Statistik för bioteknik SF1911 CBIOT3
Föreläsning 9: Modellbaserad data-analys
Timo Koski

TK

21.11.2017

Lärandemål

- Konfidensintervall för väntevärdet i $\mathcal{N}(\mu, \sigma^2)$.
  i) σ känt
  ii) σ okänt –¿ t-fördelning
- Konfidensintervall för $u_1 - \mu_2$ med $\mathcal{N}(\mu_1, \sigma_1^2)$, $\mathcal{N}(\mu_2, \sigma_2^2)$.
  i) $\sigma_1 = \sigma_2$ kända, $\sigma_1 = \sigma_2$ okända
  ii) $\sigma_1 \neq \sigma_2$ okända
- Bootstrap

Estimating a Population Mean

Now we consider the task of estimating a population mean. These are questions like

- What is the mean amount milk obtained from cows in Skåne in 2016? (Existing population)

Here we think of proceeding by observing the amount of milk produced by cows in a randomly chosen sub-population of farms in Skåne (sample).

Estimating a Population Mean: Statistical Model

We regard the estimation of a population mean as estimating the mean $E(X)$ of a statistical distribution, like

$$\mu \equiv E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

The sample mean

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

for data/samples $x_1, x_2, \ldots, x_n$.

We are going to take $\bar{x}$ as estimate of $E(X)$.

Estimating a Population Mean: Law of Large Numbers

We are going to take $\bar{x}$ as estimate of $\mu (= E(X)$. Why ?

$X_1, X_2, \ldots, X_n$ are independent have the same distribution (=equally distributed) with mean μ och standard deviation σ. If n is very large we have

The Law of Large Numbers

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \approx \mu$$

where μ is the common mean of $X_1, X_2, \ldots, X_n, \ldots$

Limits

$X_1, X_2, \ldots, X_n$ are independent have the same distribution (=equally distributed) (e.g., $X_1 \sim \mathcal{N}(\mu, \sigma^2), X_2 \sim \mathcal{N}(\mu, \sigma^2), \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$) with mean μ och standard deviation σ. Then

$$E(\bar{X}) = \mu, \quad Var(\bar{X}) = \frac{\sigma^2}{n} \quad och \quad D(\bar{X}) = \frac{\sigma}{\sqrt{n}}.$$

Law of Large Numbers

$X_1, X_2, \ldots, X_n$ are independent have the same distribution (=equally distributed) with mean μ och standard deviation σ.
If n is very large we have

The Law of Large Numbers

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i \approx \mu$$

where μ is the common mean of $X_1, X_2, \ldots, X_n, \ldots$
This is because

$$Var(\bar{X}) = \frac{\sigma^2}{n} \to 0$$

as $n \to \infty$

We are going to take $\bar{x}$ as estimate of μ. Why ?

By the Law of Large Numbers $\bar{x}$ is close to the true value μ !
There is no systematic error (a.k.a. bias) either, since $E(\bar{X}) = \mu$.

$X_1, X_2, \ldots, X_n$ are all $\mathcal{N}(\mu, \sigma^2)$, then $\bar{x}$ is the maximum likelihood estimate of μ, too. (Details omitted).

Estimating a Population Mean

We are going to follow the same general procedure as above, i.e,

- Find the critical value for chosen level of confidence α.
- Find the margin of error E
- Form the confidence interval (confidence level $1 - \alpha$)

$$\bar{x} - E < \mu < \bar{x} + E$$

Estimating a Population Mean: the critical value

When finding the critical value we have to consider two cases (a) and (b):

(a) $\sigma = \sqrt{Var(X)} = \sqrt{\int_{-\infty}^{\infty}(x - E(X))^2 f(x)dx}$ *is* **known**

(b) σ *is* **not known**

Estimating a Population Mean (a) : the critical value when σ **known**

The critical value $\lambda_{\alpha/2}$ is found as in estimation of population proportion.

Estimating a Population Mean (a)

This is justified by the central limit theorem applied to arithmetic mean

$$\bar{X} \text{ approximatively} \sim \mathcal{N}\left(\mu, \sigma^2 / n\right)$$

Or, with with the cumulative density $\Phi(x)$

$$P(a < \bar{X} \leq b) \approx \Phi\left(\frac{b-\mu}{\sigma / \sqrt{n}}\right) - \Phi\left(\frac{a-\mu}{\sigma / \sqrt{n}}\right)$$

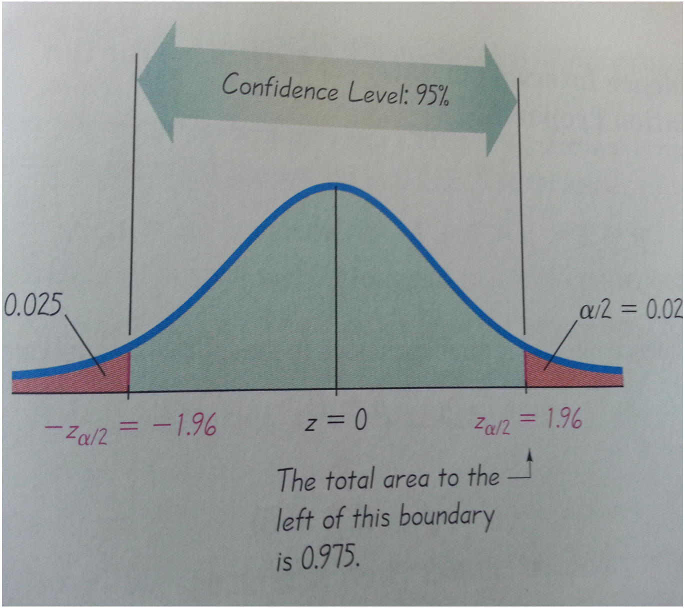
Z Score

Therefore the population Z Score is standard normal

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

and we have for Z the same picture as above.

Critical value: read $z_{\alpha/2}$ as $\lambda_{\alpha/2}$



Estimating a Population Mean (a): σ **known**

We can therefore to follow exactly the same procedure as above, i.e,

- Find the critical value $\lambda_{\alpha/2}$ for chosen level of confidence $1-\alpha$.
- Find the margin of error E
- Form the confidence interval (confidence level $1-\alpha$)

$$\bar{x} - E < \mu < \bar{x} + E$$

Estimating a Population Mean (a): the margin of error

When σ is **known**, the margin of error E is

$$E \stackrel{\text{def}}{=} \lambda_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

This reflects the fact that $\frac{\sigma}{\sqrt{n}}$ is standard deviation in $\mathcal{N}\left(\mu, \sigma^2/n\right)$.

The confidence interval (a): σ is **known**

Confidence Interval or the Interval Estimate for the population mean μ

$$\bar{x} - E < \mu < \bar{x} + E, \quad where\ E = \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Other equivalent expressions are

$$\bar{x} \pm E$$

or

$$(\bar{x} - E, \bar{x} + E)$$

Confidence Interval or the Interval Estimate for the population mean μ

Längden av intervallet

$$\bar{x} - \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}},$$

är $= 2 \cdot \lambda_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

- Intervallet blir kortare med mer data (större n) för givet α.
- Intervallet blir bredare med större $1-\alpha$ (via större $\lambda_{\alpha/2}$) för givet n: större konfidensgrad men en oprecisare skattning.

Interpretation of the CI: Correct

Suppose we have, with $\alpha = 0.05$ and some data, obtained by the procedure above $98.08 < \mu < 98.32$

We are 95 % confident that the interval with endpoints 98.08 and 98.32 contain the true value of μ.

Interpretation of the CI: Correct

This means that if we were to select many different samples of the same size and construct the corresponding CIs, 95 % of them would actually contain the value of the population proportion μ. In this correct interpretation 95% refers to the success rate of the

process being used to estimate the mean.

Interpretation of the CI: Wrong !

Because μ is a fixed constant (but unknown) it would be wrong to say that " there is a 95% chance that μ will fall between 98.08 and 98.32 ". The CI does not describe the behaviour of individual sample values, so it would be wrong to say that " 95% of all data values fall between 98.08 and 98.32 ". The CI does not describe the behaviour of individual sample means, so it would also be wrong to say that " 95% of sample means fall between 98.08 and 98.32 ".

The confidence interval (b): σ **not known**

Since the standard deviation σ is not known we need first an estimate of σ.

$$s = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

is the maximum likelihood estimate, but we should in fact take (!)

$$s = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2}$$

KTH Matematik

t score

If $s = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2}$, then we have the following:

The distribution of

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

is a **t-distribution** *with $n-1$ degrees of freedom .*

t-distribution : Degrees of freedom ??

The number of **degrees of freedom** *is the number of sample values that can vary after certain restrictions have been imposed on all data values*

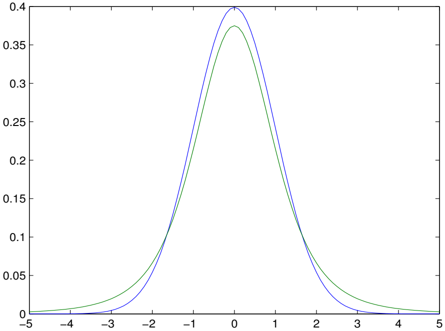
For example, if $\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n} = 5$ is imposed, we can vary, e.g., $x_1, x_2, \ldots, x_n - 1$, but x_n must be chosen so that $\bar{x}$ is equal to 5. Hence the degree of freedom is $n - 1$.

Critical value (b): σ **not known**

We use the t-distribution to find the critical value. These are now denoted by $t_{\alpha/2}$. The logic is the same as with $\lambda_{\alpha/2}$, since t-distribution is symmetric as is $\mathcal{N}(0,1)$.

t-distribution

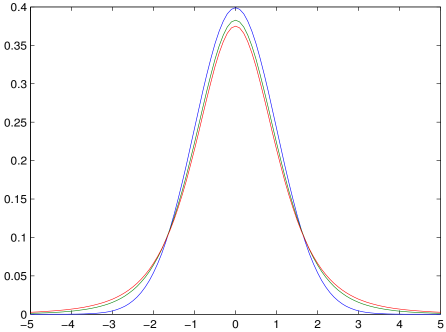
The density curves for $\mathcal{N}(0,1)$ (blue) and t with four degrees of



freedom (green)

t-distribution

The density curves for $\mathcal{N}(0,1)$ (blue) and t with four degrees of freedom (red) and t with six degrees of freedom (green)



Critical value

Here we find t_α for various values of α and f =degrees of freedom.

f	α	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
1		3.08	6.31	12.71	31.82	63.66	318.31	636.62
2		1.89	2.92	4.30	6.96	9.92	22.33	31.60
3		1.64	2.35	3.18	4.54	5.84	10.21	12.92
4		1.53	2.13	2.78	3.75	4.60	7.17	8.61
5		1.48	2.02	2.57	3.36	4.03	5.89	6.87

For example $t_{0.05} = 2.02$ for 5 degrees of freedom.

Estimating a Population Mean: the margin of error (b)

When σ is **not known**, the margin of error E is

$$E \stackrel{\text{def}}{=} t_{\alpha/2}\frac{s}{\sqrt{n}}$$

where $t_{\alpha/2}$ has $n-1$ degrees of freedom.

The confidence interval: σ **not known**

Confidence Interval or the Interval Estimate for the population mean μ σ not known

$$\bar{x} - E < \mu < \bar{x} + E, \quad \textit{where } E = t_{\alpha/2}\frac{s}{\sqrt{n}}$$

or

$$\bar{x} \pm E$$

or

$$(\bar{x} - E, \bar{x} + E)$$

KTH Matematik

Två stickprov, konfidensintervall för skillnad mellan väntevärden.

a) σ_1 och σ_2 kända

Vi vill nu skaffa oss ett konfidensintervall för $\mu_1 - \mu_2$. En naturlig skattning av $\mu_1 - \mu_2$ är $\bar{X} - \bar{Y}$. Eftersom den är en linjärkombination av oberoende normalfördelade variabler, så gäller att

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

är $\mathcal{N}(0, 1)$-fördelad.

Två stickprov, konfidensintervall för skillnad mellan väntevärden.

Av detta leds vi till

$$\underline{I_{\mu_1-\mu_2} = \bar{x} - \bar{y} \pm \lambda_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\,.$$

Om $\sigma_1 = \sigma_2 = \sigma$ reduceras detta till att

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

är $\mathcal{N}(0,1)$-fördelad och

$$I_{\mu_1-\mu_2} = \bar{x} - \bar{y} \pm \lambda_{\alpha/2}\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\,.$$

Två stickprov, konfidensintervall för skillnad mellan väntevärden.

b) $\sigma_1 = \sigma_2 = \sigma$ okänd

Vi betraktar nu fallet då $\sigma_1 = \sigma_2 = \sigma$, men där σ är okänd. Detta skattas med s där s^2 är den *sammanvägda* stickprovsvariansen. Man kan visa att man skall välja

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

och att

$$\frac{(\bar{X} - \bar{Y}) - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

är $t(n_1 + n_2 - 2)$-fördelad.

Två stickprov, konfidensintervall för skillnad mellan väntevärden.

Vi får

$$\underline{I_{\mu_1-\mu_2} = \bar{x} - \bar{y} \pm t_{\alpha/2}(n_1+n_2-2)s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\,.$$

Sammanfattning (1)

Sats

Låt $x_1, \ldots, x_{n_1}$ *och* $y_1, \ldots, y_{n_2}$ *vara slumpmässiga, av varandra oberoende stickprov från* $N(\mu_1, \sigma_1^2)$ *respektive* $N(\mu_2, \sigma_2^2)$.
Om σ_1 *och* σ_2 *är kända så är*

$$I_{\mu_1-\mu_2} = (\bar{x} - \bar{y} - \lambda_{\alpha/2} D, \bar{x} - \bar{y} + \lambda_{\alpha/2} D)$$

ett tvåsidigt konfidensintervall för $\mu_1 - \mu_2$ *med konfidensgraden* $1 - \alpha$*; här är* $D = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$.

Sammanfattning

Sats

Om $\sigma_1 = \sigma_2 = \sigma$ *där* σ *är okänt så är*

$$I_{\mu_1 - \mu_2} = (\bar{x} - \bar{y} - t_{\alpha/2}(f)d, \bar{x} - \bar{y} + t_{\alpha/2}(f)d)$$

ett tvåsidigt konfidensintervall för $\mu_1 - \mu_2$ *med konfidensgraden* $1 - \alpha$*; här är* $d = s\sqrt{1/n_1 + 1/n_2}$ *där* $f = (n_1 - 1) + (n_2 - 1)$.

Two samples = Två stickprov

In an experiment designed to test the effectiviness of paroxetine for treating bipolar depression, subjects were measured using the Hamilton Depression scale with results as follows:

Placebo group	$n_1 = 43$	$\overline{x} = 21.57, s_1 = 3.87$
Paroxenite treatment group	$n_2 = 33$	$\overline{y} = 20.38, s_2 = 3.91$

Two samples

We assume that we have independent samples and normal distributions. We must assume this, we have just summaries of data, no chance of looking at histograms or boxplots.
We insert data in pooling (sammanvägning)

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(43-1)\cdot 3.87^2 + (33-1)\cdot 3.91^2}{43+33-2}$$

$$= 15.11$$

We choose $\alpha = 0.05$. The quantiles are found from the t-distribution with $43 + 33 - 2 = 74$ degrees of freedom. This gives us $t_{0.025} = 1.993$ (by >>tinv(0.975, 74) in Matlab).

KTH Matematik

Two samples; the confidence interval computed

$$s^2 = \frac{(43-1)\cdot 3.87^2 + (33-1)\cdot 3.91^2}{43+33-2} = 15.11$$

$t_{0.025}(74) = 1.993$

$$I_{\mu_1-\mu_2} = \overline{x} - \overline{y} \pm t_{\alpha/2}(n_1+n_2-2)s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

$$= (21.57 - 20.38) \pm 1.993 \cdot \sqrt{15.11} \cdot \sqrt{\frac{1}{43}+\frac{1}{33}}$$

$$= 1.19 \pm 1.7930 = [-0.603, 2.938]$$

Två stickprov för paroxetine; Konfidensintervall med konfidens grad 0.95

$$I_{\mu_1-\mu_2} = [-0.603, 2.938]$$

Men vad säger detta för oss ? Vilken **kunskap** *om paroxetine har de givna data gett oss i form av detta konfidensintervall?*

Vi kommer att få ett svar med hjälp av teorin om hypotesprövning. Men först ett fall till.

t-distribution

The derivation of the t-distribution was first published in 1908 by William Sealy Gosset, while he worked at a Guinness Brewery in Dublin. He was prohibited from publishing under his own name, so the paper was written under the pseudonym Student.

Bootstrap: dictionary

(1) Loop of leather or cloth sewn at the top rear, or sometimes on each side, of a boot to facilitate pulling it on. (2) a means of advancing oneself or accomplishing something relying entirely on one's efforts and resources.

Bootstrap

Namnet bootstrap (på sv. stövelstropp) härstammar från uttrycket i amerikansk engelska *pull oneself up by the bootstraps*, vilket betyder att man förbättrar ens (ekonomiska/sociala o.s.v.) situation genom egna insatser utan andras hjälp. Ett uttryck med likadan betydelse är *att lyfta sig själv i håret* (jfr. baron von Münchausens äventyr).

Bootstrap

Återsampling: Vi lottar fram ett första **fiktivt** eller **bootstrap-stickprov** ur våra mätdata genom att på måfå dra n st med återläggning från $x_1, \ldots, x_n$.
Detta stickprov blir typiskt ungefär som det ursprungliga även om vissa observationer saknas och vissa finns i dubbel eller kanske tredubbel uppsättning.

Bootstrap med matlab: exempel

Vi har, till exempel,

$x_1 = -0.2746, x_2 = -1.1730, x_3 = 1.4842, x_4 = 1.1454, x_5 = -1.6248,$

$x_6 = 0.9985, x_7 = 0.4571, x_8 = -1.2315, x_9 = 0.9868, x_{10} = -0.5941$

eller

$$x = [\begin{array}{ccccc} -0.2746 & -1.1730 & 1.4842 & 1.1454 & -1.6248 \\ 0.9985 & 0.4571 & -1.2315 & 0.9868 & -0.5941 \end{array}];$$

Matlabkommandot unidrnd(10,1,10) ger tio slumptal ur den likformiga fördelningen på de tio heltalen i $1, \ldots, 10$, d.v.s $p(i) = 1/10$.

$>>$ z=unidrnd(10,1,10)

z =

5 1 6 3 2 5 6 3 7 10

Bootstrapstickprovet ges nu av $>>$ bssampl=x(z)

Bootstrap med matlab: exempel

Med >> bssampl=x(z)

fås bootstrap stickprovet

bssampl =

Columns 1 through 7

-1.6248 -0.2746 0.9985 1.4842 -1.1730 -1.6248 0.9985

Columns 8 through 10

1.4842 0.4571 -0.5941

Medelfel för en skattning & bootstrap

Ur detta nya stickprov beräknar vi skattningen $\widehat{\theta}_{1,\text{obs}}$ av en parameter θ. Detta upprepas nu B gånger (kanske $B = 1000$ eller $B = 10000$) och ur dessa B fiktiva stickprov får vi B skattningar $\widehat{\theta}_{1,\text{obs}}, \widehat{\theta}_{2,\text{obs}}, \ldots, \widehat{\theta}_{B,\text{obs}}$. Vi får då en medelfelsskattning genom att beräkna spridningen av dessa d.v.s.

$$d(\widehat{\theta}) = \sqrt{\frac{1}{B-1}\sum_{i=1}^{B}(\widehat{\theta}_{i,\text{obs}} - \widehat{\theta}_{\text{medel,obs}})^2}$$

där $\widehat{\theta}_{\text{medel,obs}} = \sum_1^B \widehat{\theta}_{i,\text{obs}}/B$ är aritmetiska medelvärdet av de B skattningarna för de B fiktiva stickproven.

Medelfel för en skattning & bootstrap

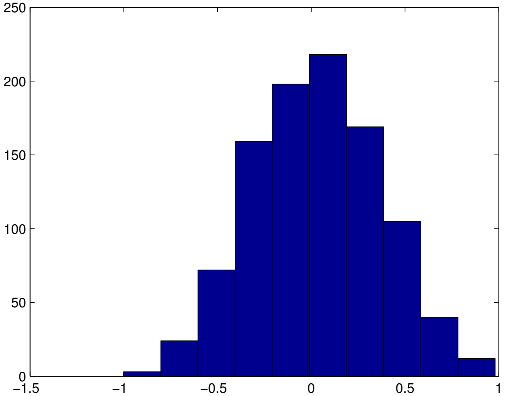
Vi hade simulerat så att

$$x = [\begin{array}{ccccc} -0.2746 & -1.1730 & 1.4842 & 1.1454 & -1.6248 \\ 0.9985 & 0.4571 & -1.2315 & 0.9868 & -0.5941 \end{array}];$$

är tio oberoende stickprov på $X_i \sim \mathcal{N}(0,1)$. Vi har att $\widehat{\theta} = \frac{1}{10}(X_1 + \ldots + X_{10}) \in \mathcal{N}(0, \frac{1}{10})$. Om vi drar B bootstrap-stickprov från x och beräknar $\overline{x}_{i,obs}$ för $i = 1, \ldots, B$, får vi enligt receptet ovan en bootstrap-skattning av $\sqrt{\frac{1}{10}} = 0.3162$. I figuren ses histogrammet för $\overline{x}_{i,obs}$ för $i = 1, \ldots, B$ med $B = 1000$, där standardavvikelsen är $d(\theta^*) = 0.339$.

Medelfel för en skattning & bootstrap

I figuren ses histogrammet för $\overline{x}_{i,obs}$ för $i = 1, \ldots, B$ med $B = 1000$, där standaravvikelsen är 0.339.



Medelfel för en skattning & bootstrap

De fiktiva stickproven är framlottade enligt den empiriska fördelningen som lägger massan $1/n$ i vardera av de n observationerna. Denna empiriska fördelning är en skattning av den sanna fördelningen F som observationerna genererats av. Vi har därmed lyckats skapa en kopia av det ursprungliga försöket (där våra data kom från F). Vi kan då undersöka egenskaper hos den ursprungliga skattningen genom undersöka skattningen i kopian.

Luria & Delbrück

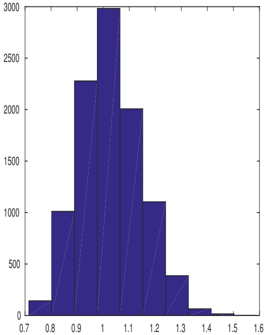
$X \sim \mathcal{P}oi(\lambda)$, $E(X) = \lambda$, $Var(X) = \lambda$.

- Simulera $x = 100$ stickprov från, t.ex., $\mathcal{P}oi(0.2)$ (poissrnd(0,2,1,100)).
- Återsampla x, $x_i^*, i = 1, \ldots, B$, t.ex. $B =$10 000 ggr. $x_i^* =$ stovelstropp(x).
- Beräkna $\bar{x}_i^*$ och std(x_i^*) för varje x_i^*
- Beräkna kvoten varians/medelvärde, $q_i = std^2(x_i^*))/\bar{x}_i^*$ för $i = 1, \ldots, B$.
- Ta fram histogram för $q_i$na, mean(q) och std(q).

Luria & Delbrück

$$X \sim \mathcal{P}oi(\lambda),\ E(X) = \lambda,\ Var(X) = \lambda \Rightarrow Var(X)/E(X) = 1.$$

- Histogram för $q_i$na, mean(q) och std(q).
- Det finns ingen matematisk formel för fördelningen av kvoten $q_i = std^2(x_i^*))/\bar{x}_i^*$. mean(q) $=$ 1.0299, std(q)$=$ 0.1164.



Bilaga: Bootstrap stovelstropp.m

function bssampl=stovelstropp(x)
% återsamplar datat i x med likformig sannolikhet (`unirnd` och
ger bssampl som ett bootstrap %stickprov. m=length(x);
z=unidrnd(m,1,m);
bssampl=x(z);