Statistik och Dataanalys I

Föreläsning 22 - Chi2-test och beslut under osäkerhet

Oskar Gustafsson

Statistiska institutionen
Stockholms universitet

Översikt

- **Chi2-test för goodness of fit**
- **Chi2-test för oberoende**
- **Beslutsfattande under osäkerhet**

Kortkampanj (Uppgift 22.2 i SDM)

- Bank har tre sorters kreditkort: Silver, Gold och Platinum.
- Marknadsföringkampanj. Skillnad i vilken kortklass kunder ansöker om?
- Undersöker $n = 200$ personers ansökningar efter kampanj.

Korttyp	Innan	Efter	Stickprov efter	Förväntat om ingen effekt av kampanj
Silver	60%	55.5%	111	$200 \cdot 0.6 = 120$
Gold	30%	29.5%	59	$200 \cdot 0.3 = 60$
Platinum	10%	15%	30	$200 \cdot 0.1 = 20$

Chi2-test Goodness-of-fit

- **Räknedata**. Antal.
- Hypoteser
  - H_0: räknedata följer fördelning med sannolikhet p_k i cell k.
  - H_A: räknedata följer annan fördelning.
- Totalt antal i hela tabellen: n
- **Förväntat antal** i cell k: $\mathrm{Exp}_k = n \cdot p_k$.
  - Exempel: $\mathrm{Exp}_{\mathrm{silver}} = 200 \cdot 0.6 = 120$
- **Observerat antal** i cell k: Obs_k
  - Exempel: $\mathrm{Obs}_{\mathrm{silver}} = 111$

Korttyp	Innan	Efter	Stickprov efter	Förväntat om ingen effekt av kampanj
Silver	60%	55.5%	111	$200 \cdot 0.6 = 120$
Gold	30%	29.5%	59	$200 \cdot 0.3 = 60$
Platinum	10%	15%	30	$200 \cdot 0.1 = 20$

Chi2-test Goodness-of-fit

- Hypoteser
  - H_0: räknedata följer fördelning med sannolikhet p_k i cell k.
  - H_A: räknedata följer annan fördelning.

- Chi2 (χ^2) test för tabell med K celler - **teststatistika**

$$\chi^2 = \sum_{k=1}^{K} \frac{(\text{Obs}_k - \text{Exp}_k)^2}{\text{Exp}_k} = \sum_{\text{all cells}} \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}}$$

- **Under H_0** - **Chi2-fördelning** med $K - 1$ frihetsgrader

$$\chi^2 \sim \chi^2_{K-1}$$

Chi2-fördelning

- χ^2-fördelningen kan ses som summan av kvadrerade $N(0,1)$ variabler.
- Om vi har $X_1, X_2, \ldots, X_n \sim N(0,1)$ så får vi att:
  - $X_1^2 \sim \chi_1^2$
  - $X_1^2 + X_2^2 \sim \chi_2^2$
  - $\vdots$
  - $X_1^2 + X_2^2 + \cdots + X_n^2 \sim \chi_n^2$
- Antar endast positiva värden (kvadrater).
- Är skev åt höger.
- Vad händer med formen när vi summerar flera slumpvariabler? (ledtråd: CGS)

Chi2-fördelning

Chi2-fördelningen

Frihetsgrader, ν 4

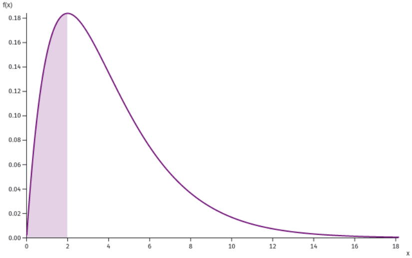
Kvantil: 2

Om $X \sim \chi^2(4)$ så gäller att

$E(X) = \nu = 4.0$

$Var(X) = 2\nu = 8.0$

$P(X \leq 2) = 0.2642$



Mattias Villani Chi2-fördelningen

Observable

Chi2-test Goodness-of-fit

- **Teststatistika**

$$\chi^2_{obs} = \sum_{\text{all cells}} \frac{(\text{Obs} - \text{Exp})^2}{\text{Exp}} = \frac{(111-120)^2}{120} + \frac{(59-60)^2}{60} + \frac{(30-20)^2}{20} = 5.6917$$

- **Under H_0** - **Chi2-fördelning** med $3 - 1 = 2$ frihetsgrader

- Kritiskt värde på signifikansnivå 5% från χ^2_2-tabell: $\chi^2_{crit} = 5.991$.

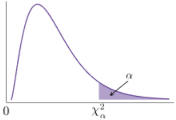
- Eftersom $\chi^2_{obs} < \chi^2_{crit}$ kan vi inte förkasta H_0.

- Finns inte stöd för att kampanjen har ändrat fördelningen över olika kortklasser.

Korttyp	Innan	Efter	Stickprov efter	Förväntat om ingen effekt av kampanj
Silver	60%	55.5%	111	$200 \cdot 0.6 = 120$
Gold	30%	29.5%	59	$200 \cdot 0.3 = 60$
Platinum	10%	15%	30	$200 \cdot 0.1 = 20$

Chi2-tabell

χ^2-fördelning



Right-tail probability:	0.100	0.050	0.025	0.010	0.005
df					
1	2.706	3.841	5.024	6.635	7.879
2	4.605	5.991	7.378	9.210	10.597
3	6.251	7.815	9.348	11.345	12.838
4	7.779	9.488	11.143	13.277	14.860
5	9.236	11.070	12.833	15.086	16.750
6	10.645	12.592	14.449	16.812	18.548
7	12.017	14.067	16.013	18.475	20.278
8	13.362	15.507	17.535	20.090	21.955
9	14.684	16.919	19.023	21.666	23.589
10	15.987	18.307	20.483	23.209	25.188

Test av oberoende - Hepatit C

	Hepatit C	Ej hepatit C	Total
Tatuering, studio	17	35	52
Tatuering, ej studio	8	53	61
Ingen tatuering	22	491	513
Total	47	579	626

- Hur skulle tabellen ovan skulle se ut om det **inte fanns något samband**?
- 47 av 626 personer testade positivt för hepatit C.
- Om hepatit C och tatueringsstatus är oberoende så borde andelen vara runt $47/626 = 0.075$ oavsett om personen var tatuerad eller inte.

Test av oberoende - Hepatit C

	Hepatit C	Ej hepatit C	Total
Tatuering, studio	17	35	52
Tatuering, ej studio	8	53	61
Ingen tatuering	22	491	513
Total	47	579	626

- $52/626 = 0.08$ har en tatuering från en tatueringsstudio. Av dessa 52 skulle vi förvänta oss att $52 * 47/626 = 3.9$ skulle ha hepatit C, om det inte finns något samband. Resten $(52 - 3.9 = 48.1)$ förväntas ej ha heptatit C.
- Av dom 61 som har en tatuering, men inte från en tatueringsstudio, förväntar vi oss $61 * 47/626 = 4.6$ med hepatit C och $61 - 4.6 = 56.4$ utan hepatit C.
- Vi beräknar dom 6 förväntade antalen om det ej finns samband och jämför med observerade data.

Test av oberoende - Hepatit C

Tatuering	Hepatit C	Obs	Exp	$\frac{(Obs-Exp)^2}{Exp}$
Studio	Ja	17	3.9	44.0
Studio	Nej	35	48.1	3.6
Ej studio	Ja	8	4.6	2.5
Ej studio	Nej	53	56.4	0.2
Ingen	Ja	22	38.5	7.1
Ingen	Nej	491	474.5	0.6

- Den totala avvikelsen är 58.0. Är denna avvikelse från vad vi förväntar oss tillräckligt stor för att vi ska förkasta antagandet om oberoende? Vi gör ett hypotestest!

Test av oberoende - Hepatit C

- H_0: dom två variablerna (tatueringsstatus och hepatit C i detta exempel) är oberoende.
- H_A : dom två variablerna är inte oberoende.
- **Teststatistika** $\chi^2 = \sum_{alla} \frac{(Obs - Exp)^2}{Exp}$.
- **Antal frihetsgrader**: $df = (n_{rader} - 1) \cdot (n_{kolumner} - 1)$.
- **Kritiskt värde**: $\chi^2_{df}(\alpha)$.
- **Förkasta** om $\chi^2_{obs} > \chi^2_{df}(\alpha)$.

Test av oberoende - Hepatit C

- H_0: dom två variablerna (tatueringsstatus och hepatit C i detta exempel) är oberoende.
- H_A : dom två variablerna är inte oberoende.
- **Teststatistika** $\chi^2_{obs} = \sum_{alla} \frac{(Obs-Exp)^2}{Exp} = 58.0$.
- **Antal frihetsgrader**:
  $df = (n_{rader} - 1) \cdot (n_{kolumner} - 1) = (3-1) \cdot (2-1) = 2$.
- **Kritiskt värde**, vi väljer $\alpha = 0.05$: $\chi^2_2(0.05) = 5.991$.
- $\chi^2_{obs} = 58 > 5.991$. Vi förkastar nollhypotesen på $\alpha = 0.05$ signifikansnivå.

Test av oberoende - antaganden

- **Räknedata**. Vi antar att vi har räknedata för individer, med värden för två variabler.
- **Oberoende**. Vi antar att observationerna är oberoende, exempelvis ett slumpmässigt urval.
- **Tillräcklig cellfrekvens**: vi antar att det förväntade antalet (Exp) är minst 5 i varje cell.

Test av oberoende - kollaps

- För hepatitexemplet så är inte det tredje antagandet uppfyllt, så vi bör vara försiktiga med våra slutsatser. Ett alternativ är att kollapsa olika kategorier. Lägsta förväntade blir då $113 * 47/626 = 8.5$, men vi tappar möjligheten att se om det är nån skillnad mellan tatuering i studio eller övrig tatuering.

	Hepatit C	Ej hepatit C	Total
Tatuering	25	88	113
Ej tatuering	22	491	513
Total	47	579	626

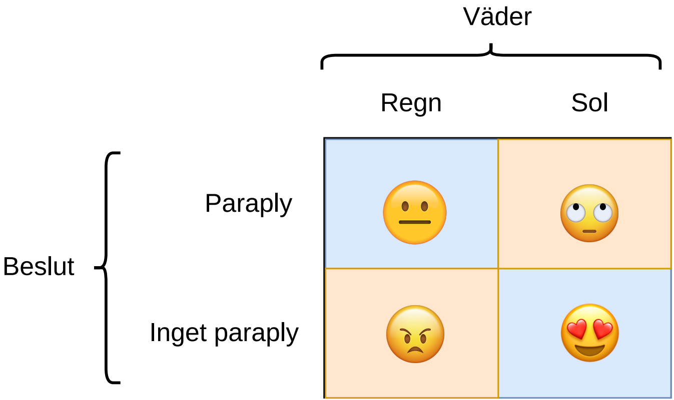
Beslut under osäkerhet

- Vi behöver ofta **fatta beslut** i en miljö med **osäkerhet**.
  - **Beslut**: Ska jag ta med ett paraply när jag går ut?
  - **Osäkerhet**: kommer det att regna?
  - **Beslut**: ska jag investera i aktier eller spara på banken?
  - **Osäkerhet**: börsens och inflationens utveckling under min placeringshorisont.
  - **Beslut**: Ska Sverige satsa på snabbtåg?
  - **Osäkerhet**: hur kommer elbilar utvecklas? klimatet? vad kommer det kosta? etc etc

Beslut och statistik

- Ett fattat beslut har **konsekvenser**.
- **konsekvenserna beror på osäkra faktorer** som vi inte vet när vi fattar beslutet.
- Vi behöver **sannolikhetfördelningen** för de osäkra kvantiteterna.
- Modellerar **osäker kvantitet** i form av en **slumpvariabel** X.
- Använder **data** (och expertkunskap) för att beräkna dessa sannolikheter. **Statistik**!

Beslut + Utfall = Konsekvens



Nyttobegreppet

- **Beslutsprocess**:
  - Du **fattar beslutet** a.
  - X **realiseras** som x.
  - Kombinationen a och x ger dig viss **nytta** (eng. **utility**):

$$U(a, x)$$

- Ibland: **förlust** $L(a, x)$ - vilket bara är negativ nytta

$$L(a, x) = -U(a, x)$$

Nytta

		Väder	
		Regn	Sol
Beslut	Paraply	0	50
	Inget paraply	-100	100

Maximin - en pessimistisk beslutsregel

- **Maximin**: välj beslut *a* som maximerar den minimala nyttan.
- **Garderar mot det värsta** som kan hända (pessimist).

		Väder: Regn	Väder: Sol
Beslut	Paraply	0	50
	Inget paraply	-100	100

- Maximin ignorerar hur sannolika utfallen är.

		Väder: 0.2 Regn	Väder: 0.8 Sol
Beslut	Paraply	0	50
	Inget paraply	-100	100

		Väder: 0.01 Regn	Väder: 0.99 Sol
Beslut	Paraply	0	50
	Inget paraply	-100	100

- I **spelteori** med intelligent **motståndare** är maximin optimal.

Maximera förväntad nytta

- Beslutsregel välj beslut a som maximerar förväntade nytta

$$\mathrm{EU}(a) = \sum_{\text{alla } x} U(a, x) \cdot P(X = x)$$

- Paraply-beslutet:

$$\begin{aligned} a_1 &= \text{Paraply}: & \mathrm{EU}(a) &= 0.2 \cdot 0 + 0.8 \cdot 50 = 40 \\ a_2 &= \text{Inget paraply}: & \mathrm{EU}(a) &= 0.2 \cdot (-100) + 0.8 \cdot 100 = 60 \end{aligned}$$

- Optimalt beslut: ta inte med paraply.

Beslut	Väder: 0.2 Regn	Väder: 0.8 Sol
Paraply	0	50
Inget paraply	-100	100

Maximera förväntad nytta

- Paraply-beslutet i Bergen:

$$
\begin{aligned}
a_1 &= \text{Paraply}: & \mathrm{EU}(a) &= 0.7 \cdot 0 + 0.3 \cdot 50 = 15 \\
a_2 &= \text{Inget paraply}: & \mathrm{EU}(a) &= 0.7 \cdot (-100) + 0.3 \cdot 100 = -40
\end{aligned}
$$

- Optimal beslut i Bergen: Paraply!

		Väder	
		0.7 Regn	0.3 Sol
Beslut	Paraply	0	50
	Inget paraply	-100	100

Credits

Dessa slides skapades för kursen statistik och dataanalys 1 av Mattias Villani HT 2023, och har modifierats av Oscar Oelrich VT 2024, och Oskar Gustafsson för VT 2025.