Föreläsning 9.

* Miniprojekt 2 - deadline $^{24}/_{5}$ kl. $16^{\underline{00}}$
  + Uppgift a-d efter laboration 3.
  + uppgift e efter laboration 4.

* Mozquizto 2 - deadline $^{28}/_{5}$ kl $23^{\underline{59}}$.
* Tenta $^{3}/_{6}$ $08^{\underline{00}}$ - $13^{\underline{00}}$; Sparta: D

Outline

* Repetition
* Icke-parametriska test
* Inferens om proportioner
* Analys av kategoridata

Repetition

Utgångspunkt: $X \in N(\mu_X, \sigma_X^2)$

$H_0: \mu_X - \mu_Y = 0$ $\quad Y \in N(\mu_Y, \sigma_Y^2)$

$H_1: \mu_X - \mu_Y \neq 0$ (eller >0 eller <0)

1) Kan man se X_k och Y_k som parvisa mätningar?

Ja! • Låt $Z_k = X_k - Y_k$ (eller $Y_k - X_k$)

$Z_k \in N(\Delta, \sigma_Z^2)$

- Sätt upp hypotes.
- Testa. Om σ_Z känd, använd normalkvantil $z_{1-\alpha}$ ($z_{1-\alpha/2}$ om dubbelsidigt)

  Om σ_Z okänd skatta med s_Z, använd t-kvantil $t_{1-\alpha, n-1}$ ($t_{1-\alpha/2, n-1}$ om dubbelsidigt).

Repetition

Nej! Alltså inte parvisa!

- σ_x och σ_y kända:

  Testa med normalkvantil och med varians

  $$\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}$$

- σ_x och σ_y okända:

  - Kan man anta att $\sigma_x = \sigma_y$?

    Ja! Testa med t-kvantil med $n_x + n_y - 2$ frihetsgrader och med varians: $s^2\left(\frac{1}{n_x} + \frac{1}{n_y}\right)$

    där $s^2 = \frac{(n_x-1)s_x^2 + (n_y-1)s_y^2}{n_x + n_y - 2}$

Repetition

Nej! (Överkurs) σ_x och σ_y antas vara olika.

- Testa med t-kvantil med

$$f = \frac{\left[s_x^2/n_x + s_y^2/n_y\right]^2}{\frac{\left[s_x^2/n_x\right]^2}{n_x - 1} + \frac{\left[s_y^2/n_y\right]^2}{n_y - 1}}$$

frihetsgrader

och varians $\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}$ -

Repetition

Hur ser konfidensintervallen ut för de olika scenarierna? Antag att variansen är okänd (annars använd normalkvantil), konfidensnivå α, samt att vi testar:

H_0: $\mu_x - \mu_y = 0$

H_1: $\mu_x - \mu_y \neq 0$

Parvis stickprov: $Z_i = X_i - Y_i \in N(\Delta, \sigma_z^2)$

$$I_\Delta = \left(\bar{z} - t_{1-\alpha/2, n-1} \cdot \frac{s_z}{\sqrt{n}}, \bar{z} + t_{1-\alpha/2, n-1} \frac{s_z}{\sqrt{n}} \right)$$

Två oberoende stickprov:

Antag $\sigma_x = \sigma_y$

$$I_{\mu_x - \mu_y} = \left(\bar{x} - \bar{y} \pm t_{1-\alpha/2, n_x + n_y - 2} \, s \sqrt{\frac{1}{n_x} + \frac{1}{n_y}} \right)$$

med $s^2 = \dfrac{(n_x - 1) s_x^2 + (n_y - 1) s_y^2}{n_x + n_y - 2}$

Repetition

Två oberoende stickprov: (överkurs)

Antag $\sigma_x \neq \sigma_y$

$$I_{\mu_x - \mu_y} = \left(\bar{x} - \bar{y} \pm t_{1-\alpha/2,\, f} \sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}} \right)$$

med

$$f = \frac{\left[s_x^2/n_x + s_y^2/n_y \right]^2}{\frac{\left[s_x^2/n_x \right]^2}{n_x - 1} + \frac{\left[s_y^2/n_y \right]^2}{n_y - 1}}$$

frihetsgrader

Repetition

* Ett bra sätt att testa om $\sigma_X = \sigma_Y$ är att ställa upp följande hypotes och testa:

$H_0 : \sigma_X^2 = \sigma_Y^2$

$H_1 : \sigma_X^2 \neq \sigma_Y^2$

Man kan visa att $\frac{S_X^2}{S_Y^2}$ tillhör F-fördelningen, dvs.

$$\frac{S_X^2}{S_Y^2} \in F_{(n_x-1),(n_y-1)}$$

Vi förkastar H_0 om $\frac{S_X^2}{S_Y^2} > F_{1-\alpha/2,(n_x-1),(n_y-1)}$

Icke-parametriska tester.

* Om vi vill testa en hypotes men inte kan göra ett normalfördelningsantagande, hur gör vi då?

* Istället för att titta på de verkliga värdena så tittar vi på tecknet (+ -)

* Hypotesen blir då:

  H_0: medianen är 0

  H_1: medianen är (större, mindre) skild från 0.

* Löser det med hjälp av p-värdesmetoden och att vi låter $X \in Bin(n, 0.5)$.

BIOSTATISTISK GRUNDKURS, MASB11
OH-BILDER 8

ICKEPARAMETRISKA TEST:

EXEMPEL: Man ville undersöka om energiförbrukning i vila är annorlunda hos personer drabbade av cystisk fibros jämfört med friska personer. Tio par matchades ihop, en i varje par hade sjukdomen medan den andre var frisk. För övrigt var personerna i varje par lika beträffande kön, ålder, vikt och längd. Resultat i energiförbrukning (kcal/dag):

Par	1	2	3	4	5	6	7	8	9	10
CF	1153	1132	1165	1460	1634	1493	1358	1453	1185	1824
Friska	996	1080	1182	1452	1162	1619	1140	1123	1113	1463
Skillnad CF-friska	157	52	−17	8	472	−126	218	330	72	361

(a) Hur gör vi om det är rimligt att tänka sig att differenserna kommer från en normalfördelning?

(b) Hur gör vi utan antaganden om normalfördelning?

EXEMPEL: I en skola ville man göra en liten pilotstudie för att se om en annorlunda idrottsträning på kort tid skulle kunna påverka skolbarnens fysiska prestationer. Man valde ut 16 barn, som var likvärdiga beträffande den fysiska kapaciteten. Barnen delades slumpmässigt in i två grupper. Under en månad följde hälften av barnen (grupp A) den normala undervisningen i ämnet Idrott och hälsa, medan de övriga barnen (grupp B) dessutom fick delta i den speciella träningen. När en månad hade gått, fick barnen vid ett gemensamt tillfälle springa en kort terrängbana och deras tider noterades. Två barn i grupp A var sjuka under testdagen. Resultat (sekunder):

Grupp A	64	62	73	54	66	71		
Grupp B	53	74	70	59	42	38	48	60

SITUATION	NORMALFÖRDELNING	ICKE-PARAMETRISKT
Ett stickprov	t-test	teckentest
Två stickprov, matchade data	t-test på differenserna	teckentest på differenser
Två stickprov, oberoende data	t-test för två ober stickprov	Wilcoxon-Mann-Whitney's test
Flera stickprov, oberoende data	ensidig variansanalys	Kruskal-Wallis test

FÖR- OCH NACKDELAR MED ICKE-PARAMETRISKA TEST

(+) Behöver inte göra antaganden om fördelning hos data

(+) Fungerar för små stickprov

(+) "Robust" mot outliers — d.v.s. påverkas inte så mycket av kraftigt avvikande värden i ett datamaterial

(−) Är inte lika "känsliga" som (har mindre styrka än) de test som baseras på normalfördelning — d.v.s. det behövs ett större stickprov för att förkasta en felaktig H_0

(−) Nollhypotesen är oftast inte lika specificerad som i "traditionella" test

(−) Utnyttjar inte all information om fördelningen som ges i data — baseras oftast på ranger, inte på de aktuella värdena i mätningarna

Interens om proportioner

* En proportion anger hur stor andel av datan som innehar en viss egenskap.

T. ex. Andelen som röstade i senaste valet eller andelen elever i kursen MASB11 som klarade mozquitto 1 på första försöket.

$$\hat{p} = \frac{x}{n}$$

(85.8% röstade i valet 2014 och 10% klarade mozquitto 1 vid första försöket.)

Inferens om proportioner

* I de flesta fallen är $X \in Bin(n, p)$ och p skattas som $p = \frac{X}{n}$.

Inferens om proportioner

* I de flesta fallen är $X \in Bin(n, p)$ och p skattas som $p = \frac{X}{n}$.

* Om vi utnyttjar detta, då kan vi skapa konfidensintervall för p samt göra hypotestest för p.

Inferens om proportioner

* I de flesta fallen är $X \in Bin(n, p)$ och p skattas som $\hat{p} = \frac{X}{n}$.

* Om vi utnyttjar detta, då kan vi skapa konfidensintervall för p samt göra hypotestest för p.

Kom ihåg att om $X \in Bin(n, p)$ då är $E[X] = np$ och $Var(X) = np(1-p)$

Inferens om proportioner

* I de flesta fallen är $X \in Bin(n, p)$ och p skattas som $\hat{p} = \frac{X}{n}$.

* Om vi utnyttjar detta, då kan vi skapa konfidensintervall för p samt göra hypotestest för p.

Kom ihåg att om $X \in Bin(n, p)$ då är $E[X] = np$ och $Var(X) = np(1-p)$

$\hat{p} = \frac{X}{n}$:

$$E[\hat{p}] = E\left[\frac{X}{n}\right] = \frac{E[X]}{n} = \frac{np}{n} = p$$

$$Var[\hat{p}] = Var\left[\frac{X}{n}\right] = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$$

Interens om proportioner

- Så om $X \in Bin(n, p)$ då är

$$E[\hat{p}] = p \quad , \quad Var[\hat{p}] = \frac{p(1-p)}{n}$$

- En rimlig skattning av p är $\frac{x}{n} = \hat{p}_{obs}$.

Alltså:

$$\widehat{E[\hat{p}]} = \hat{p}_{obs} \quad , \quad \widehat{Var[\hat{p}]} = \frac{\hat{p}_{obs}(1-\hat{p}_{obs})}{n}$$

Interens om proportioner

- Så om $X \in Bin(n, p)$ då är

$$E[\hat{p}] = p \quad , \quad Var[\hat{p}] = \frac{p(1-p)}{n}$$

- En rimlig skattning av p är $\frac{x}{n} = \hat{p}_{obs}$.

Alltså:

$$\widehat{E[\hat{p}]} = \hat{p}_{obs} \quad , \quad \widehat{Var[\hat{p}]} = \frac{\hat{p}_{obs}(1-\hat{p}_{obs})}{n}$$

---

* Om $np(1-p) \geq 10$ då kan vi normalapproximera $X \overset{\sim}{\in} N(np, np(1-p))$ vilket ger $p \overset{\sim}{\in} N\left(p, \frac{p(1-p)}{n}\right)$.

Interens om proportioner

- Så om $X \in Bin(n, p)$ då är

$$E[\hat{p}] = p \quad , \quad Var[\hat{p}] = \frac{p(1-p)}{n}$$

- En rimlig skattning av p är $\frac{x}{n} = \hat{p}_{obs}$.

Alltså:

$$\widehat{E[\hat{p}]} = \hat{p}_{obs} \quad , \quad \widehat{Var[\hat{p}]} = \frac{\hat{p}_{obs}(1-\hat{p}_{obs})}{n}$$

---

* Om $np(1-p) \geq 10$ då kan vi normalapproximera $X \overset{\sim}{\in} N(np, np(1-p))$ vilket ger $\hat{p} \overset{\sim}{\in} N\left(p, \frac{p(1-p)}{n}\right)$.

* Ett konfidensintervall för p blir då

$$I_p = \left(\hat{p}_{obs} \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}_{obs}(1-\hat{p}_{obs})}{n}} \right)$$

OBS! Ej t-fördelningen; vi har normalapprox!

Inferens om proportioner

* Vi kan också göra hypotestest.

$H_0: p = p_0$

$H_1: p > p_0$

* Vi kan testa hypotesen genom:

$$z = \frac{\hat{p}_{obs} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \underset{\sim}{\in} N(0,1)$$

Testa mot normalkvantilen

$z_{1-\alpha}$ (ensidig hypotes)

$z_{1-\alpha/2}$ (dubbelsidig hypotes).

Inferens om proportioner

* Om vi istället har två proportioner vi vill jämföra, säg

$X \in Bin(n_x, p_x)$ och $Y \in Bin(n_y, p_y)$

Vi vill testa följande hypotes:

$H_0: p_x = p_y$

$H_1: p_x \neq p_y$

Anta att $n_x p_x (1-p_x) > 10$ och $n_y p_y (1-p_y) > 10$

Inferens om proportioner

* Om vi istället har två proportioner vi vill jämföra, säg

$X \in Bin(n_x, p_x)$ och $Y \in Bin(n_y, p_y)$

Vi vill testa följande hypotes:

$H_0: p_x = p_y$

$H_1: p_x \neq p_y$

Anta att $n_x p_x (1-p_x) > 10$ och $n_y p_y (1-p_y) > 10$

Under H_0 så gäller att $p_0 = p_x = p_y$.

Så vi kan skatta p_0 gemensamt:

$$\hat{p}_0 = \frac{x+y}{n_x + n_y}$$ och testa

$$z = \frac{\hat{p}_x - \hat{p}_y}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_x} + \frac{1}{n_y}\right)}}$$ mot normalfördelningen.

Finns i formelsamlingen!

INFERENS OM PROPORTIONER:

EXEMPEL: I Stockholms län gjorde man 1990 en undersökning av förekomsten av pollenallergi bland vissa känsliga grupper. Man valde slumpmässigt ut 500 personer i åldern 20–64 år och av dessa hade 23 % pollenallergi.

- **Vad kan vi säga om andelen pollenallergiker i populationen?**
- **År 1994 gjordes motsvarande undersökning och 500 nya personer valdes ut. Då hade 29% pollenallergi. Kan man rimligen säga att det skett en förändring av benägenheten för denna typ av allergi under perioden?**

ANALYS AV KATEGORIDATA

EXEMPEL: Varje individ i en viss population hör i genetiskt hänseende till en av fyra kategorier K_1, K_2, K_3, K_4. Teoretiskt skall de fyra kategoriernas storlekar förhålla sig som $9 : 3 : 3 : 1$. Vid en undersökning av 160 slumpmässigt utvalda ur populationen fick man följande resultat:

kategori	K_1	K_2	K_3	K_4
frekvens	78	42	27	13

Talar de observerade data emot teorin?

EXEMPEL: Finns det ett samband mellan blodgrupp och risken för magsår? Blodgruppen bestämdes för 1655 magsårspatienter och för en kontrollgrupp om 10000 personer från samma stad. Resultat:

	0	A	B	AB	Totalt
Magsårspatienter	911	579	124	41	1655
Kontrollgrupp	4578	4219	890	313	10000
Totalt	5489	4798	1014	354	11655

Inferens om proportioner

Lösning: $X =$ "Antal pollenallergiker" $\in Bin(500, p_1)$

$p_1 =$ andelen pollenallergiker

$\hat{p}_1 = 23\%$

Kan vi normalapproximera?

$$n \cdot p_1(1-p_1) = 500 \cdot 0.23 \cdot 0.77 = 88.55 > 10$$

Ja!

$$I_{p_1} = \left(\hat{p}_1 \pm z_{1-\alpha/2} \sqrt{\underbrace{Var(\hat{p}_1)}_{\frac{\hat{p}_1(1-\hat{p}_1)}{n}}} \right)$$

(Låt $\alpha = 0.05$)

$$= \left(0.23 \pm 1.96 \sqrt{\frac{0.23 \cdot (1-0.23)}{500}} \right)$$

$$= (0.19, 0.27)$$

Interens om proportioner

Lösning: X = "Antal pollenallergiker 1994"

$X \in Bin(500, p_2)$

1990 $\hat{p}_1 = 0.23$ $n_1 = 500$

1994 $\hat{p}_2 = 0.29$ $n_2 = 500$

Vi vill testa

$H_0: p_1 = p_2$

$H_1: p_1 \neq p_2$

Under H_0 är $p_0 = p_1 = p_2$.

$$\hat{p}_0 = \frac{x + y}{n_1 + n_2} = \frac{\hat{p}_1 n_1 + \hat{p}_2 n_2}{n_1 + n_2} = 0.26.$$

$$z = \left| \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_0(1-\hat{p}_0)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \right| = 2.1628 > 1.96 = z_{0.975}$$

H_0 kan förkastas på nivå $\alpha = 0.05$.

Inferens om proportioner

* Om stickprovet är litet så att normalapproximation ej funkar:

Använd direktmetoden!

Exempel: $X \in Bin(8, p)$, $x = 5$

$H_0 : p = 1/2$

$H_1 : p > 1/2$

p-värde = $P(\text{"Få det vi fick eller värre"} \mid H_0 \text{ sant})$

$= P(X \geq 5 \mid p = 1/2) = 1 - P(X \leq 4 \mid p = 1/2)$

$= 1 - 0.6367 = 0.3633.$

H_0 kan ej förkastas på någon rimlig nivå.

Analys av kategoridata

Ett försök kan utfalla på "k" olika sätt.

Gör "n" oberoende försök, räkna hur många försök som hamnar i varje kategori.

Leder oftast till χ^2-test.

Vi ska gå igenom 3 situationer:

1) Test av modellanpassning
2) Homogenitetstest.
3) Oberoende-test.

Analys av kategoridata

Test av modellanpassning:

Man har

k - kategorier

O_i - antalet observationer i kategori i

samt en hypotes kring modellen: tex.

$H_0: p_1 = \frac{9}{16}, p_2 = \frac{3}{16}, p_3 = \frac{3}{16}, p_4 = \frac{1}{16}$

H_1: några av dessa sannolikheter är fel.

Man skapar sen vad man borde ha sett om H_0 var sann:

$E_i = n \cdot p_i$ för varje kategori.

	Kategori 1	2	3	4	...	k
O_i	120	48	32	107	...	99
E_i	np_1	np_2	...			np_k

Test:

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

mot $\chi^2_{(1-\alpha, k-1)}$

Frihetsgrader = antal kategorier - 1.

Analys av kategoridata

Homogenitetstest

* Vi har k kategorier och r stickprov där vi har bestämt n_r (dvs antalet mätningar i stickprov r) innan.

	Kategori				
Stickprov	1	2	...	k	Summa
1	O_{11}	O_{12}	...	O_{1k}	$n_{1.}$
2	O_{21}	O_{22}	...	O_{2k}	$n_{2.}$
⋮	⋮		⋱	⋮	⋮
r	O_{r1}	O_{r2}	...	O_{rk}	$n_{r.}$
Summa	$n_{.1}$	$n_{.2}$	...	$n_{.k}$	n

Vi testar:

$H_0: p_{11} = p_{21} = p_{31} = \dots = p_{rk}$

$H_1:$ Några p skiljer sig.

Analys av kategoridata

$H_0: p_{11} = p_{21} = p_{31} = \cdots = p_{rk}$

$H_1:$ Några p skiljer sig.

Vi skapar en ny tabell som det borde vara om H_0 var sann.

$$E_{ij} = \frac{n_{i\cdot}\, n_{\cdot j}}{n}$$

	Kategori				
Stickprov	1	2	...	k	Summa
1	E_{11}	E_{12}	...	E_{1k}	$n_{1\cdot}$
2	E_{21}	E_{22}	...	E_{2k}	$n_{2\cdot}$
⋮	⋮		⋱	⋮	⋮
r	E_{r1}	E_{r2}	...	E_{rk}	$n_{r\cdot}$
Summa	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot k}$	n

Vi testar: $X^2 = \sum_i \sum_j \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

och jämför med $\chi^2_{1-\alpha,(r-1)(k-1)}$

Med $(r-1)(k-1)$ frihetsgrader.

Analys av kategoridata

Oberoende-test: (Typ samma som homogenitetstest)

Nu testar vi

H_0: $p_{ij} = p_i p_j$ för alla i och j.

	Kategori				
Stickprov	1	2	...	k	Summa
1	O_{11}	O_{12}	...	O_{1k}	$n_{1\cdot}$
2	O_{21}	O_{22}	...	O_{2k}	$n_{2\cdot}$
⋮	⋮		⋱	⋮	⋮
r	O_{r1}	O_{r2}	...	O_{rk}	$n_{r\cdot}$
Summa	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot k}$	n

	Kategori				
Stickprov	1	2	...	k	Summa
1	p_{11}	p_{12}	...	p_{1k}	$p_{1\cdot}$
2	p_{21}	p_{22}	...	p_{2k}	$p_{2\cdot}$
⋮	⋮		⋱	⋮	
r	p_{r1}	p_{r2}	...	p_{rk}	$p_{r\cdot}$
Summa	$p_{\cdot 1}$	$p_{\cdot 2}$		$p_{\cdot k}$	1

Analys av kategoridata

Om H_0 är sann så borde vi få:

$$E_{ij} = \frac{n_{i\cdot}\, n_{\cdot j}}{n}$$

	Kategori				
Stickprov	1	2	...	k	Summa
1	E_{11}	E_{12}	...	E_{1k}	$n_{1\cdot}$
2	E_{21}	E_{22}	...	E_{2k}	$n_{2\cdot}$
⋮	⋮		⋱	⋮	⋮
r	E_{r1}	E_{r2}	...	E_{rk}	$n_{r\cdot}$
Summa	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot k}$	n

Vi testar (igen)

$$\chi^2 = \sum_j \sum_i \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

och jämför med $\chi^2_{1-\alpha,(r-1)(k-1)}$

INFERENS OM PROPORTIONER:

EXEMPEL: I Stockholms län gjorde man 1990 en undersökning av förekomsten av pollenallergi bland vissa känsliga grupper. Man valde slumpmässigt ut 500 personer i åldern 20–64 år och av dessa hade 23 % pollenallergi.

- **Vad kan vi säga om andelen pollenallergiker i populationen?**
- **År 1994 gjordes motsvarande undersökning och 500 nya personer valdes ut. Då hade 29% pollenallergi. Kan man rimligen säga att det skett en förändring av benägenheten för denna typ av allergi under perioden?**

ANALYS AV KATEGORIDATA

EXEMPEL: Varje individ i en viss population hör i genetiskt hänseende till en av fyra kategorier K_1, K_2, K_3, K_4. Teoretiskt skall de fyra kategoriernas storlekar förhålla sig som $9 : 3 : 3 : 1$. Vid en undersökning av 160 slumpmässigt utvalda ur populationen fick man följande resultat:

kategori	K_1	K_2	K_3	K_4
frekvens	78	42	27	13

Talar de observerade data emot teorin?

EXEMPEL: Finns det ett samband mellan blodgrupp och risken för magsår? Blodgruppen bestämdes för 1655 magsårspatienter och för en kontrollgrupp om 10000 personer från samma stad. Resultat:

	0	A	B	AB	Totalt
Magsårspatienter	911	579	124	41	1655
Kontrollgrupp	4578	4219	890	313	10000
Totalt	5489	4798	1014	354	11655

Test av modellanpassning.

H_0: $p_1 = \frac{9}{16}$, $p_2 = \frac{3}{16}$, $p_3 = \frac{3}{16}$, $p_4 = \frac{1}{16}$

H_1: Inte som i H_0.

$n = 160$, $\alpha = 0.05$, $k = 4$ (Antal kategorier)

Test av modellanpassning.

H_0: $p_1 = \frac{9}{16}$, $p_2 = \frac{3}{16}$, $p_3 = \frac{3}{16}$, $p_4 = \frac{1}{16}$

H_1: Inte som i H_0.

$n = 160$, $\alpha = 0.05$, $K = 4$ (Antal kategorier)

O-tabell:

Kategori	K_1	K_2	K_3	K_4
Frekvens	78	42	27	13

E-tabell: (Enligt H_0)

Kategori	K_1	K_2	K_3	K_4
Frekvens	90	30	30	10

Test av modellanpassning.

H_0: $p_1 = \frac{9}{16}$, $p_2 = \frac{3}{16}$, $p_3 = \frac{3}{16}$, $p_4 = \frac{1}{16}$

H_1: Inte som i H_0.

$n = 160$, $\alpha = 0.05$, $K = 4$ (Antal kategorier)

O-tabell:

Kategori	K_1	K_2	K_3	K_4
Frekvens	78	42	27	13

E-tabell: (Enligt H_0)

Kategori	K_1	K_2	K_3	K_4
Frekvens	90	30	30	10

Test: $\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = \frac{(78-90)^2}{90} + \frac{(42-30)^2}{30} +$

$+ \frac{(27-30)^2}{30} + \frac{(13-10)^2}{10} = 7.6$

Test av modellanpassning.

H_0: $p_1 = \frac{9}{16}$, $p_2 = \frac{3}{16}$, $p_3 = \frac{3}{16}$, $p_4 = \frac{1}{16}$

H_1: Inte som i H_0.

$n = 160$, $\alpha = 0.05$, $k = 4$ (Antal kategorier)

O-tabell:

Kategori	K_1	K_2	K_3	K_4
Frekvens	78	42	27	13

E-tabell: (Enligt H_0)

Kategori	K_1	K_2	K_3	K_4
Frekvens	90	30	30	10

Test: $\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = \frac{(78-90)^2}{90} + \frac{(42-30)^2}{30} +$

$+ \frac{(27-30)^2}{30} + \frac{(13-10)^2}{10} = 7.6$

$\chi^2_{1-\alpha,\, k-1} = \chi^2_{0.95,\, 3} = 7.815$ (tabell/datorn)

Test av modellanpassning.

H_0: $p_1 = \frac{9}{16}$, $p_2 = \frac{3}{16}$, $p_3 = \frac{3}{16}$, $p_4 = \frac{1}{16}$

H_1: Inte som i H_0.

$n = 160$, $\alpha = 0.05$, $k = 4$ (antal kategorier)

O-tabell:

Kategori	K_1	K_2	K_3	K_4
Frekvens	78	42	27	13

E-tabell: (Enligt H_0)

Kategori	K_1	K_2	K_3	K_4
Frekvens	90	30	30	10

Test: $\chi^2 = \sum_{i=1}^{4} \frac{(O_i - E_i)^2}{E_i} = \frac{(78-90)^2}{90} + \frac{(42-30)^2}{30} +$

$+ \frac{(27-30)^2}{30} + \frac{(13-10)^2}{10} = 7.6$

$\chi^2_{1-\alpha,\, k-1} = \chi^2_{0.95,\, 3} = 7.815$ (tabell/dator)

Eftersom $\chi^2 = 7.6 < 7.815 = \chi^2_{0.95,3}$

så kan vi inte förkasta H_0 på nivån $\alpha = 0.05$.

INFERENS OM PROPORTIONER:

EXEMPEL: I Stockholms län gjorde man 1990 en undersökning av förekomsten av pollenallergi bland vissa känsliga grupper. Man valde slumpmässigt ut 500 personer i åldern 20–64 år och av dessa hade 23 % pollenallergi.

- **Vad kan vi säga om andelen pollenallergiker i populationen?**
- **År 1994 gjordes motsvarande undersökning och 500 nya personer valdes ut. Då hade 29% pollenallergi. Kan man rimligen säga att det skett en förändring av benägenheten för denna typ av allergi under perioden?**

ANALYS AV KATEGORIDATA

EXEMPEL: Varje individ i en viss population hör i genetiskt hänseende till en av fyra kategorier K_1, K_2, K_3, K_4. Teoretiskt skall de fyra kategoriernas storlekar förhålla sig som $9 : 3 : 3 : 1$. Vid en undersökning av 160 slumpmässigt utvalda ur populationen fick man följande resultat:

kategori	K_1	K_2	K_3	K_4
frekvens	78	42	27	13

Talar de observerade data emot teorin?

EXEMPEL: Finns det ett samband mellan blodgrupp och risken för magsår? Blodgruppen bestämdes för 1655 magsårspatienter och för en kontrollgrupp om 10000 personer från samma stad. Resultat:

	0	A	B	AB	Totalt
Magsårspatienter	911	579	124	41	1655
Kontrollgrupp	4578	4219	890	313	10000
Totalt	5489	4798	1014	354	11655

Homogenitetstest

$H_0 : p_0 = p_A = p_B = p_{AB}$

H_1 : H_0 stämmer ej.

r - rader
c - kolumner

O-tabell

	O	A	B	AB	Totalt
Magsår	911	579	124	41	1655
Kontrollgrupp	4578	4219	890	313	10 000
Totalt	5489	4798	1014	354	11655

Homogenitetstest

$H_0 : p_0 = p_A = p_B = p_{AB}$

H_1 : H_0 stämmer ej.

r - rader
c - kolumner

O-tabell

	O	A	B	AB	Totalt
Magsår	911	579	124	41	1655
Kontrollgrupp	4578	4219	890	313	10 000
Totalt	5489	4798	1014	354	11655

E-tabell

	O	A	B	AB	Totalt
Magsår	779.4	681.3	144.0	50.3	1655
Kontrollgrupp	4709.6	4116.7	870.0	303.7	10000
Totalt	5489	4798	1014	354	11655

Homogenitetstest

$H_0 : p_0 = p_A = p_B = p_{AB}$

H_1 : H_0 stämmer ej.

r - rader
c - kolumner

O-tabell

	O	A	B	AB	Totalt
Magsår	911	579	124	41	1655
Kontrollgrupp	4578	4219	890	313	10 000
Totalt	5489	4798	1014	354	11655

E-tabell

	O	A	B	AB	Totalt
Magsår	779.4	681.3	144.0	50.3	1655
Kontrollgrupp	4709.6	4116.7	870.0	303.7	10000
Totalt	5489	4798	1014	354	11655

$$\chi^2 = \sum_{c=1}^{4} \sum_{r=1}^{2} \frac{(O_{cr} - E_{cr})^2}{E_{cr}} = \frac{(911 - 779.4)^2}{779.4} + \dots$$

$$= 49.0155.$$

Homogenitetstest

$H_0 : p_0 = p_A = p_B = p_{AB}$

H_1 : H_0 stämmer ej.

r - rader
c - kolumner

O-tabell

	O	A	B	AB	Totalt
Magsår	911	579	124	41	1655
Kontrollgrupp	4578	4219	890	313	10000
Totalt	5489	4798	1014	354	11655

E-tabell

	O	A	B	AB	Totalt
Magsår	779.4	681.3	144.0	50.3	1655
Kontrollgrupp	4709.6	4116.7	870.0	303.7	10000
Totalt	5489	4798	1014	354	11655

$$\chi^2 = \sum_{c=1}^{4} \sum_{r=1}^{2} \frac{(O_{cr} - E_{cr})^2}{E_{cr}} = \frac{(911 - 779.4)^2}{779.4} + \dots$$

$= 49.0155$. Jämförs med $\chi^2_{1-\alpha,(r-1)(c-1)}$

$\chi^2_{0.999,3} = 16.266$. Eftersom $49.0155 > 16.266$ så kan vi förkasta H_0 på nivån $\alpha = 0.001$

Provtagningstider – problemställningar från labbet

Vid vårt kemilaboratorium analyserar vi bland annat en rad prover från den näraliggande provtagningscentralen. Det har framkommit starka önskemål om att vissa patienter som tar prover ska kunna träffa en läkare vid samma besök och inte behöva boka in olika dagar för provtagning och läkarbesök. För att detta önskemål ska uppfyllas måste vi förstås veta hur lång tid det tar från det att provet tagits på patienten till anlyssvaret är klart och undersöka om det är rimligt att låta patienter vänta på provsvar samma dag. Det finns flera moment att ta hänsyn till: provet kan få vänta på provtagningscentralen tills det blir hämtat till vårt laboratorium, det behövs en viss manuell handläggningstid av provet och slutligen har vi själva processtiden i maskinen. Dessutom har vi lite olika hanteringstider beroende på vilken dag i veckan det är och om det är för- eller eftermiddag.

I datafilerna `proverfm.Rdata` och `proverem.Rdata` finns det tider (minuter) som det tog ”från patientarm till analyssvar”. Provet är på så kallad ”allmän kemi” och, som ni märker, har vi har delat upp data i två filer, en för prover tagna på förmiddagen och en för prover tagna på eftermiddagen.

Nu till våra frågor: När vi gjorde ett histogram på data slogs vi av att det inte alls liknade en normalfördelning. Det kan väl i och för sig vara rimligt, men kan vi dra några slutsatser då? Finns det andra fördelningar som kan användas för att modellera tiden? Mer specifikt:

(a) Hur sannolikt är det att en förmiddagspatient får vänta mer än två timmar på analyssvar?

(b) Vi vill kunna säga att ”95 % av förmiddagspatienterna kommer att ha sitt provsvar snabbare än x minuter”. Vad är då x? Om vi vill göra samma uttalande för eftermiddagspatienterna, vad är x då?

(c) Vi beräknar att ca 40 % av proverna kommer på förmiddagen och resten på eftermiddagen. Hur troligt är det att ett patientprov, taget någon gång under dagen, tar mer än två timmar att analysera?

(d) Om vi har 50 patientprover på förmiddagen, vad är sannolikheten att **genomsnittstiden** för dessa 50 prov överstiger en timme?

Tips på arbetsgång

- Titta på data (histogram, empirisk fördelningsfunktion). beräkna enkla mått (medelvärde, standardavvikelse). Jämför förmiddags- och eftermiddagstider.
- Anpassa en lämplig standardfördelning till förmiddagstiderna och skatta parametrarna i den fördelningen.
- Gör samma sak för eftermiddagstiderna.
- Svara på frågorna (a)–(c) genom att utnyttja de anpassade fördelningarna.
- Fundera på vad medelvärdet av 50 förmiddagstider har för fördelning. Använd de ”enkla mått” du beräknade tidigare för att hitta rätt parametrar i fördelningen.