så heter det i det statistiska fackspråket att mätvärdena är observationer på en Poissonfördelad variabel. För Poissonfördelade data borde värdena på $\bar{x}$ och s^2 vara rätt nära varandra.

Ordningsstatistika (Order statistic)
För en datamängd $\{x_1, \ldots, x_n\}$ är $x_{(k)}$ dess **k:te ordningsstatistika** om

$$x_{(k)} = \text{ k:te minsta värdet i } \{x_1, \ldots, x_n\}$$

Härmed är $x_{(1)}$ det minsta värdet i $\{x_1, \ldots, x_n\}$ och $x_{(n)}$ det största värdet i $\{x_1, \ldots, x_n\}$. Med andra ord, har vi den **ordnade datamängden**

$$\{x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n-1)} \leq x_{(n)}\}$$

Till exempel, om vi har $\{x_1 = 53, x_2 = 41, x_3 = 6.5, x_4 = 65, x_{(5)} = 19\}$ så är ordningsstatistikorna

$$\{x_{(1)} = 6.5, x_{(2)} = 19, x_{(3)} = 41, x_{(4)} = 53, x_{(5)} = 65\}$$

Problem 2.2.8.

För 22 personer beräknades kariesindex, d.v.s. antalet kariesangripna ytor bland de 100 tandytor man får om man bortser från visdomständerna samt den linguala ytan (= ytan mot tungan) på alla tänder. Resultat:

41	47	66	73	48	52	49	54	61	62	47
52	65	61	69	31	54	53	50	47	36	69

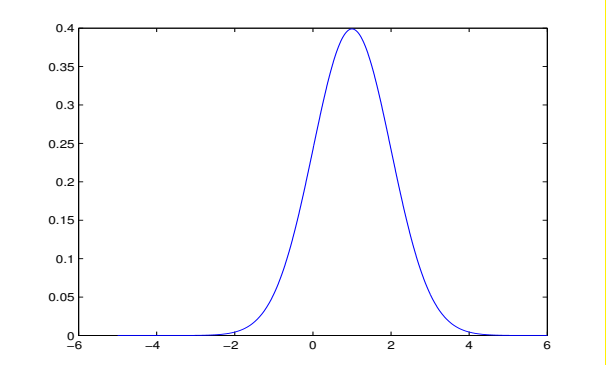
a) Beräkna ordningsstatistikorna.

b) Bestäm variationsbredd och median utifrån ordningsstatistikorna.

2.3 Normalkurvan eller klockkurvan (bell curve)

Låt $\sigma > 0$ och betrakta (den s.k.) klockkurvan eller normalkurvan

$$f(x) = \frac{1}{\sqrt{2\sigma^2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$



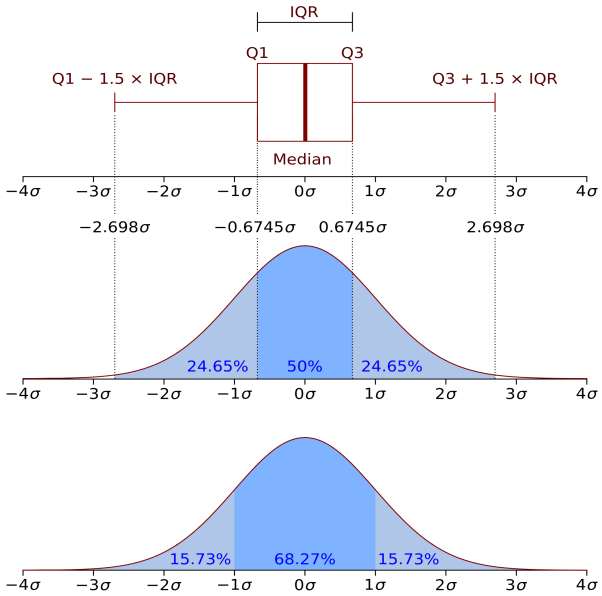
$f(x)$ med $\mu = \sigma = 1$ i figuren.
Benämningen klockkurva följer från funktionskurvans utseende, vilket har en viss likhet med en kyrkklocka.

Det gäller (se, t.ex. Robert A. Adams, Christopher Essex: Calculus - A Complete Course, 7th Ed. Section 7.8) att arean under kurvan är lika med ett, d.v.s.

$$\int_{-\infty}^{+\infty} f(x)dx = 1.$$

Problem 2.3.1.

I figuren för denna uppgift har vi klockkurvan med $\mu = 0, \sigma > 0$ och en boxplot. Vi ser t.ex. att arean under klockkurvan mellan $-\sigma$ och $+\sigma$ är 68.27 % (av totalarean =1).



Jämför boxplotten och klockkurvan ($\mu = 0, \sigma > 0$) med varandra. **Vad** kan Du säga om klockkurvan på basis av denna jämförelse?

Genom att derivera $f(x)$ en gång fås att

$$f'(x) = \frac{-(x-\mu)f(x)}{\sigma^2}.$$

Genom att sätta $f'(x) = 0$ inses att det största värdet för $f(x)$ är

$$f(\mu) = \frac{1}{\sqrt{2\sigma^2\pi}}.$$

Således är μ typvärdet för denna kurva. En derivering till med produktregeln och en hyfsning ger

$$f''(x) = -\frac{f(x)}{\sigma^2} + \frac{(x-\mu)^2 f(x)}{\sigma^4}.$$

En punkt x_o kallas en **inflektionspunkt** (se, t.ex. Robert A. Adams, Christopher Essex: Calculus - A Complete Course, 7th Ed. sid. 240), för kurvan $f(x)$, om

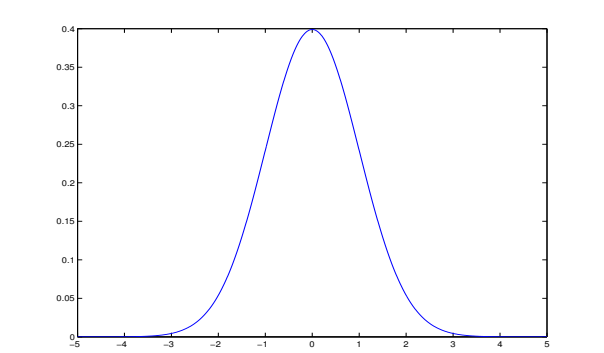
$$f''(x_o) = 0.$$

Problem 2.3.2.

Bestäm de två inflektionspunkterna till klockkurvan

$$f(x) = \frac{1}{\sqrt{2\sigma^2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

När $\mu = 0$ och $\sigma = 1$ som i figuren nedan,



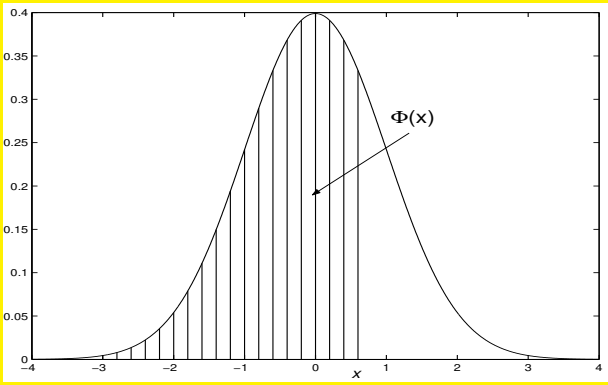
talar vi om den standardiserade klockkurvan som har den egna beteckningen

$$\varphi(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}.$$

Vi ser att kurvan är symmetrisk kring origo, $\varphi(x) = \varphi(-x)$. Vi sätter

$$\Phi(x) \stackrel{\text{def}}{=} \int_{-\infty}^{x} \varphi(t)dt.$$

Adams och Essex: Calculus - A Complete Course har redan lärt ut att den primitiva funktionen $\Phi(x)$ inte kan bestämmas annat än medelst mjukvara eller tabeller.



Det fås

$$\int_a^b \varphi(t)dt = \Phi(b) - \Phi(a).$$

Eftersom totalarean $= 1$ och är $\varphi(x)$ är symmetrisk kring origo, gäller därtill att

$$\Phi(-x) = 1 - \Phi(x),$$

d.v.s.

$$\int_{-b}^b \varphi(t)dt = \Phi(b) - \Phi(-b) = 2\Phi(b) - 1$$

Problem 2.3.3.

Medianen för den standardiserade normalkurvan är ett värde x_M sådant att $\Phi(x_M) = 0.5$.

Bestäm x_M. *Ledning:* Kolla vad som händer i $\Phi(-x) = 1 - \Phi(x)$ om Du sätter $x = 0$.

Problem 2.3.4.

$$f(x) = \frac{1}{\sqrt{2\sigma^2\pi}}\, e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

a) Verifiera att

$$\int_a^b f(x)dx = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right).$$

(Ledning: variabelbyte $z = \frac{x-\mu}{\sigma}$ i integralen i vänstra ledet.)

b) Vad är medianen för $f(x)$?

c) Enligt **a)** gäller alltså att

$$\int_{\mu-\sigma}^{\mu+\sigma} f(x)dx = \Phi(1) - \Phi(-1).$$

I tabellen nedan ges några värden av $\Phi(x)$ för $x \geq 0$. (T.ex. $\Phi(0.51) = 0.6950$).

x=0.0	0.00	0.01	0.02	0.03	0.04
0.0	0.5000	0.5040	0.5080	0.5120	0.5160
0.5	0.6915	0.6950	0.6985	0.7019	0.7054
1.0	0.8413	0.8438	0.8461	0.8485	0.8508
1.5	0.9332	0.9345	0.9357	0.9370	0.9382
2.0	0.9772	0.9778	0.9783	0.9788	0.9793
2.5	0.9938	0.9940	0.9941	0.9943	0.9945
3.0	0.9987	0.9987	0.9987	0.9988	0.9988

Dessa värden ges i Matlab av funktionen normcdf, t.ex.

```
>> normcdf(0.51,0,1)

ans =

      0.6950
```

Verifiera att

$$\int_{\mu-\sigma}^{\mu+\sigma} f(x)dx = 0.6827.$$

Problem 2.3.5.

a) Verifiera att

$$\int_{\mu-2\sigma}^{\mu+2\sigma} f(x)dx = 0.95.$$

b) Verifiera att

$$\int_{\mu-3\sigma}^{\mu+3\sigma} f(x)dx = 0.997.$$

Detta resultat kallas **three-sigma rule of thumb** (of empirical science). På engelska heter det att: *'nearly all"values are taken to lie within three standard deviations of the mean.*

Problem 2.3.6.

Sammanfatta rönen (observationerna) i de föregående uppgifterna i detta avsnitt och jämför med de motsvarande utsagorna i Tjebysovs regel.

Problem 2.3.7.

En enkel kontroll av, om det är rimligt med data-analys med hjälp klockkurvan, är att räkna, huruvida 0.68 % (ungefär 2/3) av datamängden ligger mellan $\bar{x} - s$ och $\bar{x} + s$, jfr. det ovanstående.

Nedan har vi **grupperade** data, som härstammar från en klassisk studie av Adam Quetelet, och ger frekvenserna på omfånget av bröstkorg i cm för 1516 belgiska soldater.

Klass	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42
Frekvens	2	4	17	55	102	180	242	310	251	181	103	42	19	6	2

a) **Kontrollera** huruvida 0.68 % (2/3) av datamängden ligger mellan $\bar{x} - s$ och $\bar{x} + s$.

b) **Vad** avses med en normalkurva, som beskriver detta data ?

c) Hur skulle man kunna gå tillväga för att hitta den normalkurva, som beskriver detta data (bäst)? (Ledning: det rätta svaret torde framgå senast i avsnittet om punkt- och intervallskattningar nedan).

För vidare perspektiv kring detta dataset, se även
`http://www.theatlantic.com/business/archive/2016/02/the-invention-of-the-normal-person/463365/`

Problem 2.3.8.

Q-Q -kurva (quantile-to- quantile plot)

Q-Q -kurva (Q står för quantile), `qqplot` i Matlab, är en grafisk metod avsedd för att jämföra data med en fördelning eller för att jämföra två olika datamänger. Detta utförs medelst en linje, som sammanbinder $(0, 0)$ och $(1, 1)$ genom att plotta fördelningars kvantiler mot varandra. Först väljes en svit av intervall (detaljerna utelämnas). En punkt (x, y) i plotten svarar mot en kvantilen i den andra fördelningen (y -koordinaten) mot samma kvantil i den första fördelningen (x -koordinaten). Därmed är Q-Q -kurvan en s.k. parametrisk kurva, där parametern är kvantilens nummer. Om de två sannolikhetsfördelningarna är likadana, kommer Q-Q -kurvan att approximativt ligga på diagonalen $y = x$.

Från tidigare: Betrakta igen datamängden x nedan.

30 171 184 201 212 250 265 270 272 289
305 306 322 322 336 346 351 370 390 404
409 411 436 437 439 441 444 448 451 453 470
480 482 487 494 495 499 503 514 521 522 527
548 550 559 560 570 572 574 578 585 592 592
607 616 618 621 629 637 638 640 656 668 707
709 719 737 739 752 758 766 792 792 794 802
818 830 832 843 858 860 869 918 925 953 991
1000 1005 1068 1441

Helpfilen i Matlab säger följande:

> `qqplot`(x) makes an empirical QQ-plot of the quantiles of the data in the vector x versus the quantiles of a standard Normal distribution (= standardiserad normalkurva).

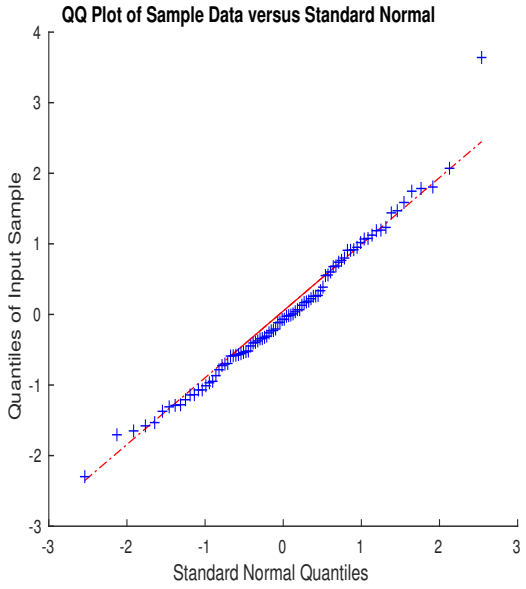
Vi omvandlar dessa data först till motsvarande z-scores

```
>>  z=(x-mean(x))./std(x);
```

och kör sedan

```
>>  qqplot}(z)
```

Resultatet återges figuren.



Vad är din slutsats?

Kapitel 3

Sannolikhet

3.1 Kombinatoriska definitioner & formler & klassisk sannolikhet

I klassisk sannolikhet bestämmer man sannolikheter genom att räkna a priori antalet gynnsamma fall och dividera med totalantalet fall i utfallsrummet. Därvid behövs ofta kombinatoriska formler och resonemang.

3.1.1 Multiplikationsprincip

Om åtgärd 1 kan utföras på a_1 sätt och åtgärd 2 på a_2 sätt, så finns det $a_1 a_2$ sätt att utföra båda åtgärderna. Generalisering till tre åtgärder blir $a_1 a_2 a_3$ o.s.v..
Vi har en mängd av n element

$$\mathcal{S} = \{e_1, e_2, \ldots, e_n\}.$$

Ur $\mathcal{S}$ tar vi ut k element. Vi vill undersöka på hur många sätt detta kan göras om olika givna villkor skall vara uppfyllda.
Följande beteckningar användes, $n = heltal \geq 0$:

$$n! = n(n-1)(n-2)\cdots 2 \cdot 1, \qquad \text{(uttalas: } n\text{-fakultet)}$$

$$\binom{a}{b} = \frac{a!}{b!(a-b)!} = \frac{a(a-1)\cdots(a-b+1)}{b!} \qquad \text{(uttalas: } a \text{ över } b\text{).}$$

Enligt konvention sätts

$$0! = 1,$$

vilket ger $\binom{a}{0} = \binom{a}{a} = 1$.

Exempel 3.1.1.

$$5! = 5 \cdot 4 \cdot 3 \cdot 2 \cdot 1 = 120.$$

3.1.2 Dragning med återläggning.

På hur många sätt kan man med återläggning dra k element ur mängden $\mathcal{S}$? Om $n = 3$ och $k = 2$ blir antalet sätt 9, nämligen

$$e_1e_1,\ e_1e_2,\ e_1e_3,\ e_2e_1,\ e_2e_2,\ e_2e_3,\ e_3e_1,\ e_3e_2,\ e_3e_3.$$

Lägg märke till att elementen anges i den ordning i vilken de dragits.
Allmänt blir antalet sätt lika med n^k. Det finns nämligen n möjliga element på första plats, n på andra plats o.s.v., och dessa kan kombineras med varandra hur som helst. Enligt multiplikationsprincipen ger detta

$$\underbrace{n \cdot n \cdots n}_{k \text{ st}} = n^k$$

olika sätt. Dragning med återläggning av k element ur n med hänsyn till ordning kan ske på n^k olika sätt.

Problem 3.1.1.

Hur många olika DNA sekvenser med 8 nukleotider finns det?

Problem 3.1.2.

I den genetiska koden svarar en viss följd av tre nukleotidbaser mot en viss aminosyra. En sådan följd av tre baser kallas för ett **kodon**.

Hur många olika följder av tre baser finns det?

Problem 3.1.3.

Tre av kodon är stoppkodon och avbryter proteinsyntesen i cellen.

Plocka på måfå ett kodon. Vad är sannolikheten att Du får ett stoppkodon?

3.1.3 Dragning utan återläggning av k element ur n.

På hur många sätt kan man utan återläggning dra k element ur mängden $\mathcal{S}$? Om $n = 3$ och $k = 2$ blir antalet sätt lika med 6, nämligen

$$e_1e_2,\ e_1e_3,\ e_2e_1,\ e_2e_3,\ e_3e_1,\ e_3e_2.$$

Allmänt kan man bestämma antalet så här: Först tar man ut ett element och sätter det på första plats, vilket kan ske på n sätt, sedan ett element bland de $n - 1$ återstående och sätter det på andra plats o.s.v.; till sist återstå r $n - k + 1$ element att välja mellan på k:te plats. Totala antalet sätt blir produkten av dessa tal, d.v.s.

$$n(n-1)\cdots(n-k+1).$$

Dragning utan återläggning av k element ur n (med hänsyn till ordning) kan alltså ske på $n(n-1)\cdots(n-k+1)$ olika sätt.
Antalet sätt brukar i detta fall kallas antalet *permutationer av k element bland n*.
Antalet permutationer av n element bland n är lika med $n(n-1)\cdots 2 \cdot 1 = n!$ eller kortare: n element kan ordnas på $n!$ olika sätt.

En viktig formel, där n-fakultet ingår, är serieutvecklingen

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}.$$

d.v.s.

$$e = \sum_{n=0}^{\infty} \frac{1}{n!}.$$

3.1.4 Dragning utan återläggning av k element ur n (utan hänsyn till ordning)

Låt oss se på specialfallet $n = 3,\ k = 2$. Eftersom vi inte skiljer på e_1e_2 och e_2e_1 etc blir det nu 3 olika delmängder, nämligen

$$e_1e_2,\ e_1e_3,\ e_2e_3.$$

Antalet fall minskar från 6 till 3, d.v.s. divideras med 2. Allmänt kan k uttagna element ordnas på $k!$ olika sätt d.v.s.

$$\frac{n(n-1)\cdots(n-k+1)}{k!} = \frac{n!}{(n-k)!k!} = \binom{n}{n-k} = \binom{n}{k}.$$

Antalet sätt brukar i detta fall kallas antalet *kombinationer av k element bland n*. *Binomialkoefficienten* $\binom{n}{k}$ kan tolkas som antalet olika (oordnade) delmängder av storlek k ur en mängd av n olika element.

Binomialsats: För varje positivt heltal n och för godtyckliga tal x och y gäller

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}.$$

En viktig tillämpning av ovanstående: $0 < p < 1$,

$$\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = (p + (1-p))^n = 1. \tag{3.1}$$

Problem 3.1.4.

A DNA primer used in the polymerase chain reaction (PCR) is a one-strand DNA fragment designed to bind (to hybridize) to one of the strands of the target DNA molecule. It was observed that primers can hybridize not only to their perfect complements, but also to DNA fragments of the same length having one or two mismatching nucleotides. We assume that the genomic DNA is so long that all possible oligonucleotides of length eight are present in the target DNA.

How many different DNA sequences may bind to an eight nucleotide long primer?

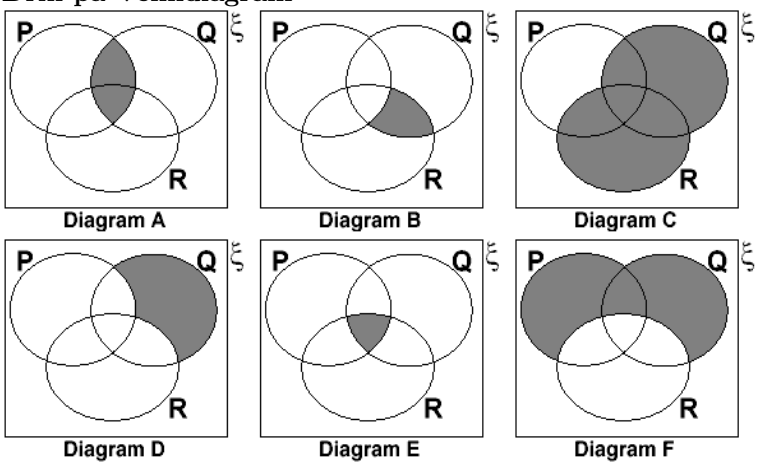
3.2 Utfallsrum, Händelser, Sannolikheter

[Engelsk-svensk Ordlista]

- Outcome= utfall
- Event = händelse
- Probability = sannolikhet

Problem 3.2.1.

Drill på Venndiagram



ξ i diagrammet svarar mot Ω.

a) I vilket diagram svarar den skuggade ytan mot $Q \cup R$?

b) I vilket diagram svarar den skuggade ytan mot $P \cap Q$?

c) I vilket diagram svarar den skuggade ytan mot $P \cap Q \cap R$?

d) Mot vilken händelse uttryckt med hjälp av P, Q, R svarar den skuggade ytan i diagram F?

Problem 3.2.2.

Rita i ett koordinatsystem mängden $\{(x, y) := 1, 2, \ldots, 6, y = 1, 2, \ldots, 6\}$. Den kan uppfattas som **utfallsrum** (=mängden av alla tänkbara utfall, på engelska **outcome space**), vid kast med ett par av tärningar en gång. Alla utfall är lika sannolika d.v.s. $P(x, y) = 1/36$.

Beräkna sannolikheten för följande händelser.

a) Poängsumman mindre än sex.

b) Samma poäng vid båda kasten.

c) Åtminstone ett av kasten ger precis två poäng.

d) Åtminstone ett av kasten ger minst fem poäng.

Problem 3.2.3.

För de två händelserna A och B gäller att $P(A \cap B) = 0.4$, $P(A) = 0.5$ och $P(A^c \cap B) = 0.1$.

Beräkna $P(A \cup B)$.

Problem 3.2.4.

Visa att sannolikheten för att exakt en av händelserna A och B inträffar är $P(A) + P(B) - 2P(A \cap B)$.

Problem 3.2.5.

För händelserna A och B gäller att $P(A \cup B^c) = 7/8$ och $P(A \cap B) = 1/9$.

a) Bestäm $P(B)$.

b) Bestäm $P(A \cup B)$ om $P(A) = 1/3$.

Problem 3.2.6.

Bestäm sannolikheten för att av 23 personer minst två har födelsedag på samma dag. Antag att året har 365 dagar och att alla födelsedagskombinationer är lika sannolika.

Ledning: Betrakta den komplementära händelsen.

3.3 Empirisk sannolikhet

Relativ frekvens för en datamängd
$\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ med n datapunkter i något Ω (alla datatyper tillåtna). x är ett värde i Ω. Låt $n(x)$ =antalet gånger x förekommer i $\mathcal{X}$. Då är $f_{\mathcal{X}}(x)$, den relativa frekvensen av x (m.a.p. $\mathcal{X}$), lika med

$$f_{\mathcal{X}}(x) = \frac{n(x)}{n},$$

Denna definition gäller för alla skalor/datatyper. Observera att $f_{\mathcal{X}}(x) = 0$ om x inte återfinns bland $\mathcal{X}$. Om $A \subseteq \Omega$, så är

$$P_{\mathcal{X}}(A) = \sum_{\text{de olika}} \; \sum_{x \in A} f_{\mathcal{X}}(x)$$

den relativa frekvensen av A eller den empiriska sannolikheten för A (m.a.p. $\mathcal{X}$).

Problem 3.3.1.

Ω = de positiva heltalen. $\mathcal{X} = \{7 \quad 5 \quad 5 \quad 9 \quad 6 \quad 6 \quad 8 \quad 8 \quad 6 \quad 2 \quad 3 \quad 8\}$.

a) Beräkna den relativa frekvensen av $A = \{2 \quad 5 \quad 8\}$.

b) Beräkna den relativa frekvensen av $B = \{6 \quad 7 \quad 10\}$.

c) Beräkna den relativa frekvensen av $A \cup B = \{2 \quad 5 \quad 6 \quad 7 \quad 8 \quad 10\}$.

Problem 3.3.2.

Relativ frekvens som sannolikhet

$A \subseteq$ och $P_{\mathcal{X}}(A) = \sum_{x \in A} f_{\mathcal{X}}(x)$, där $f_{\mathcal{X}}(x)$ beräknats utifrån $\mathcal{X}$. **Varför** är de följande utsagorna sanna:

a) $0 \leq P_{\mathcal{X}}(A) \leq 1$.

b) $P_{\mathcal{X}}(\mathcal{X}) = 1$.

c) Om A och B inte innehåller gemensamma värden, så är

$$P_{\mathcal{X}}(A \cup B) = P_{\mathcal{X}}(A) + P_{\mathcal{X}}(B).$$

Empirisk fördelningsfunktion
$f_X(x)$ utgör de relativa frekvenserna för $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$, som är kvantitativa data. Då är

$$F_{\mathcal{X}}(x) \stackrel{\text{def}}{=} \sum_{x_i \le x} f_X(x_i), \quad -\infty < x < +\infty,$$

den empiriska fördelningsfunktionen (för $\mathcal{X}$). Det gäller alltså att

$$F_{\mathcal{X}}(x) = P_{\mathcal{X}}(A), \quad \text{där } A =]-\infty, x].$$

Problem 3.3.3.

Beräkna och **rita upp** den empiriska fördelningsfunktionen $F_{\mathcal{X}}(x)$ för $\mathcal{X} = \{7 \quad 5 \quad 5 \quad 9 \quad 6 \quad 6 \quad 8 \quad 8 \quad 6 \quad 2 \quad 3 \quad 8\}$.

Problem 3.3.4.

Empirisk fördelningsfunktion

Varför är de följande utsagorna sanna:

a) $0 \le F_{\mathcal{X}}(x) \le 1$.

b) om $x_1 \le x_2 \le \cdots \le x_n$, så gäller

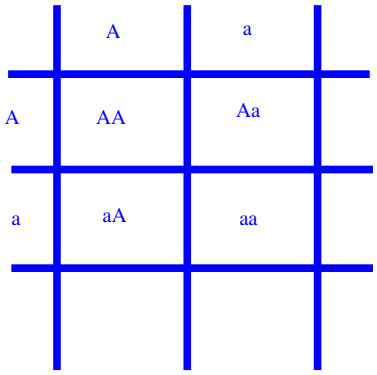
$$F_{\mathcal{X}}(x_j) - F_{\mathcal{X}}(x_{j-1}) = f_X(x_j).$$

c) om $x \le y$, så är $F_{\mathcal{X}}(x) \le F_{\mathcal{X}}(y)$.

3.4 Genetik

Problem 3.4.1.

Mendelska karaktärer (på engelska trait) styrs av ett enda locus, där genen kan ha två former, som kallas alleler.
Vi tar ett allmängiltigt exempel med blommor. Bokstaven A representerar den dominerande allelen med röda blommor och a representerar dem recessiva allelen med vita blommor. I första korsningen har de korsade plantorna, föräldrarna, två alleler AA om de är röda blommor och två alleler aa om de är vita blommor. Hybriderna i den första generationen ärver därför en allel för röda blommor och en för vita. Med genotypen Aa har de allesammans fenotypen röda blommor, ty A är dominerande i heterozygoterna, och den recessiva a allelen har ingen verkan. Den nästa generationen fås med korsningar av genotyperna Aa × Aa. Kvoten mellan plantor med röda blommor mot plantor med vita blommor är 3:1 och kallas den fenotypiska kvoten (=kvoten mellan dominanta and recessiva fenotyper). Den genotypiska kvoten framgår av s.k. **Punnetts rutschema** och är 1 AA : 2 Aa : 1 aa i figuren



Vi har alltså sannolikheterna P(röd blomma i andra generationen) $= \frac{3}{4}$ och P(vit blomma i andra generationen) $= \frac{1}{4}$.

Vad är tolkningen av sannolikhet i detta fall? Jämför även : Vidakovic sid. 61.

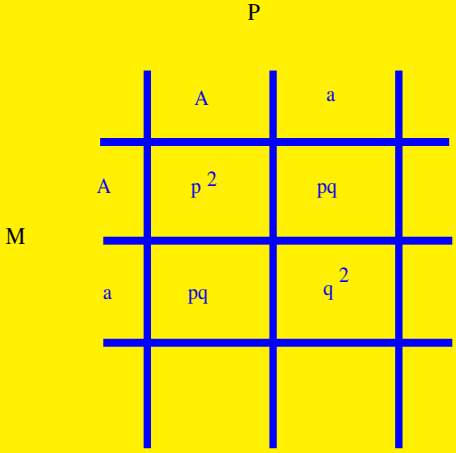
Hardy-Weinbergs ekvation
Vi betraktar en diploid population, där varje organism producerar manliga (P) och kvinnliga (M) gameter med samma sannolikhet, och har två alleler, A och a, på ett locus.
Sannolikheten för allel A är p, sannolikheten för allel a är q hos både P och M. Om vi använder korsningen

$$Aa \times Aa$$

av föräldrar, kommer hos avkomman p^2 att vara den relativa frekvensen för genotyp AA, pq är relativa frekvensen för genotyp Aa, qp är relativa frekvensen för genotyp aA, och relativa frekvensen för genotyp aa är q^2. Dessutom gäller **Hardy-Weinbergs ekvation**

$$p^2 + 2pq + q^2 = 1.$$

Detta visas i Punnetts rutschema i figuren. Genotyperna AA och aa kallas homozygota, och genotyperna aA och Aa kallas heterozygota.



Hardy-Weinbergs ekvation beskriver frekvensen (fördelningen) av de olika genotyperna under vissa förhållanden:

- en mycket stor population
- inga mutationer
- ingen migration in i eller ut ur populationen
- ingen selektion bland genotyperna
- random mating.

Hardy-Weinbergs ekvation är i själva verket **Hardy-Weinbergs jämvikt**: under förutsättningarna ovan kommer en population ha dessa samma frekvenser för alleler och genotyper i alla generationer.

Problem 3.4.2.

Den autosomalt recessiva ögonsjukdomen progressiv retinal atrofi (PRA) finns hos hundar. Om man har två kopior av den defekta allelen blir man sjuk, men anlagsbärare, som bara har en kopia av allelen är friska. PRA diagnosticeras genom en ögonundersökning och det har visat sig för rasen papillon att 1% av hundarna är sjuka.

År 2014 infördes ett DNA-test vilket möjliggör att identifiera anlagsbärarna innan avel och därmed med säkerhet undvika sjuka hundar. Låt oss anta att Hardy-Weinbergs ekvation gäller.

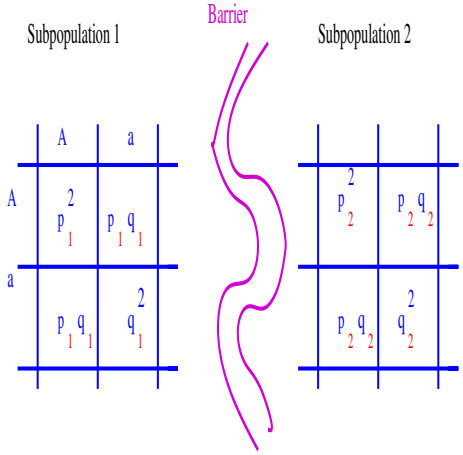
a) Vad är frekvensen av allelen för PRA?

b) Hur många anlagsbärare kan vi förvänta oss att hitta med DNA-testen?

Problem 3.4.3.

Subpopulations: The Wahlund Effect on Hardy-Weinberg equilibrium and The Wahlund Principle

Let us consider a metapopulation consisting of two reproductively separated subpopulations. To make this concrete, think of a geographical barrier to gene flow between two subgroups of the same species, followed by independent genetic drift in each subgroup. Suppose that subpopulation 1 satisfies internally Hardy - Weinberg with q_1^2 as the relative frequency of the homozygous recessive genotype. The subpopulation 2 satisfies internally Hardy - Weinberg with q_2^2 as the relative frequency of the homozygous recessive genotype.

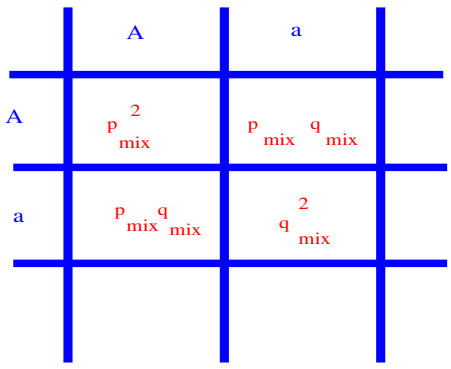


The average frequency of the homozygous recessive genotype aa in the metapopulation (ie., across the two separate subpopulations) is denoted by q_{sep} and equals

$$q_{sep} = \frac{q_1^2 + q_2^2}{2}. \tag{3.2}$$

Let us consider intermixing the two populations. This means that the barrier is removed and the new population intermingles by random mating across the whole mixed pool of individuals.

Mixed population



In the intermixed population the frequency of the allele a is

$$q_{mix} = \frac{q_1 + q_2}{2}.$$

Thus the frequency of the genotype aa in the mixed population is by Hardy-Weinberg equal to

$$q_{mix}^2 = \left(\frac{q_1 + q_2}{2}\right)^2.$$

a) Define the Wahlund variance

$$\sigma_q^2 \stackrel{\text{def}}{=} \frac{(q_1 - q_{mix})^2}{2} + \frac{(q_2 - q_{mix})^2}{2}. \tag{3.3}$$

Show (by expanding σ_q^2 in the left hand side) that

$$\sigma_q^2 = q_{sep} - q_{mix}^2.$$

Then it follows (why?) that

$$q_{sep} \geq q_{mix}^2. \tag{3.4}$$

b) The inequality in (3.4) is called the **The Wahlund Effect** for recessive (aa) homozygosity. How would you express the Wahlund Effect for homozygosity in words?

c) What happens to the incidence of rare recessive genetic diseases, when a previously isolated population comes into contact with a larger population?

d) What precaution does (3.4) suggest when applying the Hardy-Weinberg equilibrium principle to a population?

e) The Wahlund Effect applies also to AA homozygotes. In other words, set

$$p_{sep} = \frac{p_1^2 + p_2^2}{2}, p_{mix} = \frac{p_1 + p_2}{2}, p_{mix}^2 = \left(\frac{p_1 + p_2}{2}\right)^2,$$

(note that $p_{mix} + q_{mix} = 1$) and

$$\sigma_p^2 \stackrel{\text{def}}{=} \frac{(p_1 - p_{mix})^2}{2} + \frac{(p_2 - p_{mix})^2}{2}. \tag{3.5}$$

Then, as above, we get

$$\sigma_p^2 = p_{sep} - p_{mix}^2, \tag{3.6}$$

and the Wahlund principle holds for AA homozygotes. Check now that

$$\sigma_p^2 = \sigma_q^2. \tag{3.7}$$

Hence we introduce the symbol

$$\sigma_W^2 \stackrel{\text{def}}{=} \sigma_p^2 = \sigma_q^2.$$

f) Compute the total reduction of homozygosity due to mixing as

$$2 \cdot \sigma_W^2.$$

g) Who was Wahlund? (Hint: Search in Wikipedia)

h) Check that

$$p_{mix} \cdot q_{mix} = p_{mix} - p_{mix}^2. \tag{3.8}$$

i) We compute the total increase in heterozygosity. The average frequency of the heterozygous genotypes Aa and aA across the two separate subpopulations, denoted by $E(Aa)$, equals

$$E(Aa) = \frac{1}{2}\left(2p_1q_1 + 2p_2q_2\right).$$

Check that

$$E(Aa) = 2\left(\frac{p_1 + p_2}{2} - p_{sep}\right)$$

Then use (3.6), part **e)** and (3.8) to get

$$E(Aa) = 2p_{mix} \cdot q_{mix} - 2\sigma_W^2. \tag{3.9}$$

By Hardy-Weinberger (Punnett diagram above), $2p_{mix} \cdot q_{mix}$ is the expected frequency of the heterozygotes in the mixed population. Hence (3.9) shows that

$$2p_{mix} \cdot q_{mix} - E(Aa) = 2\sigma_W^2 \tag{3.10}$$

or

$$2p_{mix} \cdot q_{mix} \geq E(Aa).$$

Hence expected heterozygosity increases when mixing the populations.

Or, homozygotes lose $2\sigma_W^2$ by **f)** above, and this is balanced in (3.10) by the heterozygotic gain $2\sigma_W^2$. This is the complete **The Wahlund Principle** in the most simplified case.

3.5 Oberoende, Betingad sannolikhet, Bayes sats

Problem 3.5.1.

I en fin kursbok i genetik hittar vi följande:

If A and B are independent events, the probability that A or B will happen is the probability of A plus the probability of B minus the probability of their joint occurrence ... and the addition rule (gives)

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

a) Hur skulle Du kommentera denna utsaga?

b) Vad händer i 'addition rule", om A och B är disjunkta?

Problem 3.5.2.

Logiska resonemang & sannolikhet Anta

$$P(B|A) > P(B).$$

Läs detta som att om A inträffat (är sann), blir B mer sannolik.

a) Verifiera under antagandet $P(B|A) > P(B)$ att

$$P(A \mid B) > P(A).$$

eller: om B inträffar, blir A mer sannolik.
Till exempel, om (A=) det regnar, blir det mer sannolikt att (B=) en ytbeläggning är blöt. Nu ser vi att ytbeläggningen är blöt (B), det är alltså mer sannolikt att det regnat (A).

b) Verifiera under antagandet $P(B|A) > P(B)$ att om A^c är sant, då blir B mindre sannolik.

c) Uttryck olikheten med hjälp av ord som ytbeläggning och regn.

Problem 3.5.3.

$0 < P(A) < 1$. Vilket av följande påståenden är *felaktigt* ?

a) $P(\Omega|A) = 1$.

b) $P(A|\Omega) = P(A)$.

c) $P(A^*|A) = 0$.

d) $P(A|A) = P(A)$.

Problem 3.5.4.

För två händelserna A och B gäller att $P(A) = 0.6$, $P(A \cap B) = 0.2$ och $P(A \cup B) = 0.8$.

Bestäm $P(A|B)$.

Problem 3.5.5.

Fathers, but not mothers, with high normal alleles, sometimes referred to as intermediate alleles or mutable normal alleles, are at risk of transmitting an Huntington's disease (HD) allele with reduced or full penetrance (≥ 36) to offspring. Thus, the father is at risk of having an offspring who eventually develops HD even though the father himself will not develop the disease.

Let us make the following assumptions. (De novo means here a mutation, an alteration in a gene that is present for the first time in one family member.)

1. All de novo HD cases arise from men with high normal alleles.
2. All de novo HD cases have HD.
3. The probability that a male has a high normal allele is independent from the probability that a male has an offspring.

a) Make use of the assumptions above and the definition of conditional probability to show the following:

$$P(\text{offspring is a de novo HD case} \mid \text{male with a high normal allele has an offspring})$$

$$= \frac{P(\text{offspring develops HD})P(\text{offspring is a de novo HD case }) | \text{offspring develops HD})}{P(\text{male has an offspring}) \cdot P(\text{male has a high normal allele})}.$$

b) Under which circumstances is the formula in **a)** useful?

Hendricks, Audrey E and Latourelle, Jeanne C and Lunetta, Kathryn L and Cupples, L Adrienne and Wheeler, Vanessa and MacDonald, Marcy E et.al.: *Estimating the probability of de novo HD cases from transmissions of expanded penetrant CAG alleles in the Huntington disease gene from male carriers of high normal alleles (27–35 CAG)*, **American Journal of Medical Genetics Part A**, 149, pp. 1375–1381, 2009.

> Audrey Hendricks et.al. find (loc.cit) using various relevant databases to estimate the pertinent probabilities that the highest probability that a male with a high normal repeat has an offspring who develops HD is 1 out of 951.

Problem 3.5.6.

I en läkemedelsstudie deltar 500 slumpmässigt valda individer. En individ kan avbryta sitt deltagande i studien och detta kan ske av två orsaker: (i) till följd av en allvarlig biverkning och (ii) till följd av andra orsaker. Antag att 3% av individerna får någon form av biverkning och att 25% av de som får en biverkning avbryter sin medverkan. Av de som inte får några biverkningar kommer 0.26% av deltagarna att avbryta sitt deltagande.

a) Beräkna sannolikheten att en individ avbryter studien.

b) Beräkna sannolikheten att en person som har avbrutit studien har en biverkning.

Problem 3.5.7.

A and B are two events. *Bayes' rule* is as follows:

$$P(A \mid B) = \frac{P(A)P(B \mid A)}{P(A)P(B \mid A) + P(A^c)P(B \mid A^c)}$$

Which of the following is a correct statement about $P(A \mid B)$?

a) It is the initial, prior probability, of A given B obtained before any additional information.

b) It is the likelihood of A given B.

c) It is the probability of A revised by using additional information obtained in B.

d) It is the prevalence of A given B.

Problem 3.5.8.

A genetic hybridization experiment involved peas with purple or white flowers and green or yellow pods as summarized in the table below.

	Purple	White
Green	5	3
Yellow	4	2

Which one of the following statements is wrong?

a) You select two peas at random without replacement. The probability of a green pod and a yellow pod, in this order, is (rounded off to three decimals) = 0.264.

b) Your colleague selects one pea at random, does not show it to you, but tells that its flower is purple. Then the probability that the selected pea has a green pod is $= \frac{5}{9}$.

c) You select two peas at random with replacement. The probability of at least one pea with white flower is $= \frac{40}{196}$

d) You select two peas at random with replacement. The probability of a pea with yellow pod in the second selection $= \frac{84}{196}$.

Problem 3.5.9.

The **odds for an event** A is a number denoted by Odds(A) and defined by

$$\text{Odds}(A) \stackrel{\text{def}}{=} \frac{P(A)}{P(A^c)}.$$

Below is shown a table of a retrospective study of discharge of 6776 newborn babies and their rehospitalization.

	Rehospitalized	Not rehospitalized
Early discharge	457	3199
Late discharge	260	2860

We are interested in the odds of A = 'rehospitalization of a baby who was early discharged". Which of the following statements is correct?

a) Odds(rehospitalization when early discharged) $= \frac{1}{8}$.

b) Odds(rehospitalization when early discharged) $= \frac{1}{7}$.

c) Odds(rehospitalization when early discharged) $= \frac{1}{9}$.

d) Odds(rehospitalization when early discharged) $= \frac{1}{10}$.

Problem 3.5.10.

En genetisk sjukdom har två kända riskgener A och B som finns på olika kromosomer (d.v.s. marginellt är händelserna {en person har variant A} och {en person har variant B} oberoende). Populationsfrekvenserna (prevalence) av A resp. B är 0.1 resp. 0.04. Den marginella risken att drabbas av sjukdomen om man har riskgen A är 0.02, och motsvarande för riskgen B är risken 0.04, om en person har båda riskgenerna är risken 0.10.

a) Vad är sannolikheten att drabbas av sjukdomen om man har riskgen A men inte B?

b) Vad är den marginella populationsrisken att drabbas av sjukdomen?

c) Visa att, givet att en person är sjuk så är händelserna att den sjuke har resp. riskgen inte är oberoende.

Problem 3.5.11.

För händelserna A, B och C gäller att $P(A \cap B \cap C) = 0.1$, $P(A) = 0.5$ och $P(B \mid A) = 0.4$.

Beräkna $P(C|A \cap B)$.

Problem 3.5.12.

A och B är två händelser sådana att $P(A) = 0.4$, $P(B) = 0.7$ och $P(A \cup B) = 0.8$.

a) Beräkna $P(A|B)$.

b) Avgör om A och B är oberoende.

Problem 3.5.13.

I en viss åldersgrupp vet man av tidigare epidemiologiska studier att 1% av kvinnorna har en ännu ej diagnostiserad cancertyp. Ett visst blodtest ger ett förhöjt värde av ett protein för 90% av sådana kvinnor. För kvinnor utan cancern är motsvarande siffra 10%. Man funderar på att erbjuda alla kvinnor i den aktuella åldersgruppen ett sådant test och vill därför som underlag beräkna

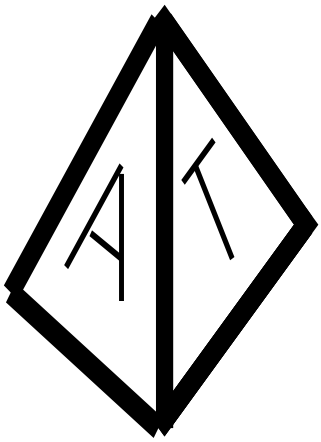
a) sannolikheten att en kvinna med förhöjt värde har cancern

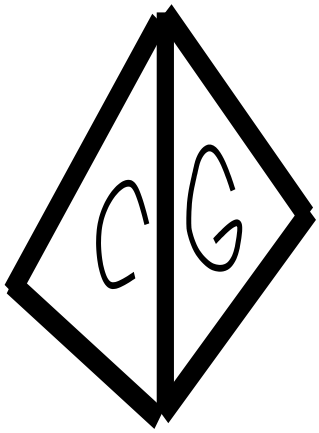
b) sannolikheten att en kvinna utan förhöjt värde har cancern.

Problem 3.5.14.

To each of the four sides of a tetrahedron there is assigned one of the letters A,T,C,G. All letters are used. We call this the **DNA die**. We toss the die in the air and note the side that it falls on. By symmetry we take

$$P(A) = P(T) = P(C) = P(G) = \frac{1}{4}$$





We assume that the tosses of the die are **independent**. Which one of the following statements is wrong?

a) The probability of getting the sequence AAT is $\frac{1}{64}$.

b) The probability of getting the sequence $AGGA$ is $\frac{1}{256}$.

c) The probability of at least one A in two tosses is $\frac{5}{16}$.

d) The probability of A in the second of two tosses is $\frac{1}{4}$.

Problem 3.5.15.

För genomet hos vissa mikrobstammar gäller att

$$P(A) = P(T) = 0.2, P(C) = P(G) = 0.3.$$

Vi är intresserade av **2-ord**, d.v.s. sekvenser av två nukleotider, t.ex. AA, AG o.s.v.. Vi antar att nukleotiderna (här även kallade bokstäverna) i ett 2-ord har valts oberoende av varandra.

a) Hur många olika 2-ord av nukleotider finns det?

b) Framställ sannolikheterna för samtliga 2-ord i en tabell. Är summan av sannolikheterna lika med ett?

c) Purinerna är $\{A, G\}$ och pyrimidinerna är $\{C, T\}$. Låt E vara händelsen att den första bokstaven är en pyrimidin. Låt F vara händelsen att den andra bokstaven är A, C eller T.

Bestäm sannolikheterna $P(E)$, $P(F)$, $P(E \cup F)$, $P(E \cap F)$.

e) Sätt $G = \{CA, CC\}$.

Bestäm sannolikheterna $P(G|E)$, $P(F|G \cup E)$, $P(F \cup G|E)$.

Problem 3.5.16.

The amino acid coding table is given below. The element found in the first (T) block in the fourth (G) column and second (C) row indicates that the DNA-triple TGC codes for the amino acid Cystine.

first base in codon	second base in codon: T	C	A	G	third base in codon
T	TTT Phe	TCT Ser	TAT Tyr	TGT Cys	T
	TTC Phe	TCC Ser	TAC Tyr	TGC Cys	C
	TTA Leu	TCA Ser	TAA stop	TGA stop	A
	TTG Leu	TCG Ser	TAG stop	TGG Trp	G
C	CTT Leu	CCT Pro	CAT His	CGT Arg	T
	CTC Leu	CCC Pro	CAC His	CGC Arg	C
	CTA Leu	CCA Pro	CAA Gln	CGA Arg	A
	CTG Leu	CCG Pro	CAG Gln	CGG Arg	G
A	ATT Ile	ACT Thr	AAT Asn	AGT Ser	T
	ATC Ile	ACC Thr	AAC Asn	AGC Ser	C
	ATA Ile	ACA Thr	AAA Lys	AGA Arg	A
	ATG Met	ACG Thr	AAG Lys	AGG Arg	G
G	GTT Val	GCT Ala	GAT Asp	GGT Gly	T
	GTC Val	GCC Ala	GAC Asp	GGC Gly	C
	GTA Val	GCA Ala	GAA Glu	GGA Gly	A
	GTG Val	GCG Ala	GAG Glu	GGG Gly	G

a) Suppose that a sequence of three independent letters is chosen by toss of the DNA die, see problem 3.5. What are the probabilities of this sequence coding for each of the possible amino acids? What is the probability that the sequence does not code for an amino acid?

b) Now suppose that an amino acid is chosen at random from the uniform distribution on the set of all 20 amino acids, and that then a DNA triplet is chosen uniformly at random from those triplets which code for this amino acid. What is the probability of getting T at position 1? At position 2? At position 3? At a randomly chosen position?

c) Sample as in part **b)**. What is the probability of getting TT at positions 1 and 2? At positions 2 and 3? Repeat for the string TG. What is the conditional probability of getting T at position 2, given that there is an T at position 1?

d) Sample as in part **b)**. Are the two events $A = \{$ T at position 1$\}$ and $B = \{$T at position 2 $\}$ independent ?

Problem 3.5.17.

Suppose now that the letters are drawn independently, but that their common distribution is C-G richer than the DNA die:

$$P(C) = P(G) = 0.275, P(A) = P(T) = 0.225.$$

a) What is the probability p_2 of getting GTTACA ? What is p_2/p_1, if p_1 is the probability of GTTACA using the DNA die ?

b) Is there any difference to the answers, if the sequence given is not GTTACA, but instead another specific sequence with A twice, C once, G once and T twice?

d) Suppose now that we have given a specific sequence of length 600, with 200 A$'$s, 100 C$'$s, 100 G$'$s and 200 T$'$s. If p_1 is the probability of getting this sequence with the DNA die, and p_2 is the probability of getting this sequence with the C-G richer distribution, what is p_2/p_1 ? Compare the value with the result of part a).

Problem 3.5.18.

Mängden A är dubbelt så sannolik som B. Hur förhåller sig $P(A \mid B)$ till $P(B|A)$?

Problem 3.5.19.

Två händelser A och B har sannolikheter skilda från noll.

a) A och B är disjunkta. Kan A och B vara oberoende?

b) A och B är oberoende. Kan A och B vara disjunkta?

Problem 3.5.20.

I en urna har man placerat tre kort, av vilka ett är rött på båda sidor, ett är vitt på båda sidor och ett är vitt på ena sidan och rött på den andra sidan. Man väljer på måfå ett av korten och tittar på den ena sidan.

a) Vad är sannolikheten att den andra sidan av kortet är röd betingat av att den betraktade sidan är det?

b) Var ligger felet i följande lösning: Om den betraktade sidan är röd så har vi fått antingen det röd-vita kortet eller det rödröda. Lika chans för båda. Sökta sannolikheten = 1/2.

Problem 3.5.21.

En tärning kastas upprepade gånger. Låt A_k vara händelsen att man inte får någon sexa i de k första kasten.

Visa att $P(A_k) = (5/6)^k, \qquad k = 1, 2, \ldots$ och att
$P(A_{k+n}|A_k) = P(A_n)$ där n och k är positiva heltal.

3.6 Kliniska prövningar och evidensbaserad biomedicinsk teknik

Bayes och evidensbaserad medicin
Den enklaste typen av sannolikhet kan illustreras med att en person väljs helt slumpmässigt ur en population. Sannolikheten att denna person skall ha en viss sjukdom är vad som kallas **prevalensen** för (förekomsten av) denna sjukdom i den aktuella populationen.
För det mesta rör det sig t.ex. i diagnostisk medicin mer om betingade sannolikheter, d.v.s. sannolikheter som påverkas av att speciella omständigheter, betingelser, föreligger. Det kan t ex vara så att det finns en viss basal (liten) sannolikhet för att en person med halsont och svullna halskörtlar har mononukleos, men om blodbilden ser ut på ett visst sätt är sannolikheten plötsligt mycket större för att den sjukdomen skall föreligga. Betingelsen här var just blodbilden.
Bayes sats ger en metod för att modifiera den första grundläggande sannolikheten för t ex att en viss sjukdom föreligger när nya betingelser kommer in i bilden. Betingelsen kan vara utfallet av en diagnostisk åtgärd, information från patienten eller kunskap om nya forskningsresultat. Uttryckt i tekniska termer innebär detta att en a priori-sannolikhet (prior probability) kan omräknas till en a posteriori-sannolikhet (posterior probability). Motsvarande termer används när sannolikhet byts ut mot odds.
A. Taube & J. Malmquist: Räkna med vad du tror. Bayes' sats i diagnostiken. *Läkartidningen* 2001, Vol. 98, Nr 24, sid. 2910−2913

Problem 3.6.1.

In 2007, the prevalence of men over the age of 50 to be diagnosed with prostate cancer (PC) in Sweden was 0.00549.

a) For a random sample of 1000 men 50 years or older, what is the probability of a PC diagnosis for no one/exactly one man/less than two men/two or more men?

b) What is the best guess for the number of men diagnosed with PC in this sample?

c) (Optional) What do we need to know in order to give a number based on the situation in Sweden (=estimate) this probability? Where does this information come from?

Problem 3.6.2.

Nasopharynxtumör (en tumörsjukdom i näshålan) är en ovanlig cancersjukdom i Sverige. Uppskattningsvis har ca 0.01% (=prevalens) av Sveriges befolkning idag en utvecklad nasopharynxtumör.

Antag att en viss diagnosmetod (baserad på blodanalys) med sannolikhet 0.999 diagnosticerar en patient som sjuk (i nasopharynx) givet att patienten i fråga verkligen har sjukdomen, samt med sannolikhet 0.005 diagnosticerar patienten som sjuk givet att patienten ej har sjukdomen.

Vad är den betingade sannolikheten att en slumpmässigt vald svensk lider av nasopharynxtumör givet att diagnosen indikerar detta?

Problem 3.6.3.

Enzyme immunoassay (EIA) tests are used to screen blood specimens for the presence of antibodies to HIV. Antibodies indicate the presence of the virus. The test is quite accurate but is not always correct.

	Test positive	Test negative
Antibodies present	0.9985	0.0015
Antibodies absent	0.0060	0.9940

Suppose that 1% (=prevalence) of a large population carries antibodies to HIV in their blood.

a) What is the probability that the test is positive for a randomly selected individual?

b) What is the probability that one individual have antibodies in his/her blood given that the test shows positive results.

Problem 3.6.4.

Bayes' rule and diagnostic tests
General terminology: $T+$ = positive test result in a diagnostic test, $T-$ = negative test result in a diagnostic test. $D+$ = has a disease, $D-$ = does not have a disease.

- **Sensitivity** $= P(T+ \mid D+)$
- **Specificity** $= P(T- \mid D-)$
- **True positive (predicted value)** $= P(D+ \mid T+)$
- **True negative (predicted value)** $= P(D- \mid T-)$
- **Overall accuracy** $= P(D+) \cdot \text{sensitivity} + P(D-) \cdot \text{specificity}$
- **False Positive** $P(\text{False positive}) = P(T+ \mid D-) = 1 - \text{specificity}$
- **False Negative** $P(\text{False negative}) = P(T- \mid D+) = 1 - \text{sensitivity}$

Bayes' rule tells amongst other things that

$$P(D+|T+) = \frac{P(T+ \mid D+)P(D+)}{P(T+)}$$

and

$$P(D-|T-) = \frac{P(T- \mid D-)P(D-)}{P(T-)}.$$

a) **Positive likelihood ratio** (LR_+) is defined as

$$\text{LR}_+ = \frac{P(T+ \mid D+)}{P(T+ \mid D-)}$$

Show that the posterior odds of $D+$ given $T+$ is

$$\frac{P(D+|T+)}{P(D-|T+)} = \text{LR}_+ \times \frac{P(D+)}{P(D-)}.$$

Here $\frac{P(D+)}{P(D-)}$ is the prior odds of $D+$.

Posterior odds = likelihood ratio × prior odds

b) Check that

$$\text{LR}_+ = \text{Sensitivity} \ / \ (1 - \text{Specificity})$$

c) **(Negative) likelihood ratio** (LR_-) is defined as

$$\text{LR}_- = \frac{P(T- \mid D+)}{P(T- \mid D-)} = \ (1 - \text{Sensitivity}) \ / \ \text{Specificity}$$

Check that

$$\frac{P(D+|T-)}{P(D-|T-)} = \text{LR}_- \times \frac{P(D+)}{P(D-)}.$$

d) Find that

$$P(\text{ True positive predicted value }) = \frac{P(\text{ D+ }) \cdot \text{ sensitivity}}{P(\text{ D+ }) \cdot \text{sensitivity } + P(\text{ D- })(1 - \text{specificity})}.$$

$$P(\text{ True negative predicted value }) = \frac{P(\text{ D- }) \cdot \text{ specificity}}{P(\text{ D- }) \cdot \text{specificity } + P(\text{ D+ })(1 - \text{sensitivity })}.$$

Problem 3.6.5.

Continuation of the preceding A new serum protein has been found that can potentially serve as a simple **biomarker** for the non-invasive diagnosis of prostate cancer. Based on a number of studies, it is estimated that this protein is present in 3.4% of all men over 50, and that it is

- present in 83% of all men diagnosed with PC (**sensitivity**),
- absent in 97% of all men not diagnosed with PC (**specificity**).

Would it make sense to use this protein for mass screening, e.g. by measuring the presence of the protein once yearly for all men over 50? In order to make a qualified judgement call or calculate

a) the probability of being diagnosed with PC, given that the protein has been found (**positive predictive value**),

b) the probability of not being diagnosed with PC given that no protein was found (**negative predictive value**).

Problem 3.6.6.

A medical research team wants to evaluate a proposed screening test for asthma, which is a chronic inflammatory disease of the airways (symptoms include wheezing, coughing, chest tightness, and shortness of breath). The test was given to a random sample of 100 patients with asthma (diagnosis based on the pattern of symptoms and response to therapy over time) and to an independent sample of 200 subjects without symptoms of the disease. All 300 were 50 years or older.

	Asthma $D+$	No asthma $D-$
Positive test $T+$	92	13
Negative test $T-$	8	187

One and only one of the following is a correct statement.

a) $P(\text{False negative}) = Pr(T- \mid D+) = 0.052$.

b) $P(T+|D+) = 0.934$.

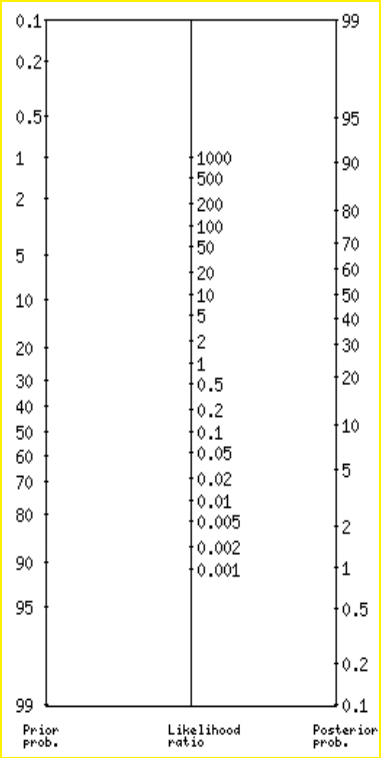
c) $P(T-|D-) = 0.935$.

d) $P(\text{False positive}) = P(T+ \mid D-) = 0.041$.

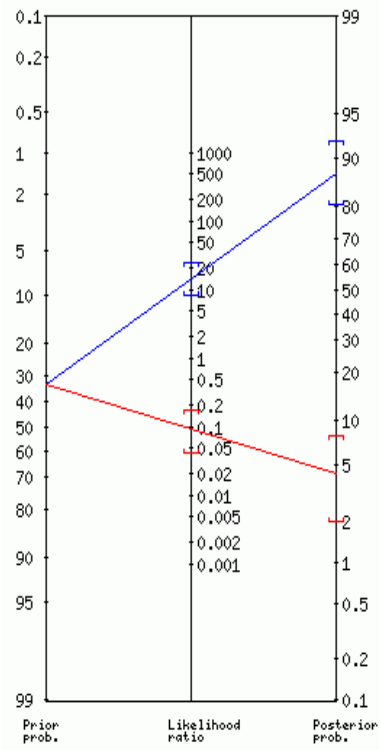
Problem 3.6.7.

There is a 10 % prevalence of being infected by a virus, here called $hiv-1$, in a certain at-risk population. A preliminary screening test for $hiv-1$ is correct in 95 % of all cases. One individual is randomly selected from the at-risk population. We want to find the sensitivity, or the probability, denoted by $P(D+|T+)$, that the selected individual is infected by $hiv-1$, if this individual has tested positive in the screening. We want to compute $P(D+|T+)$. We are going to use two different methods, the first based on the Fagan nomogram.

Fagan nomogram
The simplest of calculators of $P(D+|T+)$ is the Fagan nomogram. This was invented by T.J. Fagan, M.D.: *Letter: nomogram for Bayes theorem.*, **The New England journal of medicine**, 293, pp. 257–257, 1975. That was in the days before personal computers and handhelds. The nomogram has, according to written testimonies, an enduring popularity. The blank nomogram is shown in the Figure.



The Fagan nomogram is used as follows. Mark the prior (prevalence) $P(D+)$ in the leftmost vertical line with a dot (read in %). Mark the LR_+, i.e., $\frac{P(T+|D+)}{P(T-|D+)}$ in the middle vertical line with a dot. Join the two dots with a straight line, which, when extended to the rightmost vertical line, cuts it at $P(D+|T+)$ (read in %). An example is the blue line the same figure. The red line is for a negative test and finds $P(D+|T-)$ by the negative likelihood ratio LR_-.



The Fagan Nomogram is implemented in **Diagnostic Test Calculator** on
`http://araw.mede.uic.edu/cgi-bin/testcalc.pl`
where the desired diagram is computed and displayed after filling in the interactive fields.

a) Compute $P(D+|T+)$ by the Fagan diagram (i.e. with the diagnostic test calculator in the web address above).

b) Compute $P(D+|T+)$ by full algebra of Bayes formula.

Kapitel 4

Sannolikhetsmodeller för diskreta datatyper

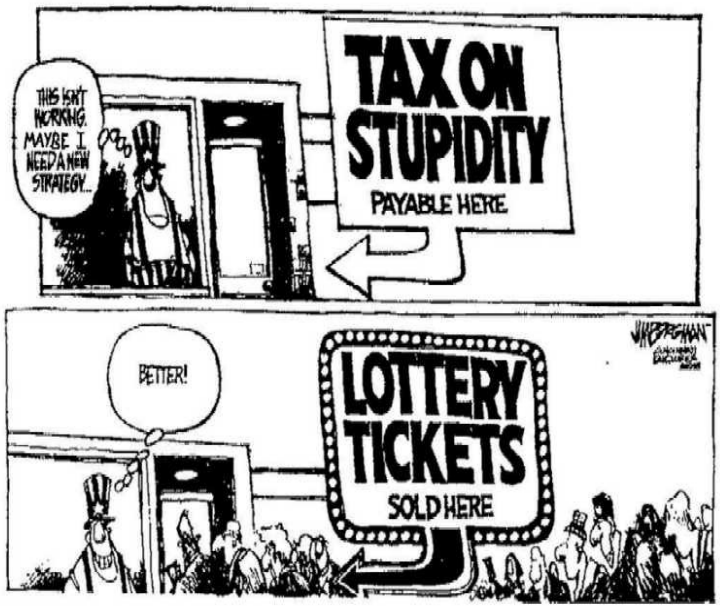
4.1 Kvantitativa antalsdata

Problem 4.1.1.

I ett lotteri ingår 1000 lotter. Det finns en vinst på 100 kr, 5 vinster på vardera 20 kr och 30 vinster på 5 kr. Varje lott kan ge högst en vinst. En person köper en lott. Låt X vara vinstbeloppet.

a) **Vilka** värden kan X anta?

b) **Bestäm** sannolikhetsfunktionen för X och rita stolpdiagram.



Problem 4.1.2.

Man har en dator till vilken tre terminaler är kopplade. Dessa används oberoende av varandra och man har sannolikheterna 3/4, 2/3 resp 1/2 för att terminal 1, 2 resp 3 används ett visst ögonblick. Låt X vara antalet terminaler som anvïnds vid ett visst ögonblick.

Bestäm sannolikhetsfunktionen för X.

Problem 4.1.3.

Den diskreta stokastiska variabeln X kan bara anta heltalsvärdena $0, 1, 2, \ldots$
$P(X = k) = p^x,\ x = 1, 2, 3, \ldots$.

a) Vad är $P(X = 0)$?

b) Vilka värden på p är möjliga?

Problem 4.1.4.

Den stokastiska variabeln X kan bara anta värdena $3, 4, 7, 8$ och 9. Man känner följande värden på sannolikhetsfunktionen: $p_X(3) = 1/3, p_X(4) = 1/4, p_X(7) = 1/6, p_X(8) = 1/6$. Beräkna

a) $p_X(9)$,

b) $F_X(5)$,

c) $P(4 \leq X \leq 8)$ och $P(X \geq 8)$.

Problem 4.1.5.

En spelare kastar en tärning efter att hen satsat 1 kr på var och en av följande tre händelser:

A) udda antal prickar

B) antal prickar högst lika med tre

C) antal prickar högst lika med två.

A) ger 3 kr i vinst om den inträffar, B) ger 2 kr medan C) ger 1 kr i vinst. Betrakta spelarens nettovinst som en stokastisk variabel X.

a) Vilka värden kan X anta?

b) Bestäm sannolikhetsfunktionen för X.

c) Beräkna $P(-2 < X < 7)$.

Oberoende stokastiska variabler/Independent random variables
X och Y är oberoende diskreta stokastiska variabler, om och endast om

$$P(X = x, Y = y) \stackrel{\text{def}}{=} P(X = x \cap Y = y) = P(X = x) \cdot P(Y = y)$$

gäller för *alla par* x, y.

Problem 4.1.6.

X och Y är två diskreta stokastiska variabler. De är oberoende och har sannolikhetsfunktionerna $p_X(x)$ och $p_Y(y)$ givna av

x	0	1	2
$p_X(x)$	0.3	0.3	0.4

y	0	1	2
$p_Y(y)$	0.25	0.4	0.35

Bestäm $P(X + Y = 2)$.

Problem 4.1.7.

De oberoende stokastiska variablerna X och Y har sannolikhetsfunktionerna $p_X(x)$ och $p_Y(y)$ givna av

x	0	1	2
$p_X(x)$	0.1	0.4	0.5

respektive

y	-1	0
$p_Y(y)$	0.2	0.8

Bestäm $P(X+Y=1)$.

Problem 4.1.8.

X har sannolikhetsfunktionen

$$p_X(x) = \begin{cases} 1/x & \text{om } x = 3,4,5,6, \\ 1/20 & \text{om } x = 8 \\ 0 & \text{annars.} \end{cases}$$

Beräkna $P(4 < X \leq 8)$.

Benfords lag Benfords lag (även känd som FDL= First Digit Law) ger fördelningen för olika siffror som förstasiffror. Lagen säger till exempel att siffran 1 bör vara förstasiffra i 30,1% av fallen, siffran 2 i 17,6% av fallen och siffran 9 i 4,6% av fallen i en mycket stor datamängd. Om en stor datamängd inte följer Benfords lag är det en indikation på att siffrorna kan (men behöver inte nödvändigtvis) vara påhittade eller manipulerade. Formeln för Benfords lag ges i nästa uppgift.
Benfords lag visar sig tillämplig inom många skilda områden.

Problem 4.1.9.

Benfords lag Benfords lag beskriver hur olika siffror är statistiskt fördelade som förstasiffror i olika material och ges av sannolikhetsfunktionen

$$p_X(x) = P(X = x) = \log_{10}\left(1 + \frac{1}{x}\right), \quad x = 1, \ldots, 9.$$

En formell kontroll (av att $p_X(k)$ är en sannolikhetsfunktion):

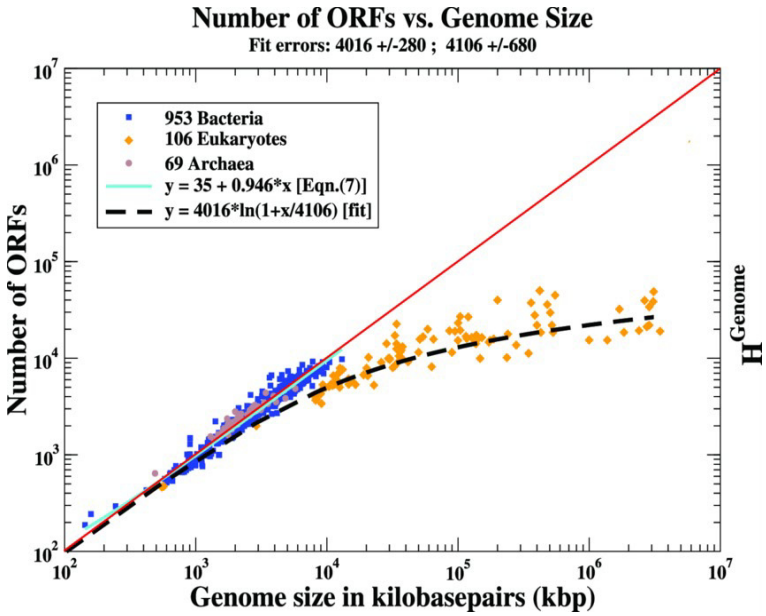
Checka att

$$\sum_{x=1}^{9} p_X(x) = 1.$$

Problem 4.1.10.

Benfords lag Data on the number of Open Reading Frames (ORFs) coded by genomes from the 3 domains of Life show the presence of some notable general features. These include essential differences between the Prokaryotes and Eukaryotes, with the number of ORFs growing linearly with total genome size for the former, but only logarithmically for the latter.

Simply by assuming that the (protein) coding and non-coding fractions of the genome must have different dynamics and that the non-coding fraction must be particularly versatile and therefore be controlled by a variety of (unspecified) probability distribution functions (pdf's), we are able to predict that **the number of ORFs for Eukaryotes follows a Benford distribution** and must therefore have a specific logarithmic form. Using the data for the 1000+ genomes available to us in early 2010, we find that the Benford distribution provides excellent fits to the data over several orders of magnitude.



The points are data from 1128 genomes available on the GOLD database in early 2010. In this log-log plot, the x-axis represents the genome size (G) in kilobasepairs. For each genome we plot on the y-axis the number of ORFs quoted for the genome in the above database. In order to facilitate comparisons, we have drawn a red diagonal line on a vertical/horizontal scale where 1 vertical axis unit corresponds to 1 kbp on the horizontal axis. The Prokaryotic genomes cluster around this (slope= 1) line. The fit to the Prokaryotes is represented here as a cyan line. The dashed line represents the best fit to the Eukaryotic ORFs and corresponds to a Benford distribution,

Friar, JL; Goldman, T; Pérez-Mercader, J: *Genome sizes and the benford distribution.* **PLOS ONE** 7 (5),(2012)

Hur ser vi Benfords lag i denna graf?

Problem 4.1.11.

X och Y är ett par av diskreta variabler. Sannolikhetsfördelningen för (X, Y) är $P(X = x, Y = y)$ och ges av tabellen nedan

	$Y = 1$	$Y = 0$
$X = 1$	0.30	0.15
$X = 0$	0.30	0.10
$X = -1$	0.10	0.05

Till exempel, $P(X = 1, Y = 1) = 0.30$, $P(X = 1, Y = 0) = 0.15$, o.s.v.. Vi kan checka att summna av siffrorna i tabellen är lika med ett.

a) Beräkna sannolikheterna $P(X = 1), P(X = 0), P(X = -1)$.

b) Beräkna sannolikheterna $P(Y = 1), P(Y = 0)$.

c) Är X och Y oberoende?

d) Beräkna $P(Y = 1|X = 1)$.

e) Beräkna $P(X = 1|Y = 1)$.

f) Jämför svaren i **d)** och **e)** med varandra. Känns de naturliga?

4.2 Väntevärden och varianser

Standardavvikelse/Standard deviation
Standardavvikelsen för en stokastisk variabel X definieras som

$$D(X) \stackrel{\text{def}}{=} \sqrt{Var(X)}.$$

Denna definition gäller både för diskreta och kontinuerliga fördelningar. Observera att $Var(X) \geq 0$ alltid.

Problem 4.2.1.

Låt X och Y vara oberoende diskreta stokastiska variabler. Båda dessa variabler kan anta värdena -1, 0, 1 med respektive sannolikheter 0.25, 0.50, 0.25. Beräkna $D(2X + Y)$.

Problem 4.2.2.

Den stokastiska variabeln X har sannolikhetsfunktionen $p_X(1) = 0.2$, $p_X(2) = 0.3$ och $p_X(4) = 0.5$. Beräkna variansen $Var(X)$.

Problem 4.2.3.

De stokastiska variablerna X och Y är oberoende och väntevärdena är $E(X) = 3$ och $E(Y) = 4$ och standardavvikelserna är $D(X) = 1.5$ samt $D(Y) = 2.5$. Beräkna standardavvikelsen $D(2X - 3Y + 7)$.

Problem 4.2.4.

Den stokastiska variabeln X antar värdena 0, 1 och 4 med sannolikheterna 0.2, 0.5 och 0.3 respektive.

a) Bestäm **variationskoefficienten** för X, som är $D(X)/E(X)$ (detta är ett sortinvariant mått på variabiliteten).

b) Vad är sambandet mellan $D(X)/E(X)$ och $CV = (s/\bar{x}) \times 100\%$?

Problem 4.2.5.

Tabellen visar fördelningsfunktionen för en diskret stokastisk variabel X. Beräkna dess standardavvikelse $D(X)$.

x	0	1	2	3	4	5
$F_X(x)$	0	0.1	0.3	0.7	0.8	1.0

4.3 Grafisk framställning av $P(X = x, Y = y)$: Eikosogram

$P(X = x, Y = y)$ och Betingad sannolikhet
P.g.a. definition gäller

$$P(X = x, Y = y) = P(X = x \mid Y = y) \cdot P(Y = y)$$

och

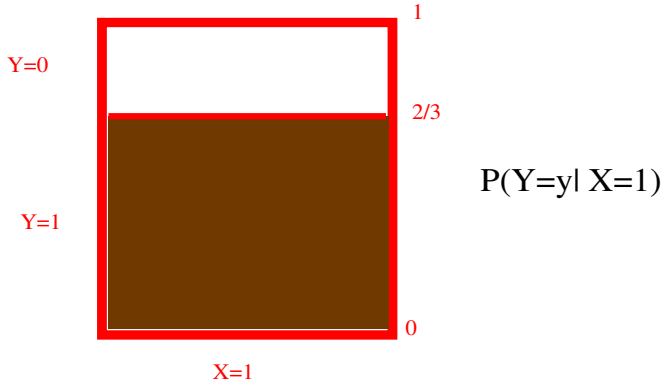
$$P(X = x, Y = y) = P(Y = y \mid X = x) \cdot P(X = x)$$

för alla x, y.
Vi skall nu ge en grafisk metod att räkna med dessa uttryck. Metoden bygger på följande formel:

Rektangelns area= bas $\times$ höjd

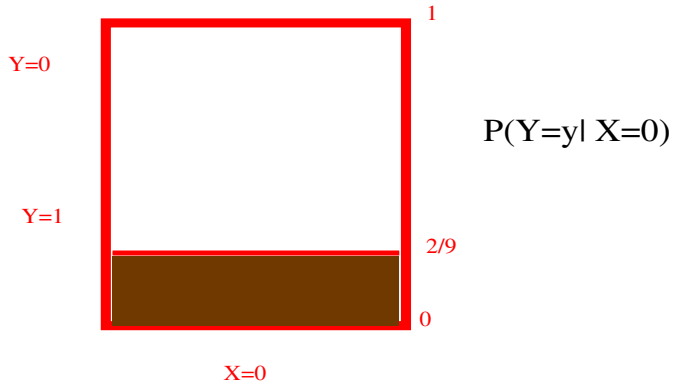
Problem 4.3.1.

X och Y är binära stokastiska variabler. I diagrammet, som även kallas **eikosogram** (?!), åskådliggörs grafiskt sannokhetsfunktionen $P(Y = y|X = 1)$ för $y = 0, 1$. Den stora kvadraten har arean=1. Sannolikheten $P(Y = 1|X = 1)$ är $2/3$ och svarar mot arean av den färgade rektangeln. Då är $P(Y = 0|X = 1) = 1/3$ och

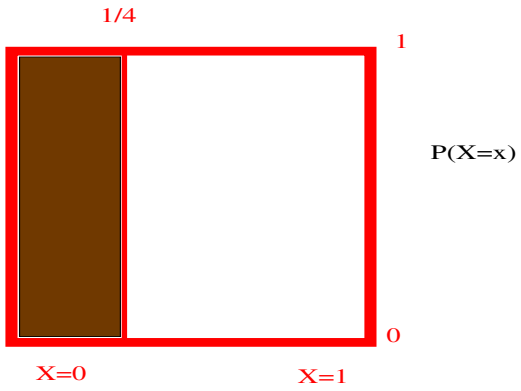


svarar mot arean av resten av kvadraten.

På samma sätt ges $P(Y = y|X = 0)$ för $y = 0, 1$ i nästa figur. Det vill säga att $P(Y = 1|X = 0) = 2/9$ är arean för den färgade rektangeln. Genom att jämföra de två diagrammen ses att Y och X är beroende, ty den betingade sannolikheten för $Y = 1$ påverkas av värdet på X.

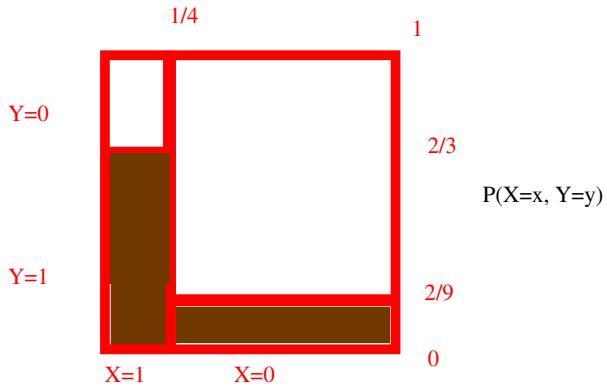


Figuren nedan ger sannolikhetsfunktionen $P(X = x)$ för X. Vi tittar på rektangeln med basen $= 1/4 (= P(X = 1))$ och arean $1/4 \times 1 = 1/4$.



Vi framställer nu $P(X = x, Y = y)$ medelst det fjärde eikosogrammet i rad, genom att rektangelns area= bas $\times$ höjd ger $P(X = x, Y = y) = P(Y = y \mid X = x) \cdot P(X = x)$. Den färgade delrektangeln till vänster representerar $P(X = 1, Y = 1)$ och har arean $1/4 \times 2/3 = 1/6$. Observera att $P(X = 1, Y = 1) = P(Y = 1 \mid X = 1) \cdot P(X = 1) = 2/3 \times 1/4$.

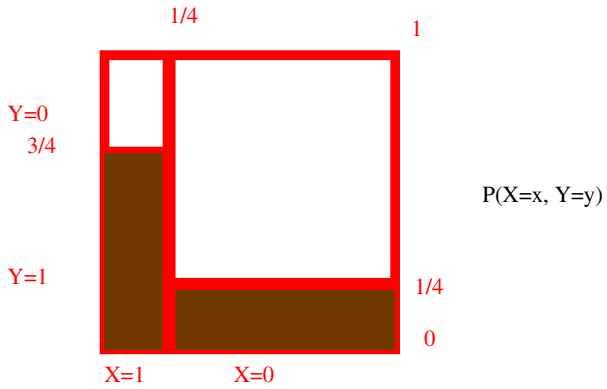
Den färgade rektangeln till höger med basen $3/4$ och höjden $2/9$ har arean $P(X = 0, Y = 1) = 3/4 \times 2/9 = 1/6$.



a) Bestäm $P(X = 1, Y = 0)$ och $P(X = 0, Y = 0)$ utifrån det fjärde eikosogrammet ovan. Checka att summan av $P(X = x, Y = y)$ blir lika med $= 1$.

b) Hur fås $P(Y = y)$ för $y = 0, 1$ ur det fjärde eikosogrammet ovan ?

c) Betrakta



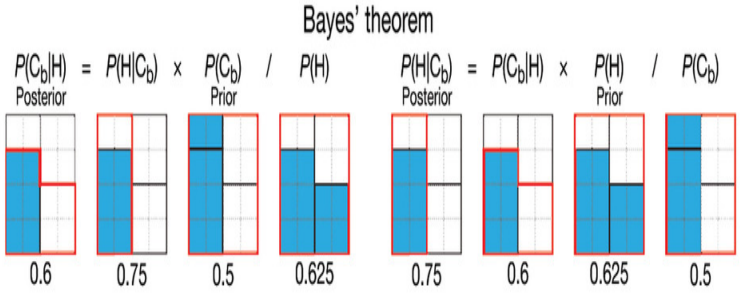
Ge $P(X = x, Y = y)$ som sannolikheterna p_{xy} i tabellen

	$Y = 1$	$Y = 0$
$X = 1$	p_{11}	p_{10}
$X = 0$	p_{01}	p_{00}

och ge $P(Y = y \mid X = x)$ och $P(X = x)$. Är X och Y oberoende?

Problem 4.3.2.

Följande eikosogram har hämtats ur J. López Puga, M. Krzywinski & Naomi Altman: *From Points of Significance: Bayes' theorem* **Nature Methods**, 12, sid. 277-278, 2015. Redogör för Bayes formel med stöd av detta. C och H_b är binära.



4.4 Binomialfördelning, Poissonfördelning, Geometrisk fördelning, Hypergeometrisk fördelning m.m

Problem 4.4.1.

The binomial distribution $\mathcal{B}\text{in}(n,p)$ results from a procedure that meets a set of conditions. Which of the following conditions is wrong, i.e., is not one of the conditions leading to $\mathcal{B}in(n,p)$?

a) The procedure has a fixed number of random events.

b) The events have outcomes in four categories.

c) The events are independent.

d) The probabilities are constant for each event.

Problem 4.4.2.

The geometric distribution $\mathcal{G}e(p)$ results from a procedure that meets a set of conditions. Which of the following conditions is wrong, i.e., is not one of the conditions leading to Ge(p)?

a) The procedure has a fixed number of random events.

b) The events have outcomes in two categories.

c) The events are independent.

d) The probabilities are constant for each event.

Problem 4.4.3.

A hypergeometric distribution $\mathcal{HG}(N,k,n)$ results from a procedure that meets the following conditions:

i) A sample of size n is randomly selected without replacement from a population of N items.

ii) In the population, k items can be classified as successes, and $N-k$ items can be classified as failures.

a) You count the number of successes in your sample of n items under the procedure **i) -ii)**. Why is this not a binomial distribution?

b) Consider the following: You have an urn of 10 marbles - 5 red and 5 green. You randomly select 2 marbles without replacement and count the number of red marbles you have selected. What is the probability distribution of the number of red marbles you have selected?

Problem 4.4.4.

The term **adventitious presence** (AP) refers to the unintentional and incidental commingling of trace amounts of one type of seed, grain or food product with another. When used in relation to plant biotechnology, the term refers to the incidental presence of biotech-derived material in food, feed or grain at levels that are consistent with generally accepted agricultural and manufacturing practices. A growing number of countries have established risk assessment procedures for approving the import of biotech crops and their derivatives.

A small part of the risk assessment procedure for the AP of genetically modified events in seed lots is by testing for AP level in a subsample of a seed lot. First a commercial from Eurofins BioDiagnostics:

> Eurofins BioDiagnostics' quantitative AP testing determines the amount of unwanted biotech traits in a seed lot. Our sensitive DNA-based tests can detect very low levels of AP in a sample. Analysis results indicate an estimate of contamination in the tested sample as well as an upper limit for the lot from which the sample was taken.

For large seed lots, a binomial distribution is assumed for the number of AP seeds, but for seed lots where the tested sample is a substantial proportion of the overall seed lot, a hypergeometric distribution is assumed.

Consider testing 3000 seeds from a 4000 overall seed lot.

a) What is the probability of finding no AP seeds in the tested lot if the overall seed lot has 4 AP seeds (= 0.1 % of seeds) using a hypergeometric distribution ?

b) What is the probability of finding no AP seeds in the tested lot if the overall seed lot has 4 AP seeds (= 0.1 % of seeds) using a binomial distribution ?

c) The estimation of AP using the hypergeometric distribution applies to the composite of the seed tested and the residual seed lot that is destined for inclusion in other studies. In the case of testing for defects in manufacturing, this type of estimation is justified if the samples tested for defects are replaced into the overall product lot after testing, but in the case of seed tested for AP, the destructive nature of testing makes this atypical.

Herman, R.A. and Robbins, K.R.: *Use of hypergeometric distribution for estimating adventitious presence of GM traits in small seed lots may be misleading*, **Seed Science Research**, 23, p. 211–212, 2013.

d) You are testing 3000 seeds from a 4000 overall seed lot. If there is one (1) AP seed, what is the probability of finding it amongst the test seeds ?

e) **How** do you estimate an upper limit of AP for the lot from which the sample was taken?

Problem 4.4.5.

Toss the DNA die independently thousand times. Let X = the number of A$'$s in the sequence obtained.

Why is it that $X \sim \mathcal{B}\text{in}(1000, 1/4)$?

Geometrisk fördelning och ffg-fördelning
Betrakta oberoende försök sådana att ett visst resultat varje gång uppträder med sannolikheten p. Försöken fortsätter tills resultatet uppträder första gången.
Låt X = antalet försök före det lyckade försöket, i vilket resultatet uppträder första gången. Då har den stokastiska variabeln X sannolikhetsfunktionen

$$p_X(x) = (1-p)^x p, \qquad x = 0, 1, 2, \ldots,$$

där $0 < p < 1$. Vi säger att X ha *geometriskt fördelning* (*geometric distribution*, se pp. 151-152, Vidakovic), och vi skriver detta förkortat som $X \sim \mathcal{G}e(p)$.
Betrakta nu igen oberoende försök sådana att ett visst resultat varje gång uppträder med sannolikheten p. Försöken fortsätter tills resultatet uppträder första gången. Om vi nu sätter X =, antalet försök som utföres *t.o.m.* resultatet uppträder, är X en stokastisk variabel med för-första-gången-fördelning. Då har X sannolikhetsfunktionen

$$p_X(x) = (1-p)^{x-1} p, \qquad x = 1, 2, \ldots,$$

där $0 < p < 1$ och X säges vara *för-första-gången-fördelad*.
På sid. 152 i Vidakovic skrives för-första-gången-fördelning förkortat som $X \sim \mathcal{G}eom(p)$.

Problem 4.4.6.

To each of the four sides of a tetrahedron there is assigned one of the letters A,T,C,G. All letters are used. A trial is the die tossed in the air with the side that it falls on registered. We have

$$P(A) = P(T) = P(C) = P(G) = \frac{1}{4}$$

We assume that the trials are independent. Set

$$X = \text{the number of times you see one of A,T or G in ten (10) trials.}$$

Which one of the following statements is correct?

a) $X \sim \mathcal{P}oi(0.75)$.

b) $X \sim \mathcal{B}in(10, 0.25)$.

c) $X \sim \mathcal{G}e(0.75)$.

d) $X \sim \mathcal{B}in(10, 0.75)$.

Problem 4.4.7.

In the herpesvirus genome, the nucleotides C,G, A and T occur in frequencies 35/100, 35/100, 15/100 and 15/100, respectively. Assuming an independence model for the genome, what is the probability that a randomly selected 15 nt long DNA fragment contains eight C's or G's and seven A's and T's?

Problem 4.4.8.

Vid syntetisering av DNA förekommer fel i 15% av sekvenserna. Felen kan antas vara oberoende. Antag att 8 syntetiserade sekvenser väljs ut slumpmässigt.

Beräkna sannolikheten att högst 2 av dem är felaktiga.

Problem 4.4.9.

A Swedish county reports that five health workers in different hospitals have been exposed to Hepatitis B via needle stick accidents. Previous experience suggests that there is a 30% chance to actually develop Hep-B after such an incident. Officials are concerned with the total number X of infected health workers, which is clearly a random variable.

a) What are the possible values for X?

b) What probability distribution would you suggest for X and why?

c) What are the parameters of this distribution?

d) What is the expected number of health workers who develop Hep-B? What is the variance?

e) What is the probability that exactly one health worker develops Hep-B?

f) What is the probability that all five develop Hep-B?

g) What is the probability that at least two develop Hep-B?

h) Assuming that all five heath workers actually develop Hep-B, is this unlikely enough to go back and re-assess the original assumption of a 30% chance of infection?

Problem 4.4.10.

Kvalitetskontroll. Ett varuparti om 100 enheter innehåller 6 defekta enheter. En köpare tar på måfå, och utan återläggning ut 5 enheter och undersöker dessa.

a) **Vad är** sannolikheten att exakt 2 av dessa är defekta?

b) Köparen accepterar partiet om högst en enhet i hans urval är defekt. **Vad är** sannolikheten för detta?

Problem 4.4.11.

Kvalitetskontroll. För en tillverkningsprocess är sannolikheten för att en enhet blir felaktig p. Man ska tillverka n enheter och vill bestämma sannolikheten för att k av dessa blir felaktiga. En tänkbar slumpmodell fås genom analogin med dragning med återläggning av n kulor från en urna som innehåller en proportion p svarta och $1-p$ vita kulor.

Bestäm sannolikheten för att

a) precis k enheter blir felaktiga

b) någon blir felaktig

c) högst en blir felaktig.

d) **Beräkna** sannolikheterna i fallen **b)** och **c)** numeriskt då $n = 3$, $p = 0.1$.

Problem 4.4.12.

Ett stort företag har vid en viss tidpunkt bland sina anställda 259 gravida kvinnor. 23 av dem är VDT-arbetare d.v.s. de arbetar med en videoskärmsterminal. Totalt fyra av graviditeterna slutar i ett missfall, och två av dessa inträffar bland de VDT-arbetande kvinnorna.

Härled en hypergeometrisk sannolikhetsfördelning för antalet missfall bland gravida VDT-arbetare under antagandet att det inte finns en association mellan VDT och missfall.

Beräkna m.h.a. denna fördelning sannolikheten för att exakt två VDT-arbetare skulle ha ett missfall. Kommentera.

Problem 4.4.13.

We consider an elementary and abstracted statistical model of the DNA sequencing process. We think of a line segment representing the target segment and a process where smaller segments are 'dropped' onto random locations of the target. The target is considered 'sequenced', when adequate coverage accumulates (e.g., when no gaps remain).

Take L and G as the fragment length and target length, respectively. We assume that $L << G$. We take furthermore the ratio $\frac{L}{G}$ as the probability to hit a single given location in the fragment. This is justified in an exercise about continuous data with uniform distribution.

Let X =the number of independent drops needed to hit a single given location in the fragment for the first time, including the successful drop.

Which of the following statements is correct?

a) $X \sim \mathcal{P}oi(L/G)$.

b) $X \sim \mathcal{G}eom(1 - L/G)$

c) $X \sim \mathcal{U}(0, G)$

d) $X \sim \mathcal{G}eom(L/G)$

Problem 4.4.14.

I genomet för husmus, *Mus musculus*, finns det ungefär 24 000 gener och av dessa har 3 000 ingen känd funktion. Antag att vi väljer 1 000 av de 24 000 generna på måfå och låter X =antalet valda gener utan känd funktion.

Vad har X för fördelning?

Ledning: Tänk på två urnor.

Problem 4.4.15.

Kösystem. Antalet kunder som i ett visst tidsintervall anländer till ett betjäningsställe kan ofta approximativt beskrivas med en stokastisk variabel X med sannolikhetsfunktionen $p_X(x) = \frac{m^x}{x!}e^{-m}, \quad x = 0, 1, 2, \ldots$ där m är en viss positiv konstant (Poissonfördelning med parametern m).

Om $m = 4$ **vad är** sannolikheten att fler än 2 kunder men färre än 5 kunder anländer under tidsintervallet ifråga?

Problem 4.4.16.

I en tillverkningsprocess, med sannolikheten p för defekt enhet, undersöker man tillverkade enheter, tills man för första gången får en defekt enhet. Låt X vara antalet undersökta enheter, den defekta enheten medräknad. Anta att felen förekommer oberoende av varandra.

a) Vilka värden kan X anta?

b) Bestäm sannolikhetsfunktionen för X.

Problem 4.4.17.

You toss again the DNA die until you get G for the first time. Tosses are independent.

Which is the correct probability distribution that describes the random number of tosses you make including the last successful one?

a) $\mathcal{P}oi(1/2)$.

b) $\mathcal{B}in(n, 1/2)$.

c) $\mathcal{G}eom(1/4)$.

d) $\mathcal{G}e(1/2)$.

Problem 4.4.18.

In the United States, about 7 percent of the male population and 0.4 percent of the female population either cannot distinguish red from green, or see red and green differently from how others do (Howard Hughes Medical Institute, 2006).

Six (6) men are randomly selected from the US population for a study of traffic signal perceptions.

Which one of the following formulas is the correct expression for the probability that exactly three of these men cannot distinguish between red and green?

a) $e^{-0.07}\frac{0.07^3}{3!}$.

b) $\frac{6!}{3!3!}0.07^3 0.93^3$.

c) $0.93^2 \cdot 0.07$.

d) $\frac{3}{6}$.

Problem 4.4.19.

The *CAG repeat* is the stretch of DNA at the beginning of the Huntington disease (HD) gene, which contains the sequence CAG repeated many times, and is pathologically long in people, who will develop HD.

We think in terms of generative modelling that in the beginning of the HD gene there opens a frame for a random number of statistically independent insertions of the triplet CAG. Let $Z =$ the number of CAG triplets inserted. Let $1-p$ be the probability of inserting a single triplet CAG. Then the probability distribution of Z is

$$Pr(Z = z) = p \cdot (1-p)^z, z = 0, 1, 2, \ldots,$$

i.e., this is the probability of z successes (i.e., probability a CAG repeat of length z) before the repeat ends. ($z = 0$ means that no extra CAG were inserted starting at this site.) If $X = Z + 1$, then $X \sim \mathcal{G}eom(p)$ (as can be checked, you are allowed to take this for granted here). Assume $p = 0.01$.

Which of the following results of computing is wrong ?

a) $E[Z] = 99$.

b) $E[X] = 100$.

c) $Var[X] = 9900$.

d) $Var[Z] = 9899$.

Problem 4.4.20.

You are working in a laboratory, where the quality of air is monitored by an Environmental Particle Counter (EPC). This instrument counts the number of particles of a certain type in samples of air. There is mounting toxicological evidence appears to indicate that exposure to these particles even in small numbers is associated with negative health effects.

A new sample is taken every hour. The different samples can be seen as independent of each other. A warning is dispatched as soon as the number of particles exceeds or is equal to a certain threshold. Let $X_i =$ be the number of particles in the sample at time i. Our model is that $X_i \sim \mathcal{P}oi(4)$. Let the threshold be $= 5$. I.e., if $X_i \geq 5$ occurs at time i, a warning is signalled.

We want to find the probability of two or more warnings during a period of ten hours. Let Y be the number of warnings during a period of ten hours. We seek

$$P(Y \geq 2).$$

Two lines of computation are sent by email from a smartphone of a statistical consult. Firstly:

$$P(X_i \geq 5) = 0.37.$$

This is readily checked by a statistics calculator under the assumption $X_i \sim \mathcal{P}oi(4)$. Secondly:

$$1 - \left(\begin{array}{c} 10 \\ 0 \end{array}\right) 0.37^0(1-0.37)^{10} - \left(\begin{array}{c} 10 \\ 1 \end{array}\right) 0.37^1(1-0.37)^9 = 0.932.$$

You are asked to explain to your colleagues at the lab what this means and why this is correct. There is one and only one correct explanation among **a) - d)**. Which one?

a) Our model is $Y \sim \mathcal{G}e(0.37)$ and then $P(Y \geq 2) = 1 - P(Y < 1)$.

b) Our model is $Y \sim \mathcal{G}e(0.37)$ and then $P(Y \geq 2) = 1 - P(Y \leq 1)$.

c) Our model is $Y \sim \mathcal{B}in(10, 0.37)$ and then $P(Y \geq 2) = 1 - P(Y \leq 1)$.

d) Our model is $Y \sim \mathcal{B}in(10, 0.37)$ and then $P(Y \geq 2) = 1 - P(Y < 1)$.

Problem 4.4.21.

Statistics for the Number of Protein Molecules Produced per Transcript

The ribosome is a kind of molecular machine present that serves as the site of biological protein synthesis (translation). Ribosomes link amino acids together in the order specified by messenger RNA (mRNA) molecules.

RNA polymerase, abbreviated as RNAP, is an enzyme that produces primary transcript RNA. Prokaryotic ribosomes bind to the mRNA as soon as it is accessible behind the transcribing RNAP. Multiple ribosomes spaced about 80 nucleotides apart simultaneously translate the emerging transcript. After release of the first protein, additional proteins are completed every several seconds as successive ribosomes reach the end of the reading frame.

We note also RNase E, which is an enzyme that has been evolutionarily conserved in both Gram-positive and Gram-negative bacteria. Its binding sites on the mRNA are close to the ribosome binding sites.

The coupling between transcription, translation, and mRNA degradation has been examined experimentally and results in the following concept.

McAdams, Harley H and Arkin, Adam: *Stochastic mechanisms in gene expression*, **Proceedings of the National Academy of Sciences, USA**, 94, 3, pp. 814–819, 1997.

Because RNase E cannot bind when its binding site is occluded, a ribosome that at the ribosome binding site protects the MRNA from degradation until the site is again exposed as the ribosome translates the mRNA. Thus, at each exposure of the ribosome binding site and RNASE E sites on the RNA, *there is a competition between ribosome and RNase E binding.* This competition leads either to successful translation and production of a protein or to degradation or inactivation of the transcript.

For statistical calculations we assume that

i) successive ribosome-RNase competitions are independent trials with $p =$ probability of success for ribosome binding

ii) binding competition determines the outcome.

Let $N =$ the number of proteins produced from a transcript at each trial.

a) What is the probability of $P(N = n)$, $n = 0, 1, 2, 3, \ldots$?

b) What is $E(N)$?

c) What is $Var(N)$?

d) What is $P(N \geq n)$?

e) If $E(N) = 10$ (proteins), which percentage of transcripts will produce 25 or more proteins?

Problem 4.4.22.

The Discrete Weibull Distribution

Distributions of pathogen counts in treated water over time are highly skewed, powerlaw- like (for power law, see later in section 5.2.1), and discrete. Over long periods of record, a long tail is observed, which can strongly determine the long-term mean pathogen count and associated health effects. J.D. Englehardt and Li Ruochen suggest in *Risk Analysis*, pp. 370−381, 2011, to use the Discrete Weibull distribution for microbial counts in water. Formally, we say that the r.v. X has the discrete Weibull distribution, $X \sim \text{dWei}(q, \beta)$, if

$$p_X(x) = q^{x^\beta} - q^{(x+1)^\beta}, \quad x = 0, 1, 2, \ldots$$

where $0 < q < 1$, $0 < \beta$.

a) Find the cumulative distribution function $F_X(x)$, for $x = 0, 1, 2, \ldots$.

b) Find $E(X)$.

c) We can say that $\mathrm{dWei}(q,\beta)$ is a generalization of $\mathrm{Ge}(p)$, where $p = 1-q$, or that $\mathrm{dWei}(q,\beta)$ has $\mathrm{Ge}(p)$ as a special case. What do we mean by this?

d) The positive number β is called the *shape parameter* of $\mathrm{dWei}(q,\beta)$. Plot $p_X(x)$ for $\mathrm{dWei}(0.7, 0.5)$ and $\mathrm{dWei}(0.7, 2)$ to check the meaning of the phrase *shape parameter*.

Problem 4.4.23.

Sum of Two Independent Poisson Random Variables
$X \sim \mathcal{P}oi(\mu)$ and $Y \sim \mathcal{P}oi(\lambda)$, X and Y are independent. Let

$$Z = X + Y.$$

We see that the values of Z must be non-negative integers. We want to find the probability distribution of Z. We find for that purpose the probability mass function of Z, i.e.,

$$P(Z = k), \quad k = 0, 1, 2, \ldots.$$

We compute this in several steps.

a) Justify for yourself the following equality of events

$$\{Z = k\} = \cup_{l=0}^{k} \{X = l, Y = k - l\}.$$

Then we get

$$P\left(Z = k\right) = \sum_{l=0}^{k} P\left(X = l, Y = k - l\right).$$

Why does this hold ?

b) Give the justification of

$$P\left(X = l, Y = k - l\right) = P\left(X = l\right) \cdot P\left(Y = k - l\right).$$

c) Use next the assumptions $X \sim \mathcal{P}oi(\mu)$ and $Y \sim \mathcal{P}oi(\lambda)$ and evaluate the sum obtained by the above, i.e.,

$$P\left(Z = k\right) = \sum_{l=0}^{k} P\left(X = l\right) \cdot P\left(Y = k - l\right).$$

Aid: You will probably find it useful to observe that

$$\frac{1}{l!} \cdot \frac{1}{(k-l)!} = \frac{1}{k!} \left(\begin{array}{c} k \\ l \end{array} \right).$$

In addition the binomial theorem should be of be value here. What is the distribution of Z?

4.5 Approximation: Samband mellan fördelningar

Om X är $\mathcal{B}in(n,p)$-fördelad med $p \leq 0.1$, så är X approximativt $\mathcal{P}oi(np)$-fördelad.

Problem 4.5.1.

A **palindrome** is a word, phrase, number, or other sequence of characters which reads the same backward or forward. A restriction enzyme is cutting DNA at a palindromic site 6 nt long.

Determine the probability that a circular chromosome, a double-stranded DNA molecule of length $L = 84000$ nt, will be cut by the restriction enzyme into exactly twenty fragments.

We assume the DNA die with independent tosses as the description of the DNA sequence. Approximation of a binomial distribution by a Poisson distribution is required.

Problem 4.5.2.

A circular double-stranded DNA of $L = 3400$ nt long was cut by a restriction enzyme. A subsequent gel electrophoresis separation indicated the presence of five DNA pieces. It turned out that the absent-minded technician could not recall exact type of restriction enzyme that was used. Still, he knew that the chemical was picked up from a box containing an equal number of 4-base cutters and 6-base cutters (restriction enzymes that cut specific 4 nt long sites and specific 6 nt long sites, respectively).

What is the posterior probability that 4-nucleotide cutter was used?

We assume the DNA die with independent tosses as the description of the DNA sequence. Approximation of the number of restriction sites by a Poisson distribution is permitted.

Problem 4.5.3.

C. Darwin has formulated the following problem[1] on what might be called rare deviations:

Let it be assumed that, in a large population, a particular affection occurs ... in one out of a million, so that the á priori chance that an individual taken at random will be so affected is only one in a million. Let the population consist of sixty millions, composed, we will assume, of ten million families, each containing six members.

Darwin assumes without any doubt that a family has two parents and four children.

The Question (by Darwin)

What are the odds that there will not be even a single family in which at least one parent and two children will be affected?

Let X = the number families thus affected, i.e., the number of families with in which at least one parent and two children are affected. The total number of families is $n = 10^7$.

a) What is the probability in a certain family to have at least one parent and two children affected? Assume independent draws from the overall population.

b) What is the distribution of X?

c) What is now $P(X = 0)$? Approximation of distribution expected.

d) What is $P(X > 0)$? The approximation $P(X > 0) \approx P(X = 1)$ is applicable.

e) Find the desired odds. Darwin continues (loc.cit)

Professor Stokes [2] *has calculated for me that the odds will be no less than 8333 millions to 1.*

Compare now your answer with the result by Stokes.

f) Darwin states that such families do exist in England (he observes $X = x > 0$). What conclusion does he/do you draw from this ?

[1] *The variation of animals and plants under domestication.* London: John Murray. 1st ed, 1st issue. Volume 2., Ch XII, p. 5, c.f. `http://darwin-online.org.uk/converted/published/1868_Variation_F877/1868_Variation_F877.2.html`

[2] (maybe this is George Gabriel Stokes of the Stokes' formula fame ?)

4.6 Lander-Watermans statistik för shotgun sekvensering

Method
Frederick Sanger was one of the laureates of the Nobel Prize in Chemistry in 1980 for developing a method to sequence short regions of DNA. It was the most widely used sequencing method for approximately 40 years since its invention.

- If the sequence is larger than 500-1000 consecutive nucleotides, the rest of it will not be read (?). There is no current technology to simply read the whole genome sequence from one end to the other. The human genome is 3 billion nucleotides long. Sequencing it using the Sanger method requires breaking it into little pieces, sequencing the pieces separately, and fitting them back together.
- Break DNA at random into many smaller pieces, and randomly select a large number of these pieces to be sequenced. Approximately the first 500 nucleotides are read from one end of the pieces. These small sequenced regions are called **reads**. We do not know their location in the genome or their strand. Combine overlapping reads into **contigs**. Sequence alignment is used to detect the overlaps.
- Additional information (scaffolds) is used to place the contigs into the proper order and direction on chromosomes.

Notations and statistical assumptions

- G =genome length in nucleotides $\approx$ 3 billion in human
- L = read length in nucleotides (assume 500)
- N = number of reads sequenced
- NL = number of nucleotides in all sequenced reads
- $a = \frac{N}{G}L$ is the **coverage** (=average number of times each nucleotide in the whole genome is sequenced)
- In each chromosome, a read of length L could start anywhere except the last $L-1$ positions.
  - In a genome of length G with c chromosomes, there are $G - c \cdot (L-1)$ possible starting positions. For human, $c \cdot (L-1) = 23(499) = 11477 << G$ so we will approximate that there are G possible starting positions. (That is, we will ignore the end effects.) The probability that one of the N reads starts at any specific nucleotide is N/G.
  - Assume reads are distributed uniformly through the genome and independently of each other.

Note that the exercises to follow do not require any extensive calculations.

Problem 4.6.1.

a) Let I be any interval of L consecutive nucleotides. Let X = number of reads starting in I.

a.i) Under the preceding assumptions, why is the distribution of X is binomial? *Hint*: success= a read starts at a nucleotide.

a.ii) What are the parameters of the binomial distribution?

a.iii) What is the probability of no reads in I? What is the probability of at least one read starting in I, i.e., the probability of a contig?

b) Approximate the binomial distribution in **a)** by the appropriate Poisson distribution.

b.i) What are now the answers to **a.ii)** - **a.iii)**?

b.ii) We need to define gaps:

> The nucleotide r is in a **gap**, if no read starts within the interval $[r - L + 1, r]$. In **b.i)** we found the probability of a gap.

What is your back-of-the-envelope estimate[3] of the expected number of nucleotides in gaps? What is your estimate of the expected number of nucleotides in contigs?

b.iii) Assume that 99% of the genome is in contigs and 1 % in gaps. What is the coverage of, e.g., the human genome? Comment on your finding.

c) Each of the contigs has a unique rightmost read. The probability that a read the rightmost read equals the probability that no other read starts within that read, and has been found in **b.i)**. If you label the rightmost read as success and others as failures. With Y = the number of successes, Y has a binomial distribution.

c.i) What are the parameters of this distribution?

c.ii) What is the expected number of contigs expressed in terms of a, G and L ?

4.7 Protein Identification & Probability Models for Scoring

Mass spectrometry (MS) is a method for the rapid identification of proteins and the characterisation of post-translational modifications. The schematic Figure 4.1 is an illustration. In this process, proteins are first digested and the resulting peptides are isolated one by one in the tandem mass spectrometer, fragmented using collision-activated dissociation, and the fragment ions sorted based on their mass-to-charge ratio. Results are displayed as spectra of the relative abundance of detected ions as a function of the mass-to-charge ratio, as in Figure 4.2. The fragmentation information in a tandem mass spectrum of a peptide can be used to search against a peptide/protein sequence database to identify the amino acid sequence represented in the spectrum, see Figure 4.1.

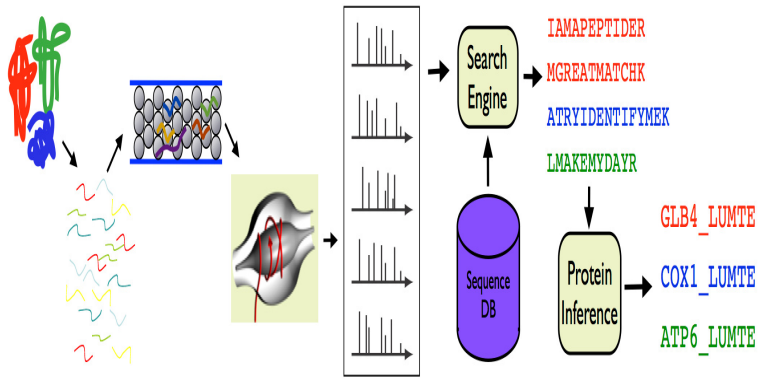
Sequence database search engines compare spectra acquired from a sample with the theoretically predicted fragmentation spectra of peptides derived from a protein database. The list of predicted peptides is ideally derived from all of the protein sequences that could be expressed in the experiment sample.

A common element of comparison in all database-searching algorithms is the assignment of a **score** that gives the quality of the match between the experimental sequence and theoretical spectrum. If a protein sequence in the reference list gives rise to a significant number of predicted masses that match the experimental values, there is some evidence that this protein was present in the original sample. The score is a measure of the closeness of fit between the spectrum and the peptide sequence retrieved from the database.

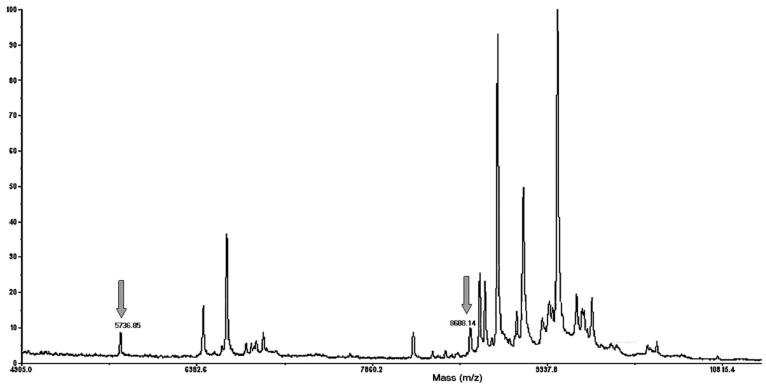
A database-searching program will always find a database sequence fitting up to some degree with the experimental mass spectrum. It has been found that searching a dataset generated from a separation of a eukaroytic tryptic digest against an in silico generated random protein database generated a significant number of positive matches.

Thus, only a part of all database search results lead to a true protein/peptide identification. The rest of the database search results correspond to assignment of tandem mass spectra to protein databases at random. It is important, therefore, that a scoring scheme provides a means of estimating the *statistical significance of each peptide match* (statistical significance: to be elucidated in the sequel). Therefore, it is required to model the protein assignment process by a probabilistic model attaching a probability to each protein assignment.

[3] A back-of-the-envelope calculation is a rough calculation, typically jotted down by an engineer on any available scrap of paper such as the actual back of an envelope.



Figur 4.1: The Process of MS-based proteomics & Protein Identification by courtesy of Lukas Käll



Figur 4.2: Mass spectrum: y-axis gives relative frequency

Many different models have been suggested in the literature, some of these adapted to special platforms of database search and MS instruments. The following problems study some basic models underlying scores of protein identification. One way to devise a score is to find the probability disribution of random matches.

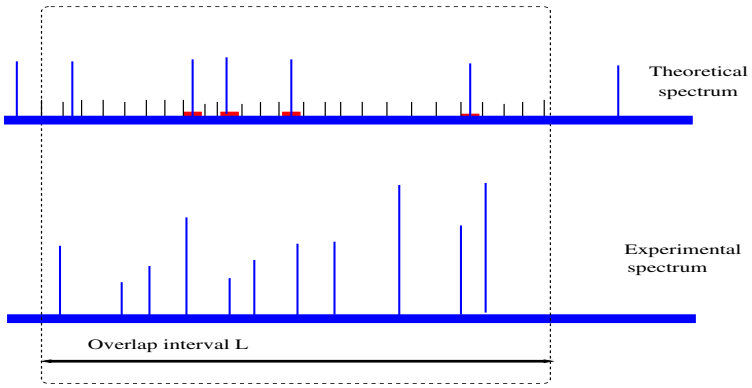
Problem 4.7.1.

RMET-distribution
due to Fridman, T., Razumovskaya, J., Verberkmoes, N., Hurst, G., Protopopescu, V., & Xu, Y: The probability distribution for a random match between an experimental-theoretical spectral pair in tandem mass spectrometry. *Journal of bioinformatics and computational biology*, 3(02), 455−476, (2005).

We consider an interval of length L on the m/z -axis, where the experimental and theoretical spectra have been aligned to overlap, thus to be called the overlap interval L, see Figure 4.3. The theoretical peaks (=peaks in the theoretical spectrum) are considered as a pattern fixed in advance. The experimental peaks (=peaks in the experimental spectrum) randomly fall onto the overlap interval.

Next we partition the overlap interval L into N slots of length $2{\cdot}\sigma$, where σ equals the experimental accuracy, so that each theoretical peak is the midpoint of one slot. We need possibly to adjust L for this. These slots in L are depicted by walls $|\quad|$ in Figure 4.3. If we allow for the vagaries of drawing, then $N = 24$ in Figure 4.3.

Each experimental peak is either a *match*, if it falls within a $\pm\sigma$ (=matching distance) around its closest peak in the theoretical spectrum, or otherwise it is a miss. The slots with a match are indicated by red colour in Figure 4.3. We are going to derive the probability distribution for the number of Random Matches for a pair



Figur 4.3: The framework in RMET illustrated

of Theoretical and Experimental spectra or the RMTE distribution.

Let k be the number of theoretical peaks in the overlap interval, $k = 5$ in Figure 4.3. Since we can on good physical grounds assume that there is at most one match in every slot, or that we count two or more matches in the same slot as one, the number of matches cannot exceed k.

There are $N - k$ slots with no peak in the theoretical spectrum. These empty slots in L are depicted without colour in Figure 4.3. .

Let there be n peaks in the experimental spectrum in the overlap interval, $n = 11$ in Figure 4.3.

We set
X = the number of matches for n peaks in an experimental spectrum to a theoretical spectrum with k peaks in an overlap interval with N slots, $k \leq N$.
We say that X has the RMET distribution. In Figure 4.3 we have $X = 4$.

a) Find the probability distribution $p_X(x) = P(X = x)$, $x = 0, 1, \ldots, \min(\mathrm{n}, \mathrm{k})$.

b) How could the probability $p_X(x)$ in **a)** be implemented in a score?

Problem 4.7.2.

The Human Proteome Organization (HUPO) considers false peptide matches at protein identification level, see Omenn, G. S., Blackwell, T. W., Fermin, D., Eng, J., Speicher, D. W., & Hanash, S. M. : Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study. *Nature biotechnology*, 24(3), 333−338, 2006. They start from the observation that protein identifications are mostly false positives, if identification is based on only one peptide. By false positive one means the claim that an experimental peptide matches a protein, when it in fact does not.

Let $\mu > 0$ be the intensity of false positive peptide matches per protein. We let X = the number of false positive peptide matches in a data base of proteins. We assume that $X \sim \mathcal{P}\mathrm{oi}(\mu K)$, where K is the number of protein identifications in a data base.

a) If 80 % of the peptide matches in a data base with $K = 5210$ are false, what is μ?

b) Let μ be as in **a)**. If there are 5210 entries in a protein data base identified with one peptide, how many false positive peptide matches per protein are to be expected ?

4.8 Bilaga 1: Momentgenerande funktion för diskreta variabler

Om X är en diskret stokastisk variabel med sannolikhetsfunktion $p_X(x)$, så definieras dess **momentgenererande funtion** $m_X(t)$ som

$$m_X(t) = E\left[e^{tX}\right] = \sum_x e^{tx} p_X(x).$$

Om denna funktion är redan bekant kan man lätt beräkna $E(X)$ och $Var(X)$ genom att beräkna de två första derivatorna på $m_X(t)$ och utvärdera dessa i noll. Detta följer av att

$$\frac{d}{dt} m_X(t) = \sum_x \frac{d}{dt} e^{tx} p_X(x) = \sum_x x e^{tx} p_X(x),$$

så att

$$\frac{d}{dt} m_X(0) = \sum_x x e^{0x} p_X(x) = \sum_x x p_X(x) = E(X).$$

På samma sätt fås andra derivatan

$$\frac{d^2}{dt^2} m_X(t) = \frac{d}{dt} \underbrace{\sum_x x e^{tx} p_X(x)}_{= \frac{d}{dt} m_X(t)}$$

$$= \sum_x x \frac{d}{dt} e^{tx} p_X(x) = \sum_x x^2 e^{tx} p_X(x),$$

och således

$$\frac{d^2}{dt^2} m_X(0) = \sum_x x^2 e^{0x} p_X(x) = \sum_x x^2 p_X(x) = E\left(X^2\right).$$

Variansen fås nu som bekant med $Var(X) = E\left(X^2\right) - (E(X))^2$.

Sammanfattning:

$$E(X) = \frac{d}{dt} m_X(0).$$

$$E\left(X^2\right) = \frac{d^2}{dt^2} m_X(0).$$

Problem 4.8.1.

Om $X \sim \mathcal{P}oi(\lambda)$, så är den momentgenererande funktionen

$$m_X(t) = \exp(\lambda(e^t - 1)).$$

Du behöver inte härleda denna funktion.

Checka att $E(X) = Var(X) = \lambda$ m.h.a. den momentgenererande funktionen.

Problem 4.8.2.

Om $X \sim \mathcal{B}in(n, p)$, så är den momentgenererande funktionen

$$m_X(t) = (1 - p + pe^t)^n$$

Du behöver inte härleda denna funktion.

Checka att $E(X) = np$ och att $Var(X) = np(1-p)$ m.h.a. den momentgenererande funktionen.

Problem 4.8.3.

I en studie om 'estimating the number of protein folds and families', framförs följande antaganden:

- There exists a **universal population of M protein (domain) families and N folds**. Each family belongs to exactly one fold; each fold includes at least one family.

- Both the database of protein structures (hereinafter structure database) and a genome are generated by a **random, independent sampling of families from the fold/family population**. In the sampling process, each of the M families has an equal probability of being drawn from universal population into the sample.

- In the fold/family population the distribution of the number of protein families in a fold is best approximated by a **logarithmic series distribution**.

En stokastisk variabel X säges ha en *logaritmisk fördelning* (logarithmic series distribution) med parameter p om $0 < p < 1$ dess sannolikhetsfunktion är

$$p_X(k) = \frac{-1}{\ln(1-p)} \frac{p^k}{k}, \quad k = 1, 2, 3, \ldots$$

a) Varför är det rätt att kalla $p_X(k)$ en sannolikhetsfördelning?

b) Den momentgenererande funktionen är

$$m_X(t) = \frac{\ln(1 - p\,e^t)}{\ln(1-p)} \text{ för } \quad t < -\ln p$$

Du behöver inte härleda denna funktion. **Checka** att $E(X) = \frac{-1}{\ln(1-p)} \frac{p}{1-p}$.

4.9 Bilaga 2: Väntevärde och varians för linjära kombinationer

Räknereglerna i denna bilaga gäller för både diskreta och kontinuerliga stokastiska variabler.

Vi ser på en linjärkombination

$$a_1X_1 + a_2X_2 + \cdots + a_nX_n + b$$

av n s.v. $X_1, \ldots, X_n$. Konstanterna $a_1, \ldots, a_n$ och b kan vara positiva eller negativa tal.
För alla s.v. $X_1, \ldots, X_n$ gäller att

$$E\Big(\sum_{i=1}^{n} a_iX_i + b\Big) = \sum_{i=1}^{n} a_iE(X_i) + b \tag{4.1}$$

För oberoende s.v. $X_1, \ldots, X_n$ gäller att

$$Var\Big(\sum_{i=1}^{n} a_iX_i + b\Big) = \sum_{i=1}^{n} a_i^2Var(X_i).$$

Om $X_1, \ldots, X_n$ är s.v. med samma väntevärde μ så gäller att

$$E\Big(\sum_{i=1}^{n} X_i\Big) = n\mu. \tag{4.2}$$

Om $X_1, \ldots, X_n$ är oberoende och har samma standardavvikelse σ gäller även att

$$Var\Big(\sum_{i=1}^{n} X_i\Big) = n\sigma^2 \quad \text{och} \quad D\Big(\sum_{i=1}^{n} X_i\Big) = \sigma\sqrt{n}. \tag{4.3}$$

När fler och fler oberoende s.v. läggs ihop, ökar alltså väntevärdet proportionellt mot antalet n, men standardavvikelsen proportionellt mot $\sqrt{n}$, alltså långsammare. Man har ofta anledning att undersöka aritmetiska medelvärdet av flera oberoende s.v.
Om $X_1, \ldots, X_n$ är oberoende s.v., var och en med väntevärdet μ och standardavvikelsen σ, och

$$\bar{X} = \sum_{i=1}^{n} X_i/n$$

är deras aritmetiska medelvärde, så gäller att

$$E(\bar{X}) = \mu, \quad Var(\bar{X}) = \sigma^2/n \quad \text{och} \quad D(\bar{X}) = \sigma/\sqrt{n}.$$

Problem 4.9.1.

SEM= standard error of the mean och medelfel
I biomedicin och bioteknik rapporteras ofta siffran SEM definierad som

$$\text{SEM} \stackrel{\text{def}}{=} \frac{\sigma}{\sqrt{n}}.$$

(SEM är altså i detta sammanhang inte Scanning Electron Microscope.)
På svenska talar man om medelfel, när vi för okänd σ insätter standardavvikelsen s, d.v.s

$$\text{medelfel} \stackrel{\text{def}}{=} \frac{s}{\sqrt{n}}.$$

Ett och endast ett av följande påståenden är falskt. Vilket?

a) SEM kan beräknas för både kontinuerliga och diskreta kvantitativa datatyper.

b) SEM är ett tal som visar hur mycket varje stickprovsmedelvärde i medeltal avviker från populationsmedlet.

c) SEM är en statistika som sammanfattar variabiliteten i en datamängd.

d) SEM är alltid mindre än standardavvikelsen.

Problem 4.9.2.

X, Y och Z är oberoende stokastiska variabler med $D(X) = D(Y) = 2$ och $D(Z) = 4$.

Beräkna $D(X + 2Y - Z + 1)$.

Problem 4.9.3.

$X_1 \sim \mathcal{P}oi(2)$, $X_2 \sim \mathcal{P}oi(2)$ and $X_3 \sim \mathcal{P}oi(2)$ are three independent random variables. Set

$$\overline{X} = \frac{1}{3}\left(X_1 + X_2 + X_3\right).$$

Which of the following statements is wrong ?

a) $E(\overline{X}) = 2$.

b) $Var(\overline{X} - 2) = \frac{2}{3}$.

c) $E((-2) \cdot \overline{X}) = -4$.

d) $Var((-2) \cdot \overline{X}) = \frac{4}{3}$.

Problem 4.9.4.

$X \sim \mathcal{G}e(p)$ och därmed gäller att $E(X) = \frac{(1-p)}{p}$ och $Var(X) = \frac{(1-p)}{p^2}$, se sid. 152 i Vidakovic. Sätt $Y = X+1$.

a) Härled sannolikhetsfunktionen för Y.

b) Bestäm $E(Y)$ och $Var(Y)$.

Kapitel 5

Sannolikhetsmodeller för kontinuerliga datatyper

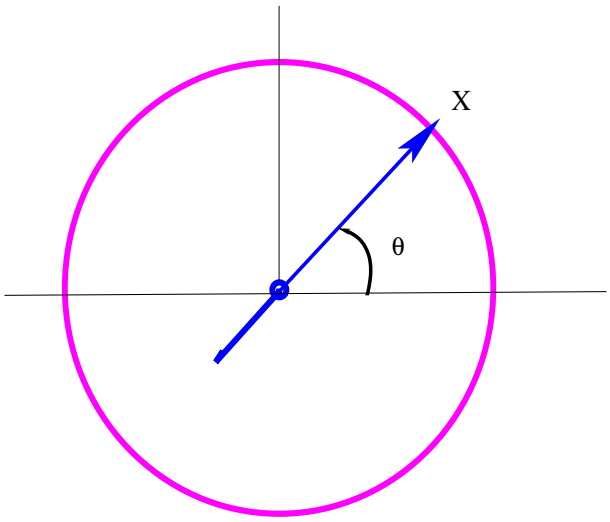
5.1 Sannolikhetsberäkningar, väntevärde och varians

Problem 5.1.1.

I figuren ser vi ett schematiskt avbildat s.k. lyckohjul. En pil roterar ett slumpmässigt antal varv, stannar i X och vi avläser rotationsvinkeln θ räknad moturs från x-axeln.

Vilken sannolikhetsfördelning skulle θ ha?

WHEEL OF FORTUNE



Problem 5.1.2.

Den stokastiska variabeln X har täthetsfunktionen $f_X(x) = \frac{24}{x^3}$ för $3 \leq x \leq 6$ och $f_X(x) = 0$ för övrigt.

Bestäm $P(X > 5.5)$.

p-kvantilerna
p-kvantilen (eller $100 \times p$ percentilen) för en kontinuerlig stokastisk variabel X är ett reellt tal x_p sådant att

$$F(x_p) = p,$$

där $F(x)$ är fördelningsfunktionen för X.

Problem 5.1.3.

Den kontinuerliga stokastiska variabeln X har en täthetsfunktion (sannolikhetskurva) $f_X(x) = Ke^{x-1}$, $1 \leq x \leq 3$ där K är en konstant, vars värde skall bestämmas.

a) Bestäm K så att f_X blir en giltig täthetsfunktion.

b) Beräkna väntevärdet för X.

c) Beräkna **medianen** för X.

Problem 5.1.4.

Vid tillverkning av kex kan avvikelsen från den tänkta tjockleken (felet) beskrivas enligt en stokastisk varibel X med täthetsfunktion

$$f_X(x) = \begin{cases} 1+x & -1 < x < 0 \\ 1-x & 0 \leq x < 1. \end{cases}$$

Värdena är givna i millimeter.

a) Beräkna sannolikheten att tjockleken avviker 0.5 mm, dvs beräkna

$$P(|X| \geq 0.5).$$

b) Bestäm fördelningsfunktionen $F_X(x)$.

c) Beräkna väntevärde och varians för X.

Problem 5.1.5.

De stokastiska variablerna X_1 och X_2 är oberoende och har varianserna $V(X_1) = k$ respektive $V(X_2) = 2$. Variansen för den stokastiska variabeln $Y = 3X_2 - X_1$ är $Var(Y) = 25$.

Bestäm k.

Problem 5.1.6.

$X \sim \mathcal{E}(1)$, dvs X har täthetsfunktionen

$$f_X(x) = \begin{cases} e^{-x} & \text{om } x \geq 0, \\ 0 & \text{annars.} \end{cases}$$

Beräkna $P(4 < X \leq 8)$.

Problem 5.1.7.

En kontinuerlig stokastisk variabel X har fördelningsfunktionen

$$F_X(x) = \begin{cases} 0 & \text{om } x < 0, \\ c \cdot x^2 & \text{om } 0 \leq x < 2, \\ 1 & \text{om } x \geq 2. \end{cases}$$

Bestäm konstanten c samt beräkna $E(X)$ och $D(X)$.

Problem 5.1.8.

Den stokastiska variabeln X har fördelningsfunktionen

$$F_X(x) = \begin{cases} 0 & \text{för } x < 1 \\ \frac{1}{12}(x^2 - x) & \text{för } 1 \leq x \leq 4 \\ 1 & \text{för } x > 4 \end{cases}$$

Beräkna väntevärde för X.

Problem 5.1.9.

Den stokastiska variabeln X har täthetsfunktionen

$$f_X(x) = \begin{cases} 8(1-x) & \text{för } 0.5 \leq x \leq 1 \\ 0 & \text{för övrigt} \end{cases}$$

Beräkna $P(X > 0.7)$.

Problem 5.1.10.

$X \sim \mathcal{E}(\lambda)$, $\lambda > 0$, dvs X har täthetsfunktionen

$$f(x) = \begin{cases} \lambda \cdot e^{-\lambda \cdot x} & \text{om } x \geq 0, \\ 0 & \text{annars.} \end{cases}$$

a) Beräkna medianen för X.

b) Jämför medianen med väntevärdet samt typvärdet och kommentera skevhet.

Problem 5.1.11.

$X \sim U(1,2)$. $Y = 3 \cdot X$.

a) Bestäm Ys fördelning.

b) Bestäm $E(Y)$.

c) Bestäm $Var(Y)$.

Problem 5.1.12.

$X \sim U(-1,1)$.

Bestäm $P(X^2 \geq 0.5)$.

Problem 5.1.13.

In previous exercise we considered a DNA covering process, where smaller segments are 'dropped' onto random locations of the target.

Take L and G as the fragment length and target length, respectively. We assume that $L << G$. Let $L = b - a$. Assume that U, a single location in the fragment, is chosen from $U(0, G)$. Then, we think of first dropping the fragmetn and chooisng the site. Show that

$$\text{The probability for a fragment to hit } U = P(a \leq U \leq b) = \frac{L}{G}.$$

Note that the single location may lie in any part of the target without this probability being altered, that is, this probability depends only on the lengths of the fragment and the target.

Problem 5.1.14.

In a probabilistic approach to chemical kinetics one can use an event-based approach. The idea is to represent dynamics in terms of discrete events occuring at times determined randomly by their respective rates.

Consider any interval $[0, t]$ subdivided into intervals $\left[\frac{it}{n}, \frac{(i+1)t}{n}\right]$, $i = 0, \ldots, n-1$ i.e.

$$[0,t] = \cup_{i=0}^{n-1} \left[\frac{it}{n}, \frac{(i+1)t}{n}\right].$$

Simplest illustration is to draw a picture.

Suppose that $\lambda \cdot \frac{t}{n}$ is the probability of one or more events occurring in $\left[\frac{it}{n}, \frac{(i+1)t}{n}\right]$ for any $i = 0, \ldots, n-1$. Then $\lambda > 0$ is the rate of events occurring. The number of events occurring in any subinterval $\left[\frac{it}{n}, \frac{(i+1)t}{n}\right]$ is independent of the number of events occurring in any other interval.

Let now T *be the time of the first occurrence of an event during* $[0, +\infty]$.

a) Find $P(T > t)$.

b) Let

$$\lim_{n \to +\infty} P(T > t).$$

When $n \to +\infty$, the length of the intervals $\left[\frac{it}{n}, \frac{(i+1)t}{n}\right]$ turns to zero for any fixed t. Hence the limit in **b)** would seem to determine a probability distribution for continuous data on $[0, +\infty)$. Check that in the limit $T \sim \mathcal{E}(\lambda)$.

c) Let $U \sim \mathcal{U}(0, 1)$. Find δ such that

$$P(T > \delta) = U,$$

where $T \sim \mathcal{E}(\lambda)$. Then δ is a random variable. Determine the distribution of δ.

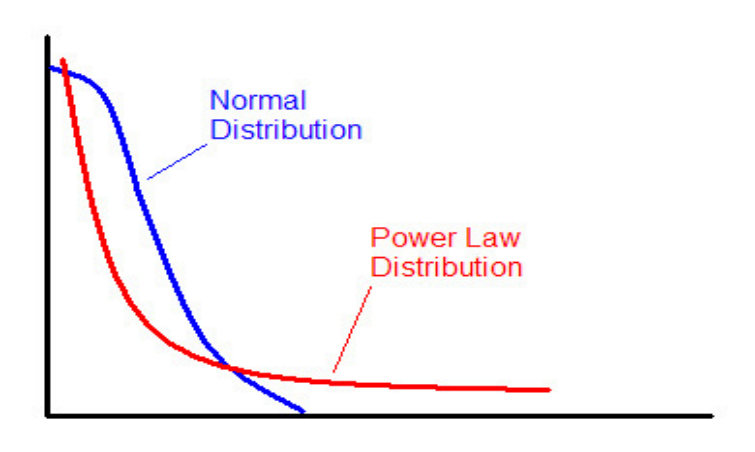
5.2 Speciella sannolikhetstäthetsfunktioner/kurvor

5.2.1 Paretofördelningen

Power Law
A probability density function $f_X(x)$ has a power-law tail or is a **Power law**, if it holds that ($\gamma > 0$)

$$f_X(x) \approxeq x^{-\gamma}, \quad \text{as } x \to \infty. \tag{5.1}$$

The Figur compares graphically a power law with the bell curve (normal curve).



The notation $f(x) \approxeq g(x)$ (at $x = a$) means the following

$$\lim_{x \to a} \frac{f(x)}{g(x)} = 1.$$

This means that the functions grow at the same rate at a. For example, if

$$f(x) = x^2, g(x) = x^2 + x,$$

then

$$\lim_{x \to \infty} \frac{f(x)}{g(x)} = \lim_{x \to \infty} \frac{1}{1 + \frac{1}{x}} = 1,$$

but at the same time $f(x) - g(x) = x$.

It has been found empirically that the power law behaviour applies to the distribution of a wide range of genome-associated quantities. These include the frequency distribution of gene family sizes, the number of transcripts per gene, the number of interactions per protein, the number of genes or pseudogenes in paralogous families, the occurrence of DNA words (short base sequences), as well as distributions of the connections of enzymes and metabolites in metabolic networks, the frequencies of distinct DNA and protein domains.
The power law behaviour is a mathematical-statistical expression of an important biological feature: **the dominance of a few members over the overall population**.
To quote one example of many, out of the 247 distinct protein folds currently assigned in the worm genome, just 10 account for the over half of the 7805 assigned domains.

Problem 5.2.1.

Pareto-fördelning Om en stokastisk variabel X har fördelningsfunktionen

$$F_X(x) = \begin{cases} 1 - \left(\frac{x_{\mathrm{m}}}{x}\right)^{\alpha} & x \geq x_{\mathrm{m}}, \\ 0 & x < x_{\mathrm{m}}, \end{cases}$$

så säger vi att X har en Pareto-fördelning med parametrarna α och x_m, och skriver $X \sim \mathcal{P}a(x_m, \alpha)$.

a) **Bestäm** den motsvarande sannolikhetstäthetsfunktionen

$$f_X(x) = \frac{d}{dx} F_X(x).$$

b) Ge en beskrivning i ord av $\mathcal{P}a(x_m, \alpha)$ utifrån $f_X(x)$.

Problem 5.2.2.

Cell Concentration Recorded at a Random Time The cell concentration, $N(t)$, in a cell culture at time $t > 0$ is given by

$$N(t) = N_0 e^{Kt}, \quad K > 0, t \geq 0.$$

where N_0 is the cell concentration in the culture at the beginning. We read the cell concentration at an exponentially distributed time $T \sim \mathcal{E}(\nu)$. This means that we choose first a value of T from an exponential distribution with parameter ν and then read the cell concentration at that random time.

a) Find

$$F_{N(T)}(x) = P\left(N(T) \leq x\right).$$

b) Find the probability density

$$f_{N(T)}(x) = \frac{d}{dx} F_{N(T)}(x).$$

c) Which distribution with a name attached to it does one recognize here ?

5.2.2 Mer

Problem 5.2.3.

Skevhet för en stokastisk variabel X är ett mått på asymmetrin i dess fördelning. Skevheten betecknas med κ och kan beräknas enligt formeln

$$\kappa \stackrel{\text{def}}{=} \frac{E[X^3] - 3\mu\sigma^2 - \mu^3}{\sigma^3},$$

där μ är väntevärdet och σ är standardavvikelsen för X.

a) Låt Y_1 vara likformigt fördelad på intervallet $(-1, 1)$, dvs $Y_1 \sim \mathcal{U}(a, b)$ med parametrar $a = -1$ och $b = 1$.
Beräkna skevhet κ för Y_1.

b) Låt Y_2 vara fördelad enligt sannolikhetstäthetsfunktionen

$$f_{Y_2}(y) = \frac{1}{2}(1 + y), -1 < y < 1.$$

Beräkna skevhet κ för Y_2.

Problem 5.2.4.

Pearson Sk_2 skevhet för en stokastisk variabel X är ett mått på asymmetrin i dess fördelning. Skevheten betecknas med $\mathcal{S}k_2$ och beräknas enligt formeln

$$\mathcal{S}k_2 \stackrel{\text{def}}{=} 3 \cdot \frac{\mu - \text{median}(X)}{\sigma},$$

där μ är väntevärdet, median(X) är medianen och σ är standardavvikelsen för X.
Låt Y vara fördelad enligt sannolikhetstäthetsfunktionen

$$f_Y(y) = \frac{1}{2}(1 + y), -1 < y < 1.$$

Beräkna skevhet $\mathcal{S}k_2$ för Y.

5.3 Bilaga: Momentgenerande funktion för kontinuerliga variabler

Om X är en kontinuerlig stokastisk variabel med täthetsfunktion $f_X(x)$, så definieras dess **momentgenererande funtion** $m_X(t)$ som

$$m_X(t) = E\left[e^{tX}\right] = \int_{-\infty}^{\infty} e^{tx} f_X(x)dx.$$

Om denna funktion är redan bekant kan man, precis som tidigare för diskreta variabler, beräkna $E(X)$ och $Var(X)$ genom att beräkna de två första derivatorna på $m_X(t)$ och utvärdera dessa i noll. I själva verket

$$\frac{d}{dt}m_X(t) = \int_{-\infty}^{\infty} \frac{d}{dt}e^{tx} f_X(x)dx = \int_{-\infty}^{\infty} xe^{tx} f_X(x)dx,$$

och

$$\frac{d}{dt}m_X(0) = \int_{-\infty}^{\infty} xe^{0x} f_X(x)dx = \int_{-\infty}^{\infty} x f_X(x)dx = E(X).$$

På samma sätt fås andra derivatan

$$\frac{d^2}{dt^2}m_X(t) = \frac{d}{dt}\int_{-\infty}^{\infty} xe^{tx} f_X(x)dx =$$

$$= \int_{-\infty}^{\infty} x\frac{d}{dt}e^{tx} f_X(x)dx = \int_{-\infty}^{\infty} x^2 e^{tx} f_X(x)dx,$$

och således

$$\frac{d^2}{dt^2}m_X(0) = \int_{-\infty}^{\infty} x^2 e^{0x} f_X(x)dx = \int_{-\infty}^{\infty} x^2 f_X(x)dx = E\left(X^2\right).$$

Variansen fås nu som bekant med $Var(X) = E\left(X^2\right) - (E(X))^2$.

Sammanfattning:

$$E(X) = \frac{d}{dt}m_X(0).$$

$$E\left(X^2\right) = \frac{d^2}{dt^2}m_X(0).$$

Problem 5.3.1.

Om $X \sim \mathcal{E}(\lambda)$, så är den momentgenererande funktionen

$$m_X(t) = \frac{\lambda}{\lambda - t}, \text{ för } t < \lambda.$$

Du behöver inte härleda denna funktion.

Checka att $E(X) = \frac{1}{\lambda}$ och $Var(X) = \frac{1}{\lambda^2}$ m.h.a. den momentgenererande funktionen.

Problem 5.3.2.

Om X har den dubbelexponentiella fördelningen, $X \sim \mathcal{DE}(0, \lambda)$, så är den momentgenererande funktionen

$$m_X(t) = \frac{\lambda^2}{\lambda^2 - t^2}, \text{ för } |t| < \lambda.$$

Du behöver inte härleda denna funktion.

Checka att $E(X) = 0$ och $Var(X) = \frac{2}{\lambda^2}$ m.h.a. den momentgenererande funktionen.

Kapitel 6

Normalfördelningen

6.1 Normalfördelade stokastiska variabler

Problem 6.1.1.

Om $X \sim \mathcal{N}(\mu, \sigma^2)$, så är dess momentgenerande funktion

$$m_X(t) = e^{\mu t + \frac{t^2\sigma^2}{2}}.$$

Du förväntas inte att härleda detta uttryck.

Checka att $E(X) = \mu$ och $Var(X) = \sigma^2$ m.h.a. $m_X(t)$.

Problem 6.1.2.

$Z \sim \mathcal{N}(0, 1)$. Which of the following statements is wrong ?

a) $P(Z \geq 0.35) = 0.3630$.

b) $P(Z \geq 1) = 0.1590$.

c) $P(2 \cdot Z \leq 1) = 0.7$.

d) $P(Z \leq -1) = 0.1590$.

Problem 6.1.3.

X är $\mathcal{N}(0, 1)$. **Bestäm**

a) $P(X \leq 1.82)$

b) $P(X \leq -0.35)$

c) $P(-1.2 < X < 0.5)$

d) a så att $P(X > a) = 5\%$

e) a så att $P(|X| < a) = 95\%$

Problem 6.1.4.

X är $\mathcal{N}(5, 2^2)$. **Bestäm**

a) $P(X \leq 6)$

b) $P(1.8 < X < 7.2)$

c) a så att $P(X \leq a) = 5\%$

Svaren skall ges både med en formel som innehåller $\Phi(x)$ och med ett numeriskt värde.

Problem 6.1.5.

$X \sim \mathcal{N}(2, 2^2)$. One and only one of the following probabilities is wrong, which one?

a) $P(X \leq 3) = 0.691$.

b) $P(3 \cdot X \leq 8) = 0.629$.

c) $P(X > 1) = 0.309$.

d) $P(X > E[X]) = 0.5$.

6.2 Summor av normalfördelade stokastiska variabler & andra kontinuerliga stokastiska variabler

Summor och normalfördelning

Om X är $\mathcal{N}(\mu, \sigma^2)$ är $\dfrac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$.
Om X är $\mathcal{N}(\mu, \sigma^2)$ så är $Y = aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$
Om X är $\mathcal{N}(\mu_x, \sigma_x^2)$ och Y är $\mathcal{N}(\mu_y, \sigma_y^2)$ så är $X + Y \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$ om X och Y är oberoende.
Om X är $\mathcal{N}(\mu_x, \sigma_x^2)$ och Y är $\mathcal{N}(\mu_y, \sigma_y^2)$ så är $X - Y \sim \mathcal{N}(\mu_x - \mu_y, \sigma_x^2 + \sigma_y^2)$ om X och Y är oberoende.
Om $X_1, X_2, \ldots, X_n$ är oberoende, alla normalfördelade $\mathcal{N}(\mu, \sigma^2)$ och $\bar{X} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$, så är $\bar{X} \sim \mathcal{N}\left(\mu, \sigma^2/n\right)$.

Problem 6.2.1.

X och Y är oberoende normalfördelade där X är $\mathcal{N}(3, 5^2)$ och Y är $\mathcal{N}(2, 1)$.

a) Beräkna $E(2X - Y - 7)$ och $D(2X - Y - 7)$.

b) Vilken fördelning har $2X - Y - 7$?

c) Beräkna $P(2X - Y - 7 > 0)$. **Svaren skall ges både med en formel som innehåller $\Phi(x)$ och med ett numeriskt värde.**

Problem 6.2.2.

I ett lager av kaffesäckar vilkas innehåll i kg anses vara $N(35, 0.5^2)$ tar man ut en säck på måfå och portionerar ut innehållet i burkar så att varje burk innehåller i det närmaste exakt 1 kg.
Hur stor är sannolikheten att

a) säcken räcker till minst 36 burkar?

b) säcken räcker till 34 burkar men inte till 36?

Problem 6.2.3.

X är diametern för axlar som tillverkas. Vid kvalitetskontroll sorterar man bort de axlar som är tjockare än 1.01 mm eller smalare än 0.99 mm. Man har funnit att $P(X > 1.01) = 8\%$ och $P(X < 0.99) = 2\%$. Antag att X är $\mathcal{N}(m, \sigma^2)$.

Bestäm m och σ.

Problem 6.2.4.

Ett IT-projekt består av två faser av vilka den andra inte kan igångsättas förrän den första avslutats. Låt X_1 och X_2 vara de tider som erfordas för respektive fas. Antag att X_1 och X_2 är oberoende och $\mathcal{N}(200, 20^2)$ resp. $\mathcal{N}(100, 15^2)$ (enhet: dagar).

Beräkna sannolikheten att den totala projekttiden $Y = X_1 + X_2$ överstiger 310 dagar.

Problem 6.2.5.

En verkstad producerar axeltappar och hylsor, anpassade till varandra. En hylsa och en tapp "passar" om skillnaden mellan hylsans inre diameter och axeltappens yttre diameter är högst 1 mm och lägst 0.2 mm. Antag nu att inre diametern i mm hos en slumpvis utvald hylsa är $\mathcal{N}(10, 0.2^2)$ och yttre diametern i mm hos en slumpvis utvald axeltapp är $\mathcal{N}(9.5, 0.1^2)$, och att tappar och hylsor tas ut oberoende av varandra.

Vad är sannolikheten för att de "passar" till varandra.

Problem 6.2.6.

Man har två neuroner som är inställda för utlösning 1 resp. 1.5 sekunder efter en impuls. Utlösningstiderna är ej konstanta utan $\mathcal{N}(1, 0.1^2)$ resp. $\mathcal{N}(1.5, 0.2^2)$ och oberoende.

Bestäm sannolikheten för att den andra neuronen utlöses före det första om de samtidigt utsätts för en impuls.

Problem 6.2.7.

En hiss i ett varuhus är markerad med "högst 10 personer eller 800 kg". Personvikten i kg hos en slumpvis uttagen vuxen varuhuskund kan antas vara $\mathcal{N}(70, 10^2)$.

Vad är sannolikheten för att 10 vuxna personer överlastar hissen enligt det andra av de två belastningskriterierna?

Problem 6.2.8.

Vid ett bioraffinaderi väges pappersrullar i buntar om tio stycken. I en sådan bunt är respektive rullars vikter $X_1, X_2, \ldots, X_{10}$ där X_i är $\mathcal{N}(\mu, \sigma^2)$, $i = 1, 2, \ldots, 10$. Buntens totalvikt är $Y = X_1 + X_2 + \cdots + X_{10}$. Av erfarenhet vet man att Y:s väntevärde är 1000 kg och dess standardavvikelse 20 kg. Man vill nu bestämma en "garantivikt" x kg/rulle sådan att 90 % av alla rullar i produktionen överskrider denna vikt.

Beräkna x. Variablerna $X_1, X_2, \ldots, X_{10}$ antas oberoende.

Problem 6.2.9.

In a model of microarrays we have observed intensities of a feature, X_1 and X_2, so that

$$X_1 = S_1 + N_1, \quad X_2 = S_2 + N_2,$$

where $S_1 \sim \mathcal{E}(2)$, $N_1 \sim \mathcal{N}(1, 1)$, $S_2 \sim \mathcal{E}(2)$ and $N_2 \sim \mathcal{N}(1, 1)$ are all independent of each other. Then X_1 and X_2 are also independent of each other. Set

$$\overline{X} = \frac{1}{2}\left(X_1 + X_2\right).$$

Which of the following statements is wrong ?

a) $E(\overline{X}) = \frac{3}{2}$

b) $Var(\overline{X} - \frac{3}{2}) = \frac{5}{8}$.

c) $E((-16) \cdot \overline{X}) = 12$.

d) $Var((-2) \cdot \overline{X}) = \frac{5}{2}$.

Problem 6.2.10.

$S \sim \mathcal{E}(2)$, $B \sim \mathcal{N}(1, 2^2)$. We set $X = S + B$. Which of the following statements is wrong ?

a) $E(X + \frac{1}{2}) = 2$.

b) $Var(-\frac{1}{2}X + \frac{1}{2}) = \frac{17}{16}$, if S and B are independent.

c) $Var(S - X) = 4$.

d) $E(S - X) = -1$.

Problem 6.2.11.

Average adult height in Sweden is 180 cm in men and 166 cm in women, with standard deviations of 7 cm and 6 cm respectively.

a) What are factors that contribute to a person's height? Does it make sense that av- erage heights within a homogenous population are generally normally distributed?

b) What percentage of men can we expect to be taller than 190 cm? Taller than 200 cm?

c) The same for women taller than 180 cm and taller than 190 cm?

d) What percentage of women can we expect to be between 160 cm and 172 cm?

e) What percentage of men can we expect to be between 166 cm and 194 cm?

6.3 Centrala gränsvärdessatsen

Problem 6.3.1.

Låt X vara en stokastisk variabel med täthetsfunktion

$$f_X(x) = 1 - \frac{x}{2}, 0 < x < 2.$$

a) Beräkna väntevärdet för X.

b) Beräkna standardavvikelsen för X.

c) Låt $X_1, \ldots, X_{100}$ vara ett oberoende stickprov från fördelningen med täthetsfunktion $f_X(x)$ i **a)**. Låt

$$Y = \frac{X_1 + \ldots + X_{100}}{100}$$

Beräkna approximativt $P(\mid Y - E[Y] \mid > 0.1)$.

Problem 6.3.2.

Vid syntetisering av DNA förekommer fel i 15% av sekvenserna. Felen kan antas vara oberoende. Antag att 800 syntetiserade sekvenser väljs ut slumpmässigt.

Beräkna approximativt sannolikheten att högst 140 av dem är felaktiga.

Problem 6.3.3.

The proportion of asthmatics amongst school children in the elementary schools of a region is 25%. A regional school board asks you to find the probability that the proportion of asthmatics in a randomly chosen class with 30 pupils from this region is between 15 % and 20%.

You reason as follows. Let $X_i = 1$ if the school children number i is asthmatic, and $X_i = 0$ if not, $i = 1, \ldots, 30$. Then

$$\bar{X} \stackrel{\text{def}}{=} \frac{1}{30}(X_1 + \ldots + X_{30})$$

is the proportion of asthmatics in the chosen class. The central limit theorem gives now that

$$\bar{X} \text{ approximatively} \sim \mathcal{N}\left(0.25, \frac{0.25 \cdot 0.75}{30}\right).$$

There is one and only one among the formulas in **a)- d)** below that correctly computes the desired probability using $\bar{X}$ and the normal distribution above. Please give your answer as one of **a)- d)**.
$\Phi(x)$ is the distribution function of the standard normal distribution.

a)

$$\Phi\left(0.20\right) - \Phi\left(0.15\right)$$

b)

$$\Phi\left(-0.05\right) - \Phi\left(-0.1\right)$$

c)

$$\Phi\left(\frac{-0.05}{0.0063}\right) - \Phi\left(\frac{-0.1}{0.0063}\right)$$

d)

$$\Phi\left(\frac{0.1}{0.079}\right) - \Phi\left(\frac{0.05}{0.079}\right)$$

Problem 6.3.4.

In biology and ecology one often studies the spatial distribution of a population of some species. Some plants can release a toxin that suppresses the growth of nearby competing plants, which often results in a uniform dispersion pattern. Let us assume that the distance between a pair of nearest neighbouring individuals is a random variable $X_i \sim \mathcal{U}(0,1)$. We compute the average distance of hundred independent pairs of plants

$$\frac{1}{100}\left(X_1 + \ldots X_{100}\right).$$

By the *central limit theorem* we know that $\frac{1}{100}\left(X_1 + \ldots X_{100}\right)$ is approximately $\sim \mathcal{N}(\mu, \sigma^2)$. Which of the following is a <u>correct</u> statement of the values for μ and σ?

a) $\mu = 50$ $\sigma = 0.2887 \cdot \sqrt{100}$.

b) $\mu = 50$, $\sigma = 0.2887/\sqrt{100}$.

c) $\mu = 0.5$, $\sigma = \frac{1}{12}/\sqrt{100}$.

d) $\mu = 0.5$ $\sigma = 0.2887/\sqrt{100}$.

Problem 6.3.5.

Sekvensering är en process för att med biokemiska metoder bestämma ordningen på nukleotiderna i DNA. I processen kan en nukleotid sekvenseras felaktigt med sannolikheten $p = 0.01$. Felen kan antas vara oberoende.

a) Antag att ett 100 nukleotider långt DNA-fragment sekvenseras. Vad är sannolikheten att fragmentet är inte innehåller några fel.

b) Antag att 10000 fragment bestående av 100 nukleotider var sekvenseras. Låt Y vara antalet fragment som helt felfria. Beräkna en approximativ approximativ normalfördelning för Y. Motivera dina approximationer!

c) Beräkna approximativt sannolikheten att 37.5% av de 10000 fragmenten är felfria, d.v.s. $P(Y \geq 3750)$.


MinION
MinION

Kapitel 7

Data-analys Del III: Population och stickprov

7.1 Beskrivning

En **population** är mängden av alla mätningar av intresse in någon kontext. Rent språkligt är befolkning ett synonym för population, men en mänsklig befolkning är inte enda sorts population, som kan förekomma i statistik. En intressant egenskap kan i allmänhet inte mätas för hela populationen, men vi önskar att dra slutsatser om denna egenskap på basis av en ändlig observerad datamängd, som är en delmängd av populationen.

Låt oss återge ett stycke ur Larry Winner: *Introduction to Biostatistics*, Department of Statistics University of Florida July 8, 2004.

1. Populations can be thought of as **existing** or **conceptual**.

   Existing populations are well-defined sets of data containing elements that could be identified explicitly. Examples include:

   (a) CD 4 counts of every American diagnosed with AIDS as of January 1, 1996.

   (b) Amount of active drug in all 20-mg Prozac capsules manufactured in June 1996.

   (c) Presence or absence of prior myocardial infarction in all American males between 45 and 64 years of age.

   Conceptual populations are non-existing, yet visualized, or imaginable, sets of measurements. This could be thought of characteristics of all people with a disease, now or in the near future, for instance. It could also be thought of as the outcomes if some treatment were given to a large group of subjects. In this last setting, we do not give the treatment to all subjects, but we are interested in the outcomes if it had been given to all of them. Examples include:

   (a) Bioavailabilities (bioavailability is the fraction of an administered dose of unchanged drug that reaches the systemic circulation, a pharmacokinetic property of a drug. When a medication is administered intravenously, its bioavailability is 100%.) of a drug's oral dose (relative to intravenous dose) in all healthy subjects under identical conditions.

   (b) Presence or absence of myocardial infarction in all current and future high blood pressure patients who receive short-acting calcium channel blockers.

2. **Samples** (stickprov) are observed sets of measurements that are subsets of a corresponding population. Samples are used to describe and make inferences concerning the populations from which they arise. Statistical methods are based on these samples having been taken at random from the population. However, in practice, this is rarely the case. We will always assume that the sample is representative of the population of interest.

   Examples include:

   a) CD 4 counts of 100 AIDS patients on January 1, 1996.

b) Amount of active drug in 2000 20-mg Prozac capsules manufactured during June 1996.

c) Prior myocardial infarction status (yes or no) among 150 males aged 45 to 64 years.

a) Bioavailabilities of an oral dose (relative to intravenous dose) in 24 healthy volunteers. Presence or absence of myocardial infarction in a fixed period of time for 310 hypertension patients receiving calcium channel blockers.

b) Test results (positive or negative) among 50 pregnant women taking a home pregnancy test.

7.2 Biomedicinska studier

I en **fall-kontrollstudie, case-control study** undersöks sambandet mellan en viss exponering och ett utfall genom att förekomsten av exponeringen bland fall och kontroller jämförs. Källpopulationen definieras som den existerande population som gett upphov till fall och kontroller. En sådan källpopulation kan till exempel vara 2008 års befolkning i Sverige. Individer som bodde i Sverige 2008 och som drabbades av det utfall man vill studera utgör fall till studien medan individer som inte drabbades av utfallet är potentiella kontroller.

Fall-kontrollstudier kan också genomföras i en existerande **kohortstudie**, såsom Women's Health Initiative (WHI), där över 100 000 kvinnor följts under många år. Det är ett vanligt förfaringssätt när man vill studera en exponering som är dyr att mäta (exempelvis en biomarkör), eftersom exponeringen då endast behöver mätas hos ett representativt urval av deltagarna i kohortstudien.

Problem 7.2.1.

Selektionsfel, selection bias Selektionsfel är ett exempel på ett systematiskt fel. I en fall-kontroll studie innebär det att kontrollerna (och/eller fallen) har valts på ett sådant sätt att de inte är representativa för individer med, respektive utan utfallet i källpopulationen med avseende på exponeringen.

Vi förklarar selektionsfel med ett exempel. I en svensk studie ville man undersöka, om långvarig frekvent använding av mobiltelefon ökade risken för hjärntumörer. Som kontrollgrupp användes ett slumpmässigt urval ur den del av Sveriges befolkning som inte hade diagnosticerats med hjärntumör. Alla skulle intervjuas med frågor om mobilanvändning och andra riskfaktorer. Det visade sig (som alltid) att ett visst antal av de som valts i kontrollgruppen tackade nej till att delta. Detta bortfall utgör en källa till selektionsfel.

På vilket sett kan bortfallet leda till selektionsfel?

Problem 7.2.2.

Self-selected entry to epidemiological studies and surveys
Ett citat av Niels Keiding & Thomas A. Louis:

> Faced with the declining participation rates in traditional epidemiological studies and in surveys, and with the developments in popular use and technical possibilities of the Internet, it has become increasingly attractive to try to recruit and accommodate willing and careful respondents by meeting them on the web, where they already spend much good time and energy. Such studies will often be wholly or at least mostly self-selected.
>
> A group of researchers used the internet to collect information on symptoms, number and intensities of the usually relatively harmless episodes of acute respiratory and gastrointestinal infections in a large human population.

Finns här en chans för selektionsfel?

Problem 7.2.3.

Recall bias
är en form av fel som kan uppstå när exponeringsdata samlats in retrospektivt och fallen och/ eller kontrollerna minns en och samma exponering olika just på grund av att de drabbats eller inte drabbats av utfallet.

På vilket sätt kunde recall bias uppstå i den svenska studien om långvarig frekvent använding av mobiltelefon och risken för hjärntumörer.

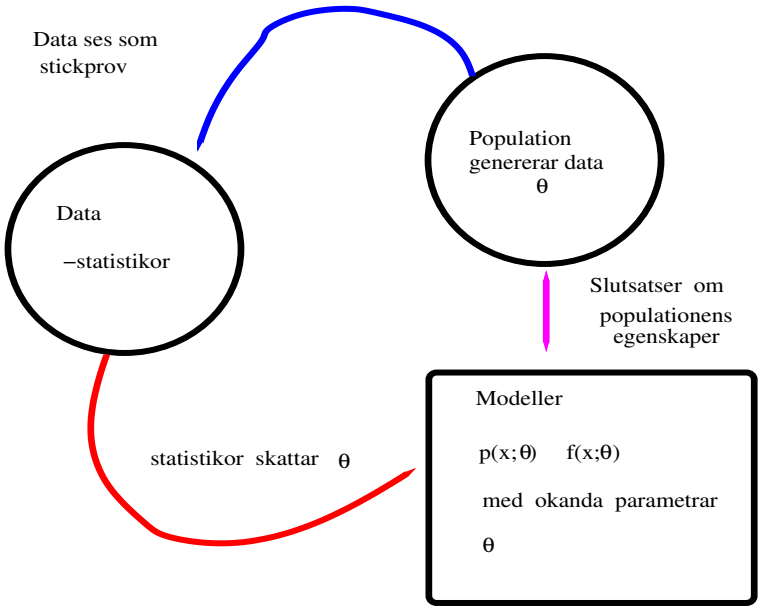
Problem 7.2.4.

Mer om selektionsfel, selection bias
Citat ur Matthew D Young, Matthew J Wakefield, Gordon K Smyth and Alicia Oshlack: *Gene ontology analysis for RNA-seq: accounting for selection bias.* **Genome Biology**, 4 February 2010.

> In RNA-seq experiments the expression level of a transcript is estimated from the number of reads that map to that transcript. The expected read count for a transcript is proportional to the gene's expression level multiplied by its transcript length. Therefore, even when two transcripts are expressed at the same level, differences in length will yield differing numbers of total reads. One consequence of this is that longer transcripts give more statistical power for detecting differential expression between samples. Similarly, more highly expressed transcripts have a greater number of reads and greater power to detect differential expression. Hence, **long or highly expressed transcripts are more likely to be detected as differentially expressed compared with their short and/or lowly expressed counterparts**. The fact that **statistical power increases with the number of reads is an unavoidable property of count data**, which cannot be removed by normalization or re-scaling. Consequently, it is unsurprising that **this selection bias has been shown to exist in a range of different experiments performed using different analysis methods, experimental designs and sequencing platforms**. When performing systems biology analyses, failure to account for this effect will lead to biased results.

Hur kan man åtgärda detta?

7.3 Parametrar i sannolikhetsmodeller och statistisk data-analys



Parametrar är numeriska deskriptiva tal som svarar mot populationer. Parametrarna betecknas gärna med grekiska bokstäver, t.ex. med μ, θ, λ, σ.

Läge och spridning är två typiska parametrar för numeriska datatyper. Medelvärde och median är lägesmått. Spridningsmått är varians, standardavikelse och variationsvidd. En annan typisk parameter är en okänd **proportion** p, $0 < p < 1$ av en egenskap i en population, ofta uttryckt som $p \times 100\%$.

Medelvärde, median, varians, standardavikelse, variationsvidd m.m. har förlänats två definitioner. Den första definitionen handlar om respektive statistikor på datamängder. Den andra definitionen är given genom matematiska parametrar i en sannolikhetsmodell för populationen.

Eftersom vi oftast inte kan (eller har råd att) observera/mäta allt av betydelse i en hel population, betraktas parametrarna som okända konstanter. Statistiska metoder av data-analys kan användas för att göra påståenden, statistiska inferenser som konfidensintervall, hypotesprövning om okända parametrar på basis av stickprov (sample data). En statistika är oftast avsedd att skatta värdet på en okänd parameter som ingår i en specifik modell för populationen.

Statistikorna är numeriska värden som beräknas från stickprov (samples). Eftersom stickprov är slumpmässigt valda ur populationen, blir statistikorna **stokastiska variabler** i den meningen att olika stickprov kommer att ge olika värden för statistikan. Statistikornas fördelning beror oftast på de okända populationparametrarna.

Problem 7.3.1.

Ett konsultuppdrag

Vi är ett branschföretag som tillhandahåller ett retursystem med lådor/pallar i plast för den svenska dagligvaruhandeln, de största kunderna är Coop, ICA, Axfood. I och med att produkterna tvättas på våra anläggningar och återanvänds ger det en stor besparing i CO2 utsläpp i jämförelse med användning av engångsemballage.

Uppgift. I nuläget har vi totalt 17 161 000 stycken returlådor. Vi har inte kontroll på hur länge dessa håller, en uppskattning är 10-15 år. Genom att inventera ett urval kommer vi kunna kategorisera det totala beståndet m.a.p. inköpsår 2000-2016. Under varje låda står datumet angivet och kan endast läsas av manuellt. Vi skulle behöva hjälp med att **fastställa hur stor andel som krävs för att urvalet ska vara representativt**? Genom att sedan även gå igenom kasserade enheter kommer vi kunna få klarhet i livslängden.

Problem 7.3.2.

In 1988, the New York State Cervical Cancer Screening Program registry included 16529 women (who live in New York State, do not have health insurance, have health insurance with a cost share that may prevent

a person from obtaining screening and/or diagnostic services, meet income eligibility requirements, meet age requirements (40 yrs and older)) with a baseline screening. Upon review of their medical records, 528 were found to have a history of breast cancer. The **prevalence** (förekomsten) of breast cancer in this 1988 selected cohort was then taken as 3.19%.

a) What is the population in this case and is it conceptual or existing?

b) What is the sample in this case?

c) What is the population parameter under discussion ?

d) What is in this case the statistic (statistika)?

Problem 7.3.3.

Skevheten för en stokastisk variabel X har tidigare definierats som

$$\kappa = \frac{E[X^3] - 3\mu\sigma^2 - \mu^3}{\sigma^3},$$

där μ är väntevärdet och σ är standardavvikelsen för X. **Skevheten för en datamängd** har tidigare definierats som

$$g_1 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^{3/2}}\ .$$

Vad är sambandet mellan κ och g_1?

Problem 7.3.4.

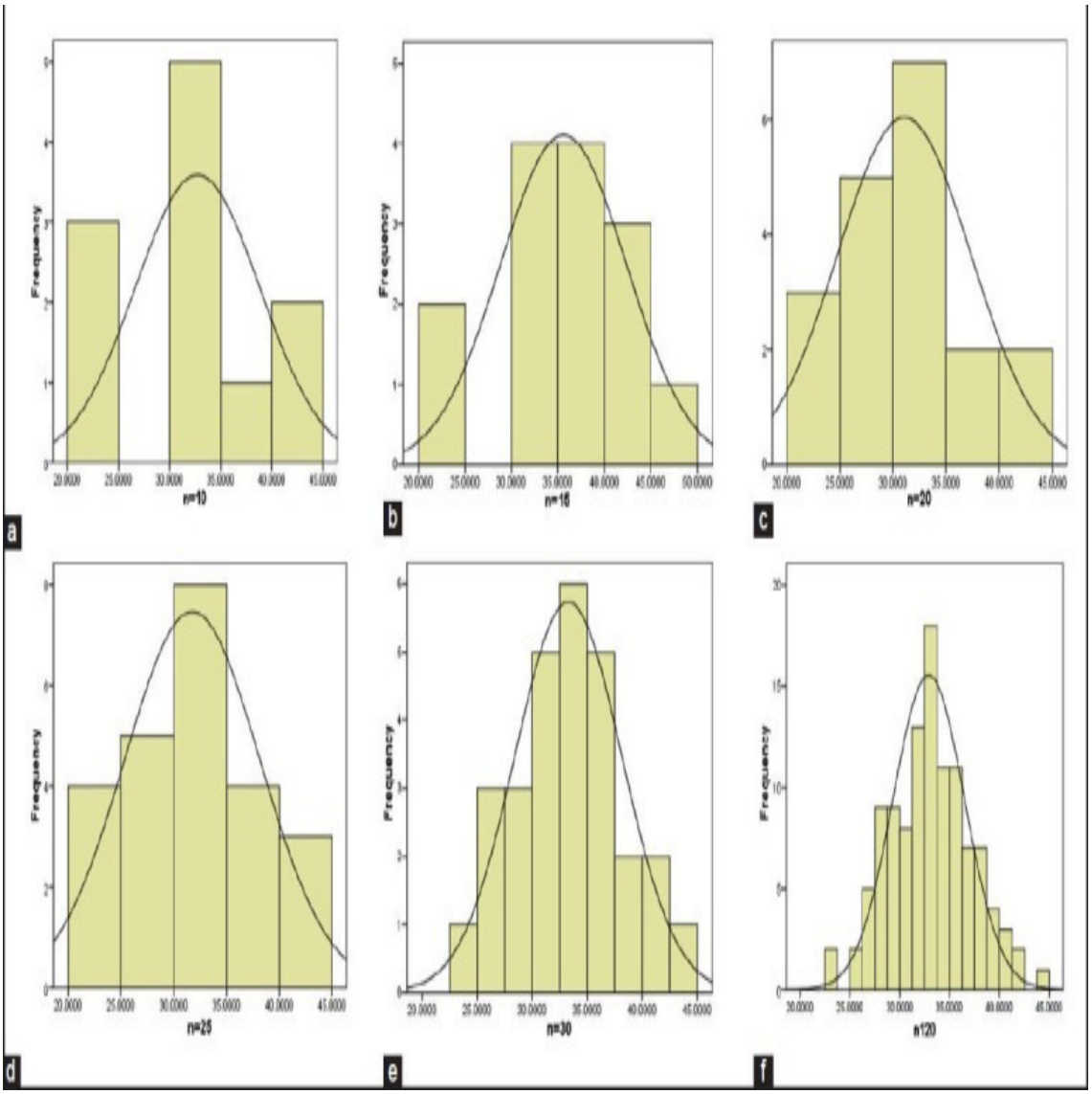
Pearsons Sk_2 mäter skevhet av en datamängd (=ett antal stickprov) och är given av

$$Sk_2 = 3 \cdot \frac{\bar{x} - \text{median(x)}}{s}.$$

Vilket matematiskt uttryck svarar mot Sk_2 i en population?

Problem 7.3.5.

Vilket samband mellan stickprov med växande storlek n och en population (existerande eller konceptuell) framställs i figuren?



Problem 7.3.6.

Consider the first four positive digits

$$\Omega = \{1, 2, 3, 4\}$$

and regard Ω as the population. The population parameter is the mean

$$\mu = \frac{1}{4}(1 + 2 + 3 + 4) = 2.5.$$

μ is the expectation of the uniform distribution on Ω.

a) A random sample of size two is to be selected from this population without replacement and taking into account the order. There are twelve (how is this number obtained?) possible samples:

$$\begin{array}{lll} 1,2 & 1,3 & 1,4 \\ 2,1 & 2,3 & 2,4 \\ 3,1 & 3,2 & 3,4 \\ 4,1 & 4,2 & 4,3 \end{array}$$

Compute the sample mean $\bar{x}$ of each of these possible samples. **Find** the *sampling distribution* of $\bar{x}$, i.e, the probabilities of the values of $\bar{x}$.

b) A random sample of size two is to be selected from the same population but with replacement and taking into account the order. **How** many possible different samples are there now? **Compute** the sample mean $\bar{x}$ of each of these possible samples. **Find** the sampling distribution of $\bar{x}$, i.e., the probabilities of the values of $\bar{x}$.

c) In what ways are the two sampling distributions in **a)** and **b)** different? In what ways are the two sampling distributions in **a)** and **b)** similar ? What do they reveal about the underlying population parameter?

d) Use, e.g., the Matlab command

```
>> x=unidrnd(4,[1000,2]);
```

to take thousand samples of size two as in **c)** and plot the bar graph of relative frequencies of the values of $\bar{x}$. **What** do we observe here?

Kapitel 8

Data-analys Del IV: Modellbaserad data-analys

8.1 Maximumlikelihood och momentskattning

Maximum–likelihood–metoden Låt x_i vara oberoende observationer på X_i, $i = 1, 2, ..., n$, där fördelningen för X_i beror på en okänd parameter θ. Det värde $\widehat{\theta}_{mle}$ som maximerar likelihoodfunktionen $L(\theta)$ (diskreta data)

$$L(\theta) = p_X(x_1; \theta) \cdots p_X(x_n; \theta)$$

(kontinuerliga data)

$$L(\theta) = f_X(x_1; \theta) \cdots f_X(x_n; \theta)$$

kallas **maximum–likelihood–skattningen (ML–skattningen)** av θ.

Momentmetoden Låt x_i vara oberoende observationer på X_i, $i = 1, 2, ..., n$, där fördelningen för X_i beror på en okänd parameter θ. Momentmetoden är en metod för estimering av θ. Vi behöver en ekvation som relaterar ett populationsmoment (t.ex., väntevärde) till θ, som vi vill skatta, t.ex

$$E[X] = g(\theta)$$

Med observationerna tar ett stickprovsmoment den okända populationsmomentets plats och ekvationen löses m.a.p. θ, t.ex.

$$\bar{x} = g(\theta) \Leftrightarrow \widehat{\theta}_{mm} = g^{-1}(\bar{x}).$$

vilket ger momentskattningen $\widehat{\theta}_{mm}$.

Problem 8.1.1.

För ett växande antal observationer i ett slumpmässigt stickprov är ett och endast ett av följande påståenden garanterat sant:

a) Sannolikhetsfördelningen för observationerna i stickprovet är approximativt normalfördelad.

b) Variansen för observationerna i stickprovet kommer att minska.

c) Medelvärdet för observationerna kommer att närma sig ett populationsmedelvärde.

d) Medelvärdet för observationerna kommer att bli approximativt normalfördelat.

Problem 8.1.2.

Antalet anrop X till en stordator under den brådaste timmen på dagen är $\mathcal{P}oi(\lambda)$. Under n dagar har man registerat $x_1, \ldots, x_n$ anrop.

a) **Använd maximumlikelihoodmetoden** för att härleda $\widehat{\lambda}_{mle}$ av λ.

b) **Använd momentmetoden** för att härleda $\widehat{\lambda}_{mm}$ av λ.

c) **Skatta** λ då man under 8 dagar fått följande antal anrop:

115 82 108 106 118 87 99 92

Problem 8.1.3.

Låt $(x_1, x_2, \ldots, x_m)$ vara ett stickprov från en binomialfördelning $\mathcal{B}in(n, p)$.
Antag att parametern n är känd medan p är okänd.

a) Använd maximumlikelihood-metoden för att härleda en punktskattare $\widehat{p}_{mle}$ av p.

b) Antag att $n = 10$, $m = 5$ och vi observerar $x_1, \ldots, x_5$ till 4, 7, 6, 9, 6. Beräkna värdet på $\widehat{p}_{mle}$ samt dess standardfel.

Problem 8.1.4.

Under n skilda tidsintervall med längder $t_1, t_2, \ldots, t_n$ har man registerat antalet partiklar
$x_1, x_2, \ldots, x_n$ från ett radioaktivt preparat. x_i kan ses som en observation på en stokastisk variabel som är $\mathcal{P}o(\lambda \cdot t_i), \qquad i = 1, 2, \ldots, n.$

a) **Härled** maximumlikelihoodskattaren av λ.

b) **Beräkna** maximumlikelihoodskattaren då $t_1 = t_2 = 10$ min. och $t_3 = 50$ min. och man registerat respektive 122, 137, 761 pariklar.

Problem 8.1.5.

Man har gjort n mätningar $x_1, \ldots, x_n$ av en fysikalisk storhet. Pga mätfel så kan dessa betraktas som oberoende observationer på en stokastisk variabel som är $\mathcal{N}(\mu, 1)$

Ta fram maximumlikelihoodskattaren av μ.

Problem 8.1.6.

Ett mätinstrument ger mätfel som är $\mathcal{N}(0, \sigma^2)$. Man gör n mätningar av kända storheter och får mätfelen $x_1, \ldots, x_n$.

a) **Härled** maximumlikelihoodskattaren $\widehat{\sigma}_{mle}$ av σ.

b) **Beräkna** $\widehat{\sigma}_{mle}$ då man fått mätfelen

$$1.2 \cdot 10^{-3} \quad -0.7 \cdot 10^{-3} \quad 3.1 \cdot 10^{-3} \quad 2.0 \cdot 10^{-3} \quad -1.8 \cdot 10^{-3} \quad -2.2 \cdot 10^{-3} \quad 0.2 \cdot 10^{-3} \quad -1.9 \cdot 10^{-3}$$

Problem 8.1.7.

Låt $(x_1, \ldots, x_n)$ vara observerade kölängder i ett kösystem. Vi antar att observationerna kommer från respektive oberoende stokastiska variabler $(X_1, \ldots, X_n)$. Varje $X_i \sim \mathcal{G}e(p)$, där p är okänd.

a) Ta fram maximum likelihood-skattningen av p.

b) Ta fram momentmetod-skattningen av p.

c) Diskutera maximum likelihood och momentmetod, när $\bar{x} = 0$ d.v.s., om $x_1 = \ldots = x_n = 0$.

Problem 8.1.8.

I molekylär evolutionsteori (fylogenetik, phylogenetics) mäter man ofta avståndet d mellan två arter (species) med hjälp av s.k. **Jukes - Cantor avstånd** (Jukes-Cantor distance). Här gör man en parvis jämförelse av n baser i DNA av de två arterna i en domän efter s.k. alignment av dessa två sekvenser. Därefter räknar man antalet gånger dessa sekvenser skiljer sig från varandra i domänen.
Låt X = antalet positioner bland n, där de två sekvenserna skiljer sig från varandra. Jukes och Cantors matematiska/statistiska modell för evolution säger att

$$X \sim \mathcal{B}in\left(n, \frac{3}{4}\left(1 - e^{-\frac{4d}{3}}\right)\right).$$

Här är d = Jukes - Cantor avståndet.

Bestäm en skattning av d med **maximumlikelihoodmetoden**, om Du observerar k skillnader (mutationer) i en domän av längd n.

8.2 Intervallskattning

$$\bar{x} \pm \lambda_{\alpha/2}\sigma/\sqrt{n}$$

är ett *konfidensintervall* för μ med *konfidensgrad* $1-\alpha$.

σ är oftast okänt och då fungerar inte ovanstående, konfidensintervallet skall ju kunna beräknas då data är kända. Vi får då använda t-fördelningen istället. På ungefär samma sätt som ovan kan man nämligen visa följande.

$$\bar{x} \pm t_{\alpha/2}(n-1)s/\sqrt{n}$$

är ett *konfidensintervall* för μ med *konfidensgrad* $1-\alpha$. $t_{\alpha/2}(n-1)$ är $\alpha/2$-fraktilen för en t-fördelning med $n-1$ frihetsgrader. s är standardavvikelsen för stickprovet, datamängden.

Problem 8.2.1.

Den stokastiska variabeln X är $\sim \mathcal{N}(0,1)$.

Bestäm a och b så att $P(|X| \leq a) = 95\%$ och $P(|X| \leq b) = 99\%$

Problem 8.2.2.

Den stokastiska variabeln X är χ^2-fördelad med 10 frihetsgrader.

Bestäm a, b, c så att $P(X < a) = 95\%$ och $P(b < X < c) = 95\%$.

Problem 8.2.3.

Den stokastiska variabeln X är t-fördelad med f frihetsgrader.

a) **Bestäm** för $f = 9$ tal a, b så att $P(|X| \leq a) = 99\%$ och $P(X > b) = 5\%$

b) Låt $t_\alpha(f)$ satisfiera $P(X > t_\alpha(f)) = \alpha$, där $X \sim t(f)$.

Kolla med mjukvara eller räknedosa eller tabell att $t_\alpha(f) \to \lambda_\alpha$ då $f \to \infty$.

Hur stort behöver f vara för att $t_{0.05}(f) = \lambda_{0.05}$ om båda talen avrundas till en decimal?

Problem 8.2.4.

Vad menas med att konfidensintervallet för väntevärdet μ har konfidensgraden 95 % ? Pricka för de korrekta svaren.

a) I det långa loppet innehåller intervallet μ i 95 % av försöken.

b) I genomsnitt över många försök innehåller intervallet 95 % av observationerna.

c) Minst 95 % av observationerna faller alltid inom intervallet.

d) Sett före försöket är det 95 % chans att intervallet kommer att hamna så att det innehåller μ.

CI for population proportion (andel)
The standard error $d(\widehat{p})$ of the proportion estimator $\widehat{p}_{mle}$ ($\widehat{q}_{mle} = 1 - \widehat{p}_{mle}$) is

$$d(\widehat{p}_{mle}) \stackrel{\text{def}}{=} \sqrt{\frac{\widehat{p}_{mle}\widehat{q}_{mle}}{n}}$$

Since $\lambda_{\alpha/2}$ separates an area of $\alpha/2$ in the right tail of the standard normal distribution, we find $\alpha/2$ such that

$$\Phi\left(\lambda_{\alpha/2}\right) = 1 - \alpha/2$$

The **margin of error** E of the proportion estimator $\widehat{p}$ is

$$E \stackrel{\text{def}}{=} \lambda_{\alpha/2}\sqrt{\frac{\widehat{p}_{mle}\widehat{q}_{mle}}{n}}$$

Confidence Interval or the Interval Estimate for the population proportion (andel) p with degree of confidence $1 - \alpha$

$$\widehat{p}_{mle} - E < p < \widehat{p}_{mle} + E, \quad \text{where } E = \lambda_{\alpha/2}\sqrt{\frac{\widehat{p}\widehat{q}}{n}}$$

Other equivalent expressions are

$$\widehat{p}_{mle} \pm E$$

or

$$(\widehat{p}_{mle} - E, \widehat{p} + E)$$

Problem 8.2.5.

You have a confidence interval $(m - E, m + E)$ for a population mean μ with confidence level 0.95 %. Which of the following statements is correct ?

a) The probability that μ is in $(m - E, m + E)$ is 0.95.

b) The proportion of times is 0.95 that the confidence interval actually does contain μ, assuming the process is repeated a large number of times.

c) 95 % of your samples must lie in $(m - E, m + E)$, when the process is repeated a large number of times.

d) 95 % of your sample means must lie in $(m - E, m + E)$, when the process is repeated a large number of times.

Problem 8.2.6.

I en maskin för sekvensering av DNA uppstår oberoende avläsningsfel med felsannolikheten p.

a) Antag att p = 0.02 och att ett 100 nukleotider långt DNA-fragment ska sekvenseras. Låt X vara det totala antalet fel. Bestäm fördelningen för X och beräkna därefter väntevärde och standardavvikelse.

b) Vid sekvensering av ett 5000 nukleotider långt DNA-fragment identifierades 79 fel. Uppskatta felsannolikheten p med hjälp av momentmetoden.

c) Beräkna därefter ett approximativt dubbelsidigt konfidensintervall med konfidensgrad 99%. Kan vi utesluta att $p = 0.02$?

Problem 8.2.7.

Koncentrationen av DNA i ett prov mäts med hjälp av en spektrofotometer. Eftersom mätningen är osäker upprepas den 10 gånger $(x_1, \ldots, x_{10})$, där medelvärdet beräknas till $\bar{x} = 4.87$ (μg/mL). Antag att observationerna från mätningen är oberoende och slumpmässigt dragna från en normalfördelning med okänt väntevärde μ och känd varians $\sigma^2 = 4$.

a) Beräkna ett dubbelsidigt konfidensintervall för väntevärdet av mängden DNA (μ). Konfidensgraden ska vara 99 %.

b) Om du ändrar stickprovsstorleken från 10, hur många mätningar måste göras om längden på konfidensintervallet inte ska överstiga 1 (μg/mL)?

Problem 8.2.8.

För att undersöka effekten av ett rostskyddsmedel gör en person 20 mätningar. Dessa mätningar uppfattas som utfall av oberoende $\mathcal{N}(\mu, \sigma^2)$-fördelade stokastiska variabler där μ och σ är okända. Din vän, kemisten, gör, efter konstens alla regler, ett exakt 95%-igt konfidensintervall för μ och erhåller intervallet $I_\mu = (30.24, 31.96)$.

a) Tyvärr har de ursprungliga mätresultaten kommit bort när Din vän får i uppdrag att även beräkna variationskoefficienten $s/\bar{x}$, där s och $\bar{x}$ är standardavvikelse respektive medelvärde av data. Hen vill inte erkänna sitt slarv, utan ber sin statistiskt kunnige vän teknologen (dvs Dig) om hjälp. **Ge henom den hjälpen**, dvs beräkna denna variationskoefficient.

b) **Gör** en lämplig skattning av sannolikheten att en kommande observation av rostskyddseffekten kommer att vara mindre än 30.0.

Problem 8.2.9.

Vid en undersökning av njurvolymen hos 10 möss erhölls följande värden:

1.01 1.15 1.17 0.96 0.91 1.09 1.17 0.87 0.91 1.06

som uppfattas som oberoende observationer på en stokastisk variabel X.

Man kan anta att X är **log-normalfördelad**, dvs att $\ln X$ är normalfördelad, säg $N(m,\ \sigma^2)$.

Beräkna ett 95 % konfidensintervall för parametern e^m.

Problem 8.2.10.

Copy number variations and Digital PCR with nanofluidic biochip

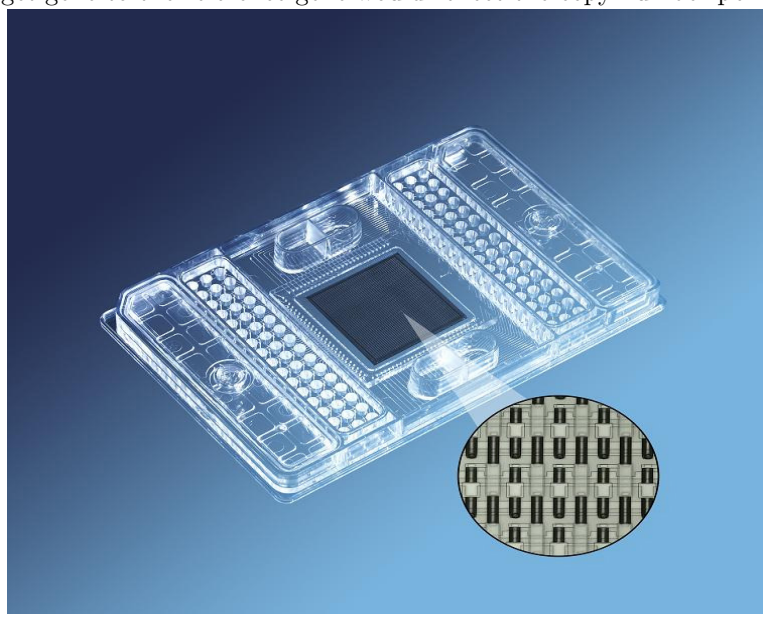
Simant Dube, Jian Qin, Ramesh Ramakrishnan: *Mathematical Analysis of Copy Number Variation in a DNA Sample Using Digital PCR on a Nanofluidic Device*, **PLoS ONE**, 3(8), 2008.

Copy number variation (CNV) is a phenomenon in which regions of the genome are repeated and the number of repeats in the genome varies between individuals in the human population. Copy number variation is a type of duplication or deletion event that affects a considerable number of base pairs. Perhaps the best-defined and most widely known CNVs are the trinucleotide repeats, which consist of three nucleotides repeating in tandem. Recent research indicates that approximately two thirds of the entire human genome is composed of repeats and 4.8-9.5% of the human genome can be classified as copy number variations. CNVs are often linked to genetic diseases in humans.

Digital PCR conventionally utilizes sequential limiting dilutions of target DNA, followed by amplification using the polymerase chain reaction (PCR). As a result, it is possible to quantitate single DNA target molecules. We utilize the digital array, which is a novel nanofluidic biochip, where digital PCR reactions can be performed by partitioning DNA molecules, instead of diluting them. This chip utilizes integrated channels and valves that partition mixtures of sample and reagents into 765 nanolitre volume reaction chambers. DNA molecules in each mixture are randomly partitioned into the 765 chambers of each panel (the total volume of the PCR mix in each panel: 6 nl$\times$ 765=4.59 μl). The chip is then thermocycled and imaged on Fluidigm's BioMark real-time

PCR system and the positive chambers that originally contained 1 or more molecules can be counted by the digital array analysis software.

CNV determination on the digital array is based upon its ability to partition DNA sequences. Given the number of molecules per panel and the dilution factor, the concentration of the target sequence in a DNA sample can be accurately calculated. In a multiplex PCR reaction with 2 or more assays, multiple genes can be quantitated simultaneously and independently, effectively eliminating any pipetting errors if separate reactions have to be set up for different genes. When a single copy reference gene is used in the reaction, the ratio of the target gene to the reference gene would reflect the copy number per haploid genome of the target gene.



Here we consider a mathematical framework to calculate the true concentration of molecules from the observed positive reactions in a panel. We show how to perform statistical analysis to find the 95 % confidence intervals of the true concentrations in a CNV experiment using the digital array with multiplex PCR.

The copy number variation problem can be stated as follows. Given two counts h_1 and h_2 of positive chambers for two genes in a digital array panel, how can one estimate a ratio of true concentrations $r = \frac{\lambda_1}{\lambda_2}$ of the two genes and a confidence interval $[r_{Low}, r_{High}]$ on the estimation?

The problem has two parts.

i) Given a count h of positive chambers, how can one estimate the true concentration λ_1 of target molecules in the DNA sample and a confidence interval $[\lambda_{Low}, \lambda_{High}]$ on this estimation?

ii) Given estimated true concentrations λ_1 and λ_2 of the reference gene and the target gene, respectively, in the DNA sample and their respective confidence intervals, how can one estimate the ratio $r = \frac{\lambda_1}{\lambda_2}$ and a confidence interval $[r_{Low}, r_{High}]$ on this estimation?

The first question can be answered by applying statistics and probability, and the second question can be answered by a numerical algorithm based on generalization of a mathematical theorem.

We are interested in estimating the true concentration of the molecules in the DNA sample from which we extracted 6 nl$\times$765=4.59 μl of sample for each panel.

Consider a conceptual population or the universe of infinite number of the digital array chambers filled with an infinite amount of the DNA sample where the true concentration of the target molecules is λ per chamber (per 6 nl). The true concentration is an unknown population parameter of this infinite DNA sample. If a chamber gets no molecule then it constitutes *failure* in the sense of Bernoulli random variables. If it gets one or more molecules, that is, if it gets a hit and is therefore positive, then it constitutes *success*. Let the probability of success be p. Note that p *is an unknown population parameter.* $\widehat{p}_{mle}$ will denote the estimator of p.

a) There are M molecules independently and uniformly distributed in C chambers. $M = \lambda C$. The probability of a molecule being in any given chamber is $\frac{1}{C}$. Show that if p= the probability of a given chamber having

at least one molecule, then

$$1 - p = \left(1 - \frac{\lambda}{M}\right)^M$$

b) M is very large. Mathematically speaking, let $M \to +\infty$, and show that

$$\lambda = -\ln(1-p).$$

c) With $\widehat{p}_{mle}$ we estimate λ by $\widehat{\lambda} = -\ln(1-\widehat{p}_{mle})$. Give now a 95 % confidence interval $[\lambda_{Low}, \lambda_{High}]$ for λ and justify your answer.

The rest of the questions in **i)** and **ii)** are left for the interested.

Problem 8.2.11.

Rule of three in statistics
Konfidensintervall för population proportion p med konfidensgrad $1-\alpha$ är

$$\widehat{p}_{mle} - E < p < \widehat{p}_{mle} + E, \quad \text{where } E = \lambda_{\alpha/2}\sqrt{\frac{\widehat{p}_{mle}\widehat{q}_{mle}}{n}}$$

där $\widehat{p}_{mle}$ är maximumlikelihodskattningen av p och $\widehat{q}_{mle} = 1 - \widehat{p}_{mle}$. Vi har binomialfördelade data. Alternativa uttryck för intervallet är

$$\widehat{p}_{mle} \pm E \qquad (\widehat{p}_{mle} - E, \widehat{p}_{mle} + E).$$

a) Vad går fel om $\widehat{p}_{mle} = 0$, d.v.s. Du har observerat noll lyckade utfall bland n utfall?

b) Förklara varför intervallet

$$(0, 3/n)$$

är ett approximativt konfidensinterval för p med konfidensgraden 95%, när Du har observerat noll lyckade utfall bland n utfall. Detta resultat kallas '**rule of three in statistics**' eller **Hanleys formel**.
Ledning Det gäller för $X \sim \mathcal{B}in(n,p)$ att $P(X=0) = (1-p)^n$.

c) I en dansk studie upptäcker man man att av 3119 kvinnliga och 1856 manliga medlemmar av danska sjundedagsadventistkyrkan har ingen diagnosticerats med gonorré under 1975-2009 enligt danska nationella patientregistret. Utifrån statistiken för hela danska befolkningen hade man väntat sig 3.4 fall av gonorré. Vad är den statistiska slutsatsen?

> Kørup AK, Thygesen LC, Christensen R, Johansen C, Søndergaard J, Hvidt NC: *Association between sexually transmitted disease and church membership. A retrospective cohort study of two Danish religious minorities*, **BMJ Open** 2016 Mar 25 (BMJ är en förkortning för British Medical Journal).

d) If nothing goes wrong, is everything all right? Hur bör rule-of-three tolkas?

Charles Darwin har uttryckt i ett av sina brev[1] att han har 'no Faith in anything short of actual measurement and the Rule of Three". Vad kan han ha menat med detta?

Problem 8.2.12.

Lomax-fördelningen används för att beskriva livslängden hos komponenter av olika slag och beskrevs ursprungligen av statistikern K.S. Lomax 1954. Täthetsfunktion för Lomax-fördelning är

$$f(x;\alpha) = \alpha(1+x)^{-(\alpha+1)}, x \geq 0$$

där $\alpha > 1$ är en okänd parameter.

[1] quoted from O.B. Sheynin: On the History of the Statistical Method in Biology. *Archive for History of Exact Sciences*, vol. 22, 1980 p. 342

a) Använd **maximumlikelihoodmetoden** för att härleda en punktskattare $\widehat{\alpha}_{mle}$ för α.

b) Väntevärdet för Lomax- fördelningen är

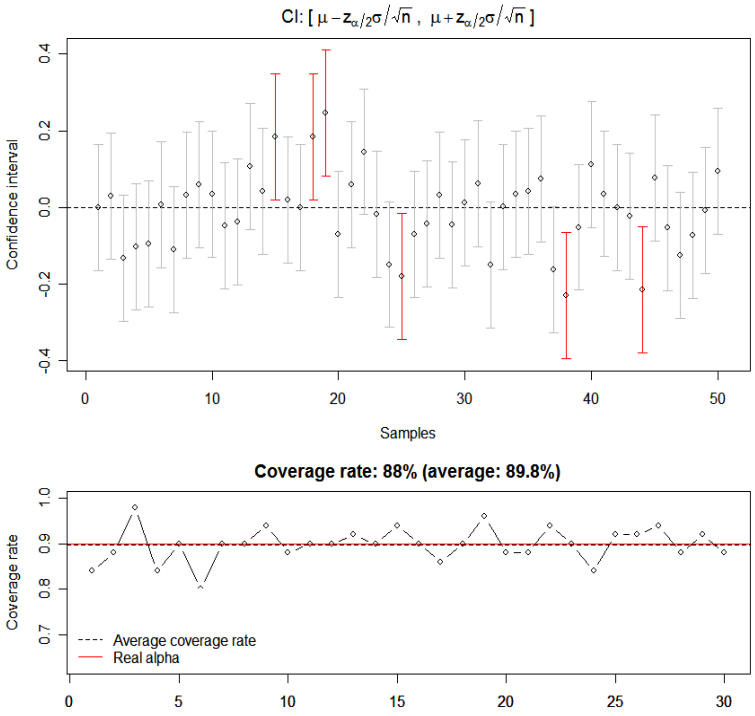
$$E(X) = \frac{1}{\alpha - 1}, \quad \text{om } \alpha > 1, \text{ annars odefinierad}$$

Använd **momentmetoden** för att härleda en punktskattare $\widehat{\alpha}_{mm}$ för α.

c) Låt 0.31, 0.59, 1.01, 0.17, 0.54 vara ett stickprov från en Lomax- fördelning med okänd parameter α. Punktskatta α genom att beräkna $\widehat{\alpha}_{mm}$ och $\widehat{\alpha}_{mle}$.

Problem 8.2.13.

I figuren ser Du 50 konfidensintervall, beräknade på basis av 50 olika datamängder med n stickprov i var, för ett väntevärde (vars sanna värde = 0) med känd varians. Läs $z_{\alpha/2}$ som $\lambda_{\alpha/2}$ och konfidensgraden som $1 - \alpha = 0.90$ (ej = α som i bilden). Vilken viktig egenskap hos processen att bilda konfidensintervall framgår av figuren?



8.3 Konfidensintervall för skillnad mellan väntevärden

8.3.1 Jämförelse mellan två väntevärden: formler

Antag att vi nu har två stickprov från två normalfördelningar, $x_1, x_2, \ldots, x_{n_1}$ respektive $y_1, y_2, \ldots, y_{n_2}$. De två stickproven antas vara $\mathcal{N}(\mu_1, \sigma_1^2)$ respektive $\mathcal{N}(\mu_2, \sigma_2^2)$ och oberoende.

Kända varianser

Ett konfidensintervall för skillnaden $\mu_1 - \mu_2$ ges av

$$\bar{x} - \bar{y} \pm \lambda_{\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$$

Okända men lika varianser

Vi antar nu att σ_1 och σ_2 är okända men lika. Den bästa skattningen av denna okända varians ges av

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

där s_1^2 och s_2^2 är stickprovsvarianserna för de två stickproven. Man kan då visa att

$$t = \frac{(\bar{x}-\bar{y}) - (\mu_1-\mu_2)}{s\sqrt{1/n_1 + 1/n_2}} \tag{8.1}$$

är t-fördelad med n_1+n_2-2 frihetsgrader. Av detta följer att ett konfidensintervall för skillnaden $\mu_1-\mu_2$ ges av

$$\bar{x}-\bar{y} \pm t_{\alpha/2}(n_1+n_2-2)s\sqrt{1/n_1+1/n_2}. \tag{8.2}$$

Parade observationer (matched pairs)

Vi vill jämföra två metoder på ett antal prov som har helt *olika* värden. På varje prov görs en analys med var och en av metoderna. Vi antar att skillnaden mellan de två metodernas resultat har samma förväntade värde, oavsett prov. Data:

Prov	1	2	3	...	n
Metod 1	x_1	x_2	x_3	...	x_n
Metod 2	y_1	y_2	y_3	...	y_n
Skillnad	z_1	z_2	z_3	...	z_n

Här är alltså $z_i = x_i - y_i, i = 1, 2, \ldots, n$.

Antag att den konstanta förväntade skillnaden är Δ. Det är då inte så svårt att visa att z-observationerna är normalfördelade med förväntat värde Δ, och att testa om metoderna är likvärdiga, d.v.s. om $\Delta = 0$, kan vi göra utgående från z-data. Vi har återfört problemet till fallet ett stickprov. Konfidensintervallet: $\bar{z} \pm t_{\alpha/2}(n-1)s/\sqrt{n}$ eller om $\mid t \mid > t_{\alpha/2}(n-1)$ där $t = \frac{\bar{z}}{s_z/\sqrt{n}}$.

Problem 8.3.1.

För att undersöka effekten av ett rostskyddsmedel behandlade man 10 järnstavar med detta. På var och en av 10 olika platser grävdes därefter en av de behandlade stavarna jämte en obehandlad stav ner. Efter 3 månader togs alla stavar upp och rostgraden mättes.

Resultat (i lämplig enhet):

Plats	1	2	3	4	5	6	7	8	9	10
Obehandlade	32.3	38.0	40.1	28.4	35.9	36.3	25.1	28.2	39.8	32.6
Behandlade	31.5	37.5	40.2	28.0	34.8	36.0	25.1	27.5	39.1	32.4

Resultatet från plats nr. k för obehandlad resp. behandlad stav är oberoende observationer på $\mathcal{N}(\mu_k,\ \sigma^2_{\text{Obeh.}})$- resp. $N(\mu_k+\Delta,\ \sigma^2_{\text{Beh.}})$-fördelade stokastiska variabler. Δ (eller egentligen $-\Delta$) är ett mått på rostskyddsmedlets effekt.

a) Beräkna ett 95 % konfidensintervall för Δ.

b) Anser Du att rostskyddsmedlet har effekt? Motivera utgående från Ditt resultat på a)-delen!

Problem 8.3.2.

Höga koncentrationer av det hormonstörande ämnet etinylestradiol är skadliga för fiskar och andra vattenlevande organismer. Vid en miljöriskbedömning mättes koncentrationen av etinylestradiol i vattenprover tagna från 10 förorenade $(x_1, \ldots, x_{10}\ )$ områden och 12 rena $(y_1, \ldots, y_{12}\ )$ områden längs den svenska västkusten. Medelvärden och standardavvikelser för de två stickproven beräknades till $\bar{x} = 23.9$, s_x= 18.2, $\bar{y} = 4.22$, s_y = 2.63 (mätvärdena är i nanogram etinylestradiol per liter vatten). Mätvärdena kan antas vara oberoende och normalfördelade med lika varianser.

Beräkna ett approximativt dubbelsidigt konfidensintervall for den sanna skillnaden i koncentration av etinylestradiol $\mu_X - \mu_Y$ mellan de två stickproven. Konfidensgraden ska vara 0.99.

Tolka konfidensintervallet. Finns det en högre koncentration etinylestradiol i de förorenade områdena?

Problem 8.3.3.

I ett laboratorium ville man bestämma vikten av en viss vätskemängd. Man gjorde därför först 4 vägningar av en tom kolv, fyllde därefter vätskan i kolven och gjorde 4 vägningar av kolven med vätska. Resultatet blev:

Vikt utan vätska	15.03	15.05	15.03	15.06
Vikt med vätska	26.66	26.65	26.64	26.66

Beräkna ett 95 % konfidensintervall för vikten av vätskemängden.

Problem 8.3.4.

Till ett laboratorium har inkommit två stora kemikaliepartier. Man har tagit n_1 prover från det första och n_2 från det andra och bestämt halten av ett visst ämne i proverna. Mätvärdena $x_1, \ldots, x_{n_1}$ är observationer på X som är $\mathcal{N}(\mu_1, \sigma)$ och $y_1, \ldots, y_{n_2}$ är observationer på Y som är $\mathcal{N}(\mu_2, \sigma)$. Här är μ_1, μ_2 och σ okända tal.

a) Härled ett konfidensintervall för medelhalten $\frac{\mu_1+\mu_2}{2}$

b) Beräkna ett 90 % konfidensintervall då

$$\begin{array}{lll} \bar{x} = 24.2 & s_x = 1.6 & n_1 = 5 \\ \bar{y} = 28.5 & s_y = 2.3 & n_2 = 10 \end{array}$$

Problem 8.3.5.

a) För att undersöka om en viss medicin har bieffekten att höja blodtrycket mättes blodtrycket dels på 50 personer som ej behandlats med medicinen (mätvärden $x_1, \ldots, x_{50}$), dels på 25 patienter som behandlats med medicinen (mätvärden $y_1, \ldots, y_{25}$). Man erhöll

$$\begin{array}{ll} \bar{x} = 148.2 & \bar{y} = 151.7 \\ s_x = 10.0 & s_y = 8.0 \end{array}$$

Bestäm ett 95 % konfidensintervall för skillnaden mellan förväntat blodtryck för de två grupperna. **Ange** alla antaganden om fördelning och oberoende.

b) Resultatet av undersökningen i **a)** blev dåligt såtillvida att konfidensintervallet blev alldeles för brett för att man skulle kunna dra några intressanta slutsatser. En konsulterad statistiker föreslog att man skulle göra ett nytt försök, i vilket man mätte blodtrycket före och efter behandling på 25 patienter (mätvärden x_i resp. $y_i,\ i = 1, \ldots, 25$). Man erhöll

$$\begin{array}{ll} \bar{x} = 149.0 & \bar{y} = 150.9 \\ s_z = 1.6 & \text{där } z_i = y_i - x_i,\ i = 1, \ldots, 25 \\ s_x = 8.1 & s_y = 9.5 \end{array}$$

Bestäm ett 95 % konfidensintervall för skillnaden mellan förväntat blodtryck före och efter behandlingen. **Ange** noga alla antaganden om fördelning och oberoende.

Problem 8.3.6.

Vid ett jordbruksförsök vill man veta om en ny vetesort ger bättre skörd än den existerande sorten. Var och en av 6 åkrar delades i två lika stora delar, där var sin sort odlas. Vilket sort, som odlas på vilken del av åkern, lottas ut. åkrarna skiljer sig i bördighet och klimat! Följande data erhölls (enhet: vikt/ytenhet) :

åker nr	1	2	3	4	5	6
Skörd sort 1	723	872	612	635	664	771
Skörd sort 2	752	891	653	650	690	802

Hur mycket större (systematisk) skörd ger sort 2 jämfört med sort 1? Ange detta i form av ett lämpligt konfidensintervall. Sedvanliga normalfördelningsantaganden och antaganden om oberoende kan göras och konfidensgraden väljes till 95%.

Problem 8.3.7.

Ett läkemedelsföretags forskningsavdelning har tagit fram ett nytt preparat avsett att sänka halten i blodet av det s.k. dåliga kolesterolet (LDL). Ett kliniskt försök skall genomföras för att avgöra om preparatet har någon effekt.

a) Antag att försöket utförs på följande sätt. Av 30 slumpvis utvalda personer behandlas 15 med preparatet. De återstående 15 utgör en kontrollgrupp och behandlas ej. Halten av LDL i blodet mäts därefter för samtliga personer. Mätvärdena för den obehandlade resp. behandlade gruppen är $x_1, \ldots, x_{15}$ resp. $y_1, \ldots, y_{15}$, och antas vara observationer av oberoende stokastiska variabler $X_1, \ldots, X_{15}$ resp. $Y_1, \ldots, Y_{15}$ sådana att $X_i \sim \mathcal{N}(\mu, \sigma^2)$ och $Y_i \sim \mathcal{N}(\mu - \Delta, \sigma^2)$ för $i = 1, \ldots, 15$, där parametrarna μ, Δ och σ är okända.

Beräkna ett tvåsidigt 95% konfidensintervall för Δ. Numeriska värden: $\overline{x} = \frac{1}{15}\sum_{i=1}^{15} x_i = 5.122$, $\overline{y} = \frac{1}{15}\sum_{i=1}^{15} y_i = 4.713$, $\sum_{i=1}^{15}(x_i - \overline{x})^2 = 7.881$, och $\sum_{i=1}^{15}(y_i - \overline{y})^2 = 8.136$.

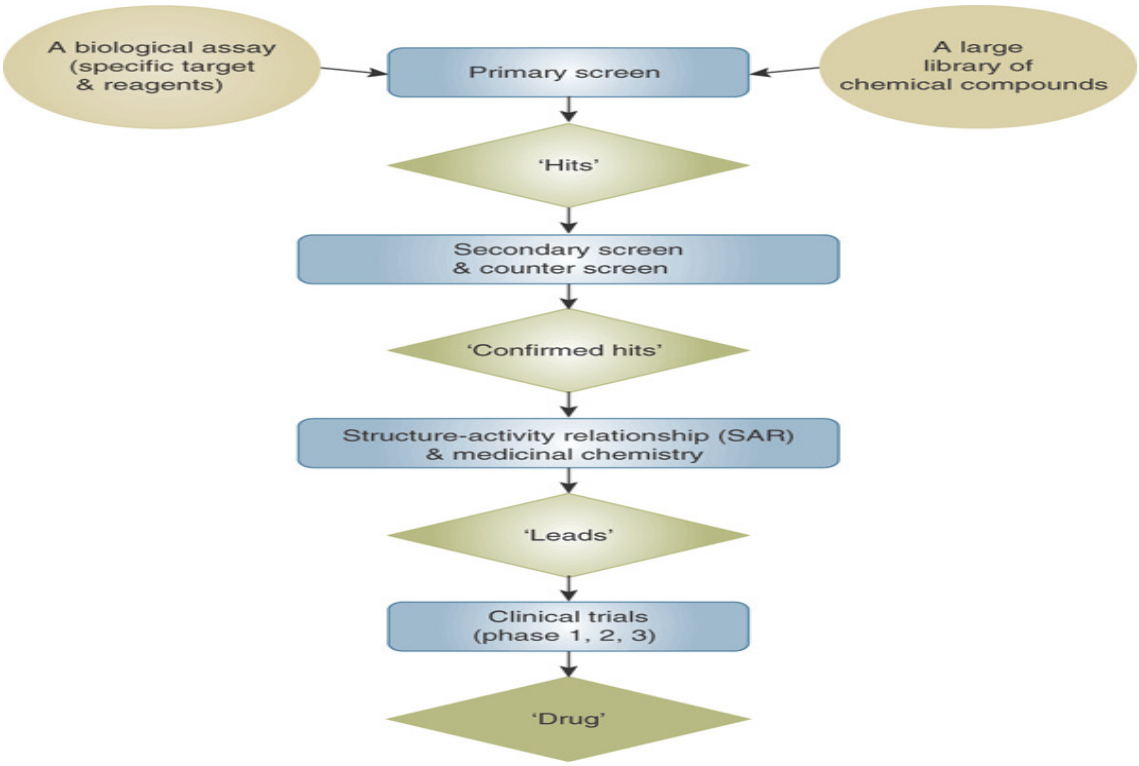
b) Försöksuppläggningen i **a)** är olämplig, därför att den förväntade halten av LDL i **blodet** troligen varierar med individen, beroende på ärftliga faktorer, diet m.m. Vi tar hänsyn till detta genom att i stället utföra försöket på följande sätt. Halten av LDL i blodet mäts hos 15 slumpvis utvalda personer *före* och *efter* behandling med preparatet. De observerade mätvärdena före resp. efter behandling är $x_1, \ldots, x_{15}$ resp. $y_1, \ldots, y_{15}$, och antas vara observationer av oberoende stokastiska variabler $X_1, \ldots, X_{15}$ resp. $Y_1, \ldots, Y_{15}$ sådana att $X_i \sim \mathcal{N}(\mu_i, \sigma^2)$ och $Y_i \sim \mathcal{N}(\mu_i - \Delta, \sigma^2)$ för $i = 1, \ldots, 15$, där parametrarna $\mu_1, \ldots, \mu_{15}$, Δ och σ är okända.

Beräkna ett tvåsidigt 95% konfidensintervall för Δ. Samma numeriska värden som i **a)** kan användas, med tillägget att $\sum_{i=1}^{15}(x_i - y_i - (\overline{x} - \overline{y}))^2 = 1.834$.

8.4 Statistical Parameters for High-Throughput Screening (HTS)

Next follows a mix of some lines of thought from Malo, Nathalie and Hanley, James A and Cerquozzi, Sonia and Pelletier, Jerry and Nadon, Robert: Statistical practice in high-throughput screening data analysis, *Nature biotechnology*, 24, 2, 167−−175,2006, with other statements:

> High-throughput screening (HTS) is the backbone of drug discovery within the pharmaceutical industry. The combination of robotic methods, parallel processing and miniaturization of biological assays has made it possible to quickly conduct millions of chemical, genetic, or pharmacological tests. Through this process one can rapidly identify active compounds, antibodies, or genes that modulate a particular biomolecular pathway.
>
> Screening is about making decisions on the modulating activity of one particular compound on a biological system. When a compound testing experiment is repeated under the same conditions or as close to the same conditions as possible, the observed results are never exactly the same, and there is an apparent random and uncontrolled source of variability in the system under study. Nevertheless, randomness is not haphazard: the usage of statistical tools in the analysis of screening experiments is the right approach to the interpretation of screening data, with the aim of making them meaningful and converting them into valuable information that supports sound decision making.



HTS is a large-scale process (see the Figure) that screens many thousands of chemical compounds in order to identify potential lead candidates rapidly and accurately. Whereas the plating format and number of compounds per plate can vary, typically just a single measurement of each compound's activity is obtained in an initial primary screen. The automated process allows the testing of several hundred plates over a period of weeks.

A compound with a desired size of effects in an HTS is called a **hit**. The process of selecting hits is called hit selection.

Two kinds of inference or decision error can occur at the primary screen step: 'false positives' and 'false negatives'. We advance the view that improving hit specificity and sensitivity cannot be met by technological and organizational improvements alone and that improvements in data analysis methods are needed to fulfill the promise of HTS.

The analytic methods for hit selection in screens without replicates (usually in primary screens) differ from those with replicates. For example, the **z-score** method is suitable for screens without replicates whereas the t-statistic is suitable for screens with replicates. The calculation of **SSMD = strictly standardized mean difference**, see below, for screens without replicates also differs from that for screens with replicates.

Hits are evaluated for biological relevance by a counter screen and confirmed as bona fide hits by a secondary screen.

Secondary screens test many fewer compounds (e.g., the 1% most active compounds from the primary screen and typically use at least duplicate measurements. Compounds with the highest measured activity levels on a primary screen will on average be less extreme on a secondary screen because of the familiar **statistical phenomenon known as 'regression toward the mean'**. Accordingly, marginal hits on the first run may fail to validate on the second run merely because of random measurement error, although the size of the statistical artifact can be minimized by improving measurement precision (e.g., by obtaining replicate measurements). Confirmed hits with an established biological activity according to a structure-activity relationship (SAR) series and medicinal chemistry are termed 'leads' that can develop into drug candidates for clinical testing.

In an HTS, quality control (QC) is of great importance. A QC characteristic in an HTS assay is how much

the positive controls, test compounds, and negative controls differ from one another in the assay.

XHD Zhang proposed **strictly standardized mean difference (SSMD)** to evaluate the difference between a positive control and a negative control in HTS assays in:

XHD Zhang: A pair of new statistical parameters for quality control in RNA interference high-throughput screening assays, *Genomics* 89 (4): 552−61, 2007.

A paper about SSMD by Zhang written for a statistically minded audience is:

Zhang XHD: Strictly standardized mean difference, standardized mean difference and classical t-test for the comparison of two groups. *Statistics in Biopharmaceutical Research* 2 (2), 292−99, 2010.

SSMD
As a statistical (population) parameter, SSMD (denoted as β) is defined as the ratio of mean to standard deviation of the difference of two populations, i.e., positive controls, and negative controls. If the populations are independent,

$$\beta \stackrel{\text{def}}{=} \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}.$$

SSMD has a probability foundation due to its link with the probability that the difference between the two sample means is positive.

Problem 8.4.1.

Let X_1 and X_2 are independent random variables. Assume that $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$.

Find

$$P(X_1 > X_2)$$

as a function of the SSMD population parameter β.

d^+ -probability is the name often used in biotechnical and pharmaceutical studies for a probability $P(X > Y)$.

Problem 8.4.2.

Let $X_1, X_2, \ldots, X_{n_1}$ and $Y_1, Y_2, \ldots, Y_{n_2}$ be independent random variables and mutually independent. Assume that $X_i \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y_i \sim \mathcal{N}(\mu_2, \sigma_2^2)$ for all i. Zhang has found the maximum likelihood estimate of β as

$$\hat{\beta} = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{n_1-1}{n_1}s_1^2 + \frac{n_2-1}{n_2}s_2^2}},$$

Zhang loc.cit has also established that $\hat{\beta}$ has an approximate normal distribution $\mathcal{N}(\beta, \sigma_{\hat{\beta}}^2)$ for large n_1 and n_2, where

$$\sigma_{\hat{\beta}}^2 \stackrel{\text{def}}{=} \frac{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}{\sigma_1^2 + \sigma_2^2} + \frac{\frac{\sigma_1^4}{n_1} + \frac{\sigma_2^4}{n_2}}{2(\sigma_1^2 + \sigma_2^2)^3} (\mu_1 - \mu_2)\,.$$

Find the confidence interval for the SSMD population parameter β with the approximate confidence level 0.95.

8.5 Normalfördelad linjär modell

Problem 8.5.1.

I följande tabell är y_i observerade värden på en stokastisk variabel

$$Y_i = \alpha + \beta x_i + \epsilon_i,$$

där $\epsilon_1, \ldots, \epsilon_n$ är oberoende och $\mathcal{N}(0, \sigma^2)$.

x_i	1	2	3	4	5	6	7	8	9	10
y_i	11.50	9.8	7.0	5.5	3.9	2.0	0.4	-1.4	-3.1	-5.0

a) Vilken fördelning har Y_i?

b) Beräkna punktskattare av α, β och σ med hjälp av minimering av

$$Q(\alpha, \beta) = \sum_{i=1}^{n} (y_i - E[Y_i])^2$$

som funktion av α och β. *Ledning:* $Q(\alpha, \beta)$ är i själva verket minsta-kvadrat kriteriet från tidigare.

c) Vad är skillnaden mellan residualen $e_i = y_i - (\widehat{\alpha} + \widehat{\beta} x_i)$ och ϵ_i?

d) Ge konfidensintervallen för α och β.

Kapitel 9

Hypotesprövning

9.1 Sammanfattning av teori: Konfidensmetoden för hypotesprövning

Konfidensmetoden
Antag att man vill undersöka om ett visst väntevärde μ (eller andel/proportion p) antar ett givet värde som vi kallar μ_0 (eller p_0). Bilda ett konfidensintervall med konfidensgrad $1-\alpha$ på vanligt sätt. Om nu μ_0 (eller p_0) ligger utanför detta intervall är det troligt att μ_0 (eller p_0) inte är det rätta värdet eftersom konfidensintervallet med sannolikhet $1-\alpha$ omfattar det sanna värdet ($1-\alpha$ skall väljas stort, ofta 95 %). Om däremot μ_0 (eller p_0) ligger i konfidensintervallet är μ_0 (eller p_0) ett av de troliga värdena och vi kan inte förkasta hypotesen att $\mu = \mu_0$. Observera att vi på intet sätt har visat att $\mu = \mu_0$ (eller $p = p_0$), utan bara att det inte finns tillräckligt starka skäl att avvisa hypotesen. Denna metod ger ett *signifikanstest* på *signifikansnivå* α.

Om det är så att $\mu = \mu_0$, är sannolikheten att konfidensintervallet inte täcker μ_0 lika med α. I så fall förkastas ju hypotesen. Signifikansnivån α är alltså **sannolikheten att förkasta en sann hypotes**.

Ett annat sätt att utföra testet är följande: Villkoret att inte förkasta hypotesen $\mu = \mu_0$ är

$$\bar{x} - \lambda_{\alpha/2}\sigma/\sqrt{n} \leq \mu_0 \leq \bar{x} + \lambda_{\alpha/2}\sigma/\sqrt{n} \tag{9.1}$$

vilket är detsamma som att

$$-\lambda_{\alpha/2} \leq \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \leq \lambda_{\alpha/2} \tag{9.2}$$

Sätt $u = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$. Villkoret blir då att

$$\mid u \mid \leq \lambda_{\alpha/2} \tag{9.3}$$

eller ekvivalent: Förkasta hypotesen $\mu = \mu_0$ om

$$\mid u \mid > \lambda_{\alpha/2} \tag{9.4}$$

Variabeln u kallas *teststorhet* eller *t*eststatistika. Om σ är okänt byts σ ut mot stickprovsstandardavvikelsen s (beräknad från data) och z_α-kvantilen mot t-kvantilen $t_{\alpha/2}(n-1)$, d.v.s. hypotesen $\mu = \mu_0$ förkastas om

$$\mid t \mid > t_{\alpha/2}(n-1), \tag{9.5}$$

där t är teststorheten $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$, eller ekvivalent om μ_0 inte tillhör konfidensintervallet $\bar{x} \pm t_{\alpha/2}(n-1)s/\sqrt{n}$.

9.2 Övningar i begreppen

	H_0 True	H_0 False
Reject H_0	Type I Error	Correct Rejection
Fail to Reject H_0	Correct Decision	Type II Error

Problem 9.2.1.

We consider a hypothesis testing problem of the following kind.

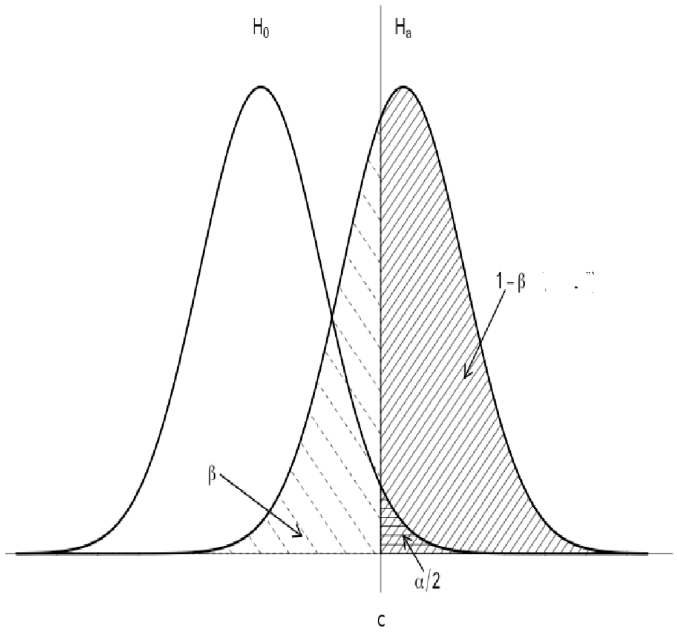
$$H_o : X \sim \mathcal{N}(\mu_o, 1)$$

$$H_1 = H_a : X \sim \mathcal{N}(\mu_1, 1),$$

where $\mu_o < \mu_1$. In other words, there are two alternatives, two different normal distributions, as populations for the observed data.

The test has the critical value is at c and the level of significance is depicted as $\alpha/2$, which you can read as α, as we have here a one-sided test. This means that if $x < c$, then H_o is accepted (not rejected), and if $x > c$, then H_1 is accepted, or, equivalently, H_o is rejected.

The figure below depicts graphically certain probabilities connected to the decisions in this hypothesis testing. Which of the following statements is the correct description of the area $1 - \beta$?



a) It is the probability of rejecting H_0, when H_0 is true.

b) It is the probability of not rejecting H_0, when H_0 is false.

c) It is the probability of rejecting a false H_0.

d) It is the probability of accepting a false H_a.

Problem 9.2.2.

You have a confidence interval $(\widehat{p}-E, \widehat{p}+E)$ for a population proportion p with confidence level 0.95 %. You are testing

$$H_o : p = 0.75 \quad H_1 : p \neq 0.75$$

Which of the following statements is correct ?

a) Type II error is the mistake of rejecting the null hypothesis when it actually is true.

b) Type II error is the mistake not to reject the alternative hypothesis when it actually is false.

c) The probability of type I error is the probability that p is outside $(\widehat{p}-E, \widehat{p}+E)$ and this probability is equal to 0.05.

d) The probability of type I error is the probability that 0.75 lies outside $(\widehat{p}-E, \widehat{p}+E)$, when H_o is in fact true and this probability is equal to 0.05.

Problem 9.2.3.

Låt $X_1, \ldots, X_n$ vara ett stickprov från en normalfördelning $\mathcal{N}(\mu, 1)$. Nollhypotesen $H_0 : \mu = \mu_0$ testas mot den alternativa hypotesen $H_1 : \mu > \mu_0$ med hjälp av teststatistikan

$$u = \frac{\bar{x} - \mu_0}{\sqrt{1/n}}.$$

H_0 förkastas om $u > c$ (c har valts för att nå en given signfikansnivå, som inte bekymrar oss i denna uppgift).

Förklara vad som menas med ett typ-II-fel. När inträffar ett typ-II-fel i testet ovan?

p-värde
I statistisk hypotesprövning är **p-värdet** sannolikheten för få ett värde på en teststorhet/teststatistika som är minst lika extremt eller extremare än det värde på denna teststorhet som faktiskt observerades, under antagandet att nollhypotesen är sann.
Nollhypotesen förkastas, om the p-värdet är mindre än en på förthand fastställd signifikansnivå α (= alpha i grekiskan). α är ofta 0.05 eller 0.01. Detta indikerar att det observerade resultatet är mycket osannolikt under nollhypotesen, d.v.s. om nollhypotesen antas gälla.
p-värdet är INTE sannolikheten för att nollhypotesen är sann.

Problem 9.2.4.

A study shows that 77 cell phone users developed leukemia while 420 018 cell phone users did not. The question is, whether cell phone users have leukemia at a rate that is greater than 0.0190 %. Let p be the population rate of leukemia among cell phone users. Your hypotheses are

$$H_o : p = 0.000190 \quad H_1 : p > 0.000190.$$

Your level of significance is $\alpha = 0.01$. The maximum likelihood estimate of p is $\widehat{p}_{mle} = 0.000183$. The test statistic has the value

$$Z = \frac{\widehat{p}_{mle} - p}{\sqrt{\frac{p\cdot(1-p)}{420095}}} = -0.33.$$

The test statistic Z is approximately $\sim \mathcal{N}(0,1)$. Which of the following statements is correct ?

a) The **p -value** is 0.629 and the null hypothesis is rejected.

b) The **p -value** is 0.005 and we fail to reject the null hypothesis.

c) The **p -value** is 0.629 and we fail to reject the null hypothesis.

d) The **p -value** is 0.005 and the null hypothesis is rejected.

Problem 9.2.5.

You are using an algorithm of **pairwise sequence alignment** (You are trying to find segments of similarity in two proteins or two DNA). The associated software at your disposal produces an alignment as well as an alignment score, which is denoted by S. It has been established, due to an ingenious combination of mathematical effort and pragmatic approximations, that $S \in \mathcal{E}(0.5)$. This means that the probability distribution function of S is

$$F_S(s) = \int_0^s 0.5e^{-0.5x}dx = 1 - e^{-0.5 \cdot s}, s > 0.$$

For a certain alignment You get $S = 4.0$. Which of the following is the correct statement?

a) The **p -value** is the probability of $S \leq 4.0$ and is in this case $= 0.86$.

b) The **p -value** is the probability of obtaining an alignment score equal to or ''more extreme'' than what was actually observed, and is in this case $= 0.14$.

c) The **p -value** is the probability of obtaining an alignment score equal to or ''more extreme'' than what was actually observed, and is in this case $= 0.86$.

d) The **p -value** is the observed level of significance of the alignment score, and is in this case $= 4.0/100 = 0.04$.

Problem 9.2.6.

Significance of correlation coefficient between telomere length and hTERT expression.

Telomeres constitute the ends of eukaryotic chromosomes In somatic cells, they progressively shorten during each cell cycle by replication-dependent loss of DNA ends. In humans, average telomere length declines from about 11 kilobases (kb) at birth to less than four kilobases in old age with average rate of decline being greater in men than in women. Ongoing shortening finally prevents telomeres from adequately protecting chromosome ends from further degradation. Cells with shortened telomeres eventually succumb to proliferative aging: Consequently, tumor cells need to compensate for replicative telomere losses to preserve their ability to proliferate indefinitely. In 90% of human cancers, maintenance of telomeres is achieved by human telomerase reverse transcriptase (hTERT) expression and activation of telomerase.

In a study the hTERT-encoding mRNA was found in 57 noncancer mucosa and colorectal carcinoma tissue samples.

Gertler, Ralf et. al.: *Telomere length and human telomerase reverse transcriptase expression as markers for progression and prognosis of colorectal carcinoma*, **Journal of Clinical Oncology**, 22, 2004.

The positive correlations between telomere length and hTERT expression were found in noncancer colorectal mucosa $r = 0.54$ and in colorectal carcinoma $r = 0.52$.

We want to test with this data $H_0 : \rho = 0$ against $H_1 : \rho \neq 0$. ρ is the unknown population value of the correlation.

a) Consider the test statistic (Vidakovic pp. 572−575)

$$t = \frac{r}{\sqrt{(1 - r^2)/(n - 2)}}.$$

It can be shown that t is approximately $\sim t(n - 2)$, here $n = 57$. Find the ***p*-values** of the two observed correlations. You may need the function `tcdf` of Matlab. If the level of significance is chosen as $\alpha = 0.001$, **what** are your conclusions?

b) Median telomere lengths in noncancer mucosa and cancer tissue of all 57 patients were 6.8 kb (range, 5.5 to 8.6 kb) and 5.7 kb (range, 4.1 to 7.6 kb), respectively. (Cancer tissue had significantly ($p < 0.001$) shorter telomeres than matched adjacent mucosa.)

The mathematics behind the fact that t is approximately $\sim t(n-2)$ assumes that the underlying data has a (bivariate) normal distribution, which also implies that telomere length should have normal distribution. **Does this assumption seem justified**?

c) Consider the test statistic (sf1911 Formelsamling)

$$w = \sqrt{\frac{(n-3)}{2}} \log\left(\frac{1+r}{1-r}\right),$$

which is approximately $\sim N(0,1)$. Repeat **a)**.

d) Which statistical model for the pairs of data (hTERT Expression, telomere length) is applicable?

Problem 9.2.7.

Du ansvarar för design av kliniska prövningar vid ett stort och välrenommerat universitetssjukhus och fått i uppdrag att analysera statistiskt en ny experimentell terapi för en viss sjukdom. Det finns en standardbehandling som har visat sig kunna besegra sjukdomen i 35% av fallen. Det finns ett godkänt etiskt tillstånd för att testa den nya terapin.
För den tänkta (konceptuella) population, som består av alla som behöver nu eller kommer senare att behöva en terapi för denna sjukdom, finns den sanna **populationsandelen population proportion** av framgång för denna nya terapi betecknad med p. Denna parameter är okänd och kan ej mätas direkt. Du har naturligtvis en nollhypotes

$$H_0 : p = 0.35.$$

och alternativ hypotes

$$H_1 : p > 0.35.$$

Nollhypotesen påstår att standardterapin redan funkar så fint som möjligt. Du har två olika design för Dina kliniska test.

a) *Design 1*: Tio patienter behandlas med den nya terapin.
Du får utfallet SSFSSFSSSF, framgångar kodas som S och misslyckanden kodas som F. De enskilda utfallen är oberoende. Vad är **p-värdet** av detta resultat?
Ledning: X = antalet lyckade (S) försok av tio. Med vilken fördelning modelleras X, om H_0 är sann?

b) Vad är maximumlikelihoodskattningen $\widehat{p}_{mle}$ av p på basis av data SSFSSFSSSF?

c) *Design 2*: Man fortsätter att behandla med den nya metoden tills man har misslyckats (F) tre gånger.
Du observerar igen SSFSSFSSSF. De enskilda utfallen är oberoende.

Vad är **p-värdet** av detta resultat?
Ledning: X = antalet lyckade (S) försok tills Du har fått tre misslyckade (F) försök. Då är $X \sim \mathcal{NB}(3, 0.65)$ (negativ binomialfördelning), om H_0 är sann (Varför?). Det gäller för $X \sim \mathcal{NB}(3, 0.65)$ att

$$f_X(x) = \Pr(X = x) = \binom{x+3-1}{x} 0.35^x 0.65^3 \quad \text{för } x = 0, 1, 2, \ldots$$

De numeriska beräkningarna görs helst med Matlabfunktionen `nbincdf`.

d) Analysera p-värdena Du fick ovan utifrån att det var samma utfall, SSFSSFSSSF, för båda design.

Hur påverkas **p-värdet** och därmed signifikansen av design ?

Problem 9.2.8.

In a machine for sequencing of DNA the proportion of reading errors is according to the manufacturer equal to 2%. You want to check this claim by means of a statistical test. When a DNA fragment with 5000 nucleotides was sequenced, 79 errors were identified. Let p be the population rate of errors for the sequencing machine. Your hypotheses are

$$H_o : p = 0.02 \quad H_1 : p \neq 0.02.$$

Your level of significance is $\alpha = 0.01$. The maximum likelihood estimate of p is $\widehat{p}_{mle} = 0.016$.
The approximate confidence interval (CI) for p is of the form

$$(\widehat{p}_{mle} - E, \widehat{p}_{mle} + E)$$

With $Z \in N(0,1)$ we have

α	λ_α	α	λ_α
0.10	1.2816	0.001	3.0902
0.05	1.6449	0.0005	3.2905
0.025	1.9600	0.0001	3.7190
0.010	2.3263	0.00005	3.8906
0.005	2.5758	0.00001	4.2649

Which of the following statements is correct ? Aid: You may need to find the endpoints of the CI in each case.

a) $E = 2.5758\sqrt{\frac{p(1-p)}{5000}}$ and we fail to reject the null hypothesis.

b) $E = 2.3263\sqrt{\frac{p(1-p)}{5000}}$ and we fail to reject the null hypothesis.

c) $E = 2.3263\sqrt{\frac{p(1-p)}{5000}}$ and the null hypothesis is rejected.

d) $E = 2.5758\sqrt{\frac{p(1-p)}{5000}}$ and the null hypothesis is rejected.

Problem 9.2.9.

För att säkerställa att pH-värdet på utgående vatten från en reningsverk inte är för lågt genomförs regelbundna provtagningar. Vid ett tillfälle togs 5 prover och pH-värdet mättes med hjälp av en pH-meter $(X_1, \ldots, X_5)$. Mätvärdena antas vara oberoende och dragna från en normalfördelning med okänt väntevärde μ och känd varians $\sigma^2 = 1.65$. Därefter testas nollhypotesen

$$H_0 : \mu = 7$$

mot den alternativa hypotesen

$$H_A : \mu < 7$$

med hjälp av teststorheten (teststatistikan)

$$U = \frac{\bar{X} - 7}{\sigma/\sqrt{n}}.$$

Härvid förkastas H_0 om $u < -1.64$.

a) Förklara vad som menas med testets styrka.

b) Beräkna testets styrka om det sanna pH-värdet är 6.3 (dvs $\mu = 6,3$).

c) Vilken stickprovsstorlek hade behövts för att testets styrka skulle överstiga 0.90 om det sanna pH-värdet var 6.3?

Problem 9.2.10.

Efter ett miljöutsläpp av stora mängder antibiotika undersöks förekomsten av motståndskraftiga bakterier. Därför samlades två stickprov in, ett med 35 jordprov från den förorenade miljön $(x_1, \ldots, x_{35})$ och ett med 43 jordprov från en kontrollmiljö $(y_1, \ldots, y_{43})$. Mängden motståndskraftiga bakterier uppmättes sedan med hjälp av PCR och följande medelvärden och standardavvikelser beräknades $\bar{x}$= 0.110, s_x = 0.109, $\bar{y}$= 0.0479 och s_y = 0.011 (mätvärdena beskriver relativ mängd motståndskraftiga bakterier). Av tidigare erfarenhet vet man att variationen i de olika stickproven ej kan antas vara lika.

a) Formulera lämpliga hypoteser och fördelningsantaganden. Genomför sedan ett enkelsidigt test för att undersöka om mängden motståndskraftiga bakterier är högre i den förorenade miljön. Signifikansnivån skall vara 0.01.

b) Beräkna p-värdet för testet i **a)**.

Problem 9.2.11.

På misstänkta rattfyllerister gör man tre bestämningar av alkoholhalten i blodet. Resultaten x_1, x_2, x_3 antas utgöra ett slumpmässigt stickprov från $\mathcal{N}(\mu, 0.05^2)$, där μ är den verkliga alkoholhalten i blodet. Om $\mu > 0.2$ har personen gjort sig skyldig till rattonykterhet. Låt oss anta att den domstol som skall döma tar hänsyn till osäkerheten i mätningarna genom att beräkna det aritmetiska medelvärdet $\bar{x}$ av de tre analysresultaten och därefter förklara personen skyldig om

$$\bar{x} > 0.2 + \lambda_{0.01} \cdot 0.05/\sqrt{3}$$

men oskyldig annars. Med statistisk terminologi kan man säga att domstolen prövar

$$H_0 : \mu = 0.2$$

mot

$$H_1 : \mu > 0.2$$

på nivån 0.01. Vilka av följande påståenden ger en någorlunda korrekt beskrivning av vad som kommer hända i det långa loppet?

a) högst 1% av alla frikända är skyldiga

b) högst 1% av alla oskyldiga blir dömda

c) högst 1% av alla skyldiga blir frikända

d) högst 1% av alla dömda är oskyldiga

9.3 Jämförelse mellan två väntevärden: formler

Antag att vi nu har två stickprov från två normalfördelningar, $x_1, x_2, \ldots, x_{n_1}$ respektive $y_1, y_2, \ldots, y_{n_2}$. De två stickproven antas vara $\mathcal{N}(\mu_1, \sigma_1^2)$ respektive $\mathcal{N}(\mu_2, \sigma_2^2)$ och oberoende. Vi vill testa hypotesen $\mu_1 = \mu_2$ vilket ofta kan tolkas som att två metoder är likvärdiga om de två stickproven har uppkommit genom användning av två metoder.

9.3.1 Kända varianser

Ett tvåsidigt konfidensintervall för skillnaden $\mu_1 - \mu_2$ ges av

$$\bar{x} - \bar{y} \pm \lambda_{\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$$

Hypotesen förkastas om 0 inte tillhör intervallet ifråga eller, ekvivalent, om

$$| u | > \lambda_{\alpha/2}$$

där u är testvariabeln $u = \frac{\bar{x}-\bar{y}}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}$.

9.3.2 Okända men lika varianser

Konfidensmetoden
Vi antar nu att σ_1 och σ_2 är okända men lika. Den bästa skattningen av denna okända varians ges av

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$$

där s_1^2 och s_2^2 är stickprovsvarianserna för de två stickproven. Man kan då visa att

$$t = \frac{(\bar{x}-\bar{y}) - (\mu_1-\mu_2)}{s\sqrt{1/n_1+1/n_2}} \tag{9.6}$$

är t-fördelad med n_1+n_2-2 frihetsgrader. Av detta följer att ett konfidensintervall för skillnaden $\mu_1-\mu_2$ ges av

$$\bar{x}-\bar{y} \pm t_{\alpha/2}(n_1+n_2-2)s\sqrt{1/n_1+1/n_2}. \tag{9.7}$$

Hypotesen $\mu_1=\mu_2$ förkastas om 0 inte tillhör konfidensintervallet (9.7). Alternativt förkastas hypotesen om $\mid t \mid > t_{\alpha/2}(n_1+n_2-2)$ där t är testvariabeln

$$t = \frac{\bar{x}-\bar{y}}{s\sqrt{1/n_1+1/n_2}}$$

9.3.3 Parade observationer (matched pairs)

Konfidensmetoden
Antag att man vill jämföra två metoder på ett antal prov som har helt *olika* värden. På varje prov görs en analys med var och en av metoderna. Vi antar att skillnaden mellan de två metodernas resultat har samma förväntade värde, oavsett prov. Data:

Prov	1	2	3	...	n
Metod 1	x_1	x_2	x_3	...	x_n
Metod 2	y_1	y_2	y_3	...	y_n
Skillnad	z_1	z_2	z_3	...	z_n

Här är alltså $z_i = x_i - y_i, i = 1,2,\ldots,n$.

Antag att den konstanta förväntade skillnaden är Δ. Det är då inte så svårt att visa att z-observationerna är normalfördelade med förväntat värde Δ, och att testa om metoderna är likvärdiga, d.v.s. om $\Delta = 0$, kan vi göra utgående från z-data. Vi har återfört problemet till fallet ett stickprov. Hypotesen $\Delta = 0$ förkastas om 0 inte tillhör intervallet $\bar{z} \pm t_{\alpha/2}(n-1)s/\sqrt{n}$ eller om $\mid t \mid > t_{\alpha/2}(n-1)$ där $t = \frac{\bar{z}}{s_z/\sqrt{n}}$.

9.4 Jämförelse av två väntevärden

Problem 9.4.1.

Koncentrationen av en aktiv ingrediens i ett flytande tvättmedel tros påverkas av vilken katalysator som används i processen. Standardavvikelsen av den aktiva koncentrationen är känd till att vara 3.0 g/l oberoende av katalysatorteknik. Tio observationer av koncentrationen tas för var och en av två katalysatorer med följande resultat:

Katalysator 1:	57.9	66.2	65.4	65.4	65.2	62.6	67.6	63.7	67.2	71.0
Katalysator 2:	66.4	71.7	70.3	69.3	64.8	69.6	68.6	69.4	65.3	68.8

Finns det någon anledning att tro att den aktiva koncentrationen beror på valet av katalysator? Basera ditt svar på beräkning av ett lämpligt 95%-konfidensintervall. Observationerna antas vara normalfördelade.

Problem 9.4.2.

I en vetenskaplig studie undersöks genuttrycket för genen p53 i olika former av tumörer. I studien användes PCR för att mäta genuttrycket hos 8 slumpmässigt utvalda individer med elakartade tumörer $(x_1, \ldots, x_8)$ och hos 7 slumpmässigt utvalda individer med godartade tumörer $(y_1, \ldots, y_7)$. Medelvärden och standardavikelser beräknades till $\bar{x} = 13.9$, $s_x = 1.63$, $\bar{y} = 9.2$, $s_y = 1.34$ (mätvärdena beskriver den relativa mängden mRNA för p53). Mätvärdena kan antas vara normalfördelade med lika varians.

a) Formulera lämpliga hypoteser och fördelningsantaganden för ett enkelsidigt test som undersöker om genuttrycket av p53 är högre i de elakartade tumörerna jämfört med de godartade tumörerna. Genomför sedan testet. Signifikansnivån ska vara 0.05.

b) Beräkna **p-värdet** för testet i **a)**.

Problem 9.4.3.

Ett läkemedelsföretag eftersträvar att utveckla en nytt läkemedel mot högt blodtryck. Ett steg i denna långa och mycket kostnadskrävande process, som sällan resulterar i en produkt på apoteksdisken, är de s.k. kliniska fas-II studierna. I en sådan studie behandlas ett litet antal patienter med högt blodtryck med det nya läkemedlet. Studien vill undersöka om läkemedlet har en positiv eller negativ effekt, d.v.s. om sjuka patienter har ett ändrat blodtryck.

I tabellen nedan har blodtrycket (övertryck i mm Hg) hos åtta patienter uppmätts före behandlingen, x_i, och efter densamma, y_i, $i = 1, 2, \ldots, 8$.

Person	1	2	3	4	5	6	7	8
Blodtrycket före x	146	147	143	149	164	140	151	141
Blodtrycket efter y	153	141	145	138	151	128	136	129

a) Formulera nu en lämplig statistisk modell (Du får anta normalfördelade data).

b) Testa hypotesen att det inte är någon ändring i blodtrycket mot hypotesen att blodtrycket har ändrats. Signifikansnivån har av den europeiska läkemedelsmyndigheten (EMA) bestämts som 5%. Din slutsats bör framgå tydligt.

Problem 9.4.4.

The amount of mRNA in the gene PROM1 is suspected to be associated with a malicious form of a tumour. To study this the level of mRNA at PROM1 was measured for ten patients $(x_1, \ldots, x_{10})$ with the malicious form and eight patients $(y_1, \ldots, y_8)$ with a less malicious form. Then it was found that

$$\overline{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = 7.76, s_x = \sqrt{\frac{1}{9}\sum_{i=1}^{10}(x_i - \overline{x})^2} = 2.02.$$

$$\overline{y} = \frac{1}{8}\sum_{i=1}^{8} y_i = 5.82, s_y = \sqrt{\frac{1}{7}\sum_{i=1}^{8}(y_i - \overline{y})^2} = 0.90.$$

The level of significance is 0.05. We make suitable assumptions about the normal distribution for xs and y's. Then we have with $t_{0.025}(10 + 8 - 2) = 2.12$ the confidence interval

$$\overline{x} - \overline{y} \pm s \cdot 2.12.$$

where

$$s = \sqrt{\frac{s_x^2}{10} + \frac{s_y^2}{8}}.$$

We obtain the interval

$$[0.43, 3.45]$$

Which of the following is the correct description of the statistical procedure we are using AND of our finding based on this this interval:

a) We are using a two-sided test for matched pairs and we reject the null hypothesis at the level of significance 0.05.

b) We are using a two-sided test for the difference between two means and we reject the null hypothesis at the level of significance 0.05.

c) We are using a two-sided test for the difference between two means and we fail to reject the null hypothesis at the level of significance 0.05.

d) We are using a two-sided test for matched pairs and we fail to reject the null hypothesis at the level of significance 0.05.

Problem 9.4.5.

För att jämföra två gödselmedel lät man 5 lantbrukare gödsla hälften av sin veteareal med medel A och den andra hälften med medel B. Man fick följande skördar per hektar:

Lantbrukare	1	2	3	4	5
Medel A	115.8	123.0	122.8	209.8	226.3
Medel B	121.2	122.1	128.0	214.9	229.2

För att få en enkel statistisk modell antog man att samtliga skördeutfall kan ses som utfall av oberoende normalfördelade stokastiska variabler med samma varians. Observera dock att lantbrukarna har gårdar med lite olika odlingsförutsättningar för vete.

a) Beräkna ett 95%-igt konfidensintervall för skillnaden i förväntad skörd mellan arealer som gödslats med A respektive B.

b) Testa hypotesen att gödselmedlen är lika bra mot alternativet att de skiljer sig år på 5% signifikansnivå. Slutsatsen skall klart framgå.

Problem 9.4.6.

Man gör undersökningar vid badplatser för att få reda på om halten av gift från giftalger överstiger vissa gränsvärden. Om man kan visa att den förväntade niv ån överstiger 0.6 sätter man upp en varningsskylt som rekommenderar folk att ej bada där och om man kan visa att den förväntade nivån överstiger 0.8 utfärdar man badförbud. Man observerar nivåerna

$$1.18 \quad 0.99 \quad 0.83 \quad 0.71 \quad 1.27 \quad 0.49 \quad 1.58 \quad 1.05$$

Antag att nivåernas värden är oberoende och varierar enligt normalfördelning. Följande info kan vara användbar: stickprovsmedelvärdet=1.0125 och stickprovsstandardavvikelsen=0.341.

a) Ska man sätta upp varningsskylten? Genomför lämplig undersökning på nivå 0.01.

b) Ska man utfärda badförbud? Genomför lämplig undersökning på nivå 0.01.

Problem 9.4.7.

Vid tillverkning av öl gäller för ett visst bryggeri att alkoholhalten i en slumpvis uttagen flaska är normalfördelad med standardavvikelsen 0.10 %. Ett stickprov på 10 flaskor tages från dagens tillverkning och alkoholhalten mätes. Antag nu att man råkar ut för kännbara ”påföljder” om alkoholhalten i genomsnitt är 3 % eller större och man är villig att taga högst 1 % risk för dessa påföljder.

Formulera lämpliga noll- och mothypoteser, **giv** signifikanskriterium och
tillämpa detta slutligen om stickprovsmedelvärdet visar sig vara 2.98 %. Obs! Tänk efter noga vid nollhypotesvalet.

Problem 9.4.8.

I ett laboratorium har man gjort 6 mätningar för att bestämma halten m av ett visst ämne i en råvara. Standardavvikelsen är känd, $\sigma = 0.2$ [%]. Man erhöll $\bar{x} = 5.575$ [%]. Halten skall helst vara $m = 5.5$ [%].

Kan denna hypotes förkastas på nivån 10 % ? Om i stället $\bar{x} = 5.675$ **vad blir** då svaret? (Man kan använda konfidensmetoden och utnyttja konfidensintervallen).

Problem 9.4.9.

Man har anledning förmoda att molekylvikterna för två kemiska föreningar A och B är lika. För att undersöka detta har man gjort 6 molekylviktsbestämningar på A och 8 på B. Den använda mätmetoden ger ett mätfel som är $\mathcal{N}(0, \sigma^2)$. Mätfelen i olika mätningar kan anses vara oberoende. Följande värden erhölls:

Molekylvikt för A	Molekylvikt för B
174.18	174.19
174.30	174.40
174.23	174.20
174.29	174.35
174.36	174.32
174.25	174.14
	174.27
	174.34

a) **Konstruera** ett test på nivån 5 % av hypotesen
H_0: A och B har samma molekylvikt mot alternativet
H_1: A och B har olika molekylvikt.

b) **Gör** numeriska beräkningar och **ange** testets utfall.

Problem 9.4.10.

I ett miljöövervakningssystem studeras övergödningen av våra vattendrag. I en viss å har man under en längre period gjort mätningar av bl a total fosforhalt. Under denna period införde man i avrinningsområdet en kemisk-biologisk rening av hushållens och industriernas avloppsvatten. För att undersöka vilken effekt dessa åtgärder haft på fosformängden i vattendraget beräknas årsmedelvärdena av total fosforhalt (mg/l) före och efter införandet av ny rening:

Fosforhalt (mg/l) före införandet:	0.12	0.14	0.07	0.09	0.15	0.09	0.10	
Fosforhalt (mg/l) efter införandet:	0.09	0.03	0.07	0.09	0.07	0.08	0.11	0.07

a) Gör ett 95% konfidensintervall för den genomsnittliga effekten av den nya reningen. Redogör för dina modellantaganden.

b) Gav åtgärderna upphov till en signifikant förändring av total fosforhalt i vattendraget? Motivera ditt svar!

Problem 9.4.11.

Industrier, sjukhus och andra organisationer som är beroende av provningsresultat från olika laboratorier, utför ofta undersökningar för att testa om laboratorierna mäter likvärdigt (s.k. proficiency testing). Vid en sådan undersökning sändes ett prov till två laboratorier, som vart och ett fick göra 5 oberoende mätningar på provet. Man kan anse att mätresultaten är normalfördelade med en och samma varians σ^2. Resultat:

Laboratorium	Mätvärden					$\sum_i x_i$	$\sum_i x_i^2$
1	41.33	41.35	41.24	41.24	41.32	206.48	8526.8090
2	41.24	41.27	41.28	41.27	41.30	206.36	8516.8918

a) Ställ upp en lämplig statistisk modell och testa hypotesen att det inte är någon skillnad i de två laboratoriernas förväntade mätresultat. 5 % signifikansnivå.

b) Antag att laboratoriernas förväntade resultat verkligen är lika. Under denna förutsättning, ge ett 95% konfidensintervall för detta väntevärde μ.

Problem 9.4.12.

You are studying the effectiveness of a medicine called Prilosec for treating heartburn by measuring gastric acid secretion in patients before and after the drug treatment. The data consists of before/after measurements for each patient.

Prilosec	Patient			
	1	2	...	n
before	x_1	x_2	...	x_n
after	y_1	y_2	...	y_n

You assume that x_j for the jth subject is a sample from $N(\mu_j, \sigma_1)$ and y_j a sample from $N(\mu_j + \Delta, \sigma_2)$. Δ is the population parameter for the effectiveness of Prilosec. Your hypotheses are

$$H_o : \Delta = 0$$

against

$$H_1 : \Delta \neq 0$$

Which is the correct test to use for this purpose?

a) t-test with matched pairs.

b) Bland-Altman-test.

c) χ^2-test.

d) t-test for two means.

9.5 Type I error, Type II error, Power, Accuracy, Sensitivity and Specificity

Types of Error and Specificity and Sensitivity
$T+$ = positive test result in a diagnostic test or method of identifying a disease , $T-$ = negative test result in a diagnostic test. $D+$ = has a disease, $D-$ = does not have a disease.

False Positives = FP , True Positives = TP , False Negatives = FN , True Negatives= TN.

	$D+$	$D-$
$T+$	TP	FP
$T-$	FN	TN

We assume now that these are observed frequencies, when diagnosis has been applied to a sampled population.
One often encounters one or several of the following criteria of performance evaluation:

- **Accuracy (A)**: $= \frac{\text{TP+TN}}{\text{TP+TN+FP+FN}}$
- **Precision (P)**: $= \frac{\text{TP}}{\text{TP+FP}}$
- **Recall (A)**: $= \frac{\text{TP}}{\text{TP+FN}}$

Problem 9.5.1.

Let N=TN+FP, P=FN+TP. It is claimed that

- FP/N = probability of type I error = 1−Specificity
- TP/P= 1-probability of type II error =Power =Sensitivity

Explain why we can talk about the notions probability of type I error, probability of type II error, Power, Sensitivity and Specificity connected in this special manner.

Problem 9.5.2.

Breast Cancer Identification
Nondestructive preoperative breast imaging techniques are widely used for breast cancer testing and diagnosis. A study
Han, F and Liang, CW and Shi, GL and Wang, L and Li, KY *Clinical applications of internal heat source analysis for breast cancer identification*, **Genetic Molecular Research**, 14, 2015.
constructs and evaluates three rules for the feasibility and efficacy of quantitative diagnosis via the thermal analysis of abnormal metabolism. Nine hundred forty-eight women who underwent breast biopsy from 2009 to 2013 were investigated. The purpose is to find the optimal separation rule amongst the three constructed for deciding between breast cancer and benign disease.

	Sensitivity (%)	False-positive rate (%)	Specificity (%)	Accuracy (%)
Rule 1	99.6	34.9	65.1	74.6
Rule 2	94.2	7.1	92.9	93.2
Rule 3	76.2	2.5	97.5	91.7

Which of the rules of identification is in view of these numbers to prefer and why?

9.6 Low Pre-Study Odds and Positive Predictive Value

The critical questions and issues in this section are due to the most downloaded technical paper from the journal PLoS Medicine, John PA Ioannidis: **Why Most Published Research Findings Are False**. *PLoS medicine* 2, 8, 2005.

There is an increasing concern that most current published research findings are false, as there is a lot of lack of replication in research findings, most notably in the field of genetic associations but also in clinical trials and traditional epidemiological studies to the most modern molecular research.

For a recent survey of this there is an article in Marie Curie, the webjournal of Swedish Research Council, `https://www.tidningencurie.se/en/nyheter/2017/11/22/credibility-affected-by-researchers-decisions/?`

For example, a survey of 600 positive associations between gene variants and common diseases showed that out of 166 reported associations studied, three or more times, only six were replicated consistently. Lack of replication results from a number of factors such as publication bias, selection bias, Type I errors, population stratification (the mixture of individuals from heterogeneous genetic backgrounds), and lack of statistical power.

Research findings are defined in this section as any relationship reaching formal statistical significance.

The probability that a claim of research finding is true may depend on study power and the number of other studies on the same question, and, importantly, the ratio of true to no relationships among the relationships probed in each scientific field, as discussed here.

Think of a team of geneticists performing a **whole genome association study** to test whether any of 100,000 gene polymorphisms are associated with susceptibility to schizophrenia.

Let tr =the number of true relationships between, e.g., gene polymorphisms and schizophrenia, and fa = the number of no relationships between gene polymorphisms and schizophrenia.

Then we have the prevalence

$$P(D+) = \text{Pre-study probability of true relationship} = \frac{tr}{fa+tr},$$

and

$$P(D-) = 1 - P(D+) = \frac{fa}{fa+tr}$$

Now we define an odds ratio of prevalences

$$R \stackrel{\text{def}}{=} \frac{P(D+)}{P(D-)} = \frac{tr}{fa}.$$

Then we can write

$$P(D+) = \frac{R}{R+1}.$$

The conditional probability of a research finding $T+$ i.e., any relationship reaching formal statistical significance, when the relationship is true, a.k.a. sensitivity, is

$$P(T+ \mid D+)$$

We call this, like Ioannides, PPV= positive predictive value or the probability of true positive. We think of hypotheses for testing at a polymorhism for association with schizophrenia as

$$\mathrm{H}_o\text{: a false relationship}$$

$$\mathrm{H}_a\text{: a true relationship}$$

The geneticists use a statistical test with Type I error rate α and with power $1-\beta$. For a finding of a true relationship between a gene polymorphism and schizophrenia, we have

$$PPV = \frac{(1-\beta)R}{(1-\beta)R+\alpha}. \tag{9.8}$$

Then the probability of false positive is

$$NPV = \frac{\alpha R}{(1-\beta)R+\alpha}. \tag{9.9}$$

The research finding of a gene polymorphism associated with schizophrenia is thus more likely to be true than false if $(1-\beta) > \alpha$.

Problem 9.6.1.

Derive (9.8) and (9.9). *Hint:* Express $P(T+ \mid D+)$ and $P(T- \mid D+)$ by means of the power of the test and of type I error and use Bayes.

Problem 9.6.2.

Based on what is known about the extent of heritability of the disease, it is reasonable to expect that probably around ten gene polymorphisms among those to be tested will be truly associated with schizophrenia. Recall that all in all 100,000 gene polymorphisms were studied. Assume that the study has power 60% and $\alpha = 0.05$.

a) Find the numerical value of PPV.

b) Compare the numerical value of PPV with that of R and reflect on the result.

c) If $\alpha = 0.05$, then 5000 of the gene polymorphisms among those tested will in the average yield a significant outcome. We assume that the ten correct ones are among these. What is the probability of finding a true associated with schizophrenia?

9.7 Bilaga: *p*-värden för teststorheter/teststatistikor med $\mathcal{N}(0,1)$ och $t(n)$

9.7.1 $\mathcal{N}(\mu, \sigma^2)$, σ känd

Vi har $X_1, \ldots, X_n$, oberoende och alla $\mathcal{N}(\mu, \sigma^2)$. Standardavvikelsen $\sigma > 0$ är känd. Vi har en nollhypotes

$$H_o : \mu = \mu_o.$$

Med data (slumpmässigt stickprov) $x_1, \ldots, x_n$ beräknar vi teststorheten

$$z = \frac{\bar{x} - \mu_o}{\sigma/\sqrt{n}}.$$

Då är

$$Z = \frac{\bar{X} - \mu_o}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

- Om
$$H_1 : \mu > \mu_o$$
så är p-värdet för ett test mot H_o lika med
$$p = P(Z \geq z) = 1 - \Phi(z) \overset{\text{Matlab}}{=} 1 - \texttt{normcdf(z,0,1)}$$

- Om
$$H_1 : \mu < \mu_o$$
så är p-värdet för ett test mot H_o lika med
$$p = P(Z \leq z) = \Phi(z) \overset{\text{Matlab}}{=} \texttt{normcdf(z,0,1)}$$

- Om
$$H_1 : \mu \neq \mu_o$$
så är p-värdet för ett test mot H_o lika med
$$p = 2 \cdot P(|Z| \geq z) = 2 \cdot (\Phi(-z) + 1 - \Phi(z))$$
$$\overset{\text{Matlab}}{=} 2 * (\texttt{normcdf(-z,0,1)} + 1 - \texttt{normcdf(z,0,1)})$$

9.7.2 $\mathcal{N}(\mu, \sigma^2)$, σ okänd

Vi har $X_1, \ldots, X_n$, oberoende och alla $\mathcal{N}(\mu, \sigma^2)$. Standardavvikelsen $\sigma > 0$ är okänd. Vi har en nollhypotes

$$H_o : \mu = \mu_o.$$

Med data (slumpmässigt stickprov) $x_1, \ldots, x_n$ beräknar vi teststorheten

$$t = \frac{\bar{x} - \mu_o}{s/\sqrt{n}},$$

där $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$. Då är

$$T = \frac{\bar{X} - \mu_o}{s/\sqrt{n}} \sim t(n-1).$$

- Om
$$H_1 : \mu > \mu_o \quad \text{lika med} \quad p = P(T \geq t) = 1 - \texttt{tcdf}(\texttt{t}, 0, 1)$$
så är p-värdet för ett test mot H_o lika med
$$p = P(T \geq t) \overset{\text{Matlab}}{=} 1 - \texttt{tcdf}(\texttt{t}, 0, 1).$$

- Om
$$H_1 : \mu < \mu_o \quad \text{lika med} \quad p = P(T \leq t) = \texttt{tcdf}(\texttt{t}, 0, 1)$$
så är p-värdet för ett test mot H_o lika med
$$p = P(T \leq t) \overset{\text{Matlab}}{=} \texttt{tcdf}(\texttt{t}, 0, 1)$$

- Om
$$H_1 : \mu \neq \mu_o$$
så är p-värdet för ett test mot H_o lika med
$$p = 2 \cdot P(|T| \geq t) \overset{\text{Matlab}}{=} 2 * (\texttt{tcdf}(-\texttt{t}, 0, 1) + 1 - \texttt{tcdf}(\texttt{t}, 0, 1))$$

9.7.3 Normalfördelad linjär modell

$$Y_i = \alpha + \beta x_i + \epsilon_i,$$

där $\epsilon_1, \ldots, \epsilon_n$ är oberoende och $\mathcal{N}(0, \sigma^2)$. Minstakvadratskattningarna $\widehat{\alpha}$ och $\widehat{\beta}$ har behandlats i tidigare övningar.

Det gäller det att

$$\widehat{\alpha} \sim \mathcal{N}\left(\alpha, \sigma^2\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)\right)$$
$$\widehat{\beta} \sim \mathcal{N}\left(\beta, \sigma^2 \frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right).$$

Med $e_i = y_i - (\widehat{\alpha} + \widehat{\beta} x_i)$ har vi

$$s^2 = \widehat{\sigma^2} = \frac{\sum_{i=1}^{n} e_i^2}{n-2}.$$

är $\chi^2(n-2)$-fördelad.

Vi kan konstruera konfidensintervall och test som förut, både då σ är känt och okänt.

$$I_\alpha = \widehat{\alpha} \pm t_{0.025}(n-2)s\sqrt{\left(\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)}.$$

$$I_\beta = \widehat{\beta} \pm t_{0.025}(n-2)\frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}.$$

Det kan hända att inget samband i form av en teoretisk regressionslinje finns, dvs. $\beta = 0$. Vi kan testa detta genom

$$H_0 : \ \beta = 0$$

mot

$$H_1 : \ \beta \neq 0$$

Förkasta H_0 på nivån 0.05 om

$$0 \notin I_\beta = \widehat{\alpha} \pm t_{0.025}(n-2)\frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}}\,.$$

Problem 9.7.1.

Vad är p-värdet i denna modell?

Kapitel 10

χ^2-, homogenitets- och oberoendetest för kategoridata, ickeparametriska test

Med observationer x_{ij} på tabellform

					Radsumma
	x_{11}	x_{12}	$\cdots$	x_{1r}	n_1
	x_{21}	x_{22}	$\cdots$	x_{2r}	n_2
	$\vdots$	$\vdots$		$\vdots$	$\vdots$
	x_{s1}	x_{s2}	$\cdots$	x_{sr}	n_s
Kolumnsumma	m_1	m_2	$\cdots$	m_r	N

$$\chi^2 = \sum_{i,j} \frac{(x_{ij} - \frac{n_i m_j}{N})^2}{\frac{n_i m_j}{N}}$$

10.0.1 Homogenitetstest

En hypotes om samma kolumnfördelning över s kategorier i r mätserier förkastas om $\chi^2 > \chi^2_\alpha$ med χ^2_α från $\chi^2((r-1)(s-1))$-fördelning.

10.0.2 Kontigenstabell

En hypotes om oberoende mellan rader och kolonner förkastas om $\chi^2 > \chi^2_\alpha$ med χ^2_α från $\chi^2((r-1)(s-1))$-fördelning.

10.0.3 χ^2-test av fördelning

En hypotes om fördelningen $p_1, p_2, \ldots, p_s$ över s kategorier i r mätserier förkastas om $\chi^2 > \chi^2_\alpha$ där

$$\chi^2 = \sum_{i,j} \frac{(x_{ij} - m_j p_i)^2}{m_j p_i}$$

och χ^2_α hämtas ur från $\chi^2((s-1)r)$-fördelning.

10.1 χ^2-test

Problem 10.1.1.

Du har använt programvara för att beräkna värdet av teststorheten $Q = 11.24$ i ett χ^2-test. Antalet frihetsgrader är lika med $= 6$. Vad är p-värdet?

Problem 10.1.2.

I en stor enhet för barnhälsovård (pediatri) vill man veta om antibiotika ordineras likformigt under en arbetsveckas dagar (måndag - fredag). Man tar ett slumpmässigt stickprov på 200 antibiotikarecept som utskrivits under de senaste 12 månaderna och registrerar på vilken arbetsdag de ordinerats.

Vad är H_o?

Är villkoren för ett χ^2-test uppfyllda?

Vad är antalet frihetsgrader?

En programvara beräknar värdet av teststorheten i ett χ^2-test och ger dessutom p-värdet $p < 0.01$. Vilken slutsats drar Du av detta?

Problem 10.1.3.

The soybean cyst nematode is a roundworm pest affecting the yield of soybean crops everywhere in the world. Nematode-resistant soybean lines have been used for agriculture, but nematode populations have adapted. One recent soybean line we will call A is resistant to most nematodes except the supervirulent nematode LY1.

A new soybean line we will call B has been shown to be resistant to LY1 and to be genetically different from soybean line A. Understanding the genetics of soybean resistance to LY1 is an important step in creating a genetically engineered soybean that will resist most nematodes, including the virulent LY1. Researchers crossed soybean lines A and B.

Arelli, P.R. and Young, L.D. and Concibido, V.C.: *Inheritance of resistance in soybean PI 567516C to LY1 nematode population infecting cv. Hartwig.* **Euphytica**, 165, 1, pp. 1−−4, 2009.

They obtained 105 second-generation plants, of which 5 were resistant to LY1 and 100 were not. The researchers state that the data are *a good fit to a three gene model, Rhg, rhg, rhg, (one dominant, two recessive genes) with a segregation ratio of 3(resistant):61 (susceptible)* (end of quote).

a) The researchers describe a 3:61 distribution. What is H_o and H_a for a χ^2-test here?

b) What are the expected counts under the null hypohesis? Are the conditions of a χ^2-test satisfied?

c) Find the χ^2- statistic and use Matlab or Tabell 4. to the p-value for this test. Do the data support the model proposed by the researchers?

d) Can you conclude that the genetic basis of soybean resistance to the virulent LY1 nematode is carried by the three genes, one dominant and one recessive. Justify your answer.

Problem 10.1.4.

The Frizzle fowl is a variety of chicken with curled feathers. Frizzle fowls were crossed with a Leghorn variety exhibiting straight feathers. The first generation (F1) produced all slightly frizzled chickens. When the F1 was interbred, the following characteristics were observed in F2:

Landauer, W. and Dunn, LC: *The Frizzle characters of fowls: Its Expression and Inheritance*, **Journal of Heredity**, 21, pp. 291−305, 1930.

Phenotype (feather type)	Observed counts
Frizzled	23
Slightly frizzled	50
Straight	120

The most likely genetic model for this is a single locus with two codominant alleles. Under such a model we should expect a 1:2:1 ratio in F2. Do the data support this model?

Problem 10.1.5.

För att utröna hur sällsynt växten *Salicornia ramosissima* (Purple Glaswort) är i England, skaffade en grupp botaniker en statistik över antalet *Salicornia ramosissima* i ett område med många salta vattenmarker i nordvästra England. Området delades in i 576 delområden om vardera en *ha* yta. Först beräknades utifrån datamaterialet medelantalet per delområde till 0.9288. Därefter slogs materialet ihop på följande sätt:

Antal växter	0	1	2	3	≥ 4
Antal delområden	229	211	93	35	8

Testa på risknivån 5% om antalet *Salicornia ramosissima* per delområde kan anses vara Poisson-fördelat. Din slutsats skall klart framgå. (Anmärkning: Antalet bör vara Poisson-fördelat om växtspridningen sker slumpmässigt utan påverkan av systematiska faktorer (annat än brackvatten).)

Problem 10.1.6.

Felet (x) vid mätning av DNA-koncentration misstänks vara normalfördelat. För att fastställa om så är fallet genomfördes 200 kontrollmätningar där felet observerades.

Nivå	Antal mätningar
$x \leq -1$	41
$-1 < x \leq 0$	53
$0 < x \leq 1$	59
$1 < x$	47

Genomför ett χ^2-test för att undersöka om felet kommer från en normalfördelning med väntevärde $\mu = 0$ och standardavvikelse $\sigma = 1.5$. Använd en signifikansnivå på 0.05.

Problem 10.1.7.

A field biologist recorded events of inbreeding in a colony of of prairie dogs. Based on a sample of 44 estruous females and 17 sexually mature males, the researcher computed the coefficient of relatedness of all possible male-female pairs in this sample. If individuals avoided inbreeding, we would expect to observe breeding mostly between individuals with little genetic relatedness. On the other hand if breeding occurred randomly without regard for genetic relatedness of mating couple, we would expect breeding observations among individuals of various relatedness to be proportional to the number of possible pairs with a given coeffient of relatedness. Here is the table of observed and expected breeding pairs in the study.

Hoogland, John L: *Why do Gunnison's prairie dogs have multiple mating ?* 55,2, pp. 351−359, **Animal behaviour**, 1998.

Coefficient of relatedness	Observed pairs	Expected pairs
$r \geq 1/4$ (siblings)	6	6
$1/4 > r \geq 1/8$	4	5
$1/8 > r \geq 1/16$	18	10
$1/16 > r \geq 1/32$	12	12
$1/32 > r \geq 1/64$	10	11
$1/64 > r \geq 1/128$	3	7
$1/128 > r \geq 0$	8	8
No known kinship	8	10

A software computes the value of the test statistic as $Q = 9.377$.

Do the data support the hypothesis that breeding amongst prairie dogs occurs randomly without regard to genetic relatedness?

Problem 10.1.8.

En sekvens av DNA från röd spindelapa (*Ateles geoffroyi*)
Denna uppgift sysslar med samma data som Exercise 9.25 i Vidakovic.

Ladda ned filen `dnadata.mat` från kursens hemsida och Matlabfiler3. Denna fil innehåller en sekvens av 8192 nukleotider från en röd spindelapas genom.

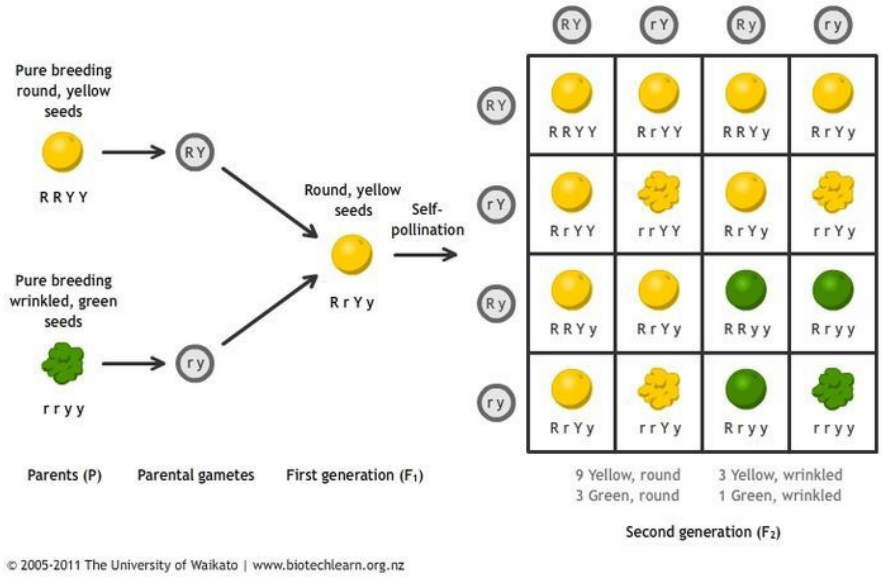
a) Bestäm frekvenserna av de fyra nukleotiderna i denna sekvens.

b) Genomför ett χ^2 -test för att undersöka om hypotesen (DNA dice)

$$\mathrm{H}_o: \quad P(A) = P(C) = P(T) = P(G) = \frac{1}{4}$$

kan anses hålla. Signifikansnivån skall vara 0.10.

Problem 10.1.9.



Man korsar (dihybrid korsning) två typer av bönor med runda gula resp. skrynkliga gröna frön. Enligt Mendels ärftlighetslära skall man i andra generationen (F_2) få fyra typer av frön med följande sannolikheter (jfr. Punnetts rutschema ovan):

$$P(\text{runda gula}) = \frac{9}{16}, P(\text{skrynkliga gula}) = \frac{3}{16}, P(\text{runda gröna}) = \frac{3}{16}, P(\text{skrynkliga gröna}) = \frac{1}{16}.$$

När man genomförde ett försök med 560 korsningar av den runda gula med den skrynkliga gröna typen fick man i andra generationen

330 runda gula, 100 skrynkliga gula,112 runda gröna, 18 skrynkliga gröna.

Testa med χ^2-metoden på den approximativa risknivån 5 % om Mendels teori verkar vara förenlig med ovanstående resultat.

10.2 Homogenitetstest, Oberoendetest

Problem 10.2.1.

I en studie undersöktes hår- och ögonfärg hos 6800 slumpmässigt utvalda tyska män. Resultatet redovisas i följande tabell.

	Mörkt hår	Ljust hår
Bruna ögon	726	131
Grå eller gröna ögon	2133	999
Blå ögon	996	1815

Utför ett test på signifikansnivån 0.1% av om det finns ett statistiskt säkerställt samband mellan ögonfärg och hårfärg hos tyska män. Ange tydligt de uppställda hypoteserna och motivera tydligt vilken slutsats som dras från testet.

Problem 10.2.2.

Inom medicinsk forskning används den så kallade VAS-skalan (*Visuell Analog Skala*) för att mäta smärta. Skalan är konstruerad som en 100 mm lång linje på vilken patienten får markera sin upplevda smärtnivå, varpå resultaten kan analyseras statistiskt. För att undersöka smärtnivån vid olika typer av kirurgi väljer en kirurg att dela upp VAS-skalan i tre kategorier: låg smärta (≤ 25) mm, acceptabel smärta 26-74 mm samt hög smärta (≥ 75) mm. Vidare valde kirurgen slumpmässigt ut 89 patienter vars smärta mättes 48 timmar efter en viss typ av operation. Av de 89 patienter som valdes ut i stickprovet var 42 som opererades med titthålskirurgi, och resten med traditionell kirurgi. Följande resultat erhölles.

	Låg smärta	Accept. smärta	Hög smärta
Titthålskirurgi	13	23	6
Traditionell kirurgi	7	28	12

Kan man hävda att smärtnivå skiljer sig åt mellan olika typ av kirurgi? Svara på frågan med hjälp av ett lämpligt statistiskt test på nivån 5%.

10.3 Ickeparametriska test

Problem 10.3.1.

Nedan finns data från två olika gruvor A och B. Data visar silverhalten i ounces per ton malm hos respektive malmprov från de 9 malmproverna från gruva A och de 14 från gruva B.

Silverhalt														
Gruva A	32	41	33	28	18	36	19	29	51					
Gruva B	22	25	28	34	28	33	28	23	20	19	42	21	20	37

Data varierar på ett sådant sätt att de inte kan anses vara normalfördelade.

Undersök på lämpligast sätt om om vi kan anta att gruvorna skiljer sig åt vad gäller silverhalten i malmen. Använd signifikansnivån 5% Ange tydligt vilka de uppställlda hypoteserna är. Din slutsats skall tydligt motiveras.

Gör en fullständig boxplott för de mätdata som kommer från gruva B. Ange även kvartilavstånd och variationsbredd för dessa mätdata.

Kapitel 11

Logistisk regression

11.1 Odds, Logistisk funktion, Logistisk fördelning, Odds ratio

Odds för en händelse A beräknas som:

$$\text{Odds (A)} = \frac{P(A)}{1 - P(A)}.$$

Oddsen för en säker händelse (d.v.s. $P(A) = 1$) är $+\infty$.

Enklare skriver vi med $p = P(A)$ och $0 \leq p \leq 1$ att

$$\text{Odds} = \frac{p}{1-p}.$$

Den naturliga logaritmen av Odds som funktion av p kallas **logit**-funktionen och skrivs som

$$\text{logit}(p) \stackrel{\text{def}}{=} \ln \frac{p}{1-p}$$

Problem 11.1.1.

Funktionen

$$\sigma(\lambda) \stackrel{\text{def}}{=} \frac{1}{1+e^{-\lambda}}$$

kallas den **logistiska** funktionen.

Verifiera att $\sigma(\lambda)$ är den inversa funktionen $\text{logit}^{-1}(p)$, d.v.s., om $\lambda = \text{logit}(p)$, så är

$$p = \text{logit}^{-1}(\lambda) = \sigma(\lambda).$$

Problem 11.1.2.

Låt $U \sim \mathcal{U}(0,1)$. Sätt

$$\epsilon \stackrel{\text{def}}{=} \log \frac{U}{1-U}.$$

Vilken fördelning har ϵ?

Problem 11.1.3.

$\epsilon \sim \text{Logistic}(0,1)$. Checka att

$$P(-\epsilon \leq x) = P(\epsilon \leq x)$$

Problem 11.1.4.

Antag att det finns endast två varandra uteslutande (mutually exclusive) hypotheser, H och icke-H, d.v.s H^c (t.ex. sjuk, frisk). Posteriori sannolikheten $p(H \mid E)$ för H givet evidensen E:

$$\text{Posterior Odds H} \stackrel{\text{def}}{=} \frac{p(H|E)}{1-p(H|E)}.$$

Verifiera att

Posterior Odds (H) = likelihood ratio × prior Odds

Oddskvot (odds ratio) betecknas med OR och är kvoten mellan oddsen från två olika villkor eller två olika populationer.

$$OR \stackrel{\text{def}}{=} \frac{\text{Odds } (A_1)}{\text{Odds } (A_2)} = \frac{\frac{P(A_1)}{1-P(A_1)}}{\frac{P(A_2)}{1-P(A_2)}}.$$

Det är viktigt att känna till den rätta tolkningen av OR.
OR mäter hur starkt närvaron/avsaknaden av egenskap A_1 i en population är associerad med närvaron/avsaknaden av egenskap A_2. Ett exempel nedan är A_1= intensiv/icke-intensiv ' användning av mobiltelefon och A_2 = en viss form avtumör i hjärnan/saknar denna tumör.

Problem 11.1.5.

X och Y är binära variabler. Sannolikhetsfördelningen för (X, Y) är $p_{xy} = P(X = x, Y = y)$ och ges av tabellen nedan

	$Y=1$	$Y=0$
$X=1$	p_{11}	p_{10}
$X=0$	p_{01}	p_{00}

Här är cellsannolikheterna p_{xy} icke-negativa och $p_{11} = P(X = 1, Y = 1)$, $p_{10} = P(X = 1, Y = 0)$, o.s.v.. Det skall gälla att

$$p_{11} + p_{10} + p_{01} + p_{00} = 1.$$

Per definition har vi att $P(Y = 1|X = 1) = \dfrac{P(Y = 1 \cap X = 1)}{P(X = 1)} = \dfrac{p_{11}}{p_{11} + p_{10}}$ och $P(Y = 0|X = 1) = \dfrac{P(X = 1 \cap Y = 0)}{P(X = 1)} = \dfrac{p_{10}}{p_{11} + p_{10}}$.

Låt oss nu tänka oss att vi vill jämföra två läkemedel A och B beträffande förmågan att förebygga recidivinfarkt (= återfall av hjärtinfarkt) hos personer som just haft infarkt. Vi har en grupp patienter som haft hjärtinfarkt. De får antingen läkemedel A eller läkemedel B. Under en viss förutbestämd period (follow-up) följes patienterna. Låt Y = hjärtinfarkt under follow-up perioden (1) eller inte (0). Läkemedel A kodas som och $X = 1$, och läkemedel B kodas som $X = 0$.

Låt nu A_1 vara händelsen $Y = 1$ givet att man fått läkemedlet A. Låt nu A_2 vara händelsen $Y = 1$ givet att man fått läkemedlet B.

a) Vad är oddsen för A_1?

b) Vad är oddsen för A_2?

c) Checka att

$$OR = \frac{p_{11}p_{00}}{p_{10}p_{01}}.$$

Om $OR > 1$ indikerar detta att recidivinfarkt är mer sannolik hos dem, som fått A än de som fått B. Om $OR < 1$ indikerar detta att recidivinfarkt är mer sannolik hos dem, som fått B än de som fått A.

Problem 11.1.6.

X och Y är binära variabler. Sannolikhetsfördelningen för (X, Y) är $p_{xy} = P(X = x, Y = y)$ och ges av tabellen nedan

	$Y=1$	$Y=0$
$X=1$	p_{11}	p_{10}
$X=0$	p_{01}	p_{00}

Bestäm OR, när X och Y är oberoende. *Ledning:* Formeln i **c)** i den föregående uppgiften.

Logoddskvot

$$\ln OR = \ln p_{11} + \ln p_{00} - lnp_{10} - lnp_{01}.$$

Ofta har vi data givna i en **kontingenstabell**.

	$Y=1$	$Y=0$
$X=1$	n_{11}	n_{10}
$X=0$	n_{01}	n_{00}

där n_{xy} är frekvensen av (x,y) i en undersökning. Då kan cellsannolikheterna p_{xy} estimeras statistiskt med

	$Y=1$	$Y=0$
$X=1$	$\hat{p}_{11}$	$\hat{p}_{10}$
$X=0$	$\hat{p}_{01}$	$\hat{p}_{00}$

där $\widehat{p}_{xy} = n_{xy}/n$ med $n = n_{11}+n_{10}+n_{01}+n_{00}$, summan cellfrekvenserna.

Logoddskvoten för kontingenstabell är

$$L = \log\left(\frac{\hat{p}_{11}\hat{p}_{00}}{\hat{p}_{10}\hat{p}_{01}}\right) = \log\left(\frac{n_{11}n_{00}}{n_{10}n_{01}}\right).$$

En liten överkurs i Logoddskvoten
Det gäller approximativt att

$$L \sim \mathcal{N}(\ln(OR), \sigma^2).$$

Standardavvikelsen för statistikan L är approximativt

$$\mathrm{s} = \sqrt{\frac{1}{n_{11}}+\frac{1}{n_{10}}+\frac{1}{n_{01}}+\frac{1}{n_{00}}}.$$

Problem 11.1.7.

I en studie utförd bland de anställda av EGAT (=Electricity Generating Authority of Thailand) undersökte man rökningens verkan på dödlighet i hjärt- och kärlsjukdomar. Följande data registerades.

X =Rökare i början	Y = Kardiovaskulär död under follow-up		
	Ja (1)	Nej (0)	Total
Ja (1)	31	1386	1417
Nej (0)	15	1883	1898
Total	46	3269	3315

Beteckningar:
A_1: Y=1 givet att man är rökare. A_2: Y=1 givet att man är icke-rökare.

a) Vad är oddsen för A_1?

b) Vad är oddsen för A_2?

c) Vad är OR? Vad säger det erhållna värdet på OR om dödlighet i hjärt- och kärlsjukdomar för rökare v.s. icke-rökare.

d) Ge ett approximativt 95 %-igt konfidensintervall för ln OR i populationen. *Ledning*: Se överkursen ovan. Vad är konfidensintervallet för OR?

e) Kan Du förkasta hypotesen att dödlighet i hjärt- och kärlsjukdomar är oberoende av rökning på signifikansnivån 0.05?

11.2 Logistisk regression

Låt oss sätta

$$\lambda \stackrel{\text{def}}{=} \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

där x_1 och x_2 är t.ex. två förklarande variabler. s
Antag att för $0 < op < 1$

$$\text{logit}(p) = \lambda = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

Då följer att

$$p = \frac{e^{\beta_0+\beta_1 x_1+\beta_1 x_2}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2}},$$

En binär stokastisk variabel Y sådan att

$$p = P(Y=1|x_1,x_2) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2}}$$

säges satisfiera en **logistisk regression**.

Problem 11.2.1.

Bestäm $P(Y=0|x_1,x_2)$ för Y som satisfierar logistisk regression.

Problem 11.2.2.

Formlerna i denna övning spelar en roll i Newton-Raphson algoritmen vid maximum likelihood estimat för parametrarna i logistisk regression. Vi betraktar den logistiska regressionen:

$$P(Y=1 \mid x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

Checka att

$$\frac{\partial}{\partial\beta_0}P(Y=1 \mid x) = P(Y=1 \mid x)(1-P(Y=1 \mid x)).$$

Här är $\frac{\partial}{\partial\beta_0}P(Y=1 \mid x)$ derivatan av $P(Y=1 \mid x)$ med avseende på β_0.

Checka att

$$\frac{\partial}{\partial\beta_1}P(Y=1 \mid x) = P(Y=1 \mid x)(1-P(Y=1 \mid x))x.$$

Här är $\frac{\partial}{\partial\beta_1}P(Y=1 \mid x)$ derivatan av $P(Y=1 \mid x)$ med avseende på β_1.

Problem 11.2.3.

Suppose we have two populations, where $x_1^{(i)} = 1$ in the population (i) and $x_1^{(ii)} = 0$ in the second population, the other predictor $x_2^{(i)} = x_2^{(ii)}$ being equal in the two populations. Then

$$p_1 = \frac{e^{\beta_0+\beta_1 x_1^{(i)}+\beta_2 x_2}}{1+e^{\beta_0+\beta_1 x_1^{(i)}+\beta_2 x_2}}$$

$$p_2 = \frac{e^{\beta_0+\beta_1 x_1^{(ii)}+\beta_2 x_2}}{1+e^{\beta_0+\beta_1 x_1^{(ii)}+\beta_2 x_2}}.$$

The logarithm of the odds ratio becomes

$$\ln OR = \ln\frac{p_1}{1-p_1} - \ln\frac{p_2}{1-p_2}$$

$$= \beta_1\left(x_1^{(i)} - x_1^{(ii)}\right).$$

Hence a unit change in x_1 corresponds to e^{β_1} change in odds and β_1 change in logodds.

Logistic regression

$$\widehat{\text{logit}} = -4.8326 + 1.0324x$$

was estimated (fitted) with $x = 1$ for smokers (i) and $x = 0$ for non-smokers (ii) with the EGAT data below.

Smoker at entry	Cardiovascular death during follow-up		
	Yes	No	Total
Yes	31	1386	1417
No	15	1883	1898
Total	46	3269	3315

a) Find the odds ratio for cardiovascular death amongst smokers and amongst non-smokers using the estimated logistic regression.

b) Find the odds for cardiovascular death amongst smokers and amongst non-smokers using the estimated logistic regression.

c) Find the probability (risk) for cardiovascular death amongst smokers and amongst non-smokers using the estimated logistic regression.

Problem 11.2.4.

> The purpose of a population-based, case-control study was to test the hypothesis that long-term mobile phone use increases the risk of brain tumors. The authors identified all cases aged 20-69 years who were diagnosed with glioma or meningioma during 2000-2002 in certain parts of Sweden. Randomly selected controls were stratified on age, gender, and residential area. Detailed information about mobile phone use was collected from 371 (74%) glioma and 273 (85%) meningioma cases and 674 (71%) controls. For regular mobile phone use, the odds ratio (using logistic regression) was 0.8 (95% confidence interval: 0.6, 1.0) for glioma and 0.7 (95% confidence interval: 0.5, 0.9) for meningioma.
>
> Lönn, Stefan and Ahlbom, Anders and Hall, Per and Feychting, Maria: *Long-term mobile phone use and brain tumor risk.* **American journal of epidemiology**, 166, pp. 526–535 2005.

What is your conclusion about mobile phone use and brain tumor risk based on these ORs?

Problem 11.2.5.

Physical and hormonal factors contribute to bone weakness in aging women. A study examined the number of bone fractures among a random sample of 1469 elderly men and women. We set the binary response variable Y =fracture to be equal to 1 for one or more fractures and 0 for no fracture in the previous three years. In this context, $p = P(Y = 1|x)$ is the probability of having a bone fracture in the past three years for an elderly person. We consider the logistic regression of fracture as a function of x =gender (0 for men and 1 for women). We estimate by suitable software the following

$$\text{logit}(p) = -4.758 + 1.089\text{gender}.$$

The same software outputs also that standard error (medelfelet) for $\beta_0 = 0.409997$ and standard error for $\beta_1 = 0.471244$

Note that here

$$\text{logit}(p) = \beta_0 + \beta_1 x.$$

and

$$p = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}$$

so that

$$odds(\text{fracture}|\text{x}) = \frac{p}{1-p} = \frac{\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}}{1 - \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}}$$
$$= e^{\beta_0+\beta_1 x}.$$
$$OR = \frac{odds(\text{fracture}|\text{x} = 1)}{odds(\text{fracture}|\text{x} = 0)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}$$

Which of the following statements is **true**?

a) The odds of a fracture are not significantly different for men and for women.

b) The odds of a fracture are significantly higher for women than for men, about 3 to 7.5 times with 95% confidence.

c) The odds of a fracture are significantly higher for women than for men, about 1.2 to 7.5 times with 95% confidence.

11.3 'Liability threshold"-modellen i etiologi

Problem 11.3.1.

Det antas att ett antal olika gener och miljöfaktorer, som alla har var för sig en liten effekt, verkar som riskfaktorer ellet skyddande faktorer för utveckling av en sjukdom som t.ex. schizofreni. När tillräckligt mycket av riskfaktorer ackumulerat sig och överväger de skyddande faktorerna, kan sjukdomen uppstå. Eftersom många olika faktorer bidrar till detta, har denna liabilitet (belastning, predisposition, sårbarhet) i verkligheten en kontinuerlig kvantitativ skala. Såsnart som liabilitet passerar en viss punkt eller tröskel (threshold), då uppstår en diskret fenotyp (t.ex. drabbad av schizophreni versus ej drabbad av schizofreni).

Denna s.k. **liability threshold** modell gör antagandet att den underliggande liabiliteten att drabbas av en sjukdom är normalfördelad i en befolkning. X =liability är en stokastisk variabel och

$$X \sim \mathcal{N}(\mu, \sigma^2).$$

Tröskeln (d.v.s. liability threshold) t är ett tal sådant att

$$P(X \geq t) = p.$$

Med ord, om liabilitet överstiger tröskeln t, så är sannolikheten att bli sjuk lika med p. Vi kombinerar nu detta med logistisk regeression genom att ta

$$p = \text{logit}^{-1}(\lambda) = \sigma(\lambda),$$
$$\lambda = \beta_0 + \beta_1 x_1 + \beta_1 x_2,$$

där x_1 och x_2 är t.ex. två faktorer som påverkar förekomsten av en sjukdom (eller att en individ har exponerats för dessa variabler). p vara sannolikheten att en individ i befolkningen drabbas av sjukdomen.
Detta innebär att om två individer har samma värden på x_1 och x_2, så har de samma sannolikhet för sjukdomen. d.v.s.

$$p = \frac{e^{\beta_0+\beta_1 x_1+\beta_1 x_2}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2}},$$

Vår uppgift är att bestämma λ, μ och σ, om p är givet.

Sham, Pak C and Walters, EE and Neale, Michael C and Heath, Andrew C and MacLean, CJ and Kendler, Kenneth S : Logistic regression analysis of twin data: estimation of parameters of the multifactorial liability-threshold model, 24, pp. 229−238, **Behavior genetics**, 1994.

a) Checka att om $X \sim \mathcal{N}(\mu, \sigma^2)$ och $P(X \geq t) = p$, så gäller för t, μ, σ och λ sambandet

$$\Phi\left(\frac{t-\mu}{\sigma}\right) = \frac{1}{1+e^{\lambda}}.$$

b) Vi kan, vid närmare eftertanke, sätta $t = 0$ utan att ändra på analysens allmängiltighet. Vi väljer även $\mu = \lambda$ utan ändra på analysens allmängiltighet. Lös då ekvationen

$$\Phi\left(\frac{-\lambda}{\sigma}\right) = \frac{1}{1+e^{\lambda}}.$$

m.a.p. σ. Du behöver den inversa funktionen Φ^{-1}, som är även känd som **probit** function.

c) Låt oss återvända till den logistiska regressionen för EGAT data d.v.s.

$$\lambda = \widehat{\text{logit}} = -4.8326 + 1.0324x$$

som skattades för dessa data med $x = 1$ för rökare och $x = 0$ för icke-rökare. Bestäm nu σ_x för $x = 1$ och $x = 0$ och sedan liabilitysannolikheterna,

$$P(X > 0) \quad x = 0, x = 1,$$

där $X \sim \mathcal{N}(-4.8326 + 1.0324x, \sigma_x^2)$. Kommentera resultatet.

Hjälp : Med Matlab som hjälpmedel kommer Du att behöva här probit implementerad som `norminv`:

```
X = norminv(P,MU,SIGMA)}  returns the inverse cdf for the normal
   distribution with mean MU and standard deviation SIGMA, evaluated at
   the values in P.
```

Kapitel 12

Data-analys Del V: Bayesiansk data-analys

12.1 Inledning

Bayes: Räkna med vad du tror
I den klassiska (frekventistiska) statistiken förknippas sannolikhetsbegreppet med en situation som kan upprepas - man gör slumpmässiga dragningar av individer, observationer, ur en viss population och uttalar sig sedan om den bakomliggande populationen på grundval av dessa observationer.
I den bayesianska statistiken används sannolikheter även för att mäta graden av tilltro, och man integrerar därvid information från nya observationer med tidigare tillgänglig kunskap eller gissningar. Fundamental i denna metodik är Bayes sats.

A. Taube & J. Malmquist: Räkna med vad du tror. Bayes' sats i diagnostiken. *Läkartidningen* 2001, Vol. 98, Nr 24, sid. 2910−2913

Vi behöver i övningarna i detta kapitel en ny form av Bayes' sats. Vi betraktar ett exempel, $X \sim \mathcal{B}er(p)$ (Bernoullifördelning med parameter p, $0 \leq p \leq 1$). Detta innebär med andra ord att

x	0	1
$p_X(x)$	$1-p$	p

Nu ger vi funktionen $p_X(x)$ i tabellen ovan en alternativ symbol som $p_X(x \mid p)$ (betingad sannolikhet för $X = x$ givet p) för att explicit lyfta fram p som parameter. Låt dessutom $\pi(p)$ vara en sannolikhetskurva som funktion av p.

Bayes' sats ger $\pi(p \mid x)$:
Sannolikhetskurvan för p givet x är

$$\pi(p \mid x) = \frac{p_X(x \mid p)\pi(p)}{m(x)},$$

där

$$m(x) = \int_0^1 p_X(x \mid p)\pi(p)dp.$$

- $\pi(p)$ **a priori** fördelning/täthet för p.
- $p_X(x \mid p)$ **likelihood**, **sannolikhetsmodell**
- $\pi(p \mid x)$ **posteriori** fördelning/för p givet x.
- $m(x)$ marginalfördelning.

Vi har att

$$\int_0^1 \pi(p \mid x)dp = 1.$$

12.2 Bayes och klinisk prövning

Problem 12.2.1.

Vi fortsätter med de kliniska prövningarna från uppgift XXXX och gör en data-analys med Bayes' sats ovan som hjälpmedel. Vi väljer en a priori fördelning $\pi(p)$ för den okända proportionen p. Denna a priori fördelning ger uttryck för tidigare tillgänglig (expert)kunskap eller gissningar eller osäkerhet om populationsparametern p. Vi 'räknar med vad du tror' och väljer $\pi(p)$ som

$$\pi(p) = \begin{cases} 1 & 0 \leq p \leq 1 \\ 0 & \text{för övrigt.} \end{cases}$$

Därmed säger vi att $p \sim \mathcal{U}(0,1)$ (standard uniform, Vidakovic sid. 161-162). Observera därtill att $\mathcal{U}(0,1) = \mathcal{B}e\,(1,1)$, **Betatätheten** (Beta density, se bilagan nedan). Detta uttrycker (fullständig) brist på förhandskunskap eller tro om p. Efter valet av prior satsar vi på Design 1, varmed tio patienter behandlas efter varandra med den nya terapin.
Antag att det första utfallet är en framgång (S). Vad är posteriorin $\pi(p|S)$? Vi har data av kategorityp, så vi skriver

x	F	S
$p_X(x\|p)$	$1-p$	p

så att likelihood $\times$ prior är

$$p_X(S \mid p)\pi(p) = p \cdot 1.$$

Därför fås marginalfördelningen som

$$m(S) = \int_0^1 p_X(S \mid p)\pi(p)dp = \int_0^1 pdp = \left[\frac{p^2}{2}\right]_0^1 = \frac{1}{2}.$$

Således ger Bayes' sats

$$\pi(p \mid S) = \begin{cases} 2p & 0 \leq p \leq 1 \\ 0 & \text{för övrigt.} \end{cases}$$

Figuren nedan (från Berry, Donald A: *Bayesian clinical trials*, **Nature reviews Drug discovery**,5, pp. 27-36, 2006) är uppdelad i tolv fält. I fältet högst till vänster (rubrik: Prior) plottas $\pi(p)$ och i följande fältet (After S) till höger plottas $\pi(p \mid S) = 2p$.

Antag att det andra utfallet är igen en framgång (S). Vad är posteriorin $\pi(p|SS)$? Vi antar att utfallen är oberoende. Då fås

$$p_X(SS \mid p) = p_X(S \mid p) \cdot p_X(S \mid p) = p^2$$

och

$$m(SS) = \int_0^1 p_X(SS \mid p)\pi(p)dp = \int_0^1 p^2 dp = \left[\frac{p^3}{3}\right]_0^1 = \frac{1}{3}$$

och Bayes' sats ger

$$\pi(p \mid SS) = \begin{cases} 3p^2 & 0 \leq p \leq 1 \\ 0 & \text{för övrigt.} \end{cases}$$

Det andra fältet (Another S) till höger om Prior i figuren visar grafiskt denna sannolikhetsfördelning. De två lyckade terapierna har flyttat massan i den likformiga a priori sannolikheten för p åt höger så att fördelningen är skev åt vänster.

Nästa utfall blir ett första F. Vad blir posteriorin $\pi(p|SSF)$? Med oberoende utfall har vi

$$p_X(SSF \mid p) = p_X(S \mid p) \cdot p_X(S \mid p) \cdot \underbrace{p_X(F \mid p)}_{=(1-p)} = p^2(1-p).$$

Då följer

$$m(SSF) = \int_0^1 p_X(SSF \mid p)\pi(p)dp = \int_0^1 p^2(1-p)dp =$$
$$= \int_0^1 p^2 dp - \int_0^1 p^3 dp = \frac{1}{3} - \frac{1}{4} = \frac{1}{12}$$

och Bayes' sats ger

$$\pi(p \mid SSF) = \begin{cases} 12p^2(1-p) & 0 \leq p \leq 1 \\ 0 & \text{för övrigt.} \end{cases}$$

a) Beräkna $\int_0^{0.35} \pi(p \mid SSF)dp$.
Ledning: $\pi(p \mid SSF)$ är i själva verket **Betatätheten** $\mathcal{B}e\,(3,2)$, så i Matlab kan man använda `betacdf(0.35,3,2)`. Det går att räkna med penna och kalkylator, gör gärna detta som kontrollräkning (av Matlab).

Vad säger **a)** om (posteriori)sannolikheten för att den nya terapin är sämre än standardterapin?

b) Bestäm $\pi(p \mid SSFSSFSSSF)$ och plotta denna som funktion av p i $[0,1]$ och jämför med fältet 'Final" i nedersta raden. Obs! Du behöver inte jobba genom hela sekvensen från utfall till nästa utfall.
Hjälp: Om Du tillämpar räknegången ovan, har Du utdelning av bilagan nedan med **Betaintegralen**, som för de positiva heltalen n och k ger

$$\int_0^1 p^n(1-p)^k dp = \frac{\Gamma(n+1)\cdot\Gamma(k+1)}{\Gamma(n+k+2)}.$$

och att $\Gamma(r+1) = r!$.

c) Beskriv m.h.a. fältet Final maximumlikelihoodskattningen $\widehat{p}_{mle}$ på basis av data SSFSSFSSSF i förhållande till $\pi(p \mid SSFSSFSSSF)$.

d) Beräkna $\int_0^{0.35} \pi(p \mid SSFSSFSSSF)dp$. Vad säger denna (posteriori)sannolikhet om huruvida den nya terapin är sämre än standardterapin?

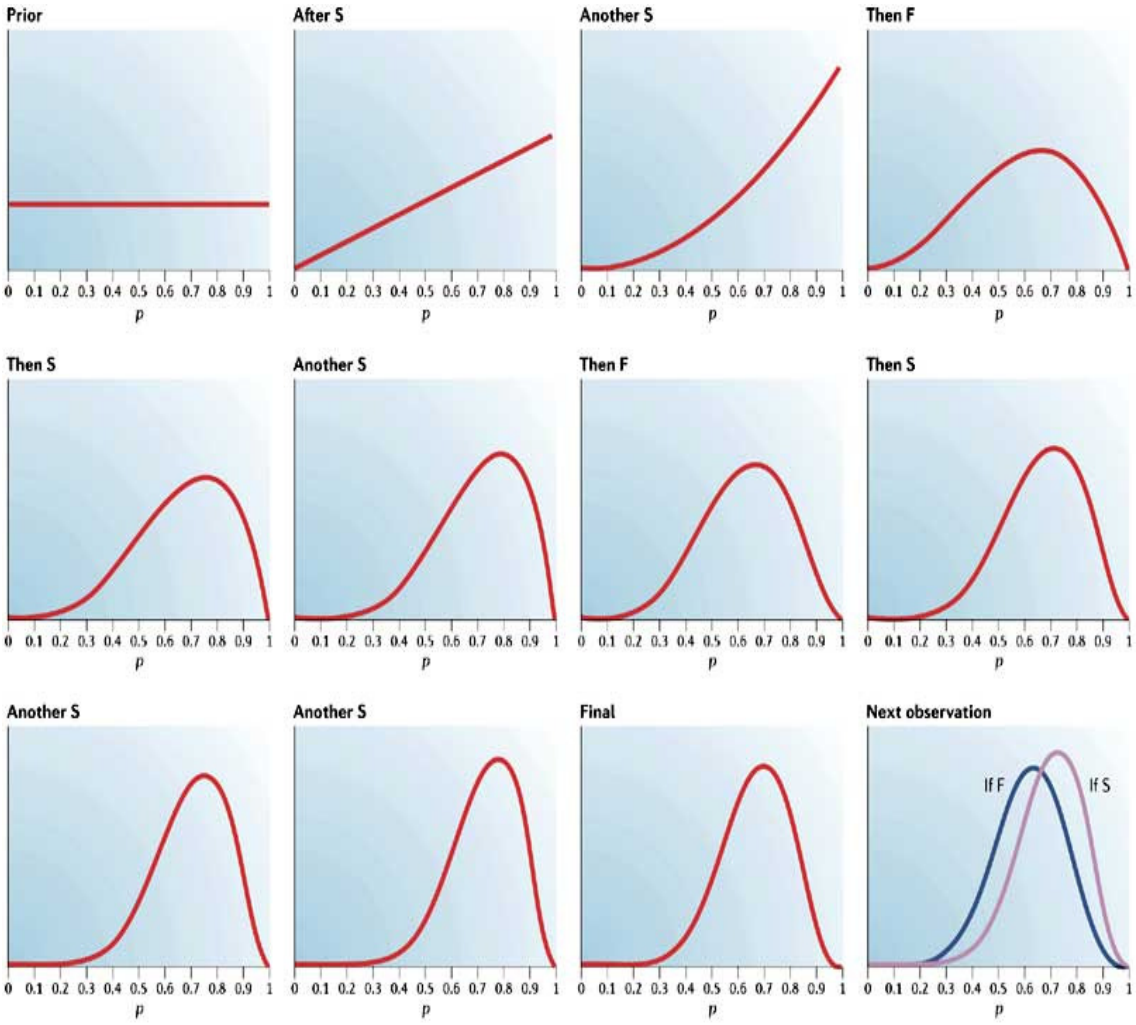
f) Vad kommer till uttryck i de 11 fälten från Prior till Final? Vad uttrycker fältet Next observation?

e) **Extra** Vad är den betingade sannolikheten för att en följande patient, som får den nya terapin, är missllyckande? Obs! Denna sannolikhet ges inte i fältet Next observation!
Två förslag:

i) $P(F \mid SSFSSFSSSF) = \int_0^1 p\pi(p \mid SSFSSFSSSF)dp$.

ii) $p_X(F \mid \widehat{p}_{mle})$.

iii) Jämför svaren **i) -ii)**.



D.A. Berry: *Bayesian clinical trials.* **Nature Reviews Drug Discovery** 5, 27-36 (January 2006)

Problem 12.2.2.

If nothing goes wrong, is everything all right? Rule of Three and Bayes

Vi vill göra en jämförelse av Rule-of-Three med posteriorisannolikhet. Vi har

$$\begin{array}{c|cc} x & \text{F} & \text{S} \\ p_X(x|p) & (1-p) & p \end{array}$$

så att med tio oberoende utfall $P(FFFFFFFFFF) = (1-p)^{10}$, nothing goes wrong. I detta får F stå för ett försök inte har gått fel). Detta innebär att vi har posterioritätheten `betapdf(p,1,11)` i Matlab för $\pi(p \mid \text{FFFFFFFFFF})$.

a) Ge den explicita formeln för $\pi(p \mid \text{FFFFFFFFFF})$, om Du använder den likformiga $\pi(p)$ som prior.

b) Plotta $\pi(p \mid \text{FFFFFFFFFF})$ för $p \in [0,1]$.

c) Beräkna

$$\int_0^{3/10} \pi(p \mid \text{FFFFFFFFFF}) dp$$

och diskutera denna sannolikhet m.a.p. konfidensintervallet i Rule-of-Three.

Problem 12.2.3.

Diskutera följande korta anförande:

Practitioners usually seek the 'best solution' implied by the data, but observations should only be used to falsify possible solutions, not to deduce any particular solution.

12.3 Bilaga: Betaintegral

Betaintegral
Samtliga övningarna ovan innehåller exempel på följande integral som kallas *Betaintegral*. n och m är positiva heltal.

$$\int_0^1 x^n (1-x)^m dx = \frac{n! \cdot m!}{(n+m+1)!}. \tag{12.1}$$

J.f.r Adams & Essex: Calculus. A Complete Course 7th Ed., p.818.

The core of the **BayesMendel** working group collaborates from Dana Farber Cancer Institute at Harvard University and The Johns Hopkins Sidney Kimmel Comprehensive Cancer Center. The BayesMendel Lab is dedicated to the development of methodologies, models, and open source software for predicting who may carry a cancer susceptibility gene. We use statistical ideas that go back to Bayes (pictured on the left) and genetic models that go back to Mendel (pictured on the right).
`http://bcb.dfci.harvard.edu/bayesmendel/index.php`

Kapitel 13

Variansanalys

13.1 Ensidig variansanalys

Problem 13.1.1.

A microbiologist is studying the bacterial contamination of Swedish one krona (1 SEK) coins. The microbiologist collects four coins at random from each of three diffrent kinds of premises: a fast food street stand, a shop selling sandwiches, and from a shop selling newpapers and magazines.

The data on the number of bacteria isolated from the twelve randomly chosen 1 SEK coins is recorded in the table.

Shop	Observations					
Fast food stand	140	108	76	400	$\overline{y}_1 = 181$	$s_1 = 150$
Sandwich shop	2	21	0	42	$\overline{y}_2 = 16.25$	$s_2 = 20$
Newspaper shop	40	5	5	0	$\overline{y}_3 = 12.5$	$s_3 = 18.5$

Here $\overline{y}_i$ and s_i are the sample mean and standard deviation for $i = 1$ fast food stand, $i = 2$ Sandwich shop, $i = 3$ Newspaper & Magazine shop.

The microbiologist wants to test the hypothesis that there is no significant difference between the unknown population means of the bacterial counts from the three shops. The level of significane is chosen in advance as 5% .
A statistical software package for one-way ANOVA produces the following:

Source	df	SS	MSS
Between shops	2	74065.167	37032.583
Within shops	9	68173.75	7574.861

Test statistic is $F = \frac{\text{MS (between shops)}}{\text{MS(within shops)}} = 4.89$. The p-value is $= 0.037$. F has in this case the Fisher distribution $F(2, 9)$.

Which one (and the only one) of the following is an instance of a correct statement based on ANOVA?

a) As the critical value is $F_{0.037}(2, 9) = 4.86$, we reject the null hypothesis and conclude that all the means are different.

b) Since $F_{0.037}(2, 9) = 4.86$, the p-value is the probability that F is larger than 4.86.

c) Since $p = 0.037 < 0.05$, there is less than 5% chance of obtaining F of the magnitude 4.89. We reject the null hypothesis that all means are equal at significance level 5%.

d) $\overline{y}_1 = 181$ is clearly much bigger than both $\overline{y}_2 = 16$ and $\overline{y}_3 = 13$. Therefore all the population means cannot be equal.

Problem 13.1.2.

Tre termometrar används regelbundet i ett laboratorium. För att kontrollera den relativa noggrannheten hos termometrarna placerades dessa i slumpmässig ordning i en cell som höll 0 ^{o}C. Varje termometer placerades i cellen fyra gånger, med följande resultat.(^{o}C).

Termometer	1	2	3
	0.10	-0.20	0.90

0.90	0.80	0.20
-0.80	-0.30	0.30
-0.20	0.60	-0.30

Hjälpsumma:$\sum_i \sum_j (y_{ij} - \bar{y}_{..})^2 = 3.3267$

Vi gör en ensidig variansanalys. Låt y_{ij}=observation nr j från termometer nr i.
Då fås $\bar{y}_{1.} = 0 \quad \bar{y}_{2.} = 0.225 \quad \bar{y}_{3.} = 0.275$ och $\sum_i (\bar{y}_{i.} - \bar{y}_{..})^2 = 0.0429$. Härav fås variansanalystabellen

Källa	Fg	Kvs	Mkvs
Mellan termometrar	2	0.1717	0.0858
Inom termometrar	9	3.1550	0.3506
Totalt	11	3.3267	

Testa på 5%-nivån om termometrarna mäter likvärdigt.

Problem 13.1.3.

Vid ett laboratorieförsök tillverkades papper under fem olika tryck. Hos de tillverkade pappersarken registrerades den s k slitfaktorn. För två av de fem trycken undersöktes sju prover pappersark medan man för de tre övriga trycken nöjde sig med fyra prover vid vardera trycket. Resultat:

tryck	slitfaktor							$\bar{y}_i$	stickpr.var s_i^2
35	112	119	117	113	108	111	115	113,5714	13,9524
50	108	99	112	118				109,25	63,5833
71	120	106	102	109				109,25	59,5833
100	110	101	99	104				103,5	23,0000
141	100	102	96	101	102	101	105	101	7,3333

$\bar{y}_{..} = 107{,}3077 \quad \sum_i \sum_j (y_{ij} - \bar{y}_{..})^2 = 1207{,}5385$. Modell: Tryckbestämningarna sker utan fel. Slitfaktorbestämningarna har oberoende normalfördelade fel med väntevärde 0 och konstant varians.

Vi har ensidig variansanalys där $k = 5$ och $n_1 = n_5 = 7$ och $n_2 = n_3 = n_4 = 4$ och $N = 2 \cdot 7 + 3 \cdot 4 = 26$. Vi får kvs inom stickprov till
$(7-1)13.9524 + (4-1)63.5833 + (4-1)59.5833 + (4-1)23.0000 + (7-1)7.3333 = 566.2$
som ger variansanalys-tabellen

Källa	fg	kvs	mkvs
Mellan betingelser	4	641.3	160.3
Inom betingelser	21	566.2	27.0=$\widehat{\sigma}^2$
Total	25	1207.5	

där vi fått kvs "Mellan betingelser" genom subtraktion.

a) Undersök huruvida trycket har signifikant inverkan på slitfaktorn.

b) Ge ett 95% konfidensintervall för förväntade slitfaktorn vid trycket 35.

Problem 13.1.4.

Till ett laboratorium hade inkommit vattenprover från fyra sjöar. På var och en av proverna gjordes två analyser av pH-värdet. Analyserna är behäftade med slumpmässiga fel med samma varians.

Sjö	Analysresultat	
1	4.5	4.7
2	4.8	4.9
3	5.7	5.9
4	6.1	6.2

I en ensidig variansanalys antar man alltså att med y_{ij}=j:te värdet för sjö nr i med $i = 1, 2, 3, 4$ och $j = 1, 2$ så är dessa utfall av oberoende stokastiska variabler Y_{ij} som är $\mathcal{N}(\mu_i, \sigma^2)$. Vi har alltså modell A (systematisk faktor) och har $n_i = 2$ för $i = 1, 2, 3, 4$ samt $k = 4$ och $N = 4 \cdot 2 = 8$.

Vi får

$$\sum_{i=1}^{4} n_i (\bar{y}_{i.} - \bar{y}_{..})^2 =$$

$$= 2\left((4.6-5.35)^2+(4.85-5.35)^2+(5.8-5.35)^2+(6.15-5.35)^2\right) = 3.31.$$

Vi kombinerar detta med den givna kvadratsumman som svarar mot

"total-kvadratsumman" och erhåller för kvs "Inom sjöar" värdet 3.36-3.31=0.05. Detta ger följande variansanalystabell

Källa	Fg	Kvs	Mkvs
Mellan sjöar	3.31	3	1.1033
Inom sjöar	0.05	4	$0.0125=\widehat{\sigma}^2$
Totalt	3.36	7	

Testa huruvida sjöarnas pH-halt skiljer sig åt.

Problem 13.1.5.

Vi citerar utförligt indelningen i Nielsen, Lars K and Smyth, Gordon K and Greenfield, Paul F: *Hemacytometer Cell Count Distributions: Implications of Non-Poisson Behavior*, **Biotechnology progress**, 1991.

> Enumeration of cells via a hemacytometer is a standard technique. Accurate estimates of cell numbers are essential for developing and validating quantitative models of cell growth and product expression. The two main parameters determining the accuracy of cell enumeration using the hemacytometer technique are the number of hemacytometers that are counted (i.e., number of fillings) and the number of cells counted within each hemacytometer. A priori knowledge of the underlying distribution of measured cell counts can be used (a) to determine how many cells and sides have to be counted to obtain a particular level of reliability and (b) to increase the reliability for a given number of sides and cells counted. The latter is due to the fact that if the variance is known, no reliability is lost through estimation of this parameter. Under the assumptions that samples are fully suspended, that the volume over the counting area on the hemacytometer is constant, and that cells are distributed randomly over the hemacytometer, cell counts can be described by
>
> $$X = \text{cell count } \sim \mathcal{P}oi(VN),$$
>
> where V is the volume over the counting area and N is the cell concentration in the sample. if the cell counts are distributed according to a Poisson distribution, the variance is equal to the mean. In our laboratory, however, we have generally observed a variance significantly higher than the mean, suggesting that cell counts do not follow a Poisson distribution. This has occurred despite the application of very careful analytical procedures. We investigate possible reasons for the higher variance and the use of alternative statistical distributions to describe hemacytometer cell counts more accurately.
>
> The data set consists of 360 observations from 20 samples, where the two sides of a hemacytometer have been counted and the cell count has been recorded for each of the nine major squares in the grid on the hemacytometer.
>
> $$X_{ijk} = \text{cell count sample i, side j , square k} \sim \mathcal{P}oi(\lambda_{ij}).$$

Det kan visas att approximativt

$$Y_{ijk} = \sqrt{X_{ijk}+\frac{3}{8}} \sim \mathcal{N}(\sqrt{\lambda_{ij}}, 0.25).$$

Y_{ijk} är en s.k. half power transformation av X_{ijk}.

Då vill vi testa om variansen i half power transformerade data har varians som signifikant skiljer sig från 0.25. Detta går att ordna med en variansanalys:

$$Y_{ijk} = \mu + a_j + S_{j(i)} + E_{k(ij)}$$

där μ är totalmedelvärdet, a_j är ett tal som adderas för att kompensera för den kända differensen mellan stickproven, $S_{j(i)}$ är normalfördelad addition till j:te sidan av det i:te stickprovet och $E_{k(ij)}$ är en normalfördelad addition till k:te kvadraten.

Resultatet visas i ANOVA-tabellen nedan.

Source	df	SS	MSS
Between sides	20	10.30	$0.50/0.27 \approx 0.84$
Within sides	320	87.35	$87.35/0.25 \approx 349.4$

Vad är slutsatsen?

13.2 Tvåsidig variansanalys

Problem 13.2.1.

Vid ett s.k. ringtest skulle man jämföra fyra laboratorier (B_1, B_2, B_3 och B_4) med varandra. De fick mäta halten av ett visst enzym i tre stycken prover ($A_1, A_2,$ och A_3). Data (i lämplig enhet):

	B_1	B_2	B_3	B_4	$\bar{y}_{i.}$
A_1	6.43	7.36	7.03	6.22	6.76
A_2	9.35	9.86	9.08	9.27	9.39
A_3	6.03	6.90	6.69	6.50	6.53
$\bar{y}_{.j}$	7.27	8.04	7.60	7.33	$\bar{y}_{..} = 7.56$

$\sum_i \sum_j y_{ij}^2 = 707.6222$. Man antar att det kan finnas systematiska skillnader mellan laboratorierna, men att dessa skillnader i så fall är desamma oberoende av prov. Följande kvadratsummor beräknades

Mellan prov	20.1992
Mellan laboratorier	1.1070
Totalt	21.7790

Vi gör här en två-sidig variansanalys med en replikation/cell och vi har en additiv modell dvs $Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$ där Y_{ij}=resultatet för prov nr i vid laboratorium j, $i = 1, 2, 3$, $j = 1, 2, 3, 4$. Vi får följande variansanalystabell där $p = 3$, $q = 4$ och där kvs för residualen fås med subtraktion till $21.7790 - 20.1992 - 1.1070 = 0.4728$.

Källa	fr.gr.	kvs	mkvs
Mellan prov	3-1=2	20.1992	10.0996
Mellan laboratorier	4-1=3	1.1070	0.3690
Residual	(3-1)(4-1)=6	0.4728	$0.0788 = \widehat{\sigma}^2$
Totalt	12-1=11	21.7790	

a) Testa hypotesen att laboratorierna ej skiljer sig åt. 5 % signifikansnivå.

b) Testa med konfidensmetoden på 5% nivå om det föreligger någon skillnad mellan laboratorium B_1 och laboratorium B_4.

Problem 13.2.2.

Torrvikten (i procent) bestämdes för sex lika stora partier bryggjäst, a, b, c, d, e och f, av tre analytiker A, B och C med resultat.

	Parti					
Analytiker	a	b	c	d	e	f
A	20.1	14.7	13.1	17.8	16.0	14.9
B	20.0	14.9	13.0	17.7	16.2	15.1
C	20.2	15.1	12.9	17.9	16.1	15.0

$\sum_i \sum_j (y_{ij} - \bar{y}_{..})^2 = 93.585 \qquad \bar{y}_{..} = 16.15$

En lämplig statistisk modell är tvåsidig variansanalys $Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$, där
$\alpha_i = i$-te analytikereffekten;
$\beta_j = j$-te partieffekten.
Vi har

$$\bar{y}_{1.} = 16.1 \quad \bar{y}_{2.} = 16.15 \quad \bar{y}_{3.} = 16.2$$

$$\bar{y}_{.1} = 20.1 \quad \bar{y}_{.2} = 14.9 \quad \bar{y}_{.3} = 13.0 \quad \bar{y}_{.4} = 17.8 \quad \bar{y}_{.5} = 16.1 \quad \bar{y}_{.6} = 15.0$$

Detta ger $\sum_i (\bar{y}_{i.} - \bar{y}_{..})^2 = 0.0050$ och $\sum_j (\bar{y}_{.j} - \bar{y}_{..})^2 = 31.135$.

Variansanalystabell:

	Fg.	Kvs.	Mkvs.
Mellan analytiker	2	$6 \cdot 0.005 = 0.03$	0.015

Mellan partier	5	$3 \cdot 31.135 = 93.405$	18.681
Residual	10	0.15	$0.015 = \hat{\sigma}^2$
Total	17	93.585	

a) Testa på 5 % signifikansnivå om analytikerna mäter likvärdigt.

b) Testa på 5 % nivån att partierna är likvärdiga. Om det är en signifikant skillnad mellan partierna, ange vilka som skiljer sig signifikant åt genom att bilda 95 % konfidensintervall för de systematiska skillnaderna.

13.3 Variansanalys och mikromatriser

Mikromatriser, genchips, DNA-chips eller microarrays är en molekylärbiologisk metod för att samtidigt mäta de relativa koncentrationerna av mRNA för tusentals olika gener i ett prov, exempelvis odlade celler eller en vävnad. (Unionpedia)

After image processing, a microarray dataset is a set of Cy3 and Cy5 intensity values for the set of arrays that were hybridized and for the genes spotted on the arrays. For every gene g spotted on the arrays used in an experiment, the data contain a Cy3 and Cy5 intensity measurement.

Let y_{ijkg} be the intensity for gene g on array i from dye j . The subscript k indicates which RNA sample the measurement represents. By the experimental design chosen by the investigator, i and j determine the variety k. In other words, the investigator has chosen which RNA sample to label with dye j for hybridization to array i. Thus, the subscripts i, j and g suffice to identify a data value in the data array. The y_{ijkg} are assumed to be on log or similar scale and any pre-processing is assumed to be complete. Informally, the y_{ijkg} are normalized log intensities.