

## Matematisk statistik

⚠ Ej i formelblad. Förväntad kunskap.

❗ Viktigt att kunna.

❗ Mycket viktig att kunna. En "hint" om förekomst på tenta har förekommit.

# Kapitel 1

## Sannolikhet

Sannolikhet förkortas vanligtvis *s/h*.

Ange alltid sannolikhet som ett tal mellan **0** och **1**. Skriv enbart i procentform i svar. Detta då alla formler nyttjar det förstnämnda.

Uttryck	Engelskt uttryck	Förklaring
Utfallsrummet $\Omega$	Sample space	Mängden av alla tänkbara fall. Ex) Tärning: {1, 2, 3, 4, 5, 6}
Population	Population	
Stickprov	Sample	
Händelse	Event	En delmängd av utfallsrummet. Ex) Udda tal på en tärning: {1, 3, 5}
Oberoende händelser		Sannolikheten att båda inträffar ( $A$ )
Disjunkt	Disjoint	Två mängder som är disjunkta har <b>inget</b> gemensamt
Partition	Partition	En uppdelning av $\Omega$
Binomialfördelning		En statistisk fördelning. Den talar om hur sannolikheten ser ut för olika utfall. 1 fördelas på olika utfall.
Likformig sannolikhet		$p_x(x)$ har identisk sannolikhet oavsett $x$

### Den klassiska sannolikhetsdefinitionen

$$\underbrace{P}_{\text{Probability (sannolikhet)}} = \frac{\text{antal gynnsamma fall}}{\text{totalt antal fall}}$$

## Mängdlära

Den tomma mängden betecknas  $\emptyset$ .

$A \subseteq B$  betecknar att  $A$  är en **delmängd** till  $B$ . Det vill säga att alla element i  $A$  finns i  $B$ .

$x \in A$  betecknar att  $x$  finns i  $A$ .

$|A|$  betecknar **kardinaliteten** hos  $A$ , det vill säga antalet element.

$A \cap B$  betecknar **snittet** av  $A$  och  $B$ . Det vill säga alla element som finns i både  $A$  och  $B$ . För **oberoende** händelser gäller att  $P(A \cap B) = P(A) * P(B)$ .

$A \cup B$  betecknar **unionen** av  $A$  och  $B$ . Det vill säga alla element som finns i antingen  $A$  eller  $B$ . Sannolikheten för unionen av två händelser:  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$ .

$A \setminus B$  betecknar mängden  $a \in A | a \notin B$ . Det vill säga alla element i  $A$  som inte finns i  $B$ , "A inte B".

$u$  betecknar **universalmängden**. I den finns alla element i sammanhanget.

$A^c$  eller  $u \setminus A$  betecknar **komplementet** till  $A$ . Det vill säga alla element som inte finns i  $A$ .

## Permutationer (*Permutations*)

En permutation är ett sätt att välja och arrangerar element i en viss ordning. Hänsyn tas till ordningen, d.v.s.  $ACB \neq ABC$ .

$$\frac{n!}{(n-r)!} = {}^n P_r$$

### Exempel

Fem personer ingår i en grupp. Tre väljs ut. På hur många sätt kan detta ske?

$5 * 4 * 3 = 60$  olika sätt.

### Exempel

$n$  av  $r$  väljs.

$$n(n-1)(n-2)\dots(n(r-1)) = \frac{n!}{(n-r)!}$$

## Kombinationer (*Combinations*)

Hänsyn tas till inte till ordningen, d.v.s.  $ACB = ABC$ .

$$\frac{n!}{(n-r)!r!} = {}^n C_r = \binom{n}{r}$$

## ! Betingad sannolikhet (*Conditional probability*)

$$\frac{P(A \cap B)}{P(A)} = P(B | A)$$

### Exempel

Både  $A$  och  $B$  inträffar.

$$P(A \cap B) = P(A) * \underbrace{P(B | A)}_{\text{"B givet A"}}$$

Division med  $P(A)$  ger definitionen:

$$\frac{P(A \cap B)}{P(A)} = P(B | A)$$

Om  $A$  och  $B$  är oberoende gäller:

$$P(A \cap B) = P(A) * P(B)$$

Insättning i betingad sannolikhet ger:

$$\frac{P(A) * P(B)}{P(A)} = P(B | A) \iff P(B) = P(B | A)$$

## Satsen om total sannolikhet

Om  $B_1, \dots, B_n$  är en partition av  $\Omega$ , då gäller:

$$\begin{aligned} P(A) &= P(A \cap B1) + P(A \cap B2) + \dots \\ &\iff P(A) * P(B_1 | A) + P(A) * P(B_2 | A) + \dots \end{aligned}$$

## ! Bayes sats

Om  $B_1, \dots, B_n$  är en partition av  $\Omega$  och  $P(B_i) \neq 0, \forall i$  gäller för varje händelse  $A$  att

$$P(B_j | A) = \frac{P(B_j) * P(A | B_j)}{P(A)} = \frac{P(B_j) * P(A | B_j)}{\sum_{i=1}^n P(B_i) * P(A | B_i)}$$

## Exempel

Anta tillverkning av elektroniska prylar. Komponenter fås av diverse underleverantörer.

$$\begin{aligned} A &= \text{fel} \\ P(A | \text{underlev. 1}) &= 0.05 \\ P(A | \text{underlev. 2}) &= 0.10 \\ P(A | \text{underlev. 3}) &= 0.02 \\ P(\text{underlev. 1}) &= 0.5 \\ P(\text{underlev. 2}) &= 0.2 \\ P(\text{underlev. 3}) &= 0.3 \\ P(\text{underlev. 1} | A) &= \frac{0.5 * 0.05}{(0.5 * 0.05) + (0.2 * 0.10) + (0.3 * 0.02)} = \\ &= \frac{0.025}{0.025 + 0.02 + 0.006} = \frac{0.025}{0.051} \approx 0.5 \approx 50\% \end{aligned}$$

Har man en söndrig komponent är det alltså 50% chans att det är underleverantör 1 som levererat komponenten.

## Dragning utan återläggning (sample without replacement)

Kallas även för stickprov utan återläggning.

$$\frac{\text{gynnsamma utfall}}{\text{möjliga fall}} = \frac{\binom{A}{\text{tagna ur A}} \binom{B}{\text{tagna ur B}} \dots}{\binom{A+B+\dots}{\text{tagna}}}$$

Detta gäller även då vi har flera grupper, exempelvis A, B och C. Notera att summan av de tagna i täljaren måste vara lika med de tagna i nämnaren.

### Exempel

Det finns i en grupp studenter 50 svenska elever och 10 indiska. Två väljs ut slumpmässigt.

$$P(\text{en svensk, en indie}) = \frac{\text{gynnsamma utfall}}{\text{möjliga fall}} = \frac{\binom{50}{1} \binom{10}{1}}{\binom{60}{2}} = \frac{50 * 10}{1770} \approx 0.28$$

## Dragning med återläggning (sample with replacement)

Kallas även för stickprov med återläggning. Följer binomialfördelningen.

För beroende händelser:

$$p^x (1-p)^{n-x}$$

För oberoende händelser:

$$p^x (1-p)^{n-x} \binom{n}{x}$$

$\binom{n}{x}$  blir då och då 1, så som exemplet nedan och kan därmed försummas.

Där  $p$  är sannolikheten för ett lyckat utfall,  $x$  är antalet lyckade försök och  $n$  antalet utförda försök.

### Exempel

$$\begin{aligned} P(\text{vit kula}) &= p \\ P(\text{svart kula}) &= 1 - p \\ P(\text{två vita kolor}) &= p * p = p^2 \\ P(\text{en vit, en svart kula}) &= 2p(1-p) \\ P(\text{två svarta kolor}) &= (1-p)^2 \\ p^2 + 2p(1-p) + (1-p)^2 &= 1 \end{aligned}$$

### Exempel

En tenta består av 10 kryss-uppgifter. Det finns två alternativ. 5 frågor medför godkänt betyg. Vilken är chansen att man får godkänt genom att chansa svar?

$$Bin(\underbrace{10}_n, \underbrace{0.5}_p) \Rightarrow$$

$$P(10) + P(9) + \dots + P(5) = 0.5^{10} + 0.5 * 0.5^9 * 10 + \dots + 0.5^{10} \binom{n}{x} + \dots$$

$$0.623 \Rightarrow 0.377 \approx 38\%$$

Det är alltså 38% chans att få godkänt genom att slumpa svar.

## Stokastiska variabler

En stokastisk variabel är en slumpmässig variabel som antar olika, slumpmässiga värden. Tecknas vanligtvis som  $X, Y$  eller  $Z$ .

### Diskreta stokastiska variabler

Dessa variabler antar vissa fixa värden. Anteckna ändligt (såsom en tärning) eller ett uppräknat antal. Dessa variabler adderas och subtraheras vanligtvis vid beräkningar.

Följer följande form:

$$p_X(x) = P(\underbrace{X}_{\text{abstrakt variabel}} = \underbrace{x}_{\text{ett tal}})$$

#### Exempel - en tärning

$$p_X(3) = \frac{1}{6}$$

$$p_X(x) = \frac{1}{6} \Rightarrow \text{likformig sannolikhet}$$

### Kontinuerliga stokastiska variabler

Dessa variabler består av icke-uppräknliga värden. Beräkning medför derivator och integral.

## Fördelningsfunktion (*probability function*)

Betecknas  $F_X(x) = P(X \leq x)$ .

Notera att  $F_X(\text{övre gräns}) = 1$ .

#### Exempel - en tärning

$$F_X(0.5) = 0 (X \leq 0.5 \text{ går ej})$$

$$F_X(2) = \frac{2}{6} (X \leq 2 \text{ gäller för 1 och 2})$$

$$F_X(5) = \frac{5}{6}$$

$$F_X(9) = 1 (\text{alla möjliga utfall är mindre än } 9)$$

$$P(X \geq 6) = 1 - P(X \leq 5)$$

## ! Väntevärde / genomsnitt (*mean*)

Generellt sett gäller  $\sum x * p_X(x)$  - utfallet \* sannolikheten för utfallet.

Väntevärdet för en diskret stokastisk variabel defineras som:

$$\mu = E(X) = \sum_x x p_X(x)$$

där  $P_X(x)$  är sannolikheten för utfallet  $x$  för den stokastiska variabeln  $X$ . Summeringen görs över alla  $x$  i utfallsrummet. Notera att det inte är alltid ett väntevärde existerar i utfallsrummet. Exempelvis kan en tärning aldrig ge **3.5** när den slås.

Väntefärdet för en kontinuerlig stokastisk variabel defineras som:

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

där  $f(x)$  är fördelningens täthetsfunktion / frekvensfunktion. Alltid **1!**

## Exempel - en tärning

Vi kastar en tärning  $n$  gånger. Låt  $n$  vara stort.

$$\frac{\frac{n}{6} * 1 + \frac{n}{6} * 2 + \dots + \frac{n}{6} * 6}{n} = \\ \frac{1}{6} * 1 + \frac{1}{6} * 2 + \dots + \frac{1}{6} * 6 = 3.5$$

## ! Exempel - binomialfördelning

Det är viktigt att se när man kan använda detta.

$$\sum (x * p^x) (1-p)^{n-x} \binom{n}{x} = \dots = \underbrace{np}_{\text{viktigt!}}$$

Detta används då vi har  $n$  oberoende försök och sannolikheten för att lyckas  $P(\text{lyckas}) = p$  är konstant.  $X$  är antalet lyckade  $\Rightarrow X$  är binomialfördelat,  $X \in \text{Bin}(n, p)$ .

## Exempel - binomialfördelning

Man slår en tärning. Vad är sannolikheten att tre kast ger en sexa och två övriga?

$$\frac{1}{6}^1 \frac{5}{6}^2 \binom{3}{1} = 3 * \frac{1}{6} * \frac{25}{36} = \frac{75}{216} \approx 0.347$$

## Variationsmått

Varians är ett mått på variationen. Alltid större än 0.

$$E((X - m)^2) = \\ E(X^2) - m^2 = \\ E(x^2) - E(X)^2 = \\ V(X)$$

## Exempel - mynt

$$\begin{aligned} p_X(1) &= 0.6 \\ p_X(2) &= 0.2 \\ p_X(3) &= 0.2 \end{aligned}$$

Vilket är väntevärdet?

$$E(X) = \sum p_X(x) = 1 * 0.6 + 2 * 0.2 + 5 * 0.2 = 2$$

Detta betyder att man i genomsnitt får en 2-krona vid slumpmässigt val av mynt.

$$\begin{aligned} V(X) &= E(X^2) - E(X)^2 \\ E(X)^2 &= 2^2 = 4 \\ E(X^2) &= \sum x^2 p_X(x) = 1^2 * 0.6 * 2^2 * 0.2 + 5^2 * 0.2 = 6.4 \\ E(X^2) - E(X)^2 &= 6.4 - 4 = 2.4 \end{aligned}$$

Standardavvikelsen  $\sigma = \sqrt{2.4} \approx 1.55$ .

## Hypergeometrisk fördelning

En fördelning vid dragning utan återläggning.

### Exempel

Vi har  $a$  vita kolor och  $b$  svarta kolor.

$$P(x \text{ vita}, n-x \text{ svarta}) = \frac{g}{m} = \frac{\binom{a}{x} \binom{b}{n-x}}{\binom{a+b}{n}}$$

## Kapitel 3

---

## Diskret fördelning

Vad vi berört hittills. Bland annat likformig fördelning och Poisson-fördelning.

Här tecknas täthetsfunktionen  $p_X$  istället för  $f_X$ .

### Likformig fördelning:

Det är samma chans att få alla tal. Exempelvis har en tärning en likformig fördelning där varje uppställning ögon har  $\frac{1}{6}$  chans att slås.

$$\begin{aligned} P_X^k &= \frac{1}{n} \\ F_X(x) &= \sum_{k \leq x} P_X(k) \end{aligned}$$

### Poisson-fördelning

Sannolikhet under ett intervall, exempelvis tid.

$$p_X(x) = e^{-m} * \frac{m^x}{x!}$$

$$E(X) = m$$

$$V(X) = m$$

Där  $m$  är medelvärdet per intervalls-enhet. Exempelvis fyra samtal per minut. Notera att väntevärdet och variansen har samma värde.

$$X_1 \in Po(m_1)$$

$$X_2 \in Po(m_2)$$

$$X_1 + X_2 \in Po(m_1 + m_2)$$

### Exempel

Till en telefonväxel för ett företag kommer i snitt fyra samtal per minut. Vad är sannolikheten för noll samtal per minut?

Lösning:

$$m = 4$$

$$p_X(0) = e^{-4} * \frac{4^0}{0!} = e^{-4} \approx 0.018$$

### Exempel

Antal bilar på E22 är fyra per minut. Låt  $X$  beskriva antalet bilar per minut. Vad är sannolikheten att två eller färre bilar förekommer under fyra minuter?

$$p(X \leq 2)$$

$$E(X) = V(X) = 4$$

Svar: **0.238** enligt tabell.

### Geometrisk fördelning

$$P(X = n) = (1 - p)^n p$$

$$E(x) = \frac{1 - p}{p}$$

$$V(x) = \frac{1 - p}{p^2}$$

### Binomialfördelning

Ett antal försök och en sannolikhet att ett försök lyckas. Samma chans varje gång.

$$p_X(x) = \binom{n}{x} p^x (1 - p)^{n-x}, x = 0, 1, \dots, n$$

$$E(X) = np$$

$$V(X) = np(1 - p)$$

### Exempel

Används generellt då: Ett antal försök,  $P(x \text{ lyckas})$ ? Samma sannolikhet varje gång.

Vi har **10** komponenter. Sannolikheten att en är sönder är  $P(sönder) = 0.20$ .

Vad är sannolikheten att  $P(X \leq 1)$ ?

Svar: **0.37581** enligt tabell.

### Exempel

**20** försök görs. Sannolikheten för att ett lyckas är **0.2**. Vad är sannolikheten för att  $P(X \leq 3)$ ?

Svar: **0.41145** enligt tabell.

## Kontinuerlig fördelning

Man skiljer på diskreta och kontinuerliga fördelningar då matematiken bakom uträkningar skiljer sig.

Sannolikheten att  $a \leq x \leq b$  genom integralräkning:

$$\int_a^b f_X(x) dx$$

### Täthetsfunktion (density)

$$f_X(x) = F'_X(x)$$

## Likformig fördelning

$$f_X(x) = 1$$
$$F_X(x) = \int_{-\infty}^{\infty} f_X(x) dx = \int_0^x dy$$

### Exempel

Vi slumpar tal på en miniräknare. Den har samma fördelning mellan **0** och **1** ( $[0, 1]$ ).

Låt oss beräkna sannolikheten att  $x \leq 0.3$ .

$$P_x(x \leq 0.3) = \int_0^{0.3} f(x) dx = \int_0^{0.3} 1 dx = 0.3$$
$$F_X(x) = \int_{-\infty}^{\infty} f_X(x) dx = \int_0^x dy$$

## Exponentialfördelning

Används vanligtvis för väntetider. Exempelvis "tiden till första bilen kör förbi". Aldrig negativa värden på  $x$ ,  $x \geq 0$ .  $\lambda$  står för intensiteten / händelse per tidsenhet.

$$f_X(x) = \lambda e^{-\lambda x}$$
$$F_X(x) = \int_0^x \lambda e^{-\lambda y} dy = 1 - \lambda e^{-\lambda x}, x \geq 0$$

### Exempel - väntevärde

$$E(x) = \int_{-\infty}^{\infty} x F_X(x) dx = |\text{ej negativa } x| = \int_0^{\infty} x F_X(x) dx$$

$$F_X(x) = \lambda e^{-\lambda x}$$

$$E(x) = \int_0^{\infty} x \lambda e^{-\lambda x} dx = |\text{partial integration}| = \dots = \frac{1}{\lambda}$$

$$V(X) = E((X - m)^2) = E(X^2) - E(X)^2 = \frac{2}{\lambda^2} - (\frac{1}{\lambda})^2 = \frac{1}{\lambda^2}$$

$$\text{standardavvikelsen } \sigma = \sqrt{\frac{1}{\lambda^2}} = \frac{1}{\lambda}$$

## ! Räknelagar för väntevärden, varians och standardavvikelse

$$E(aX + b) = aE(X) + b$$

$$V(aX + b) = v(aX) = a^2 V(aX)$$

$$\sigma_{aX+b} = |a|\sigma_X$$

## Normalfördelning (normal distribution / gaussian distribution)

Normalfördelningen är den vanligaste fördelningen. Exempelvis medelvärdet av många mätningar, så som "längden av 18-åringar".

### Specialfall - Standardiserad normalfördelning (standardized normal distribution)

Anta att  $m = 0$ ,  $\sigma = 1$ , d.v.s  $x$  tillhör normalfördelningen mellan **0** och **1**.

$$\phi(x) = f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

$$\Phi(y) = \int_{-\infty}^y \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

Notera termen  $\frac{1}{\sqrt{2\pi}}$ , denna kallas normeringskonstant då den får integralens värde att bli ett (krav per definition). För  $\Phi(x)$  finns ingen enkel funktion.

Då det inte finns någon enkel funktion uppskattar man istället värden. Dessa värden finns i en tabell på sidan 339 i boken.

#### Exempel

$\Phi(0.31)$ ?

Slå upp **0.3** i den vänstra kolumnen och gå till den raden. Gå sedan till kolumnen där det översta värdet är **0.01**.

#### Exempel

För att få ner  $\sigma_X$  till en tiondel krävs 100 mätningar ( $\sqrt{n} = 10$ ,  $n = 100$ ).

# Normalfördelning

$$\begin{aligned}X_1 &\in N(m_1, \sigma_1) \\X_2 &\in N(m_2, \sigma_2) \\X_1 + X_2 &\in N(m_1 + m_2, \sigma_1 + \sigma_2)\end{aligned}$$

## ! Centrala gränsvärdesatsen CGS

Medelvärdet blir mer normalfördelat när mer mätningar görs.

Om  $X_1, X_2, \dots, X_n$  är en oändlig följd oberoende och likformigt fördelade stokastiska variabler med väntevärde  $m$  och standardavvikelsen  $\sigma$  där  $0 < \sigma < \infty$  gäller för summan  $S = X_1 + X_2 + \dots + X_n$  att  $P\left(\frac{S-n*m}{\sigma\sqrt{n}} < a\right) \rightarrow \Phi(a)$  då  $n \rightarrow \infty$ .

$$X \in N(m, \sigma)$$

## Specialfall - standardiserad

$$N(0, 1)$$

### Exempel

#### Exempel 1

$$\begin{aligned}X &\in N(0, 1) \\P(X \leq 0.23) &= |\text{tabell}| = 0.5910\end{aligned}$$

#### Exempel 2

$$\begin{aligned}X &\in N(0, 1) \\P(X \leq 0.84) &= |\text{tabell}| = 0.7995\end{aligned}$$

#### Exempel 3

När negativa värden ges gör man om olikheten då tabellen inte har negativa värden.

$$\begin{aligned}P(x \geq 1) &= 1 - P(X \leq 1) \\&= 1 - \Phi(x)\end{aligned}$$

#### Exempel 4

18-åringar mönstrar. Längden är normalfördelad likt  $X \in N(179, 6)$ . Beräkna sannolikheten att en 18-åring som mönstrar är 173cm eller kortare.

$$P(X \leq 173) = P\left(\underbrace{\frac{X - 179}{6}}_Z \leq \frac{173 - 179}{6}\right) = \underbrace{P(Z \leq -1)}_{N(0,1)} = 1 - P(Z \leq 1) \approx 0.1587$$

#### Exempel 5

Vad är  $P(X \leq 162)$ ?

Notera att i fallet kontinuerliga fördelningar spelar det inte någon roll om man skriver  $P(X \leq 162)$  eller  $P(X < 162)$ .

Lösning:

$$P(X \leq 162) = |\text{transformera}| = P\left(\underbrace{\frac{X - 168}{6}}_{Z, E(Z)=0, \sigma=1}\right) \leq \frac{162 - 168}{8} = P(Z \leq -1) = \\ 1 - P(z \leq 1) = 1 - \Phi(1) = 1 - 0.8413 = 0.1587$$

## Kapitel 4

### Exempel - dubbelintegraler

$$\int_{-1}^1 \int_{-1}^1 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dy dx = \int_{-1}^1 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} * \underbrace{\int_{-1}^1 \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} dy}_{\Phi(1)-\Phi(-1)} dx = \\ (\Phi(1) - \Phi(-1)) \underbrace{\int_{-1}^1 \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx}_{\Phi(1)-\Phi(-1)} = (\Phi(1) - \Phi(-1))^2$$

På tentan räknar vi ej med dubbelintegraler, det finns viktigare saker.

### Recap

Hittills har vi haft endimensionella funktioner för täthet.

En dimension	Två dimensioner	Tre dimensioner
$F_X(x)$	$F_{X,Y}(x, y)$	$F_{X,Y,Z}(x, y, z)$
$f_X(x)$	$f_{X,Y}(x, y)$	$f_{X,Y,Z}(x, y, z)$

### Definition för fördelningsfunktionen

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y)$$

### Definition för sannolikhetsfunktionen

$$f_{X,Y}(x, y) = \frac{\delta^2 F_{X,Y}(x, y)}{\delta x \delta y}$$

$$F_{X,Y}(x, y) = \int_{-\infty}^x \int_{-\infty}^y f_{X,Y}(u, v) dv du$$

### ! Marginalfördelningen

$$F_X(x) = F_{X,Y}(x, \infty)$$

$$F_Y(y) = F_{X,Y}(\infty, y)$$

För fördelningsfunktionen för  $x$  spelar inte  $y$  roll, måste vara mindre än oändligheten.

## ! Täthetsfunktioner

Ska ej räknas på tentan men bra att veta hur formler fungerar.

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx$$

När  $dy$  används 'försätts'  $y$ .

## Väntevärde, varians, standardavvikelse för kontinuerliga intervall

$$\mu = E(g(x, y)) = \int_{-\infty}^{\infty} g(x, y) * f_{X,Y}(x, y) dx dy$$

$$V(X) = \sigma^2 = \int_{-\infty}^{\infty} f(x)(x - \mu)^2 dx$$

$$\sigma = \sqrt{\int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx}$$

### Exempel 53 från boken

$p_{X,Y}(x, y) = P(X = x, Y = y)$ . Oberoende.

x/y	1	2	3	summa
0	0.1	0.2	0.1	0.4
5	0.15	a	b	0.6
summa	0.15	0.2+a	0.1+b	1

Observera att summan alltid är ett!

$$a = (0.2 + a) * 0.6$$

$$b = 0.6(0.1 + b)$$

$$0.2 = 0.12 + 0.6a \Rightarrow 0.08 = 0.6a \Rightarrow a = 0.133$$

$$b = 0.6 - 0.16 - 0.133 = 0.317$$

## Flerdimensionell normalfördelning

För kännedom, ej nödvändigtvis på tentan.

### Specialfall

$$M_X = 0, M_Y = 0, \sigma_X = 1, \sigma_Y = 1$$

$$f_{X,Y} = \frac{1}{2\pi\sqrt{1-\rho}} * \exp\left(\frac{-1/2}{1-\rho^2}(x^2 - 2\rho xy + y^2)\right)$$

## Korrelation

### ! Beroende

**Det är viktigt att kunna tolka diagram över korrelationen.**

Anta att två variabler är beroende. Vet vi  $x$  kan vi förutspå värdet på  $y$ .

Anta två datamängder där  $x$  är längd och  $y$  vikt.

Då  $\rho$  är  $\pm 1$  är korrelationen exakt. Det vill säga att  $y$  beror på  $x$ . Då  $\rho$  är exempelvis **0.9** eller **0.8** finns ett starkt samband mellan  $x$  och  $y$ . Är  $\rho > 0$  talar man om en positiv korrelation. Större  $x$ , större  $y$ .

Om  $\rho$  är 0 är värdena helt oberoende. Är  $\rho$  nära noll finns det en mycket svag korrelation. Är  $\rho < 0$  har man en negativ korrelation. Större  $x$ , mindre  $y$ .

$$\frac{C(X, Y)}{\sigma_X * \sigma_Y} = \rho$$

Där  $C(X, Y)$  är kovariansen / samvariationen av  $X$  och  $Y$ . Notera att  $\rho$  inte har någon sort eller enhet, det är enbart en konstant. Notera också att  $\rho$  alltid är i intervallet  $-1 \leq \rho \leq 1$ .

## Kovarians

$$C(X, Y) = Cov(X, Y) = E(X * Y) - E(X) * E(Y)$$

## Stora talens lag

Kommer ej på tentan men räknas som allmänbildning.

Låt  $X_1, X_2, \dots, X_n$  vara oberoende och lika fördelade stokastiska variabler med väntevärde  $m$  och standardavvikelsen  $\sigma$ .

$$\text{Låt } m_n = \frac{X_1 + X_2 + \dots + X_n}{n}.$$

Då gäller att  $\forall \epsilon > 0$  att  $\lim_{n \rightarrow \infty} P(m - \epsilon < X_n < m + \epsilon) = 1$ .

Med andra ord: desto fler mätningar, desto mindre standardavvikelse.

## Markovs olikhet

För  $a > 0$  och den stokastiska variabeln  $Y \geq 0$  gäller att  $P(Y \geq a) \leq \frac{E(Y)}{a}$ .

## Chebyshevs olikhet

Låt  $X$  vara en stokastisk variabel där  $E(X) = m$  och standardavvikelsen är  $\sigma$ . Då gäller att  $\forall k$  är  $P(|X - m| \geq k * \sigma) \leq \frac{1}{k^2}$ .

### ! Exempel

Vad är sannolikheten att medelvärdet av två kvinnors längd är mindre eller lika med 165?

$$\overline{X} = \underbrace{\frac{X_1+X_2}{2}}_{\text{normalf. rdelat}}, E(\overline{X}) = 168, \sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{2}} = \frac{\sigma}{\sqrt{2}} \approx 4.25.$$

Läggs två normalfordelade variabler ihop blir summan normalfordelad. Det gäller ej för exponentialfordelade etc.

$$P(\overline{X} \leq 165) = P\left(\underbrace{\frac{\overline{X} - 168}{4.25}}_Z \leq \frac{165 - 168}{4.25}\right) = P(Z \leq -0.71) = \\ \Phi(-0.71) = 1 - \Phi(0.71) = 1 - 0.7611 = 0.2389$$

**! Viktigt:**

Standardavvikelsen är standardavvikelsen för ett annat  $X$ , delat med  $\sqrt{n}$ , n antal mätningar.

$$\sigma_{\overline{X}} = \frac{\sigma_X}{\sqrt{n}}$$

## Kapitel 5

---

Transformationer

### Generell metod

Gäller då invers till  $g(X)$  finns och  $g(X)$  är växande ( $y' = \text{positivt}$ ).

$$Y = g(X)$$

$$F_Y(x) = P(Y \leq x) = P(g(X) \leq x) = P(\underbrace{g^{-1}(g(X))}_X \leq g^{-1}(x)) = F_X(g^{-1}(x))$$

### Exempel

$X$  är exponentialfordelad med  $\lambda = 3$ . Vad är fördelningen för  $Y = 2X$ ?

*Obs! Börja med fördelningsfunktionen.*

$$F_Y(x) = P(Y \leq x) = P(2X \leq x) = P(X \leq \underbrace{\frac{x}{2}}_{\text{Om positivt. Annars } \leq \text{ blir } >}) = F_X\left(\frac{x}{2}\right) = F_X(x) = 1 - e^{-3x}, x \geq 0$$

### Exempel

Om  $X$  har fördelningsfunktionen  $F_X(x)$  vad är då förelningsfunktionen för  $Y = X$ ?

$$F_Y(x) = P(Y \leq x) = P(-X \leq x) = P(x \geq -x) = 1 - P(X < -x) = 1 - F_X(-x)$$

### Exempel

$Y = \sqrt{X}$  Vad är fördelningsfunktionen för  $Y$  uttryckt i  $F_X(x)$ ?

$$F_Y(x) = P(Y \leq x) = P(\sqrt{X} \leq x) = P(X \leq x^2) = F_X(x^2)$$

## Exempel

$$Y = X^2$$

$$F_Y(x) = P(Y \leq x) = P(X^2 \leq x) = P(|X| \leq \sqrt{x}) = P(-\sqrt{x} \leq X \leq \sqrt{x}) = F_X(\sqrt{x}) - F_X(-\sqrt{x})$$

Notis:  $P(a \leq x \leq b)$  är på tallinjen det område mellan  $a$  och  $b$ , d.v.s  $F_X(b) - F_X(a)$

## ! Exempel - från tenta

I en julgransbelysning är lamporna seriekopplade.  $\mathbf{Y}$  är livslängden för en seriekoppling (Då den är seriekopplad är  $\mathbf{Y}$  alltså livstiden för den lampa som först går sönder). Med andra ord är

$$\mathbf{Y} = \min(X_1, X_2, \dots, X_n).$$

Vilken fördelning har slingan?

Ofta används Weibullfördelningen för att beskriva livslängden:  $F_X(x) = 1 - e^{-\frac{x}{b}^c}$ . Ett specialfall är när  $c = 1$  då Weibullfördelningen uppför sig som en vanlig exponentialfördelning.

$$\begin{aligned} F_Y(x) &= P(Y \leq x) = P(\min(X_1, X_2, \dots, X_n) \leq x) = 1 - P(\min(X_1, X_2, \dots, X_n) > x) \\ &\stackrel{\text{oberoende}}{=} 1 - P(X_1 > x, X_2 > x, \dots, X_n > x) = 1 - P(X_1 > x) * P(X_2 > x) * \dots * P(X_n > x) = \\ &\quad 1 - (1 - F_{X_1}(x)) * (1 - F_{X_2}(x)) * \dots * (1 - F_{X_n}(x)) \end{aligned}$$

Ett specialfall är då alla har samma fördelning:

$$1 - (1 - F_X(x))^n$$

## ! Exempel

Anta ett exponentialfördelat  $\mathbf{X}$ .

$$P(a < X \leq b) = \boxed{\int_a^b f_X(x) dx} = \boxed{F_X(b) - F_X(a)}$$

Notera att  $f_X(-a) = 0$  och  $F_X(-a) = 0$  för alla  $a \in N, a > 0$ .

## Kapitel 5

---

## Exempel

$F_X(x)$  är känd. Vad har  $\mathbf{Y} = 3\mathbf{X}$  för fördelning?

Lösning:

$$F_Y(x) = \underbrace{P(Y \leq x)}_{\text{definition}} = \underbrace{P(3X \leq x)}_{\text{insätt}} = \underbrace{P(X \leq \frac{1}{3}x)}_{F_X(\frac{1}{3}x)} = F_X(\frac{1}{3}x)$$

$F_X$  är känd. Svar:  $F_X(\frac{1}{3}x)$ .

## Exempel - från tenta

$Y = \ln(X)$ . Bestäm  $F_Y(x)$  som funktion av  $F_X(x)$ .

Lösning:

$$F_Y(x) = \underbrace{P(Y \leq x)}_{\text{definition}} = \underbrace{P(\ln(X) \leq x)}_{\text{insatt}} = \underbrace{P(X \leq e^x)}_{F_X(e^x)} \stackrel{\ln är växande i punkten }{=} F_X(e^x)$$

## Livslängd för minimum (seriekoppling)

Se exempel om julgran för exempel.

$$F_Y(x) = 1 - \underbrace{(1 - F_X(x))^n}_{\text{en enhet}}$$

### Exempel

Anta att  $X_1, X_2, \dots, X_n$  är exponentialfördelade. Det vill säga att  $F_X(x) = 1 - e^{-\lambda x}$ . Då har livslängden för minimum fördelningsfunktionen:

$$\begin{aligned} F_Y(x) &= 1 - (1 - F_X(x))^n = 1 - (1 - (1 - e^{-\lambda x}))^n = \\ &= 1 - (e^{-\lambda x})^n = 1 - e^{-\lambda n x}, x \geq 0 \end{aligned}$$

Kommentar:

$$\begin{aligned} E(X) &= \frac{1}{\lambda} = |\text{exempel}| = 100 \\ E(Y) &= \frac{1}{n\lambda} = |\text{exempel}| = \frac{100}{20} = 5 \end{aligned}$$

Livslängden är livslängden för en lampa delad med antalet.

## Livslängd för maximum (parallelkoppling)

$$F_Y(x) = F_X(x)^n$$

### Exempel

Låt  $X_1, X_2, \dots, X_n$  beskriva komponenter i ett system. När fungerar systemet? Det fungerar så länge minst en komponent fungerar. Då är livslängden  $Z = \max(X_1, X_2, \dots, X_n)$ .

$$\begin{aligned} F_Z(x) &= P(Z \leq x) = P(\max(X_1, X_2, \dots, X_n)) = P(X_1 \leq x, X_2 \leq x, \dots, X_n \leq x) \stackrel{\text{beroende}}{=} \\ &= P(X_1 \leq x) * P(X_2 \leq x) * \dots * P(X_n \leq x) = F_{X_1} * F_{X_2} * \dots * F_{X_n} = \\ &\text{notera att det inte är någon omskrivning med 1-....} \quad \text{Anta } F_{X_1}(x) = F_{X_2}(x) = \dots = F_{X_n}(x) = \boxed{F_X(x)^n} \end{aligned}$$

Notera att det på tentan räcker att använda den sista biten av formeln.

# Kapitel 6

Normalfordelningen kan användas för att approximera andra fördelningar.

## Binomialfordelningen

Givet att  $X \in Bin(n, p)$  där  $n$  är antalet försök och  $p$  är sannolikheten för ett lyckat försök gäller att om variansen  $npq \geq 10$  kan binomialfordelningen skrivas om som normalfordelning likt  $\underset{\sim}{X} \in N(np, \sqrt{npq})$ .

### Exempel

$$n = 100, p = 0.2. P(X \leq 15) = ?.$$

Här har vi inte en tabell för så många försök. Approximera med normalfordelningen.

$$X = X_1 + X_2 + \dots + X_{100}$$

Notera den centrala gränsvärdessatsen (CGS). Tillräckligt många observationer i en summa går mot normalfordelning oavsett fördelning för ursprungsfördelning.

$$X \in Bin(n, p), \text{givet CGS}, X \in N(m, \sigma)$$

Där  $m$  är väntevärdet för försöket och  $\sigma$  är standardavvikelsen. För binomialfordelning gäller att  $m = n * p$ , att  $\sigma = \sqrt{npq}$  och att  $q = 1 - p$ .

$$\begin{aligned} E(X_i) &= \sum x p_X(x) = 0(1-p) + 1*p = p \\ E(X) &= 100 * p \\ \sigma &= \sqrt{100 - p(1-p)} = \sqrt{16} = 4 \end{aligned}$$

Nu kan vi utgå från normalfordelningen.

$$P(X \leq 15) = P(\underbrace{\frac{X - 20}{4}}_Z \leq \frac{15 - 20}{4}) = P(Z \leq -1.25) = \Phi(-1.25) = 1 - \Phi(1.25) = 0.1506$$

### Exempel

Studenten Lena går på ett antal tentor. Under sin studietid går hon på 40 stycken. Låt  $X$  teckna antalet klarade tentor.  $X \in Bin(40, 0.7)$ . Vad är sannolikheten att Lena klarar 25 tentor eller färre? Hon borde klara  $40 * 0.7 = 28$  stycken tentor.

$$P(X \leq 25) = P\left(\frac{X - 28}{\sqrt{8.4}} \leq \frac{25 - 28}{\sqrt{8.4}}\right) = P(Z \leq \frac{-3}{\sqrt{8.4}}) = \Phi(-1.04) = 1 - \Phi(1.04) \approx 0.15$$

Notera att  $npq = 8.4$  i detta exemplet. Vanligtvis är  $\leq 10$  inte önskvärt för omskrivning av binomialfordelningen.

Svar: Sannolikheten är 0.15.

### Exempel

Vi gör hundra försök. Sannolikheten att ett försök lyckas är trettio procent. Det vill säga  $X \in \text{Bin}(100, 0.30)$ . Vad är sannolikheten att lyckas med 35 eller färre försök?

$$npq = 100 * 0.3 * 0.7 = 21, 21 > 10 \Rightarrow \text{norm. approx ok}$$

$$P\left(\underbrace{\frac{X - 30}{\sqrt{21}}}_Z \leq \frac{35 - 30}{\sqrt{21}}\right) \Rightarrow P\left(Z \leq \frac{5}{\sqrt{21}}\right) = \Phi(1.09) \approx 0.8621$$

Svar: Sannolikheten är **0.8621**.

## Poission-fördelning

Givet att  $X \in Po(m)$  där  $m$  är väntevärdet för  $X$  gäller att om  $m \geq 15$  kan Poisson-fördelningen skrivas om som normalfördelning likt  $\tilde{X} \in N(np, \sqrt{npq})$ .

### !! Exempel

Man studerar E22 under tio minuter.  $E(X) = V(X) = 10 * 4 = 40$ . Notera att generellt gäller att om  $X_1 \in Po(m_1), X_2 \in Po(m_2)$  så är också  $X_1 + X_2 \in Po(m_1 + m_2)$ .

$m \geq 15$ , approximation av fördelningen med normalfördelning är ok.

$$P(X \leq 50) = P\left(\frac{X - 40}{\sqrt{40}} \leq \frac{30 - 40}{\sqrt{40}}\right) = P\left(Z \leq \frac{10}{\sqrt{40}}\right) \approx \Phi(1.58) \approx 0.94$$

Svar: sannolikheten är **0.94**.

### !! Exempel

En telefonväxel tar mot 5 samtal per minut. Vad är sannolikheten att det under tio minuter förekommer **45** samtal eller mindre?

Lösning:

$$\begin{aligned} X_1 &\in Po(m_1) \\ X_2 &\in Po(m_2) \\ X_1 + X_2 &\in Po(m_1 + m_2) \end{aligned}$$

$X = \#\text{samtal under tio min}$ .  $X \in Po(50)$ .  $50 \geq 15$ , ok med approximationen.

$$P(X \leq 45) = P\left(\frac{X - 50}{\sqrt{50}} \leq \frac{45 - 50}{\sqrt{50}}\right) = P\left(Z \leq -\frac{5}{\sqrt{50}}\right) = \Phi(-0.71) = 1 - \Phi(0.71) = 0.2389$$

## Exempel

$X \in N(m_1, \sigma_1), Y \in N(m_2, \sigma_2)$ .

$X + Y$  är normalfördelat:

$$\begin{aligned} E(X + Y) &= E(X) + E(Y) = m_1 + m_2 \\ V(X + Y) &= V(X) + V(Y) + 2C(X, Y) \end{aligned}$$

$X - Y$  är normalfördelat:

$$\begin{aligned} E(X - Y) &= E(X) - E(Y) = m_1 - m_2 \\ V(X - Y) &= V(X) + V(Y) - 2C(X, Y) \end{aligned}$$

Observera att det är  $V(X) + V(Y)$  i båda fall.

Detta betyder att ingen linjär operation förändrar fördelningen för stokastiska variabler!

### Exempel 92 ur boken

Låt  $X$  tecknamannens längd och  $Y$  kvinnans. Vad är sannolikheten att en mans länd är mindre än en kvinnas?  $P(X < Y)$ ?

Ett trick är att subtrahera hela olikheten med högerledet.

$$P(X < Y) \iff P(X - Y < 0)$$

Tidigare har vi sett att även  $X - Y$  är normalfördelat. Låt  $Z = X - Y$ .

$$E(Z) = 177 - 167 = 10$$

$$V(Z) = 7^2 + 7^2 = 98$$

$$\sigma = \sqrt{98}$$

Alltså  $Z \in N(10, \sqrt{98})$ .

$$P(Z < 0) = P\left(\underbrace{\frac{Z - 10}{\sqrt{98}}}_{U} < \frac{0 - 10}{\sqrt{98}}\right) = P(U < -1.01) = 1 - \Phi(1.01) = 0.15625$$

## Kapitel 7

### Intensitet

Antalet fel per tidsenhet.

$$\lambda(t) = \lim_{h \rightarrow 0} \frac{P(Y \leq t + h \mid Y > t)}{h}$$

### ! Funktionssannolikhet (Reliability)

$$\begin{aligned} R(t) &= P(Y > t) = 1 - P(Y \leq t) = 1 - F_Y(t) \Rightarrow \\ P(Y \leq t + h \mid Y > t) &= \frac{P(t < Y \leq t + h)}{P(Y > t)} \Rightarrow \\ \lambda(t) &= \lim_{h \rightarrow 0} \frac{P(t < Y \leq t + h)}{P(Y > t)h} = \lim_{h \rightarrow 0} \frac{F_Y(t + h) - F_Y(t)}{h * R_Y(t)} \Rightarrow \\ \lim_{h \rightarrow 0} \frac{1 - R_Y(t + h) - (1 - R_Y(t))}{h * R_Y(t)} &= \dots = \lim_{h \rightarrow 0} \frac{R_Y(t) - R_Y(t + h)}{h * R_Y(t)} \Rightarrow \\ &\boxed{-\frac{R'_Y(t)}{R_Y(t)}} \end{aligned}$$

$-\frac{R'_Y(t)}{R_Y(t)}$  beskriver antalet fel per tidsenhet. Detta är viktigt att kunna inför tenta.

$$R(t) = R(0) * e^{-\int_0^t \lambda(y)du} = R(0)e^{-\lambda t} \underset{\text{vanligtvis}}{=} e^{-\lambda t}$$

För övrigt är  $R(0)$  sannolikheten att systemet fungerar vid början,  $P(\text{fungerar från början})$  exempelvis  $p = 0.98 \Rightarrow R(0) = 0.98, R(t) = 0.98e^{-\lambda t}$ .

IFR: increasing failure rate / växande. DFR: decreasing failure rate.

## ! Exempel

Låt  $\lambda(t)$  vara känd och  $-\frac{R'_Y(t)}{R_Y(t)}$  vara okänd.

$$\begin{aligned} \lambda(t) &= -\frac{R'_Y(t)}{R_Y(t)} = R'(t) + \lambda(t)R(t) = 0 \Rightarrow \\ y' + \lambda(t)y &= 0 \Rightarrow y(t) = C * e^{-\int_0^t \lambda(y)du} \end{aligned}$$

Notera likheten med  $C * e^{-\lambda t}$ .

$$\begin{aligned} y' + \lambda(t)y &= 0 \Rightarrow y'(t) * e^{-\int_0^t \lambda(y)du} \\ t = 0 \Rightarrow R(0) &= C, R(t) = \boxed{R(0)e^{-\int_0^t \lambda(y)du}} \end{aligned}$$

## Exempel

Vi har en exponentialfördelad livslängd  $Y$ . Vilken är funktionssannolikheten?

$$\begin{aligned} R(t) &= R(0) * e^{-\int_0^t \lambda(y)du} \\ \lambda(t) &= -\frac{R'(t)}{R(t)} = -\frac{-\lambda e^{-\lambda t}}{1 - (1 - e^{-\lambda t})} = \frac{\lambda e^{-\lambda t}}{e^{-\lambda t}} = \lambda \end{aligned}$$

Felintensiteten av en exponentialfördelad stokastisk variabel är alltså konstant.

## Exempel

Vid vilken tidpunkt har funktionssannolikheten  $R(t)$  gått ned till 50%? Anta exponentialfördelade livslängder.  $\lambda$  är givet.

$$\begin{aligned} R(t) &= 0.5 \\ e^{-\lambda t} &= 0.5 \\ -\lambda t &= \ln 0.5 \\ t &= \frac{\ln 0.5}{-\lambda} \end{aligned}$$

## Exempel

$\lambda(t) = C * \sqrt{t}, R(t) = ?$

$$\begin{aligned} R(t) &= e^{-\int_0^t c\sqrt{u}du} \\ u^{0.5} \Rightarrow \int u^{0.5} du &= \frac{u^{1.5}}{1.5} \Rightarrow R(c) = e^{-C\frac{t^{1.5}}{1.5}} \end{aligned}$$

## Exempel

Anta weibullfördelningen:

$$\lambda(t) = C * \frac{t^{C-1}}{a^C}$$

Vad är  $R(t)$ ?

$$R(t) = R(0)e^{-\int_0^t \lambda(u)du} = e^{-\int_0^t C * \frac{u^{C-1}}{a^C} du} = e^{-\frac{t^C}{a^C}} = e^{-(\frac{t}{a})^C}$$

# Kapitel 11

---

Statistisk teori och metodik. "Ingen matematisk formel utan massvis med data".

## Skattning (estimate)

---

Som upp "uppskattning" fast utan värderingen som vanligt associeras. En skattning måste vara rätt i genomsnitt. Väntevärdet är det riktiga värdet. Gärna låg variation.

En skattning tecknas  $\theta^*$ .

### Definition punkt-skattning

#### Krav 1

$$E(\theta^*) = \theta$$

En skattning sägs vara väntevärdesriktig (*wr*) om motsvarande stickprovsvariabel har väntevärdet **theta** ( $\theta$ ). På engelska heter det un-biased estimate.

Bias för en skattning är  $E(\theta^*) - \theta$ .

Detta krav innebär i kort att man ska undvika systematiska fel.

#### Krav 2

Stora talens lag - sannolikheten att man är långt från det riktiga värdet minskar med mer data.

Om för varje fixt  $\theta$  och varje givet  $\epsilon > 0$  gäller att  $P(|\theta^*(X) - \theta| > \epsilon) \rightarrow 0$  då  $n \rightarrow \infty$  sägs skattningen vara **konsistent** (viktigt).

#### Krav 3

Variationen bör vara låg, så liten som möjligt. Om för  $\theta_1, \theta_2$  (skattningar)  $V(\theta_1) < V(\theta_2)$  sägs  $\theta_1$  vara en **effektivare skattning** (efficient estimate) än  $\theta_2$ . Det vill säga, för  $\theta_1$  är variationen mindre, då krävs mindre data.

### ! Viktiga noter

Väntevärdet skattas med medelvärdet av alla observationer:  $\frac{1}{n} \sum x_i$ .

Skattningen av standardavvikelsen,  $\sigma^*$  skattas:

$$\sigma = \sqrt{V(X)} = \sqrt{\frac{1}{n-1} \left( \sum x_i^2 - \frac{1}{n} (\sum x_i)^2 \right)}$$

Skattningen av standardavvikelsen i kvadrat,  $(\sigma^2)^* = s^2$ :

$$(\sigma^2)^* = s^2 = \frac{1}{n-1} (\sum x_i^2 - \frac{1}{n} (\sum x_i)^2)$$

Variansen,  $V(X) = E(X^2) - E(X)^2$  skattas med:

$$\frac{1}{n-1} (\sum x_i^2 - \frac{1}{n} (\sum x_i)^2)$$

Motivering för variansen: Notera att

$$\begin{aligned} E(X^2) &= \frac{\sum x_i^2}{n} = \frac{1}{n} \sum x_i^2, \\ E(X)^2 &= \left( \frac{\sum x_i}{n} \right)^2 \Rightarrow \\ E(X^2) - E(X)^2 &= \frac{1}{n} \left( \sum x_i^2 - \frac{(\sum x_i)^2}{n} \right) \end{aligned}$$

Notera att  $\frac{1}{n}$  i det sista ledet bör vara  $\frac{1}{n-1}$  för att överensstämma med skattningen för variansen. Utöver det finns samtliga likheter. Detta kommer från en härledning att uttrycket inte är väntevärdesriktigt (ett krav för att få vara en skattning). Att sätta  $\frac{1}{n-1}$  löser detta. Se appendix e) i boken för bevis.

## Standardmetoder

För väntetider gäller

$$E(X) = \begin{cases} \sum_{-\infty}^{\infty} x * p_X(x) & \text{om diskret} \\ \int_{-\infty}^{\infty} f_X(x) dx & \text{om kontinuerlig} \end{cases}$$

## Generella metoder

Exempel på de två nedanstående metoderna förekommer efter båda metoderna presenterats.

### !! Maximum likelihood-metoden (ML-metoden)

Vanlig och relativt enkel. Skattningen har mindre varians än någon annan metod. På tentan erfodras lösning steg för steg.

Se boken sida 179-180 för egenskaper hos ML-skattningen.

1. Mindre varians än någon annan skattning.
2. Asymptotiskt normalfördelad då  $n \rightarrow \infty$ . Det vill säga, mer och mer normalfördelad desto mer data. Lätt att räkna med.
3. Att skatta en funktion. Funktionen av ML-skattningen  $g(\theta) \Rightarrow g(\theta^*)$ . Exempelvis  $x^2 \Rightarrow (x^*)^2$
4. Den är asymptotiskt värderiktig då  $n \rightarrow \infty$ . Den är inte alltid detta, men med tillräckligt högt  $n$  ger skattningen alltså väntevärdesriktighet.

### Steg 1

Ställ upp likelihood-funktionen  $L(\theta)$  och förenkla den så långt som möjligt.

$$L(\theta) = \begin{cases} p_X(x_1, \theta) * p_X(x_2, \theta) * \dots * p_X(x_n, \theta) & \text{om diskret} \\ f_X(x_1, \theta) * f_X(x_2, \theta) * \dots * f_X(x_n, \theta) & \text{om kontinuerlig} \end{cases}$$

### Steg 2

Beräkna logaritmen  $\ln(L(\theta))$ .

$$\begin{aligned} (f * g)' &= fg' + f'g \\ (\ln(f * g))' &= (\ln f + \ln g)' = \frac{f'}{f} + \frac{g'}{g} \end{aligned}$$

### Steg 3

Beräkna derivatan av  $\ln(L(\theta))$  med avseende på  $\theta$ . Maximera sedan.

$$\frac{d}{d\theta} \ln(L(\theta)) = 0$$

Lös sedan ut  $\theta$ . Detta ger ML-skattningen av  $\theta, \theta^*$ .

## Minsta kvadrat-metoden (MK-metoden)

Minimera variansen som funktion av  $\theta$ .

$$\sum (x_i - m(\theta))^2 = 0$$

## Exempel

### Exempel - ML-metoden (diskret)

Vi skattar väntevärdet  $m$  i en diskret Poissionfördelning. Låt  $x_1, x_2, \dots, x_n$  teckna observationer.

### Steg 1

$$\begin{aligned} L(m) &= e^{-m} * \frac{m^{x_1}}{x_1!} * e^{-m} * \frac{m^{x_2}}{x_2!} * \dots * e^{-m} * \frac{m^{x_n}}{x_n!} = \\ &\frac{e^{-m*n} * m^{\sum x_i}}{x_1! * x_2! * \dots * x_n!} \end{aligned}$$

### Steg 2

$$\ln(L(m)) = -m * n + (\sum x_i) * \ln(m) - \ln(x_1! * x_2! * \dots * x_n!)$$

### Steg 3

$$\begin{aligned} \frac{d}{dm} \ln(L(m)) &= -n + \frac{\sum x_i}{m} - 0, \\ -n + \frac{\sum x_i}{m} = 0 &\iff \frac{\sum x_i}{m} = n \iff m = \underbrace{\frac{\sum x_i}{n}}_{\text{medelvärde}, \bar{x}} \end{aligned}$$

Svar:  $m^* = \bar{x}$

### Exempel - ML-metoden (kontinuerlig)

Vi skattar värdet  $\lambda$  för en exponentialfördelning,  $f_X(x) = \lambda e^{-\lambda x}$ .

Steg 1

$$L(\lambda) = \lambda e^{-\lambda x_1} * \lambda e^{-\lambda x_2} * \dots * \lambda e^{-\lambda x_n} == \\ \lambda^n * e^{-\lambda \sum x_i}$$

Steg 2

$$\ln(L(\lambda)) = n * \ln \lambda - \lambda \sum x_i$$

Steg 3

$$\frac{d}{d\lambda} \ln(L(\lambda)) = \frac{n}{\lambda} - \sum x_i, \\ \frac{n}{\lambda} = \sum x_i = 0 \iff \frac{n}{\lambda} = \sum x_i \iff \lambda^* = \frac{n}{\sum x_i}$$

Svar:  $\lambda^* = \frac{1}{\bar{x}}$ .

### Exempel (ML-metoden)

Betrakta följande täthetsfunktion.

$$f_X(x) = \begin{cases} a^2 x e^{-ax} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Vi har data  $x_1, x_2, \dots, x_n$  och utför en ML-skattning på  $a$  ur ovanstående funktion.

Steg 1

$$L(a) = a^2 x_1 e^{-ax_1} * a^2 x_2 e^{-ax_2} * \dots * a^2 x_n e^{-ax_n} = \\ a^{2n} (x_1 * x_2 * \dots * x_n) * e^{-a \sum x_i}$$

Steg 2

$$\ln(L(a)) = 2n \ln a + \sum \ln x_i - a \sum x_i$$

Steg 3

$$\frac{d}{da} \ln(L(a)) = \frac{2n}{a} - \sum x_i = 0, \\ a^* = \frac{2n}{\sum x_i} = 2 * \frac{1}{\bar{x}}$$

Svar:  $a^* = \frac{2}{\bar{x}}$ .

### Exempel - från tenta 01-16 (ML-metoden)

Anta följande sannolhetsfunktion.

$$f_X(x) = \begin{cases} \frac{b}{x^c} & x \geq a \\ 0 & x < a \end{cases}$$

- a) 0.4 poäng. Beskriv sambandet för  $a$ ,  $b$  och  $c$  för att ovanstående funktion ska vara en täthetsfunktion.
- b) 0.6 poäng. Bestäm ML-skattningen för  $c$ ,  $a$  och  $b$  är givna.

#### Lösning uppgift a)

1. Funktionen är alltid över noll.

2. Se följande uträkning.

$$\begin{aligned} \int_{-\infty}^{\infty} f_X(x) dx &= 1 \Rightarrow \int_a^{\infty} bx^{-c} dx = \left[ b * \frac{x^{-c+1}}{-c+1} \right]_a^{\infty} = \\ 0 - b \frac{a^{-c+1}}{-c+1} &= b * \frac{a^{-c+1}}{c-1} = 1, \\ b &= \frac{c-1}{a^{-c+1}} \end{aligned}$$

#### Lösning uppgift b)

##### Steg 1

$$\begin{aligned} L(c) &= \frac{b}{x_1^c} * \frac{b}{x_2^c} * \dots * \frac{b}{x_n^c} = b^n * \frac{1}{(x_1 * x_2 * \dots * x_n)^c} = \frac{(c-1)^n}{a^{(c-1)^n}} * \frac{1}{(x_1 * x_2 * \dots * x_n)^c} = \\ &(c-1)^n * a^{n(c-1)} * \frac{1}{(x_1 * x_2 * \dots * x_n)^c} \end{aligned}$$

##### Steg 2

$$\ln(L(c)) = n * \ln(c-1) * \ln a - c * \sum \ln x_i$$

##### Steg 3

$$\begin{aligned} \frac{d}{dc} \ln(L(c)) &= \frac{n}{c-1} + n * \ln a - \sum \ln x_i = 0 \\ \frac{n}{c-1} &= \sum \ln x_i - n * \ln a \\ \frac{c-1}{n} &= \frac{1}{\sum \ln x_i - n * \ln a} \\ c-1 &= \frac{n}{\sum \ln x_i - n * \ln a} \\ c &= 1 + \frac{n}{\sum \ln x_i - n * \ln a} \end{aligned}$$

### Exempel - uppgift 5 från tenta 2016-10-28 1.0p (ML-metoden)

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty < x < \infty$$

ML-skatta  $\sigma$ .  $\sigma$  är känt.

### Steg 1

$$\begin{aligned} L(m) &= \frac{1}{\sigma\sqrt{2\pi}} * e^{-\frac{(x-m)^2}{2\sigma^2}} * \dots = \\ &(\frac{1}{\sigma\sqrt{2\pi}})^n * e^{\frac{1}{2\sigma^2}((x_1-m)^2 + \dots)} \end{aligned}$$

### Steg 2

$$\ln(L(c)) - n * \ln\sigma - \frac{n}{2} \ln(2\pi) - \frac{1}{2\sigma^2} * \sum(x_i - m)^2$$

### Steg 3

$$\begin{aligned} \frac{d}{dm} \ln(L(m)) &= -\frac{1}{2\sigma^2} \sum 2(x_i - m) * -1 = 0 \\ \sum(x_i - m) &= 0 \\ \sum x_i &= \sum m = nm \\ m^* &= \frac{\sum x_i}{n} = \bar{x} \end{aligned}$$

Notera att vi redan tidigare sett definitionen för  $m^*$  då den är känd och bevisad. Men i detta fallet måste den härledas för att få poäng på tentan.

### Exempel (MK-metoden)

Exempel ur boken, sida 180. Anta en exponentialfördelning.  $\lambda$  ska skattas.

Lösning: Minimera  $\sum(x_i - \frac{1}{\lambda})^2$  med avseende på  $\lambda$ .

$$\begin{aligned} \frac{d}{du} \sum(x_i - \frac{1}{\lambda})^2 &= 0 \\ \sum 2(x_i - \frac{1}{\lambda}) * \frac{1}{\lambda^2} &= 0 \\ \sum(x_i - \frac{1}{\lambda}) &= 0 \\ \sum x_i &= \sum \frac{1}{\lambda} = \frac{n}{\lambda} \\ \lambda^* &= \frac{n}{\sum x_i} = \frac{1}{\bar{x}} \end{aligned}$$

## Definition av medelfel (skattning av $\sigma$ )

### Exempel

$$\begin{aligned}
X &\in \text{Bin}(n, p) \\
E(X) &= np \\
V(X) &= npq, q = 1 - p \\
\sigma &= \sqrt{npq} = \sqrt{np(1-p)} \\
\sqrt{np^*q} &= \sqrt{np^*(1-p^*)}
\end{aligned}$$

## Kapitel 12

---

Hittils har vi behandlat punktskattningar. Nu följer intervallskattningar. "Hur stor andel vill köpa X? Undersökning visar med 95% sannolikhet att 30-35% vill köpa X."

### ⚠️ Exempel

$$\begin{aligned}
E(X) &= \int_{-\infty}^{\infty} t * f_X(t) dt \underset{X \geq 0}{=} \int_0^{\infty} t f_X(t) dt \\
\int gh dt &= Gh - \int G' h dt \\
\left| \begin{array}{ll} h(t) = t & h'(t) = 1 \\ g(t) = f_X(t) & G(t) = F_X(t) + C \end{array} \right| &= \left[ t(F_X(t) + C) \right]_0^{\infty} - \int_0^{\infty} (F_X(t) - 1) dt \\
F_X(\infty) = 1 \Rightarrow \left[ t(0) \right]_0^{\infty} &= 0 \Rightarrow \int_0^{\infty} \underbrace{(1 - F_X(t))}_{R(t)} dt \Rightarrow \\
&\boxed{\int_0^{\infty} (R(t)) dt}
\end{aligned}$$

Notera att detta inte står i formelbladet.

### Exempel

Anta en exponentialfördelning. Innan skrev vi då:

$$E(X) = \int_0^{\infty} t \lambda e^{-\lambda t} dt = \dots = \frac{1}{\lambda}$$

Nu skriver vi istället:

$$\int_0^{\infty} R(t) dt = \int_0^{\infty} (1 - (1 - e^{-\lambda t})) dt = \int_0^{\infty} e^{-\lambda t} dt = \left[ \frac{e^{-\lambda t}}{-\lambda} \right]_0^{\infty} = 0 - (-\frac{1}{\lambda}) = \frac{1}{\lambda}$$

## Intervallskattning

Vanligtvis har man en gräns, en sannolikhet med vilken man vill att svaret ska stämma. Av tradition är denna **konfidensgrad 95%**. Denna konfidensgrad anges vanligtvis i uppgiften på en tenta, annars utgår man från att det är **95%** som gäller.

För **95%** gäller alltså:

$$\boxed{\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}} \text{ kant } \sigma},$$

$$\boxed{\bar{x} \pm t_{0.025}(n-1) * \frac{s}{\sqrt{n}} \text{ okant } \sigma}$$

## Exempel

Anta normalfördelningen  $X \in N(0, 1)$ . Beräkna  $a$  i  $P(-a < X < a) = 0.95$ .

$$F_X(a) - F_X(-a) \iff \Phi(a) - \Phi(-1) = \Phi(a) - (1 - \Phi(a)) = 2\Phi(a) - 1 = 0.95 \Rightarrow \\ 2\Phi(a) = 1.95 \Rightarrow \Phi(a) \approx 0.975.$$

Genom att kolla i tabellen för normalfördelning finner vi att det värde som ger  $\Phi(a) = 0.975$  är  $a = 1.96$ . Det vill säga att  $P(-1.96 < X < 1.96) = 0.95$ .

Fortsättningsvis ges följande.

$$\begin{aligned} P(-1.96 < \frac{\bar{X} - m}{\sigma/\sqrt{n}} < 1.96) = 0.95 &= |\text{mult. med } \frac{\sigma}{\sqrt{n}}| = \\ P(-1.96 \frac{\sigma}{\sqrt{n}} < \bar{X} - m < 1.96 \frac{\sigma}{\sqrt{n}}) = 0.95 &= |\text{bort med } \bar{X}| = \\ P(-1.96 \frac{\sigma}{\sqrt{n}} - \bar{x} < -m < 1.96 \frac{\sigma}{\sqrt{n}} - \bar{x}) &= P(1.96 \frac{\sigma}{\sqrt{n}} + \bar{x} > m > \bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}) \end{aligned}$$

Detta innebär att med **95%** sannolikhet ligger väntevärdet i intervallet. Alltså är  $(-1.96 \frac{\sigma}{\sqrt{n}} + \bar{x}; 1.96 \frac{\sigma}{\sqrt{n}} + \bar{x})$  ett **95% konfidensintervall**.

Formeln ges då:

$$\boxed{(\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}})}$$

*Notera att formeln antar ett symmetriskt intervall. Står på sida 2 i formelbladet, näst längst ned.*

$$\boxed{\bar{x} \pm \lambda_{a/2} \frac{\sigma}{\sqrt{n}}}.$$

Där  $\lambda_{a/2}$  är 1.96 och  $a/2 = 0.025$ . Se en graf över normalfördelningen för mer förståelse.

## Exempel

Låt felsmarginalen vara  $1.96 \frac{\sigma}{\sqrt{n}}$ .

$$\sigma^* = s = \sqrt{\frac{1}{n-1} \left( \sum x_i^2 - \frac{1}{n} \left( \sum x_i \right)^2 \right)}$$

Finns i formelblad.

## Exempel - tentauppgift (känd $\sigma$ )

Längden för en kvinnlig teknolog är normalfördelat.  $X \in N(m, \sigma^2)$ . Man får mätevärdena **165, 170, 167, 168**. Bestäm ett **95%** konfidensintervall för längden.

Lösning:

$$\bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}, \bar{x} = 167.5, \sigma = 7, n = 4$$

$$167.5 \pm 1.96 \frac{7}{\sqrt{4}} = 167.5 \pm 1.96 * 3.5 = 167.5 \pm 6.86 \Rightarrow (160.64; 174.4)$$

För konfidensintervall avrundar man den nedre gränsen nedåt och den övre gränsen uppåt. På så vis täcks större yta och sannolikheten för rätt ökar.

Svar: **(160.6; 174.4)**.

### Exempel (okänd $\sigma$ )

Skatta  $\sigma$  ur data. Ur formelblad:

$$\sigma^* = s = \sqrt{\frac{1}{n-1} \left( \sum x_i^2 - \frac{1}{n} \left( \sum x_i \right)^2 \right)}$$

1. Skatta  $\sigma$  med s.
2.  $\bar{x} \pm \underbrace{t_{0.025}(n-1)}_{95\%} * \underbrace{\frac{s}{\sqrt{n}}}_{f}$

$t_{0.025}$  är en t-kvantil. Finns tabell över värdet i formelbladet.  $f =$  frihetsgraden =  $n - 1$ .

Längden för en kvinnlig teknolog är normalfördelat.  $X \in N(m, \sigma^2)$ . Man får mätevärdena **165, 170, 167, 168**. Bestäm ett **95%** konfidensintervall för längden.

$$\sum x_i^2 = 165^2 + 170^2 + \dots + 168^2 = 112238$$

$$(\sum x_i)^2 = (165 + 170 + \dots + 168)^2 = 112225$$

$$\sqrt{\frac{13}{3}} = 2.08$$

$$167.5 \pm t_{0.025}(3) * \frac{2.08}{2} = 167.5 \pm 3.31 \Rightarrow (164.19; 170.81)$$

Svar: **(164; 171)**

### Exempel

Data: **3, 4, 5, 6**.

$$\bar{x} \pm t_{0.025}(n-1) \frac{s}{\sqrt{n}}$$

$$\bar{x} = \frac{3+4+5+7}{4} = \frac{19}{4} = 4.75$$

$$s^2 = \frac{1}{n-1} \left( \sum x_i^2 - \frac{1}{n} \left( \sum x_i \right)^2 \right) = \frac{1}{3} (99 - \frac{1}{4} * 19^2) \approx 2.92$$

$$4.75 \pm t_{0.025}(3) * \frac{\sqrt{2.92}}{\sqrt{4}} = 4.75 \pm 3.182 * 0.85 = 4.75 \pm 2.72$$

Svar: **(2.03; 7.47)**.

# Generaliseringar av intervallskattning

Det är inte alltid man har ett konfidensintervall som tidigare, **95%** utan ibland skiljer det sig. På tentan kommer endast de kvantilerna som står i tabellen.

## Exempel - 99% konfidensintervall

Ytan utanför intervallet i en normalfördelning är **1%**, det vill säga att vardera sida täcker **0.005**,  $\lambda_{0.005}$ . Denna kvantil kallas fem-promilleskvantilen. Detta står i formelblad.  $\lambda_{0.005} = 2.58$ .

Detta ger formlerna

$$\bar{x} \pm 2.58 \frac{\sigma}{\sqrt{n}}, \text{ kant } \sigma$$

$$\bar{x} \pm t_{0.005}(n-1) * \frac{s}{\sqrt{n}}, \text{ okant } \sigma$$

## Exempel - 90% konfidensintervall

Ytan på vardera sida är **5%**  $\Rightarrow \lambda_{0.05} = 1.645$ .

Detta ger formlerna

$$\bar{x} \pm 1.645 \frac{\sigma}{\sqrt{n}}, \text{ kant } \sigma$$

$$\bar{x} \pm t_{0.05}(n-1) * \frac{s}{\sqrt{n}}, \text{ okant } \sigma$$

# Andra fördelningar (normalapproximation)

Man förutsätter att man kan göra en normalapproximation. Det är det enda som gör detta enkelt. Att svara exakt blir mycket komplicerat.

"Översättning" av ovanstående formler ger:

$$p^* \pm 1.96 \frac{\sqrt{npq}}{n} = p^* \pm 1.96 \sqrt{\frac{p^* q^*}{n}} \text{ binomial}$$

$$m^* \pm 1.96 * \sqrt{\frac{m^*}{n}} \text{ poission}$$

## !!! Exempel (binomial)

Vi undersöker andelen miljöpartister (mp). Av 1000 personer är 50 mp.

$$p^* = \frac{50}{1000} = 0.05$$

$$npq = 1000 * 0.05 * 0.95 = 47.5 > 10 \Rightarrow \text{norm. approx. ok!}$$

$$p^* \pm 1.96 \frac{\sqrt{47.5}}{1000} = p^* \pm 1.96 \underbrace{\sqrt{\frac{0.05 * 0.95}{1000}}}_{0.014} \Rightarrow (0.036; 0.064)$$

Det vill säga att miljöpartiets röstandel kommer att ligga mellan **3.6** till **6.4** procent.

## !! Exempel - poisson

Till en viss telefonväxel kommer i genomsnitt 80 samtal på en 2-minutersintervall. Gör ett **95%** konfidensintervall för väntevärdet av antalet samtal på två minuter.

$$80 \pm 1.97 \sqrt{\frac{80}{1}}$$
$$80 > 15 \Rightarrow \text{norm. approx. ok :-)}$$
$$80 \pm \underbrace{1.96\sqrt{80}}_{17.5} \Rightarrow (62.5; 97.5)$$

Svar: **(62.5, 97.5).**

## !! Exempel

"Hur många behöver man fråga för felmarginal på en viss procent?"

$$p^* \pm 1.96 \sqrt{\frac{p^* q^*}{n}}$$
$$\underbrace{0.02}_{\text{felmarginal}} = 1.96 \sqrt{\frac{p^* q^*}{n}}$$
$$0.0004 = 1.96^2 \frac{p^*(1-p^*)}{n}$$
$$n = 2500 * 1.96^2 \underbrace{p^*(1-p^*)}_{\text{graf max. vid } p^*=0.5}$$

Det vill säga worst case  $p^* = 0.5 \Rightarrow n = 2401$ . Utgår man från worstcase är man alltid på den säkra sidan.

Jämför med fallet då  $p^* = 0.1 \Rightarrow n = 864$ . Dessa siffror avrundas uppåt (större chans).

## !! Exempel

En teknolog vill undersöka hur stor andel som vill köpa en viss produkt. Konstanten  $p$  är helt okänd. Hur många personer måste tillfrågas om felmarginalen  $\leq 5\%$ ?

Lösning:

Okänt  $p \Rightarrow$  worst case.  $p^* = 0.5$ .

$$0.05 = 1.96 \sqrt{\frac{p^*(1-p^*)}{n}}$$
$$0.0025 = 1.96^2 \frac{p^*(1-p^*)}{n}$$
$$n = \frac{1.96^2}{0.0025} * 0.25 = 1.96^2 * 100 = 384$$

Det vill säga att teknologen behöver fråga 384 personer.

## Stickprov i par

Bit	Före	Efter	Differens
2	...	...	$z_1 = y_1 - x_1$
1	...	...	$z_2 = y_2 - x_2$
$n$	...	...	$z_n = y_n - x_n$

Då vi tidigare hade två stickprov (före och efter) har vi nu ett stickprov, differensen.

$$\bar{z} \pm t_{0.025}(n-1) \frac{s_z}{\sqrt{n}}$$

## Oberoende stickprov

$x_1, x_2, \dots, x_{n_1}$  och  $y_1, y_2, \dots, y_{n_2}$ .

$$\bar{x} - \bar{y} \pm 1.96 \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Finns i formelbladet. "Comparison of expected values..." .

## Vanligt specialfall

$\sigma_1 \approx \sigma_2 = s$

$$\bar{x} - \bar{y} \pm t_{0.025}(f)s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$f = n_x - 1 + n_y - 1$$

Skattningen av  $s$ :

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

## Exempel

Vi har en teknolog som sommarjobbar på en fruktodling.  $s_x = 30g$ ,  $s_y = 25g$ ,  $s_x \approx s_y \approx s$ . Finns det någon skillnad mellan äpplenas och päronens vikt  $\bar{x} = 0.2$ ,  $\bar{y} = 0.17$ . Vi har tio äpplen och tio päron

$$s^2 = \frac{9s_x^2 + 9s_y^2}{18} = \frac{9 * 30^2 + 9 * 25^2}{18} = 762.5 \Rightarrow s \approx 27.6$$

$$\bar{x} - \bar{y} \pm t_{0.025}(18) * s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$200 - 170 \pm \underbrace{2.10}_{\text{ur tabell}} * 27.6 \sqrt{\frac{1}{10} + \frac{1}{10}}$$

$$30 \pm 2.1 * 27.6 \sqrt{0.2} = 30 \pm 25.9 \Rightarrow (4.1, 55.9) \Rightarrow (4, 56)$$

Det vill säga, med 95% sannolikhet ligger skillnaden mellan ett äpple och ett pärons vikt mellan 4 och 56 gram.

# Kapitel 13

## Hypotestest

$$u = \frac{\bar{x} - m}{\sigma/\sqrt{n}} \quad (\text{känt } \sigma)$$

$$t = \frac{\bar{x} - m}{s/\sqrt{n}} \quad (\text{okänt } \sigma)$$

För känt  $\sigma$  jämförs  $u$  med **1.96**. För okänt  $\sigma$  jämförs  $t$  med  $t_{0.025}(n - 1)$ .

### Exempel från förra

$$u = \frac{\bar{x} - 120}{\sigma/\sqrt{n}}, u \in N(0, 1)$$

Hypotesen ovan ( $H_0$ ) är sann om äpplena väger 120g i genomsnitt.

$$\begin{aligned} \bar{x} &= \frac{150 + 160}{2} = 155 \\ u &= \frac{155 - 120}{0/\sqrt{2}} = \frac{35}{10} = 3.5 \end{aligned}$$

Här ser vi att detta blir fel. Det som är fel är 120. Det vill säga att det finns en verlig skillnad mellan äpplen och pärons vikt och att det inte är slumpen.

### Exempel - känt $\sigma$

Data: äpplens vikt:  $x_1 : 150g, x_2 : 160g, x_3 : 140g, x_4 : 120g$ . Vi har nollhypotesen att de i genomsnitt ska väga **120g** (nollhypotes: det vi utgår ifrån). Sigma är känt.

$$\begin{aligned} m &= 120 \\ \sigma &= 10 \text{ enligt nollhypotes} \\ n &= 4 \\ \bar{x} &= \frac{150 + 160 + 140 + 120}{4} = 142.5g \\ u &= \frac{\bar{x} - 120}{10/\sqrt{4}} = \frac{22.5}{5} = 4.5 \end{aligned}$$

Med 95% chans hamnar  $u$  mellan **-1.96 < u < 1.96**. 4.5 är utanför intervallet. Det sker endast med 5%. Sannolikt är det snarare vår hypotes fel.

**4.5 > 1.96 ⇒ nollhypotes ( $H_0$ ) förkastas, äpplena väger inte 120g i snitt**

### Exempel - okänt $\sigma$

Data:  $x_1 : 1, x_2 : 3, x_3 : 4$ .

$$s^2 = \frac{1}{n-1} \left( \sum x^2 - \frac{1}{n} \left( \sum x \right)^2 \right)$$

$$\sum x^2 = 1 + 9 + 16 = 26$$

$$\sum x = 1 + 3 + 4 = 8 \Rightarrow \frac{1}{n} \left( \sum x \right)^2 = \frac{1}{3} 8^2 = 21.33 \Rightarrow$$

$$\frac{1}{2} (26 - 21.33) = 2.33 \Rightarrow s = \sqrt{2.33} \approx 1.53$$

$$H_0 : m = 1$$

$$t = \frac{\bar{x} - m}{s/\sqrt{n}} = \frac{2.67 - 1}{1.53/\sqrt{3}} = \frac{1.67}{1.33} \sqrt{3} \approx 1.89$$

jämför med  $t_{0.025}(3-1) = t_{0.025}(2) = 4.303$   
 $-4.303 < 1.89 < 4.303$

Värdet ligger i intervallet. Det är sannolikt slumpen som påverkar värdet, inget faktiskt fel.

## Normalapproximation av binomialfördelningen

$$X \in Bin(n, p)$$

$$H_0 : p = p_0$$

test:  $np_0q_0 > 10 \Rightarrow$  norm. approx. ok

$$u = \frac{p^* - p_0}{\sqrt{\frac{p_0 q_0}{n}}}, q_0 = 1 - p_0$$

jämför med  $\pm 1.96$

### !! Exempel

En ordförande i ett parti tror att 10% i Karlskrona tillhör partiet. Hans partikamrat utför en undersökning. Han intervjuar 400 personer varav 30 tillhör partiet. Kan man förkasta hypotesen på nivån **0.005** (5%)?

$$u = \frac{p^* - p_0}{\sqrt{\frac{p_0 q_0}{n}}} = \frac{0.075 - 0.10}{\sqrt{\frac{0.1 * 0.9}{400}}} = -\frac{0.025}{\sqrt{\frac{0.09}{400}}} = -\frac{0.025}{0.3} * 20 = -\frac{0.5}{0.3} \approx -1.67$$

Vi drar slutsatsen att vi inte kan förkasta hypotesen. Värdet ligger i intervallet. Mycket av osäkerheten utgår från att vi har ett litet stickprov. I vanliga fall frågar man gärna 1000-2000 vid en undersökning.

## Normalapproximation av poissonfördelningen

$$u = \frac{\bar{x} - m}{\sqrt{\frac{m}{n}}}$$

### Exempel

Man antar att det på E22 under vårrusningen kommer 20 bilar per minut. Vi mäter antalet bilar och får 12 bilar per minut. Kan man förkasta hypotesen? Gör testet på nivån **0.05**.

Lösning:

$$u = \frac{12 - 20}{\sqrt{\frac{20}{1}}} = -\frac{8}{4.47} \approx -1.79$$

Vi kan inte förkasta hypotesen,  $-1.79 \in [-1.96, 1.96] \Rightarrow H_0$  förkastas ej.

## Stickprov i par (beroende)

Före	Efter	Differens
$x_1$	$y_1$	$z_1 = y_1 - x_1$
...	...	...
$x_n$	$y_n$	$z_n = y_n - x_n$

$$H_0 : z = 0$$

$$t = \frac{\bar{z}(-0)}{s_z / \sqrt{n}}$$

*t jämförs med  $t_{0.025}(n - 1)$*

Notera att formeln i detta fall indikerar att det förväntade värdet är **0**. Man drar generellt sätt bort det förväntade värdet från medelvärdet. Vidare är här  **$t_{0.025}$**  beteckningen för ett **95%** konfidensintervall. Är  **$t$**  utanför intervallet sägs det finnas en *signifikant avvikelse*. Är  **$t$**  inom intervallet ligger värdet *inom fejmarginalen*.

## Exempel

Före	Efter	Differens
100	120	20
150	160	10
200	205	5

$$H_0 : \text{skillnaden} = 0$$

$$t = \frac{\bar{z}(-0)}{s_z / \sqrt{n}}$$

$$\bar{z} = \frac{35}{5} = 11.67$$

$$s_z^2 = \frac{1}{n-1} \left( \sum x^2 - \frac{1}{n} \left( \sum x \right)^2 \right) = \frac{1}{2} (525 - \frac{1}{3} 35^2) = 58.33$$

$$t = \frac{11.67}{\sqrt{58.33}} \sqrt{3} \approx 2.65$$

$$t_{0.025}(2) = 4.30$$

Slutsats: Hypotesen kan inte förkastas. Det kan stämma att det är samma hållfastighet.

## Två oberoende stickprov

Vi har två oberoende stickprov,  $x_1, x_2, x_{n_1}$  och  $y_1, y_2, y_{n_2}$ . Vi har även nollhypotesen  $H_0 : m_X = m_Y$  (väntevärdena för de bågge stickproven).

$$\frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}}}$$

Det finns även en annan formel som gäller då  $s_X \approx s_Y$ :

$$\frac{\bar{x} - \bar{y}}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$s = \sqrt{\frac{(n_1 - 1)s_x^2 + (n_2 - 1)s_y^2}{n_1 - 1 + n_2 - 1}}$$

Den sistnämnda formeln står i formelbladet.

### Exempel

Grupp 1: 120, 122, 124 g. Grupp 2: 123, 126, 127 g.

$$s_x^2 = \frac{1}{n_1 - 1} \left( \sum x_i^2 - \frac{1}{n_1} \left( \sum x_i \right)^2 \right) = \frac{1}{2} \left( \sum x_i^2 - \frac{1}{3} \left( \sum x_i \right)^2 \right) =$$

förenkla genom att ta bort 120

$$\sum x_i^2 = 0^2 + 2^2 + 4^2 = 20$$

$$\sum x_i = 0 + 2 + 4 = 6$$

$$s_x^2 = \frac{1}{2} (20 - \frac{1}{3} 36) = 2$$

$$s_y^2 = \frac{1}{n_2 - 1} \left( \sum y_i^2 - \frac{1}{n_2} \left( \sum y_i \right)^2 \right) = \frac{1}{2} \left( \sum y_i^2 - \frac{1}{3} \left( \sum y_i \right)^2 \right) =$$

förenkla genom att ta bort 123

$$\sum y_i^2 = 0^2 + 3^2 + 4^2 = 25$$

$$\sum y_i = 0 + 3 + 4 = 7$$

$$s_y^2 = \frac{1}{2} (25 - \frac{1}{3} 49) = 4.33$$

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{122 - 125.33}{s\sqrt{\frac{1}{3} + \frac{1}{3}}}$$

$$s = \sqrt{\frac{(n_1 - 1)s_x^2 + (n_2 - 1)s_y^2}{n_1 - 1 + n_2 - 1}} = \sqrt{\frac{2 * 4 + 2 * 4.33}{4}} = 2.04$$

$$t = \frac{-3.33}{2.04\sqrt{\frac{2}{3}}} \approx -1.99$$

Vi jämför detta värde med  $t_{0.0025}$  för att få ett test på 5% nivån med frihetsgraden  $n_1 - 1 + n_2 - 1$ .

$$t \approx -1.99$$

$$t_{0.025}(3 - 1 + 3 - 1) = t_{0.025}(4) = 2.776$$

$$-2.776 \leq -1.99 \leq 2.886$$

Slutsats: värdet ligger inom felsmarginalen. Hypotesen kan ej förkastas.

## Mer generellt om hypoteser

Beslut	$H_0$ är sann	$H_0$ är falsk
Acceptera $H_0$	Inget fel	$Typ\ II\ fel.\ p = \beta$
Förkasta $H_0$	$Typ\ I\ fel.\ p = \alpha$	Inget fel

Man vill ha en liten sannolikhet att man accepterar falskheter. Samtidigt vill man ha en liten sannolikhet att man förkastar sanningen. Den enda lösningen är att exempelvis intervju fler mäniskor. Man vill ha litet värde för både  $\alpha$  och  $\beta$ . Man brukar säga att man gör testet på femprocentsnivå, enprocentsnivå etc. I verkligheten är det vanligt att  $\alpha = 5\%$  och  $\beta = 10\%$  eller  $\beta = 20\%$ . Det är mer viktigt att  $\alpha$  är litet än att  $\beta$  är det. Generellt gäller att ett litet värde för  $\alpha \Rightarrow$  ett större värde för  $\beta$ . Likaså gäller det att ett litet värde på  $\beta \Rightarrow$  ett större värde för  $\alpha$ .

## Massignifikans

Anta att  $\alpha = 0.05$ . Vi gör  $n$  tester. Det finns en risk att upptäckter som beror på slumpen görs. Vad är då sannolikheten att vi får minst en signifikant avvikelse:

$$P(\text{minst en signifikant avvikelse}) = 1 - P(\text{ingen signifikant avvikelse}) = 1 - (1 - \alpha)^n$$

I vårt fall innebär detta:

$$1 - (1 - 0.05)^n = 1 - 0.95^n$$

## Exempel

Låt  $\alpha = 0.05$ . En tabell bildas över olika värden för  $n$ , antalet tester eller riskfaktorer.

$n$	$1 - (1 - \alpha)^n$
1	0.05
2	0.0475
5	0.23
20	0.64

För tjugo olika riskfaktorer gäller det att en av dem är signifikant med en sannolikhet av **64%**, när de egentligen bara utgörs av slumpen. Tänk på alla forskningsartiklar som säger sig hittat kopplingar mellan två saker när det i själva verket är mer troligt att det är slumpen som inverkat.

## Exempel - ett stickprov

$\sigma$  är okänt. Skatta  $s$  ur värdet.

Vikter hos killar: **75, 80, 82, 90.**

$H_0$ : en genomsnittlig kille väger **75kg**.

Testa på nivån **0.05**.

$$t = \frac{\bar{x} - m_0}{s_x \sqrt{n}}$$

$$\bar{x} = \frac{75 + 80 + 82 + 90}{4} = 81.75$$

$$t = \frac{81.75 - 75}{s_x \sqrt{4}}$$

$$s_x^2 = \frac{1}{n_1} \left( \sum x_i^2 - \frac{1}{n} \left( \sum x_i \right)^2 \right)$$

vi subtraherar med 80  $\Rightarrow$  data :  $-5, 0, 2, 10$

$$\sum x_i^2 = (-5)^2 + 0^2 + 2^2 + 10^2 = 129$$

$$\sum x_i = -5 + 0 + 2 + 10 = 7 \Rightarrow \left( \sum x_i \right)^2 = 49$$

$$s_x^2 = \frac{1}{3} \left( 129 - \frac{1}{4} * 49 \right) = 38.92$$

$$s = \sqrt{38.92} \approx 6.24$$

$$t = \frac{6.75}{6.24} * 2 = 2.16$$

$$t_{0.025}(3) = 3.18$$

$$-3.18 \leq 2.16 \leq 3.18$$

Slutsats: hypotesen kan ej förkastas.

## Sammanfattning av hypotestest

$$\text{testfunktion} = \frac{\text{matvärde} - \text{vantevärde}}{\text{standardavvikelse}}$$

Vanligtvis används ett **5%-nivå**. Det vill säga att det är **5%** risk att förkasta en korrekt hypotes. Man kan, om första nivån misslyckas, gå vidare till exempelvis **1%**.

$\alpha = 0.05$  kallas för enstjärning signifikans, \* signifikans  $\alpha = 0.01$  kallas för \*\* signifikans  $\alpha = 0.001$  kallas för \*\*\* signifikans

## Repetition kovarians

$$\rho = \frac{C(X, Y)}{\sqrt{\sigma_X^2 * \sigma_Y^2}}$$

$$\rho^* = r$$

$$C(X, Y) = E((X - m_X)(Y - m_Y)) = E(XY) - E(X) * E(Y)$$

$$\sigma_X^2 = \frac{1}{n-1} \left( \sum x_i^2 - \frac{1}{n} \left( \sum x_i \right)^2 \right)$$

$$\sigma_Y^2 = \frac{1}{n-1} \left( \sum y_i^2 - \frac{1}{n} \left( \sum y_i \right)^2 \right)$$

Notera att uttrycket för standardavvikelsen blir mycket stort. Vid skattning av kovariansen skulle detta resultera i ett allt för stort uttryck. Därför ersätter vi dessa med uttrycken kvadratsumma.

## Kvadratsumma och skattning av kovarians

$$S_{XX} = \sum x_i^2 - \frac{1}{n} \left( \sum x_i \right)^2$$

$$S_{YY} = \sum y_i^2 - \frac{1}{n} \left( \sum y_i \right)^2$$

$$S_{XY} = \sum x_i y_i - \frac{1}{n} \left( \sum x_i \right) \left( \sum y_i \right)$$

$$C^*(X, Y) = \frac{1}{n-1} S_{XY}$$

## Pearsons korrelation

Används när man har siffror till både  $x$  och  $y$ , samt man vill undersöka graden av **linjärt beroende**.

$$\rho^* = r = \frac{\frac{1}{n-1} S_{XY}}{\sqrt{\frac{1}{n-1} S_{XX} * \frac{1}{n-1} S_{YY}}} = \frac{S_{XY}}{\sqrt{S_{XX} * S_{YY}}}$$

Notera att det för  $\rho^*$ , precis som för  $\rho$  gäller att  $-1 \leq \rho^* \leq 1$ .

Finns i formelblad på blad 5,  $r_{pearson}$ .

## Förklaringsgrad

Man kan beräkna andelen av variationen i  $y$  som förklaras med ett rätlinjigt samband mellan  $x$  och  $y$ . Denna andel brukar kallas  $R^2$  (förklaringsgrad / coefficient of determinations). För en rät linje gäller att  $R^2 = r^2$ .

### Exempel 14.1 s.233

Sambandet mellan rader kod i mjukvara och tid som lagts ner på programvaran.  $X$  betecknar LoC,  $Y$  betecknar arbetstid.

$X$	$Y$
800	4
900	5
1050	6
1600	10

$$r = \frac{S_{XY}}{\sqrt{S_{XX} * S_{YY}}}$$

Man tar varje värde för  $X$  och multiplicerar med varje värde för  $Y$ .

$$\begin{aligned} S_{XY} &= 2812.5 \\ S_{XX} &= 381875 \\ S_{YY} &= 20.75 \end{aligned}$$

$$r = \frac{2812.5}{\sqrt{381875 * 20.75}} = 0.999$$

Det är en mycket stark korrelation mellan datapunkterna. En typiskt bra modell.

I detta exemplet är  $R^2 = 0.998$ . Modellen förklrar alltså **99.8%** av variationen i  $y$  med hjälp av variationen i  $x$ .

## Att anpassa en linje till data

"Hur drar man den *bästa* räta linjen?"

Man mäter avståndet i kvadrat mellan linjen och varje datapunkt lodrätt. Sedan minimerar man summan av dessa kvadrater.

$$\begin{aligned} \sum(\text{punktens } y - \text{linjens } y)^2 \\ \iff \\ \sum(y_i - (a + bx_i))^2 \end{aligned}$$

Vilka värden för  $a$  och  $b$  gör att linjen blir så liten som möjligt? Det är alltså minimering i flera variabler.

Generellt sätter man minimum genom att sätta derivatan till noll. Vi ska derivera med avseende på  $a$  och avseende på  $b$ . Båda ska vara noll.

$$\begin{aligned} \sum(y_i - (a + bx_i))^2 &= Q(a, b) \\ \frac{dQ}{da} &= 0 \\ \frac{dQ}{db} &= 0 \end{aligned}$$

Vi behöver inte kontrollera att det är minimum, maximum eller terasspunkt. Det är alltid minimum.

$$\begin{aligned}
\frac{dQ}{da} &= -2 * \sum (y_i - (a + bx_i)) = 0 \Rightarrow \\
\sum y_i &= \sum a + b \sum x_i \Rightarrow \\
a &= \frac{\sum y_i}{n} - b \frac{\sum x_i}{n} \Rightarrow \\
a^* &= \bar{y} - b^* \bar{x} \\
\frac{dQ}{db} &= -2 * \sum ((y_i - (a + bx_i)) * x_i) = 0 \Rightarrow \\
\sum x_i y_i &= a \sum x_i + b \sum x_i^2 = \\
(\frac{\sum y_i}{n} - b \frac{\sum x_i}{n}) \sum x_i &+ b \sum x_i^2 \Rightarrow \\
\underbrace{\sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i}_{S_{XY}} &= b (\underbrace{\sum x_i^2 - \frac{1}{n} (\sum x_i)^2}_{S_{XX}}) \iff \\
S_{XY} &= b * S_{XX} \Rightarrow \\
b^* &= \frac{S_{XY}}{S_{XX}}
\end{aligned}$$

Detta behöver vi inte kunna på tentan. Men viktigt för att förså hur man kommer fram till saker och ting.

## !! Sammanfattning linje-anpassning

(Även "gör en prognos")

- Beräkna skattningen av korrelationen. Beräkna först  $S_{XX}$ ,  $S_{YY}$  och  $S_{XY}$

$$r = \frac{S_{XY}}{\sqrt{S_{XX} * S_{YY}}}$$

- Testa så att korrelationen inte beror på slumpen.

$$t = \frac{r}{\sqrt{1 - r^2}} * \sqrt{n - 2}$$

$H_0$  : korrelationen från punkt 1,  $\rho = 0$

jämför med  $t_{0.025}(n - 2)$

om inte  $-t_{0.025}(n - 2) \leq t \leq t_{0.025}(n - 2)$  så förkastas  $H_0$ . Då finns ett samband och då går vi vidare

- Om vi förkastat  $H_0$ , gå vidare med regressionsanalys

$$y = a^* + b^* x$$

$$b^* = \frac{S_{XY}}{S_{XX}}$$

$$a^* = \bar{y} - b^* \bar{x}$$

- Beräkna förklaringsgraden

$$R^2 = \frac{S_{xy}^2}{S_{XX} * S_{YY}}$$

$R^2$  tecknar andelen av  $y$ -variationen som förklaras av  $x$ .

## Spearmans korrelation

Används i alla andra fall (gentemot Pearsons korrelation). Det vill säga exempelvis då man inte har siffror på alla variabler, eller då man vill undersöka ett ej linjärt samband.

Betecknas  $r_s$ . Skiljer sig på så vis att datan rangordnas.

rang x	rang y
1	2
2	1
...	...
$n$	...

Exempelvis skulle rangordningen av  $x$  vara typer av medicin och rangordningen av  $y$  vara smärtan.

Man sätter sedan in rang  $x$  och rang  $y$  istället för  $x$  och  $y$  i formeln för Pearsons korrelationskoefficient. Då får man Spearmans  $r_s$ . Man får en förenklad uträkning:

$$r_s = 1 - \frac{6 \sum d_i^2}{n^3 - n}$$

där  $d$  är differensen mellan rang  $x$  och rang  $y$ .

Notera att även  $r_s$  ligger i intervallet  $-1 \leq r_s \leq 1$ .

Testfunktionen för Spearmans korrelation är

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

### !! Exempel

Man utsätter studenter för två olika test. Någon misstänker att de i själva verket mäter samma saker och därför är onödiga.

student	test 1	test 2	$d$
Adam	1	3	-2
Bertil	2	4	-2
Calle	3	1	2
Diana	4	2	2

$$\begin{aligned} r_s &= 1 - \frac{6 \sum d_i^2}{n^3 - n} = \\ &1 - \frac{6(4+4+4+4)}{4^3 - 4} = 1 - \frac{96}{60} = -0.6 \end{aligned}$$

### Exempel olinjärt samband

Låt säga att  $y = ae^{bt}$  och Spearmans  $r_s = 1$ . Det kan tyckas konstigt, men denna ekvation är linjär efter transformation!

$$\begin{aligned} y &= ae^{bt} \\ \underbrace{\ln y}_z &= \underbrace{\ln a}_c + bt \\ z &= c + bt \end{aligned}$$

### Exempel olinjärt samband

Skriv följande olinjära uttryck på ett linjärt sätt:  $y = ae^{bx}$ .

$$\begin{aligned} y &= ae^{bx} \\ \underbrace{\ln(y)}_z &= \underbrace{\ln(a)}_c + bx \\ z &= c + bx \\ b^* &= \frac{S_{XZ}}{S_{XX}} \\ c^* &= \bar{z} - b^*\bar{x} \\ \ln a^* &= c^* \\ a^* &= e^{c^*} \end{aligned}$$

### Exempel olinjärt samband

Skriv följande uttryck på linjär form:  $y = a + bx^2$ .

$$\begin{aligned} y &= a + bx^2 \\ y &= a + bz \\ b^* &= \frac{S_{ZY}}{S_{ZZ}} \\ a^* &= \bar{y} - b^*\bar{z} \end{aligned}$$

### Exempel olinjärt samband

Skriv följande uttryck på linjär form:  $y = ax^b$ .

$$\begin{aligned} y &= ax^b \\ \underbrace{\ln(y)}_z &= \underbrace{\ln(a)}_c + \underbrace{\ln(x^b)}_{b*\ln(x)} \\ z &= c + b * \underbrace{\ln(x)}_u \\ b^* &= \frac{S_{UZ}}{S_{UU}} \\ c^* &= \bar{z} - b^*\bar{u} \end{aligned}$$

### !!!! Exempel

$y = a + bx$ . Data:  $x : 0, 3, 4$  och  $y : 2, 4, 5$ .

Beräkna kvadratsummorna

$$S_{XX} = \sum x_i^2 - \frac{1}{n} \left( \sum x_i \right)^2 = 3^2 + 4^2 - \frac{1}{3}(3+4)^2 = 25 - \frac{1}{3}49 = 8.67$$

$$S_{YY} = \sum y_i^2 - \frac{1}{n} \left( \sum y_i \right)^2 = 2^2 + 4^2 + 5^2 - \frac{1}{3}(2+4+5)^2 = 45 - \frac{1}{3}921 = 4.67$$

$$S_{XY} = \sum xy - \frac{1}{n} \left( \sum y * \sum x \right) = 32 - \frac{1}{3}77 = 6.33$$

$$r = \frac{S_{XY}}{\sqrt{S_{XX} * S_{YY}}} = \frac{6.33}{\sqrt{8.67 * 4.67}} = 0.994$$

$$R^2 = 0.994^2 = 0.988$$

$$b^* = \frac{S_{XY}}{S_{XX}}$$

$$a^* = \bar{y} - b\bar{x}$$

$$b^* = \frac{6.33}{8.67} = 0.73$$

$$a^* = 3.67 - 0.73 * 2.33 = 1.97$$

Svar:  $y = 1.97 + 0.73x$

## Exempel - anpassning

Anpassa exponentialkurvan  $y = ae^{bx}$  till följande data:

$x$	$y$	$y = ae^{bx}$	$x$	$z = \ln(x)$
0	1	$\underbrace{\ln(y)}_z = \underbrace{\ln(a)}_c + bx$	0	$\ln(1) = 0$
1	3		1	$\ln(3) = 1.10$
2	6	$z = c + bx$	2	$\ln(6) = 1.79$

$$S_{XX} = \dots = 5 - \frac{1}{3}9 = 5 - 3 = 2$$

$$S_{ZZ} = \dots = 0^2 + 1.1^2 + 1.79^2 - \frac{1}{3}2.89^2 = 1.63$$

$$S_{XZ} = \dots = 4.68 - \frac{1}{3}3 * 2.89 = 1.79$$

$$b^* = \dots = \frac{1.79}{2} = 0.895$$

$$c^* = \bar{z} - b^*\bar{x} = 2.89 - 0.895 * 1 = 0.069$$

$$c^* = \ln(a) \Rightarrow e^{c^*} = a^* = 1.07$$

Svar:  $1.07 * e^{0.895x}$

## Dummyvariabel

Variabel som är 0 om man inte tillhör en viss grupp och 1 om man tillhör gruppen.

### Exempel

1 om man är Karlskronabo, 0 om man inte är Karlskronabo.

$y = a + b * d$ ,  $d = 0$  eller  $d = 1$ .

För Karlskronabor gäller att  $y = a + b * 1 = a + b$ . För ej Karlskronabor gäller  $y = a + b * 0 = a$ .

Koefficienten  $b$  kan tolkas som skillnaden mellan grupperna

## Regression efter transformation

### Exempel

Anpassa kurvan  $y = a + \underbrace{bx^2}_z$  med minsta-kvadratmetoden. Data:

$x$	$y$
0	1
2	5
5	20
6	35

Lösning:

$x$	$y$	$z$
0	0	1
2	4	5
5	25	20
6	36	35

$$y = a + bz$$

$$b^* = \frac{S_{ZY}}{S_{ZZ}}$$

$$a^* = \bar{y} - b^* \bar{z}$$

$$S_{ZZ} = \sum z^2 - \frac{1}{n} \left( \sum z \right)^2 = \underbrace{(0^2 + 4^2 + 25^2 + 36^2)}_{1937} - \frac{1}{4} 65^2 = 880.75$$

$$S_{ZY} = \sum zy - \frac{1}{n} \left( \sum z \sum y \right) = \underbrace{520 + 1260}_{1780} - \frac{1}{4} 65 * 61 = 788.75$$

$$b^* = \frac{788.75}{880.75} \approx 0.90$$

$$a^* = \bar{y} - b^* \bar{z} = 15.25 - 0.90 * 16.25 = 0.63$$

Svar: kurvan är  $y = 0.63 + 0.90x^2$

### Exempel 223 i boken

$$y = a * x^b$$

Lösning:

$$\ln(y) = \ln(a) + b * \ln(x)$$

Approximativ tolkning: "då  $x$  är priset och  $b$  är antal % som försäljningen ökar då priset ökar med 1% (priselasticitet)".

Data:  $x(kr) : 300, 350, 370, 420$   $y(antal) : 150, 120, 110, 95$

$$\underbrace{\ln(y)}_z = \underbrace{\ln(a)}_c + b * \underbrace{\ln(x)}_y$$

$$z = c + b * u$$

$u = \ln x$	$z = \ln y$
5.70	5.01
5.86	4.79
5.91	4.70
6.04	4.55
6.11	4.25

$$S_{UU} = \underbrace{\sum u^2}_{176.14} - \frac{1}{n} \left( \underbrace{\sum u}_{176.06} \right)^2 = 0.08$$

$$S_{UZ} = \sum uz - \frac{1}{n} \left( \sum u \sum z \right) = 137.5 - \frac{1}{5} 29.62 * 23.3 = -0.17$$

$$b^* = \frac{S_{UZ}}{S_{XX}} = \frac{-0.17}{0.08} = -2.125$$

$$c^* = \bar{z} - b^* \bar{y} = 4.66 + 2.125 * 5.93 = 17.25$$

$$a = e^{c^*} = e^{17.25} = 3.1 * 10^7$$

Notera att vi avrundat frikostigt här. Detta gör att svaret diffar på flera decimaler!

Svar:  $y = 3.1 * 10^7 * x^{-2.125}$