SF1915 Sannolikhetsteori och statistik
6 hp
Föreläsning 12
χ^2-test

Jörgen Säve-Söderbergh

Anpassningstest – test av given fördelning

n oberoende försök med r möjliga olika utfall

Händelse	A_1	A_2	$\ldots$	A_r
Antal	x_1	x_2	$\ldots$	x_r
	$P(A_1)$	$P(A_2)$	$\ldots$	$P(A_r)$

$$P(A_1) + P(A_2) + \cdots + P(A_r) = 1$$

$$x_1 + x_2 + \cdots + x_r = n$$

x_1 är ett utfall av X_1

x_r är ett utfall av X_r

$(X_1, X_2, \ldots, X_r)$ multinomialfördelad s.v.

Anpassningstest – test av given fördelning

$$H_0 : P(A_1) = p_1, \quad P(A_2) = p_2, \quad \ldots, \quad P(A_r) = p_r$$

$$p_1 + p_2 + \cdots + p_r = 1$$

Varje $X_j \in Bin(n, p_j)$

$$E(X_j) = np_j$$

I varje cell (för varje händelse A_j) beräknar vi det s k χ^2-avståndet

$$\frac{(x_j - np_j)^2}{np_j} = \frac{(\text{Observerad frekvens} - \text{förväntad frekvens under } H_0)^2}{\text{förväntad frekvens under } H_0}$$

Anpassningstest – test av given fördelning

Om H_0 är sann gäller att

$$Q_{\text{obs}} = \sum_{j=1}^{r} \frac{(x_j - np_j)^2}{np_j}$$

är approximativt $\chi^2\,(r-1)$-fördelad, då $n \to \infty$.

Signifikanstest

Förkasta H_0, om $Q_{\text{obs}} > \chi^2_\alpha\,(r-1)$

Förkasta ej H_0, om $Q_{\text{obs}} \leq \chi^2_\alpha\,(r-1)$

Vi bör ha $np_j \geq 5$ för att kunna garantera att signifikansnivån är α.

Beräkningsformel för anpassningstest

Vi kan skriva om testvariabeln

$$Q_{\text{obs}} = \sum_{j=1}^{r} \frac{(x_j - np_j)^2}{np_j}$$

till

$$Q_{\text{obs}} = \sum_{j=1}^{r} \frac{x_j^2}{np_j} - n$$

Exempel – anpassningstest

Väljer bilförare att köra i vilken fil som helst på en fyrfilig motorväg?

Etttusen bilar observerades.

Fil	1	2	3	4
Antal bilar	294	276	238	192

Testa på nivån 5% att bilförarna använder vilken som helst fil.

Anpassningstest – test av fördelning med skattade parametrar

$H_0 : P(A_1) = p_1(\theta), \quad P(A_2) = p_2(\theta), \quad \ldots, \quad P(A_r) = p_r(\theta)$
för något θ.

Vi skattar θ med ML och substituerar

$$p_j(\theta^*_{\text{obs}}) = p_j^*$$

Om H_0 är sann gäller att

$$Q_{\text{obs}} = \sum_{j=1}^{r} \frac{\left(x_j - np_j^*\right)^2}{np_j^*}$$

är approximativt $\chi^2(r-k-1)$-fördelad, då $n \to \infty$.

k=antalet skattade parametrar

Tumregel $np_j^* \geq 5$.

Anpassningstest – test av fördelning med skattade parametrar – test av normalfördelning

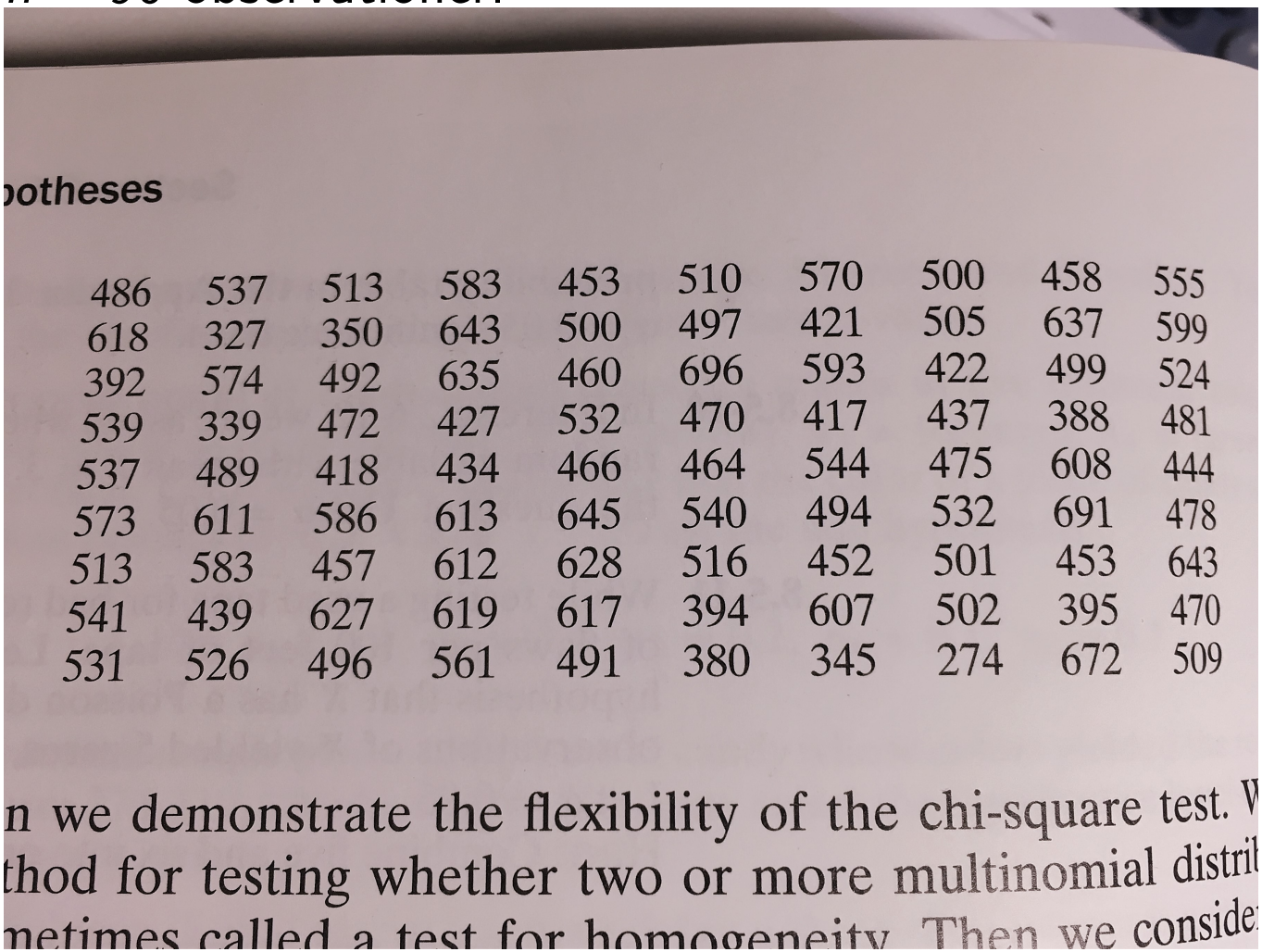
Vi ska använda det mer generella testet för att utföra ett test av normalfördelning $X \sim N(\mu, \sigma)$.

Då kommer sannolikheterna $P(A_i) = p_i(\theta)$ att bero på normalfördelningens två parametrar.

$\theta = (\mu, \sigma)$

Exempel

$n = 90$ observationer.



ɔotheses

486	537	513	583	453	510	570	500	458	555
618	327	350	643	500	497	421	505	637	599
392	574	492	635	460	696	593	422	499	524
539	339	472	427	532	470	417	437	388	481
537	489	418	434	466	464	544	475	608	444
573	611	586	613	645	540	494	532	691	478
513	583	457	612	628	516	452	501	453	643
541	439	627	619	617	394	607	502	395	470
531	526	496	561	491	380	345	274	672	509

n we demonstrate the flexibility of the chi-square test. W
thod for testing whether two or more multinomial distril
metimes called a test for homogeneity. Then we conside

min $= 274$ $\bar{y} = 511.633$ $s = 87.576$ max $= 672$

Anpassningstest – test av fördelning med skattade parametrar – test av normalfördelning

$y_1, y_2, \ldots, y_n$ är observationer på den stokastiska variabeln Y.

Dela upp Y:s variationsområde $\{y : -\infty < y < \infty\}$ i k ömsesidigt uteslutande mängder $A_1, A_2, \ldots, A_k$.

Låt

$$H_0 : Y \sim N(\mu, \sigma).$$

$$P(A_i) = p_i(\mu, \sigma) = \int_{A_i} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(w-\mu)^2}{2\sigma^2}\right] dw$$

en funktion av de okända parametrarna μ och σ.

Anpassningstest – test av fördelning med skattade parametrar – test av normalfördelning

x_i är frekvensen av y_i som befinner sig i A_i

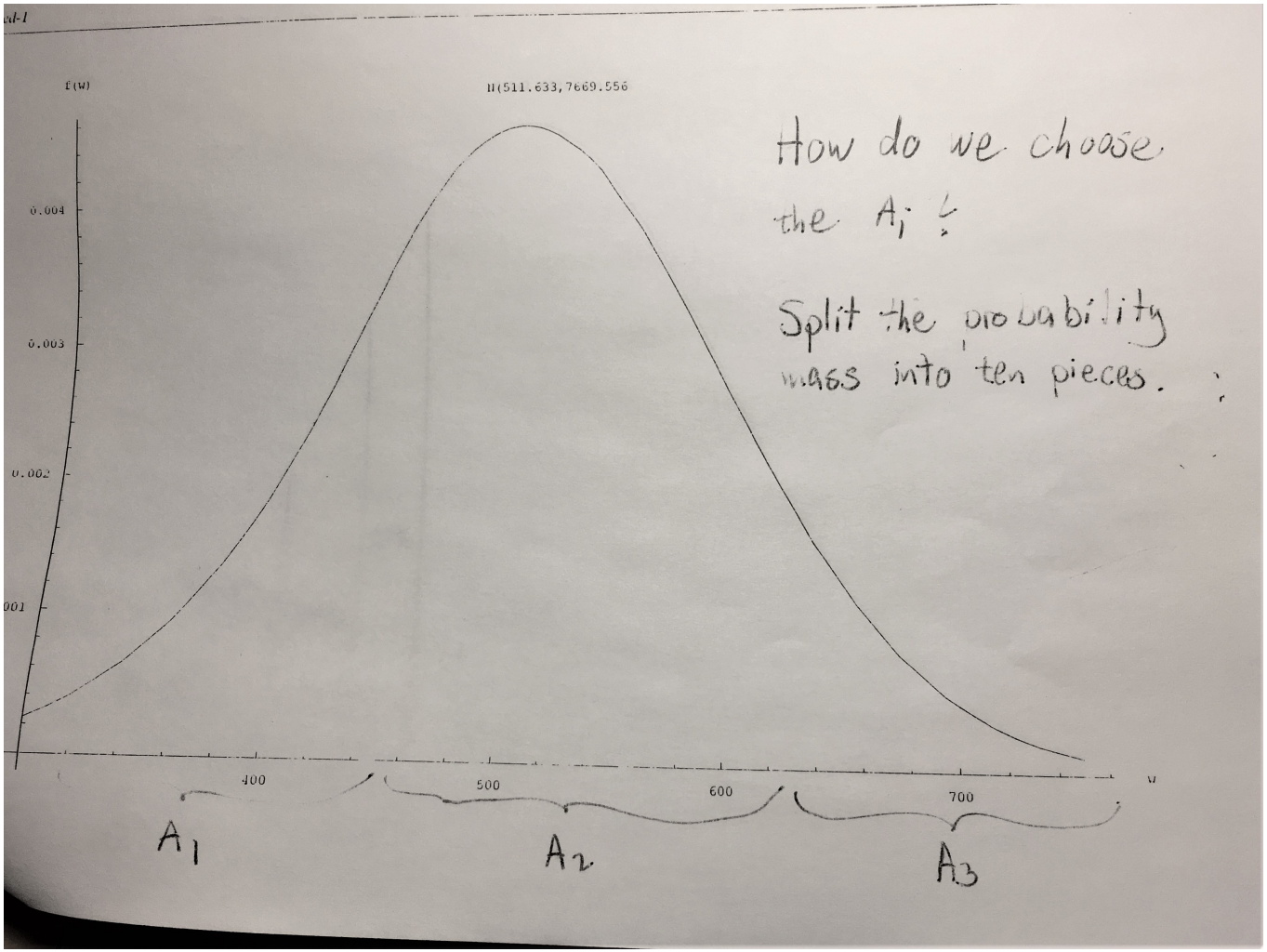
Vi skattar μ med $\overline{y}$.

Vi skattar σ med s.

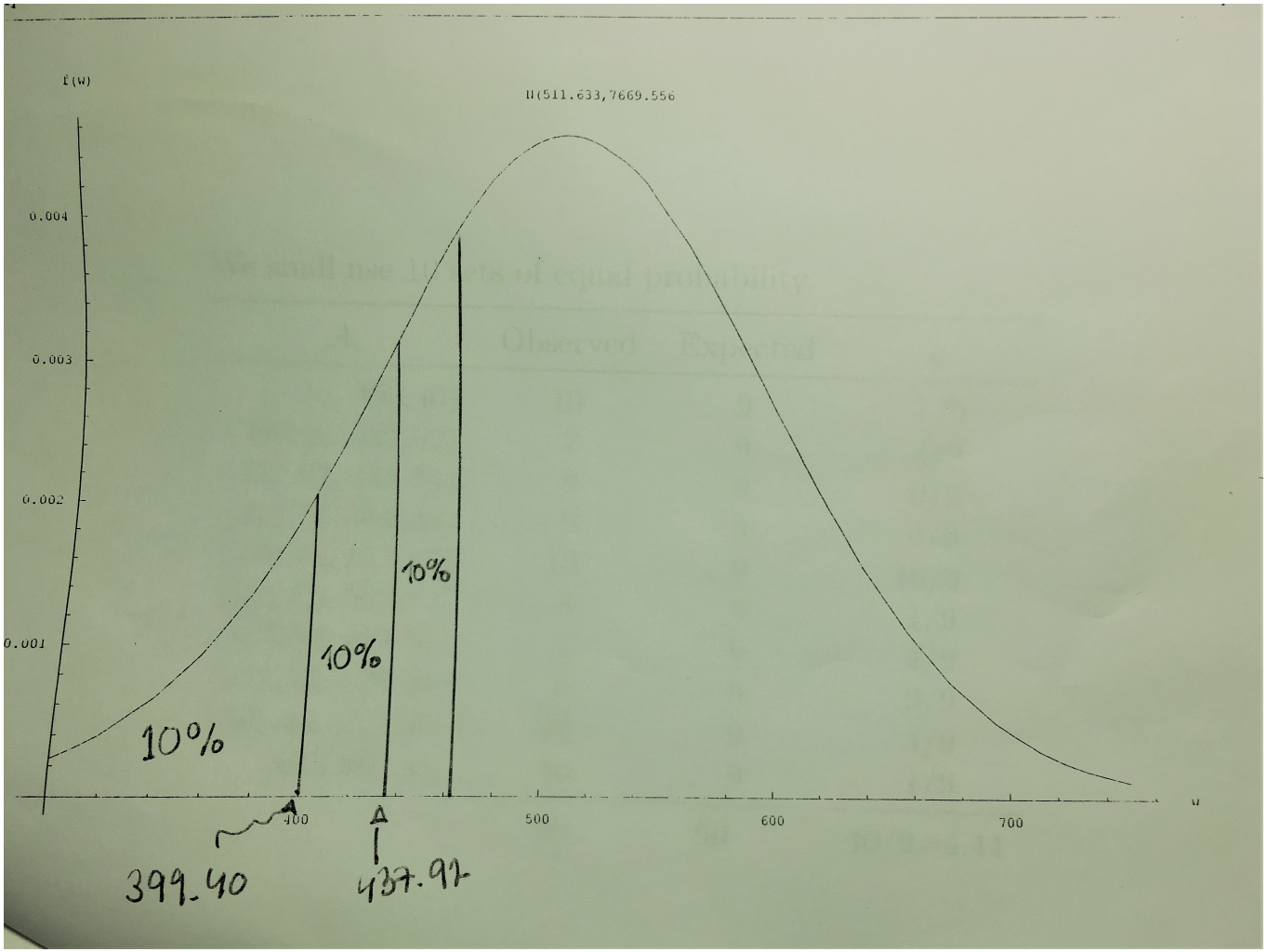
$$p_i^* (\mu, \sigma) = \int_{A_i} \frac{1}{\sqrt{2\pi}\, s} \exp \left[-\frac{(w - \overline{y})^2}{2\, s^2} \right] dw$$

$N(511.633, 87.576)$

Anpassningstest – test av fördelning med skattade parametrar – test av normalfördelning



Anpassningstest – test av fördelning med skattade parametrar – test av normalfördelning



Anpassningstest – test av fördelning med skattade parametrar – test av normalfördelning

Med detta val av klasserna A_i blir varje $p_i^* = \frac{1}{10}$

Förväntade frekvenser $np_i^* = 90\ \frac{1}{10} = 9$

$np_i^* = 9 > 5$ tumregeln uppfylld.

Anpassningstest – test av fördelning med skattade parametrar – test av normalfördelning

We shall use 10 sets of equal probability.

A_i	Observed	Expected	q
$(-\infty, 399.40)$	10	9	1/9
$[399.40, 437.92)$	7	9	4/9
$[437.92, 465.71)$	9	9	0/9
$[465.71, 489.44)$	9	9	0/9
$[489.44, 511.63)$	13	9	16/9
$[511.63, 533.82)$	8	9	1/9
$[533.82, 557.55)$	7	9	4/9
$[557.55, 585.34)$	6	9	9/9
$[585.34, 623.86)$	11	9	4/9
$[623.86, \infty)$	10	9	1/9
	90	90	40/9=4.44

Anpassningstest – test av fördelning med skattade parametrar – test av normalfördelning

$$q = \frac{\left(x_i - np_i^*\right)^2}{np_i^*} = \frac{(10-9)^2}{9} = \frac{1}{9}$$

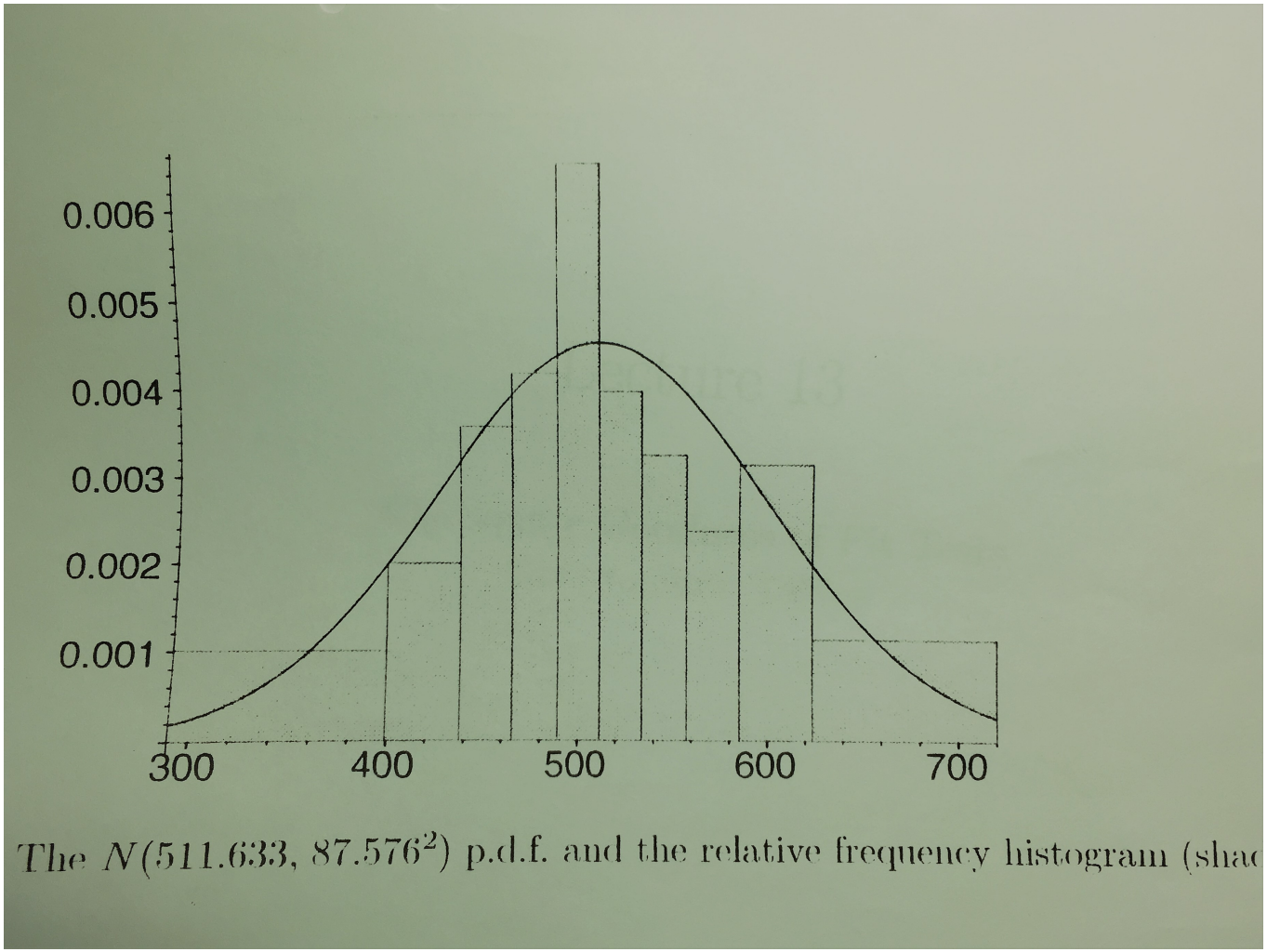
Vi har skattat två parametrar, så $k = 2$.

$r - k - 1 = 10 - 2 - 1 = 7$ frihetsgrader

Eftersom $4.44 < 14.07 = \chi^2_{0.05}(7)$ kan vi ej förkasta H_0.

Det är alltså möjligt att fördelningen för Y är normalfördelningen.

Anpassningstest – test av fördelning med skattade parametrar – test av normalfördelning



The $N(511.633, 87.576^2)$ p.d.f. and the relative frequency histogram (shad

Homogenitesttest

Flera serier av oberoende försök.

Är det samma sannolikheter

$$P(A_1), P(A_2), \ldots, P(A_r)$$

som ligger bakom samtliga serier?

Är de *homogena*?

H_0 sannolikheterna är lika.

Homogenitesttest

s serier av försök. H_0 sannolikheterna är lika.

Serie	A_1	A_2	$\dots$	A_r	Antal försök
1	x_{11}	x_{12}	$\dots$	x_{1r}	n_1
2	x_{21}	x_{22}	$\dots$	x_{2r}	n_2
$\vdots$					
s	x_{s1}	x_{s2}	$\dots$	x_{sr}	n_s
Summa	$x_{\cdot 1}$	$x_{\cdot 2}$	$\dots$	$x_{\cdot r}$	n

$x_{ij} = x_{\text{försök } i, \text{utfall nr } j}$

$$Q_{\text{obs}} = \sum_{i=1}^{s}\sum_{j=1}^{r} \frac{\left(x_{ij} - n_i p_j^*\right)^2}{n_i p_j^*} \qquad p_j^* = (p_j)_{\text{obs}}^* = \frac{x_{\cdot j}}{n}$$

$$p_j^* = (p_j)_{\text{obs}}^* = \frac{x_{\cdot j}}{n}$$

Bästa skattningen av det gemensamma $P(A_j)$-värdet som vi kan göra med de sammanlagda observationerna.

Homogenitetstest

Förkasta hypotesen om homogenitet, om

$$Q_{\text{obs}} > \chi^2 (r-1)(s-1)$$

Tumregel $n_i p_j^* \geq 5$.

Tre maskiner M_1, M_2 och M_3.

Varje tillverkas enhet kan klassificeras som bra, halvbra (kan räddas) eller oanvändbar.

Tre stickprov om 110, 90 respektive 200 uttas.

Ger de tre maskinerna M_1, M_2 och M_3 samma fördelning på kvaliteten hos de tillverkade enheterna?

	Bra	Halvbra	Dåliga
M_1	73	26	11
M_2	65	18	7
M_3	166	16	18

Undersök med ett χ^2-test om M_1, M_2 och M_3 ger samma fördelning på kvaliteten hos de tillverkade enheterna.

Oberoendetest

Homogenitetstest innebär att vi jämför flera populationer. (Varje serie är observationer på en population).

I oberoendetest har vi en population som har delats av slumpen i två eller flera kategorier. (Blom talar om egenskaper).

Utför n slumpmässiga försök.

Den första egenskapen har delats in i s ömsesidigt uteslutande kategorier $B_1, B_2, \ldots, B_s$.

Den andra egenskapen har delats in i r ömsesidigt uteslutande kategorier $A_1, A_2, \ldots, A_r$.

Sannolikheten att inneha två kategorier $B_i \cap A_j$ betecknas

$$p_{ij} = P\left(B_i \cap A_j\right), \quad i = 1, 2, \ldots, s, \quad j = 1, 2, \ldots, r.$$

Oberoendetest

Låt x_{ij} beteckna frekvensen för $B_i \cap A_j$. (sr händelser som $B_i \cap A_j$)

$\sum_{i=1}^{s} \sum_{j=1}^{r} x_{ij} = n$ $\sum_{i=1}^{s} \sum_{j=1}^{r} p_{ij} = 1.$

Om A_j ska inträffa, måste någon av händelserna $B_1 \cap A_j, B_2 \cap A_j, \ldots, B_s \cap A_j$ inträffa

$$P(A_j) = p_{\cdot j} = \sum_{i=1}^{s} p_{ij}$$

På samma sätt B_i om och endast om någon av $B_i \cap A_1, B_i \cap A_2, \ldots, B_i \cap A_r$

$$P(B_i) = p_{i\cdot} = \sum_{i=1}^{r} p_{ij}$$

Oberoendetest

Vi måste skatta sannolikheterna $p_{.j}$ och $p_{i.}$

$$p^*_{.j} = \frac{\sum_{i=1}^{s} x_{ij}}{n} = \frac{x_{.j}}{n} \qquad \text{och} \qquad p^*_{i.} = \frac{\sum_{j=1}^{r} x_{ij}}{n} = \frac{x_{i.}}{n}$$

Oberoendetest

Vi önskar testa om egenskaperna A och B är oberoende

$$H_0 : p_{ij} = P(B_i \cap A_j) = P(B_i)\,P(A_j) = p_{i.}p_{.j}, \quad \text{för alla } (i,j)$$

mot alternativet

$$H_1 : p_{ij} \neq p_{i.}p_{.j}, \quad \text{för något } (i,j).$$

Testvariabel är

$$Q_{\text{obs}} = \sum_{i=1}^{s}\sum_{j=1}^{r} \frac{(x_{ij} - np^*_{i.}p^*_{.j})^2}{np^*_{i.}p^*_{.j}}$$

som är approximativt χ^2-fördelad med $(r-1)(s-1)$ frihetsgrader, om n är stort.

Tumregel $np^*_{i.}p^*_{.j} \geq 5$

Oberoendetest

För beräkning

$$np^*_{i.}p^*_{.j} = n\left(\frac{x_{i.}}{n}\right)\left(\frac{x_{.j}}{n}\right) = \frac{x_{i.}x_{.j}}{n}$$

Vi börjar med att beräkna rad- och kolumntotaler för att kunna använda formeln ovan.

Oberoendetest – Exempel

Fyrahundra studenter vid University of Iowa studerades.

Studenterna klassificerades efter vilket college de studerade vid, samt efter kön.

	Business	Engineering	Liberal Arts	Nursing	Pharmacy	Total
Man	21	16	145	2	6	190
Kvinna	14	4	175	13	4	210
Total	35	20	320	15	10	400

Testa på 5% signifikansnivå om valet av studieinriktning är oberoende av kön.