Matematisk statistik

(Cornelia Schiebold)

Innehåll:

1. Sannolikhetsteori
2. Diskreta stokastiska variabler
3. Kontinuerliga stokastiska variabler
4. Oberoendemått, summor av stokastiska variabler och centrala gränsvärdessatsen
5. Beskrivande statistik
6. Punktskattning
7. Intervallskattning
8. Hypotesprövning
9. Linjär regression

Kursbok: G. Blom, J. Enger, G. Englund, J. Grandell, L. Holst, *Sannolikhetsteori och statistikteori med tillämpningar*, Studentlitteratur

1 Sannolikhetslära, några grundläggande begrepp

- Resultatet av ett slumpmässigt försök kallas ett utfall och brukar betecknas ω (lilla omega).
- Utfallsrummet Ω (stora omega) är mängden av alla möjliga utfall.
- En händelse A är en delmängd av Ω ($A \subseteq \Omega$), alltså en samling av utfall. Att A inträffar innebär att det utfall som inträffar tillhör A ($\omega \in A$). Händelsen $\{\omega\}$ kallas för en elementär händelse.
- **Mål:** Tillordna varje händelse A i ett utfallsrum Ω en sannolikhet $P(A)$ (ett tal p mellan 0 och 1 där $P(A) = p$ betyder att A inträffar med sannolikhet $p \cdot 100\%$).

Exempel 1: Kast med en tärning

- utfall: antal ögon (och betecknas med 1, 2, 3, 4, 5, 6)
- utfallsrummet $\Omega = \{1, 2, 3, 4, 5, 6\}$ med $|\Omega| = 6$ elementära händelser. De är $\{1\}$, $\{2\}$, $\{3\}$, $\{4\}$, $\{5\}$, $\{6\}$
- exempel på händelser: A: "antal ögon udda", B: "antal ögon högst 2" motsvarar delmängderna $A = \{1, 3, 5\}$, $B = \{1, 2\}$ av utfallsrummet

Exempel 2: Kast med två (olika färgade) tärningar

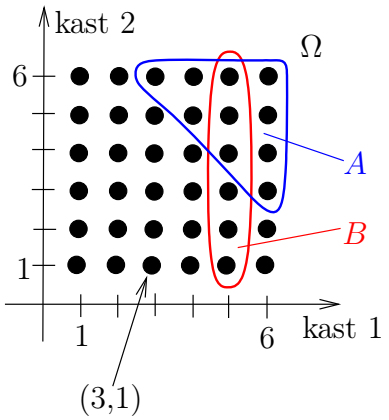
- utfall: ett par (i, j) där i är antal ögon av första tärningen och j är antal ögon av andra tärningen
- utfallsrummet $\Omega = \{(i, j) \mid i, j = 1, \ldots, 6\}$ med $|\Omega| = 6^2$
- exempel på händelser: A: "tärningssumman är större än 8", B: "första kastet är 5"

$A = \{(3,6), (4,6), (5,6), (6,6),$
$(4,5), (5,5), (6,5),$
$(5,4), (6,4),$
$(6,3)\},$

$|A| = 10$

$B = \{(5, j) \mid j = 1, \ldots, 6\},$

$|B| = 6$



Det klassiska sannolikhetsbegreppet: Antag att det bara finns ändligt många möjliga utfall som alla är lika sannolika. Då sätter man

$$P(A) = \frac{|A|}{|\Omega|}$$

i Exempel 1: $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{3}$; **i Exempel 2:** $P(A) = \frac{5}{18}$, $P(B) = \frac{1}{6}$.

Exempel 3:

a) Man kastar en tärning tre gånger och noterar hur oft man fick en sexa. Betrakta händelsen att man har fått lika ofta sexa som ej sexa. Notera att denna händelse inte kan inträffa.

$\Omega = \{0, 1, 2, 3\}$,

$A = \varnothing$ (den tomma mängden), $P(A) = 0$.

Vi återkommer till detta viktiga exempel i avsnittet om binomialfördelningen.

b) Man kastar en tärning 3 gånger och noterar kastet där man har fått sexan för första gången. Händelsen man är intresserad i är att man inte får en sexa alls. Vi betecknar denna elementära händelsen med $\{\infty\}$.

$\Omega = \{1, 2, 3\} \cup \{\infty\}$,

Observera att de elementära händelserna inte är lika sannolika:

$$P(\{1\}) = \tfrac{1}{6}, \quad P(\{2\}) = \left(1 - \tfrac{1}{6}\right)\tfrac{1}{6}, \quad P(\{3\}) = \left(1 - \tfrac{1}{6}\right)^2\tfrac{1}{6}, \quad P(\{\infty\}) = \left(1 - \tfrac{1}{6}\right)^3.$$

c) Samma spel som i **b)** men man kastar tärningen N gånger så att $\Omega = \{1, \ldots, N\} \cup \{\infty\}$. Likandant som i **b)** får man

$$P(A) = (1 - \tfrac{1}{6})^N.$$

Observera att $P(A)$ sträver mot 0 om N går mot oändligheten.

d) Man kastar en tärning tills sexa erhålls för första gången och noterar nummret av detta kast. Igen är man intresserad i händelsen att man inte får sexa alls.

$\Omega = \mathbb{N} \cup \{\infty\}$,

$A = \{\infty\}$, $P(A) = 0$.

Händelsen har sannolikhet 0 trots att den inte är omöjlig.

e) Man plockar ett tal mellan noll och ett. Händelsen man är intresserad av är att det plockade talet är större än -1.

$\Omega = [0, 1]$,

$A = [0, 1]$ (hela utfallsrummet), $P(A) = 1$.

Anmärkning: Man skiljer mellan diskreta (Exempel 3 a)-c)) och kontinuerliga (Exempel 3 d)) utfallsrum. Utfallsrummen i Exemplen 3 a), b) kallas även ändliga.

1.1 Mängdlära

En mängd är en väldefinierad samling av element. Om elementet ω tillhör mängden A skriver vi $\omega \in A$.

Anmärkning: Observera att en mängd inte förändras om man

- byter elementens ordningsföljd **Ex:** $\{a, b\} = \{b, a\}$
- upprepar element **Ex:** $\{a, a\} = \{a\}$

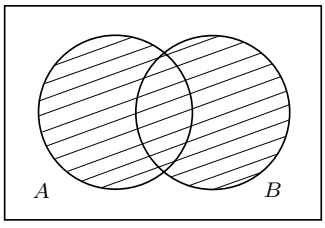
Den tomma mängden betecknas med $\varnothing$.

Låt A, B vara mängder. Vi säger att A och B stämmer överens ($A = B$) om de innehåller samma element. A är en delmängd av B ($A \subseteq B$) om varje element i A också ligger i B. A är en äkta delmängd av B ($A \subset B$) om $A \subseteq B$ och $A \neq B$.

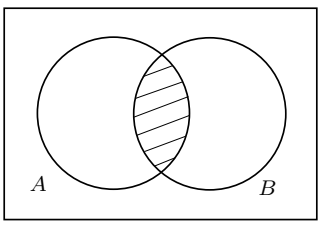
Operationer för mängder: För mängder A, B definieras

- $A \cup B = \{\omega \mid x \in A \text{ eller } \omega \in B\}$, föreningen av A och B
- $A \cap B = \{\omega \mid x \in A \text{ och } \omega \in B\}$, snittet av A och B
- $A \backslash B = \{\omega \in A \mid \omega \notin B\}$, differensmängden A utan B

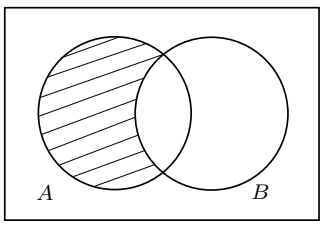
Illustration genom Venndiagram:



$A \cup B$



$A \cap B$



$A \backslash B$

Om man betraktar alla mängder som delmängder i en gemensam maximal mängd Ω, så definieras komplementet till en mängd A (inom Ω) genom

$$A^\star = \Omega \backslash A = \{\omega \mid \omega \notin A\}$$

Observera att $A \cap A^\star = \varnothing$ och $A \cup A^\star = \Omega$.

De Morgans lag:

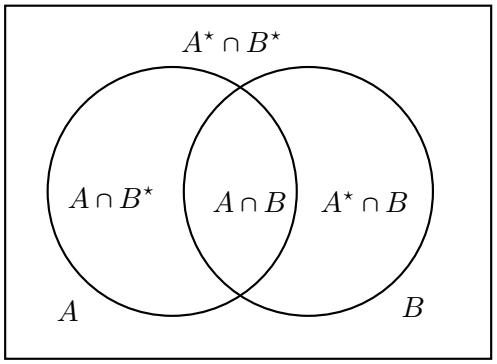
- $(A \cup B)^\star = A^\star \cap B^\star$,
- $(A \cap B)^\star = A^\star \cup B^\star$

Två mängder A, B kallas för disjunkta (åtskilda) om $A \cap B = \varnothing$. Notera att $(A \cap B)$, $(A \cap B^\star)$ delar A i två disjunkta delmängder. Vi skriver

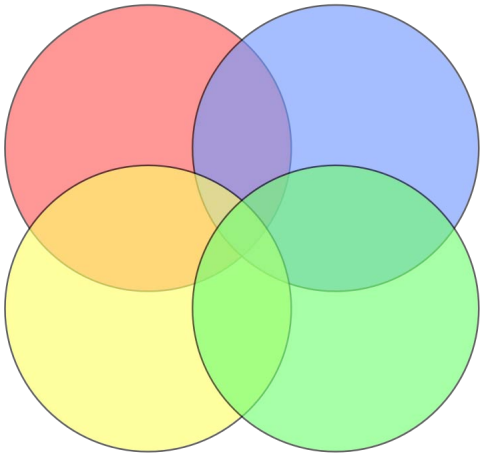
$$A = (A \cap B) \uplus (A \cap B^\star)$$

Observera $A \backslash B = A \cap B^\star$

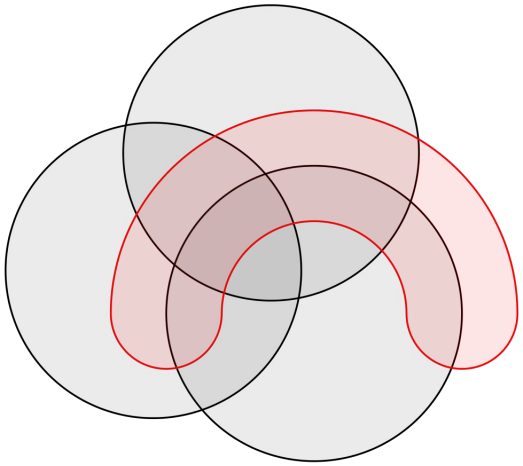
Hur delar man upp Ω i disjunkta mängder m.a.p. två givna delmängder A, B?
Vi har $\Omega = (A \cap B) \uplus (A^\star \cap B) \uplus (A \cap B^\star) \uplus (A^\star \cap B^\star)$.



Observera att Venndiagram för fler än fyra mängder blir mer komplicerade, se
en.wikipedia.org/wiki/Venn_diagram



Ej ett giltigt Venndiagram!



En möjlighet att rita ett Venndiagram för 4 mängder

1.2 Sannolikheter

En sannolikhetsmått P tillordnar varje händelse A ett tal $P(A)$ med $0 \leq P(A) \leq 1$ sådant att följande egenskaper gäller:

(1) $P(\Omega) = 1$,

(2) $P(A_1 \cup A_2 \cup \ldots) = P(A_1) + P(A_2) + \ldots$ för parvis oförenliga händelser $A_1, A_2, \ldots$ (observera att oförenliga händelser motsvarar disjunkta mängder).

Vidare egenskaper (med bevis)

(3) $P(A^\star) = 1 - P(A)$

(4) $P(\varnothing) = 0$

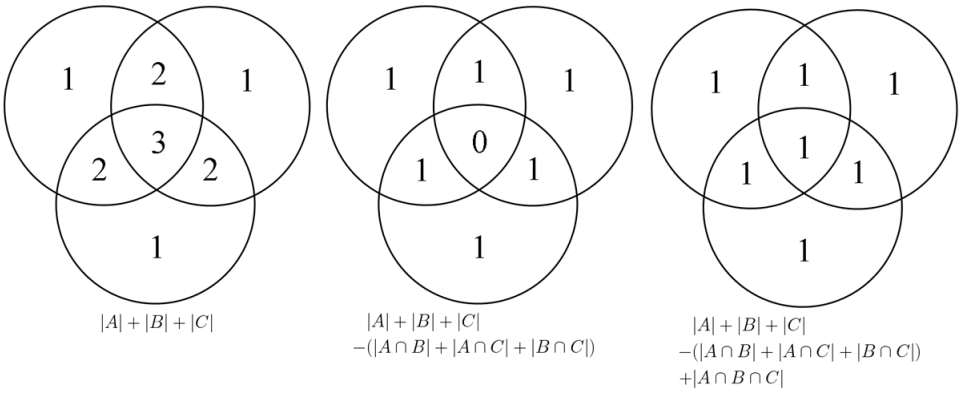
(5) $A \subseteq B \Rightarrow P(A) \leq P(B)$

Additionssats: $\boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$ (med bevis)

Exempel 4 (tas upp i föreläsningen).

Sannolikheten för en förening av tre mängder:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$



(Principen om inklusion/exklusion)

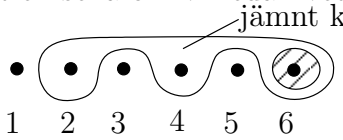
1.3 Betingad sannolikhet

Exempel 1 (forts.): Kast med en tärning

Hur stor är sannolikheten att ha kastat en sexa om vi redan vet att vi har kastat

a) ett udda tal? 0

b) ett jämnt tal? $\frac{1}{3}$



Låt $P(B) > 0$. Då definieras den betingade sannolikheten $P(A|B)$ för A givet att B har inträffat genom

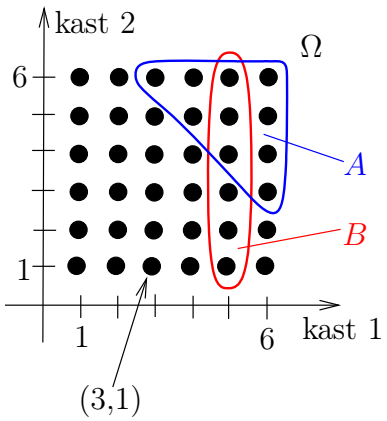
$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

Observera definitionen medför att $P(A|B) + P(A^\star|B) = 1$.

Följdsats: Om $P(A) > 0$ och $P(B) > 0$ då gäller $\boxed{P(A|B)P(B) = P(B|A)P(A)}$

Exempel 2 (forts.):

Hur stor är sannolikheten att tärningssumman är större än 8 (händelse A) om första kastet är femma (händelse B)?

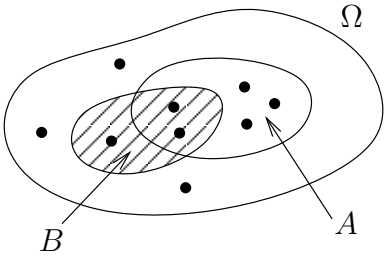


$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{3/36}{6/36} = \frac{1}{2}.$$

I allmänheten gäller för ett ändligt utfallsrum Ω där man antar att alla elementära händelser är lika sannolika att

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

förutsatt att $B \neq \varnothing$.



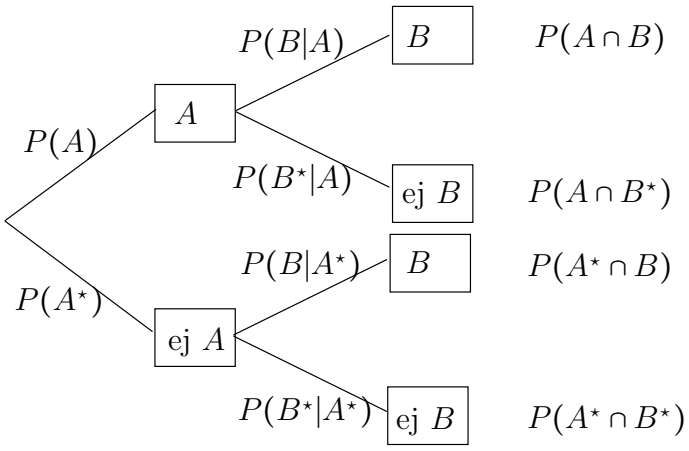
Exempel 5 (tas upp i föreläsningen).

Lagen om total sannolikhet: Om $\Omega = B_1 \cup B_2 \cup \ldots \cup B_N$ där händelserna B_j har positiv sannolikhet och är parvis oförenliga, gäller för varje händelse A

$$P(A) = \sum_j P(A|B_j)P(B_j).$$

Anmärkning: Man kan visa att $P(\cdot|B)$ är ett sannolikhetsmått på utfallsrummet B.

Träddiagram: Låt A, B vara händelser med $P(A)$, $P(B) > 0$.



Exempel: Monty Hall-problemet, se http://sv.wikipedia.org/wiki/Monty_Hall-problemet

1.4 Oberoende händelser

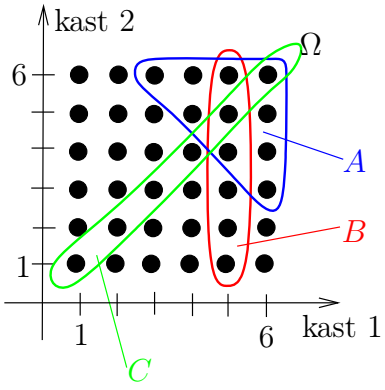
Om $P(A \cap B) = P(A)P(B)$ så kallas A, B för oberoende händelser.

Observation 1: Om A, B är oberoende händelser med $P(B) > 0$, så är $P(A|B) = P(A)$, d.v.s. vetandet om B har inträffat spelar ingen rol för A:s sannolikhet.

Exempel 2 (forts.):

A, B är inte oberoende.

Låt C vara händelsen att första och andra kastet är lika.



Då är B, C oberoende
eftersom

$$P(B \cap C) = P(\{(5,5)\}) = \frac{1}{36} = \frac{6}{36} \cdot \frac{6}{36} = P(B)P(C).$$

Observation 2: Om A, B är oberoende händelser, så är A, $B^\star$ också oberoende händelser eftersom

$$P(A) = \underbrace{P(A \cap B)}_{=P(A)P(B)} + P(A \cap B^\star) \implies P(A \cap B^\star) = P(A)\big(1 - P(B)\big) = P(A)P(B^\star)$$

Anmärkning: Om A och B är oberoende (med positiv sannolikhet) så ersätts $P(B|A)$ med $P(B)$, $P(B^\star|A)$ med $P(B^\star)$ o.s.v. i träddiagrammet.

Exempel 4 (forts) (tas upp i föreläsningen).

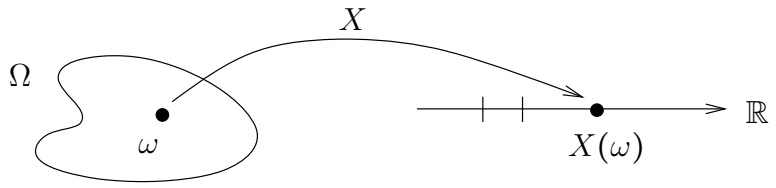
Observation 3: Två oförenliga händelser A, B med $P(A)$, $P(B) > 0$ kan inte vara oberoende eftersom

$$P(A)P(B) > 0 = P(\varnothing) = P(A \cap B).$$

Anmärkning: Definitionen av oberoende händelser kan utvidgas till fler än två händelser, se [Blom et al.], Definition 2.8, för detaljer.

2 Diskreta stokastiska variabler

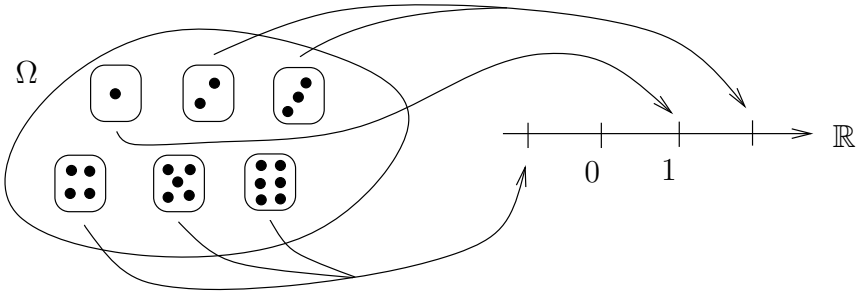
En stokastisk variabel (s.v.) X är en funktion definierad på ett utfallsrum Ω med värden i de reella talen $\mathbb{R}$.



I Kapitel 2 antar vi att X bara kan anta uppräkneligt många värden $\ldots < x_{-2} < x_{-1} < x_0 < x_1 < x_2 < \ldots$. En sådan s.v. X kallas för diskret.

Observera: $\sum_j P(X = x_j) = 1$

Exempel 1: Vid ett tärningskast får man 1 kr om ettan kommer upp, 2 kr om tvåan eller trean kommer upp, annars måste man betala 1 kr.



ω						
$X(\omega)$	1	2	2	−1	−1	−1

$$P(X = -1) = \tfrac{1}{2}, \qquad P(X = 1) = \tfrac{1}{6}, \qquad P(X = 2) = \tfrac{1}{3}.$$

Funktionen

$$p_X(x) = \begin{cases} P(X = x), & x = \ldots x_{-1}, x_0, x_1, \ldots, \\ 0, & \text{för övrigt.} \end{cases}$$

kallas sannolikhetsfunktionen till X.

Anmärkning;

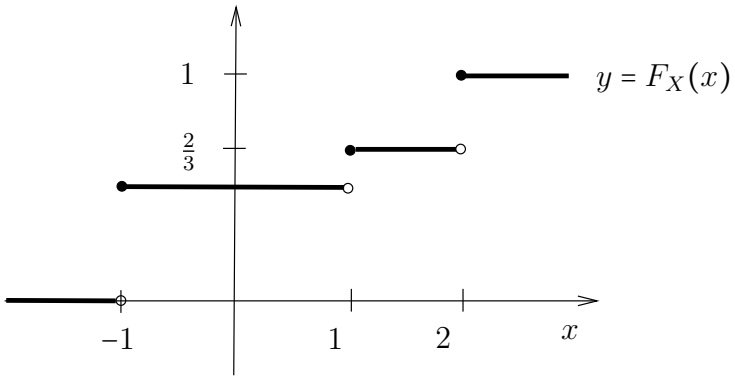
a) En stokastisk variabel modellerar hur man evaluerar ett experiment genom talvärden.

b) Från och med nu intresserar vi oss inte längre för experimentet (utfallsrummet Ω) men bara för talvärden som antas av den stokastiska variablen.

Definition: Funktionen $F_X : \mathbb{R} \to [0,1]$ given genom $F_X(x) = P(X \leq x)$ kallas för fördelingsfunktionen för X.

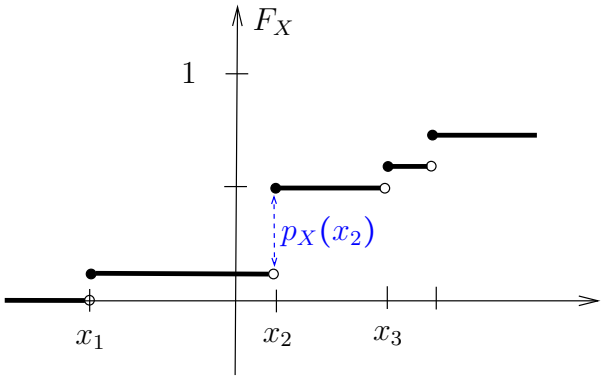
Exempel 1 (forts.):

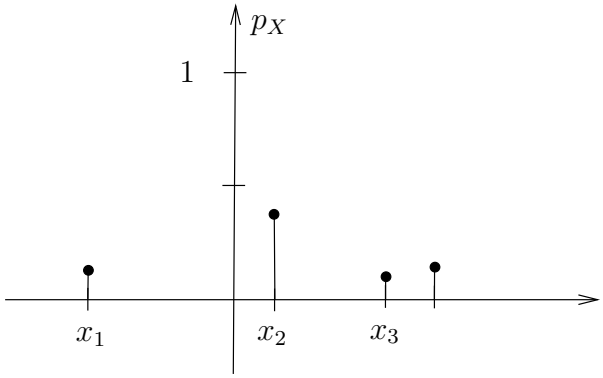
$$F_X(x) = \begin{cases} 0, & x < -1, \\ \frac{1}{2}, & -1 \leq x < 1, \\ \frac{2}{3}, & 1 \leq x < 2, \\ 1, & 2 \leq x. \end{cases}$$



Sats: $p_X(x_n) = F_X(x_n) - F_X(x_{n-1})$

Följdsats: $F_X(x_n) = \sum_{j \leq n} P(X = x_j)$





Fördelningsfunktionens egenskaper:

- Fördelningsfunktionen är en (icke avtagande) trappstegsfunktion med språngställena precis i $\ldots x_{-1}, x_0, x_1, \ldots$.
- $0 \le F_X(x) \le 1$ för alla x. Om vi har ändligt många värden $x_1, \ldots, x_n$ så gäller $F_X(x) = 1$ för $x \ge x_n$ och $F_X(x) = 0$ för $x < x_1$.

Observera att sannolikheter för godtyckliga händelser kan uttryckas genom fördelningsfunktionen.

Exempel 1 (forts.): Betrakta händelserna A: "man förlorar pengar", B: "man vinner minst 1 kr" och C: "man vinner mer än 1 kr".

$$\begin{aligned} P(A) &= P(X = -1) = F_X(-1) \\ P(B) &= P(X = 1 \text{ eller } X = 2) = 1 - P(X = -1) = 1 - F_X(-1) \\ P(C) &= P(X = 2) = F_X(2) - F_X(1) \end{aligned}$$

2.1 Läges- och spridningsmått

2.1.1 Lägesmått

Väntevärdet: $\mu := E(X) = \sum_j x_j P(X = x_j)$

2.1.2 Spridningsmått

Variansen: $\sigma^2 := V(X) = E\big((X-\mu)^2\big) = \sum_j (x_j - \mu)^2 P(X = x_j) \ge 0$

Standardavvikelsen: $\sigma = D(X) = \sqrt{V(X)}$

Sats: $V(X) = E(X^2) - \big(E(X)\big)^2$

Bevis: Med förkortningen $p_j = P(X = x_j)$ får vi

$$V(X) = \sum_j (x_j - \mu)^2 \, p_j = \underbrace{\sum_j x_j^2 \, p_j}_{=E(X^2)} - 2\mu \underbrace{\sum_j x_j \, p_j}_{=\mu} + \mu^2 \underbrace{\sum_j p_j}_{=1} = E(X^2) - \mu^2.$$

Exempel 1 (forts.):

$$E(X) = (-1)\frac{1}{2} + (1)\frac{1}{6} + (2)\frac{1}{3} = \frac{1}{3}$$
$$V(X) = (-1-\frac{1}{3})^2\frac{1}{2} + (1-\frac{1}{3})^2\frac{1}{6} + (2-\frac{1}{3})^2\frac{1}{3} = 1,\overline{8} \quad \Longrightarrow \quad D(X) = \sqrt{1,\overline{8}} \approx 1,37$$

Exempel 2 (tas upp i föreläsningen).

2.2 Några viktiga diskreta fördelningar

2.2.1 Den likformiga fördelningen över $\{x_1, \ldots, x_N\}$

- utfall $x_1 < \ldots < x_N$
- alla utfall är lika sannolika: $P(X = x_j) = \dfrac{1}{N}$

Observera:

$$E(X) = \frac{1}{N}\sum_{j=1}^{N} x_j \qquad V(X) = \frac{1}{N}\sum_{j=1}^{N}(x_j - \mu)^2$$

2.2.2 Binomialfördelningen

Betrakta N oberoende upprepningar av ett försök i vilket en viss händelse A inträffar med sannolikhet $P(A) = p$. Låt

$$X = \text{antalet gånger } A \text{ har inträffat}$$

Sats: $P(X = j) = \binom{N}{j} p^j (1-p)^{N-j}$ där $j = 0, 1, \ldots, N$.

Vi säger att X är binomialfördelad med parametrar N och p och skriver $X \in \text{Bin}(N, p)$. Man kan visa:

$$E(X) = Np, \qquad V(X) = Np(1-p).$$

Exempel 3 (tas upp i föreläsningen).

Värdena $P(X \leq x)$ finns tabellerade för $p \leq \frac{1}{2}$. I fallet $p > \frac{1}{2}$ använder vi

Sats: Låt $X \in \text{Bin}(N, 1-p)$ och $Y \in \text{Bin}(N, p)$. Då gäller $\boxed{P(X \leq j) = P(Y \geq N-j)}$.

Bevis:

$$\begin{aligned} P(X \leq j) &= \sum_{k=0}^{j} \binom{N}{k} (1-p)^k (1-(1-p))^{N-k} \\ &\overset{\kappa=N-k}{=} \sum_{\kappa=N-j}^{N} \binom{N}{N-\kappa} (1-p)^{N-\kappa} p^{\kappa} \\ &= \sum_{\kappa=N-j}^{N} \binom{N}{\kappa} p^{\kappa} (1-p)^{N-\kappa} \\ &= P(Y \geq N-j) \end{aligned}$$

Exempel 4: Bestäm $P(X \leq 3)$ om $X \in \text{Bin}(10; 0,8)$.

Låt $Y \in \text{Bin}(10; 0,2)$. Då gäller

$$P(X \leq 3) = P(Y \geq 10-3) = 1 - P(Y < 7) = 1 - P(Y \leq 6).$$

Från tabellen får vi $P(Y \leq 6) = 0,9991$, alltså $P(X \leq 3) = 0,0009$.

2.2.3 Poissonfördelningen

Låt $\lambda > 0$. En stokastisk variabel X med sannolikhetsfunktion

$$P(X = j) = e^{-\lambda} \frac{\lambda^j}{j!}, \qquad j = 0, 1, 2, \ldots,$$

kallas poissonfördelad med parameter λ och vi skriver $X \in \text{Po}(\lambda)$.

Sats: $E(X) = V(X) = \lambda$.

Bevis:

$$\begin{aligned} E(X) &= \sum_{j \geq 0} j P(X = j) = e^{-\lambda} \sum_{j \geq 0} j \frac{\lambda^j}{j!} = e^{-\lambda} \sum_{j \geq 1} j \frac{\lambda^j}{j!} \\ &= \lambda e^{-\lambda} \sum_{j \geq 1} \frac{\lambda^{j-1}}{(j-1)!} = \lambda e^{-\lambda} \sum_{k \geq 0} \frac{\lambda^k}{k!} = \lambda e^{-\lambda} e^{\lambda} \\ &= \lambda. \end{aligned}$$

Likadant visar man $E(X(X-1)) = \lambda^2$. Sen följer

$$V(X) = E(X^2) - \lambda^2 = E(X(X-1) + E(X) - \lambda^2 = \lambda.$$

Anmärkning:

(1) Förekomst: händelser A som inträffar slumpmässigt och oberoende av varandra i tiden, X betyder antalet händelser A som inträffar under ett visst tidsintervall av given längd.

Exempel: radioaktiv sönderfall, inkommande anrop till en telefonväxel.

(2) Poissonfördelning som approximation för binomialfördelningen: Man kan visa att i gränsfallet, då $p \to 0$ och $N \to \infty$ går mot oändligheten, under det att $Np = \lambda$ är fast,

$$\binom{N}{j} p^j (1-p)^{N-j} \to \frac{\lambda^j}{j!}.$$

För $X \in \text{Bin}(N, p)$ gäller alltså att X är approximativt Poissonfördelad med parameter Np.

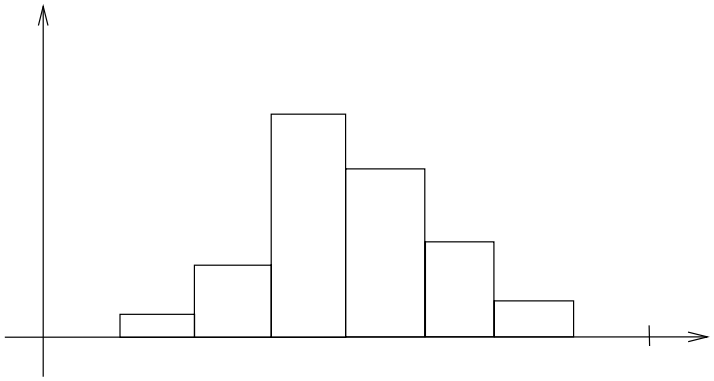
Exempel 5 (tas upp i föreläsningen).

3 Kontinuerliga stokastiska variabler

En kontinuerlig stokastisk variabel X kan anta alla värden i ett (eventuellt oändligt) reellt intervall.

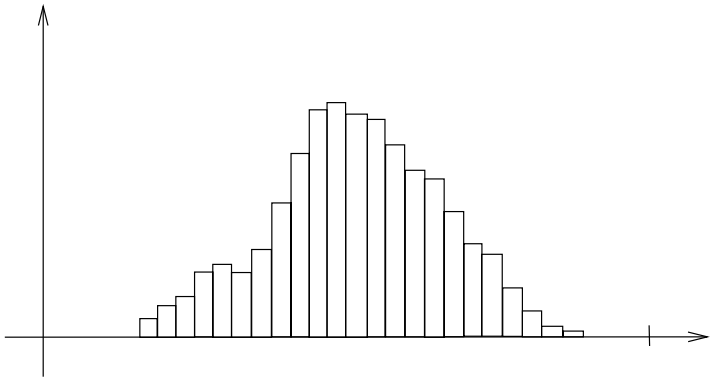
Tankeexperiment:

Antag att vi har genomfört ett visst experiment 100 gånger och att vi sedan har klassindelat vårt material. Då kan det tänkas att vi fick följande histogram:



Ett även bättre intryck av situationen får vi om vi gör

- flera experiment
- finare klassindelning



Rita alla histogram sådana att deras area är lika med 1. Då motsvarar arean som ligger över ett intervall den relativa frekvensen för motsvarande klassen.

Man kan hoppas att man får en kontinuerlig funktion i gränsfallet.

Vi säger att en funktion f är en täthetsfunktion om den uppfyller följande villkor:

(1) $f(x) \geq 0$ för alla x,

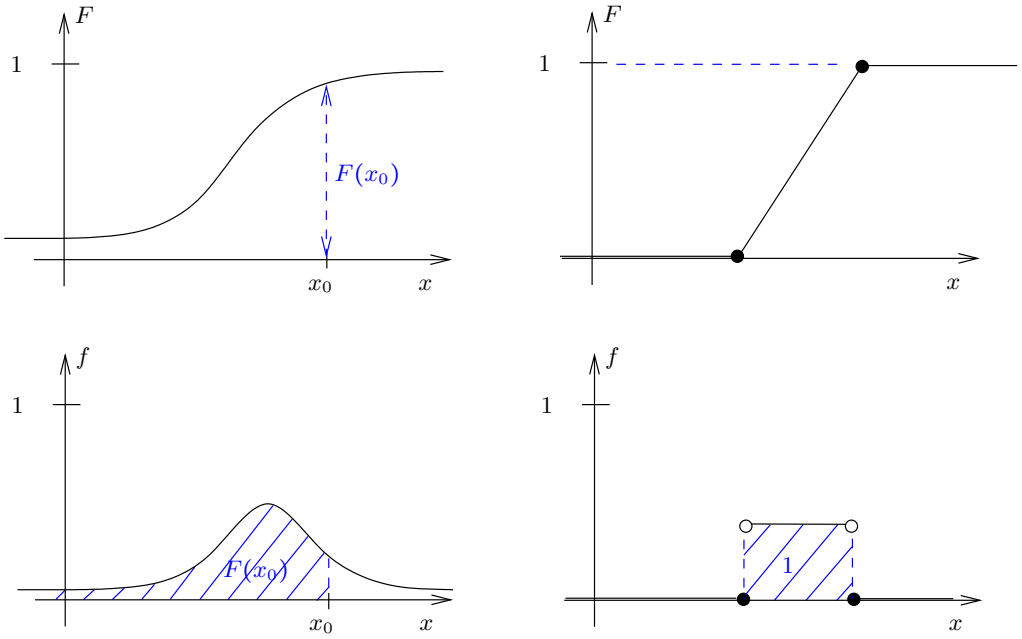
(2) $\int_{-\infty}^{\infty} f(t)\ dt = 1$.

I fortsättningen betraktar vi bara stokastiska variabler X med täthetsfunktioner f som är kontinuerliga utom i högst ändligt många språngställen.

Fördelningsfunktionen $F_X : \mathbb{R} \to [0,1]$ för X är given genom $F_X(x) = P(X \leq x)$.

Sats: $\boxed{F_X(x) = \int_{-\infty}^{x} f(t)\ dt}$

Exempel:



Exempel 1: Den s.v. X har täthetsfunktionen $f(x) = \begin{cases} kx^2, & 0 < x < 1, \\ 0 & \text{för övrigt.} \end{cases}$.

a) Bestäm konstanten k.

b) Bestäm fördelningsfunktionen till X.

Lösning: a) $1 \stackrel{!}{=} \int_{-\infty}^{\infty} f(t)\,dt = \int_0^1 kt^2\,dt = k\left[\frac{t^3}{3}\right]_0^1 = \frac{k}{3} \implies k = 3$

b) $F_X(x) = \int_{-\infty}^{x} f(t)\,dt = \begin{cases} 0, & x \le 0, \\ \int_0^x 3t^2\,dt = x^3, & 0 < x < 1, \\ 1, & 1 \le x. \end{cases}$

Fördelningsfunktionens egenskaper:

- F är kontinuerlig. F är även deriverbar utom möjligtvis i språngställena av f och det gäller att

$$F' = f.$$

- F är icke avtagande (d.v.s. $x_1 < x_2 \implies F(x_1) \le F(x_2)$) med $\lim_{x\to-\infty} F(x) = 0$ och $\lim_{x\to\infty} F(x) = 1$.

Man kan använda fördelningsfunktionen för att bestämma sannolikheten av "förnunftiga" händelser, t.ex.

$$\begin{aligned} P(X > a) &= 1 - P(X \le a) = 1 - F(a), \\ P(a < X \le b) &= P(X \le b) - P(X \le a) = F(b) - F(a). \end{aligned}$$

Observera att $P(X = a) = 0$. Detta medför bland annat att $P(X \le a) = P(X < a)$.

Exempel 1 (forts.): $P(X < \frac{1}{2}) = P(X \le \frac{1}{2}) - P(X = \frac{1}{2}) = F_X(\frac{1}{2}) - 0 = \frac{1}{8}$.

3.1 Läges- och spridningsmått

3.1.1 Lägesmått

Väntevärdet: $\mu := E(X) = \int_{-\infty}^{\infty} tf(t)\,dt$

Medianen är det värde m som uppfyller $F_X(m) = \frac{1}{2}$
(om detta värde är entydigt bestämmt).

3.1.2 Spridningsmått

Variansen: $\sigma^2 := V(X) = E\big((X-\mu)^2\big) = \int_{-\infty}^{\infty}(t-\mu)^2 f(t)\ dt \geq 0$

Standardavvikelsen: $\sigma = D(X) = \sqrt{V(X)}$

Sats: $V(X) = E(X^2) - \big(E(X)\big)^2$

Bevis:

$$\begin{aligned} V(X) &= \int_{-\infty}^{\infty}(t-\mu)^2 f(t)dt \\ &= \underbrace{\int_{-\infty}^{\infty} t^2 f(t)dt}_{=E(X^2)} - 2\mu \underbrace{\int_{-\infty}^{\infty} t f(t)dt}_{=E(X)=\mu} + \mu^2 \underbrace{\int_{-\infty}^{\infty} f(t)dt}_{=1} \\ &= E(X^2) - \mu^2 \end{aligned}$$

Exempel 1 (forts.):

$$E(X) = \int_{-\infty}^{\infty} t f(t)\ dt = 3\int_0^1 t^3\ dt = \frac{3}{4},$$

$$E(X^2) = \int_{-\infty}^{\infty} t^2 f(t)\ dt = 3\int_0^1 t^4\ dt = \frac{3}{5} \Longrightarrow V(X) = \frac{3}{5} - \left(\frac{3}{4}\right)^2 = \frac{27}{80},$$

För att hitta medianen m löser vi ekvationen $F_X(m) = \frac{1}{2}$. Från F_X:s graf vet vi att $0 < m < 1$. Ekvationen $m^3 = \frac{1}{2}$ har lösningen $m = \frac{1}{\sqrt[3]{2}} \approx 0,79$.

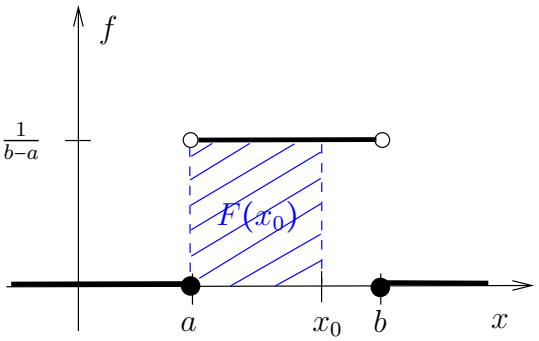
3.2 Några viktiga kontinuerliga fördelningar

3.2.1 Likformig fördelning på intervallet (a,b) (där $a < b$)

Om den s.v. X har täthetsfunktionen

$$f(x) = \begin{cases} \dfrac{1}{b-a}, & a < x < b, \\ 0, & \text{för övrigt.} \end{cases}$$

säges X vara likformigt fördelad på intervallet (a,b) och vi skriver $X \in \mathrm{U}(a,b)$.



Fördelningsfunktionen för X är $F_X(x) = \begin{cases} 0, & x \le a, \\ \frac{x-a}{b-a}, & a < x < b, \\ 1, & b \le x. \end{cases}$

Observera:

$$E(X) = \frac{a+b}{2}, \qquad V(X) = \frac{(a-b)^2}{12}$$

Exempel 2 (tas upp i föreläsningen).

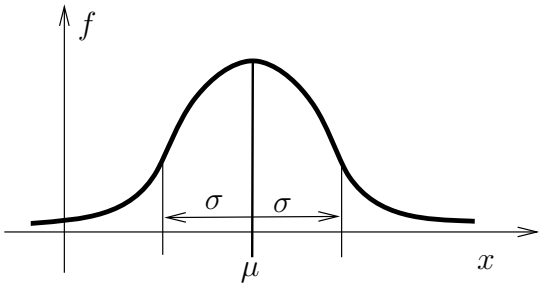
3.2.2 Normalfördelningen med parametrar $\mu \in \mathbb{R}$ och $\sigma > 0$

Om den s.v. X har täthetsfunktionen

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\Big(-\frac{(x-\mu)^2}{2\sigma^2}\Big), \quad x \in \mathbb{R}.$$

säges X vara normalfördelad med parametrar μ, σ och vi skriver $X \in \mathrm{N}(\mu, \sigma)$.

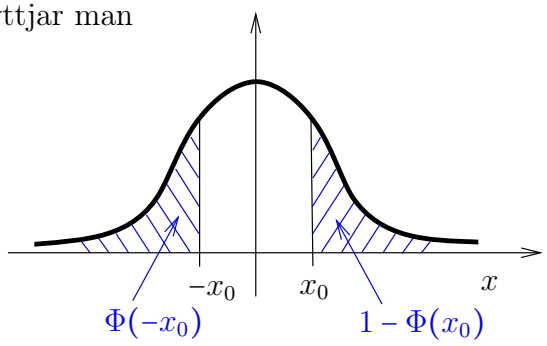
f är symmetrisk kring μ
och f:s spridning ökar med σ.



Problem: f har ingen elementär primitiv funktion!

I fallet $\mu = 0$ och $\sigma = 1$ brukar man beteckna fördelningsfunktionen med Φ. Värdena till Φ finns tabellerade för $x \ge 0$. För $x < 0$ utnyttjar man

$$\Phi(-x) = 1 - \Phi(x).$$



Exempel 3: Låt $X \in \mathrm{N}(0,1)$. Bestäm

$$\textbf{a)}\ P(X \le \tfrac{1}{4}), \qquad \textbf{b)}\ P(X \le -\tfrac{1}{4}), \qquad \textbf{c)}\ P(X > -\tfrac{1}{4}).$$

Lösning:

a) $P(X \le \frac{1}{4}) = \Phi(\frac{1}{4}) \overset{\text{tabell}}{=} 0,5987$

b) $P(X \le -\frac{1}{4}) = \Phi(-\frac{1}{4}) = 1 - \Phi(\frac{1}{4}) \overset{\textbf{a)}}{=} 1 - 0,5987 = 0,4013$

c) $P(X > -\frac{1}{4}) = 1 - P(X \le -\frac{1}{4}) = 1 - \Phi(-\frac{1}{4}) = 1 - (1 - \Phi(\frac{1}{4})) = \Phi(\frac{1}{4}) = 0,5987$

Vad gör man om $X \in \mathrm{N}(\mu, \sigma)$? Då använder man att $\boxed{P(X \le x) = \Phi(\frac{x-\mu}{\sigma})}$.

Denna relation kan man inse på följande sätt:

$$P(X \le x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}}\, dt \overset{(\star)}{=} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{x-\mu}{\sigma}} e^{-\frac{s^2}{2}}\, ds = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

där i $(\star)$ genomförs substitutionen $s = \frac{t-\mu}{\sigma}$.

Exempel 4: Låt $X \in \mathrm{N}(1,2)$. Bestäm

$$\textbf{a)}\ P(X \le 3), \qquad \textbf{b)}\ P(|X-1| < 2).$$

Lösning:

a) $P(X \le 3) = \Phi(\frac{3-1}{2}) = \Phi(1) \overset{\text{tabell}}{=} 0,8413$

b) Observera $|X-1| < 2 \iff -2 < X - 1 < 2 \iff -1 < X < 3$. Alltså

$$P(|X-1| \le 2) = P(-1 < X < 3) = P(X < 3) - P(X \le -1) = P(X \le 3) - P(X \le -1)$$
$$= \Phi(\tfrac{3-1}{2}) - \Phi(\tfrac{-1-1}{2}) = \Phi(1) - \Phi(-1) = \Phi(1) - (1 - \Phi(1)) = 2\Phi(1) - 1 \overset{\textbf{a)}}{=} 0,6826$$

Man kan visa:

$$E(X) = \mu, \qquad V(X) = \sigma^2.$$

3.2.3 Exponentialfördelningen med parameter $\lambda > 0$

Om den s.v. X har täthetsfunktionen

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0, \\ 0, & x < 0. \end{cases}$$

sägs X vara exponentialfördelad med parameter λ och vi skriver $X \in \text{Exp}(\lambda)$.

Fördelningsfunktion för X är $F(x) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0, \\ 0, & x < 0. \end{cases}$.

Man visar:

$$E(X) = \frac{1}{\lambda}, \quad V(X) = \frac{1}{\lambda^2}.$$

Typiska exempel: tiden mellan händelser som inträffar slumpmässigt och oberoende av varandra (sönderfall i ett radioaktivt preparat, inkommande anrop till en telefonväxel), livslängder av elektroniska komponenter

Exempel 5 (tas upp i föreläsningen).

Viktig exempel: Låt ξ vara livslängden hos en elektronisk komponent. Antag att $\xi \in \text{Exp}(\lambda)$. Vi vet att komponenten har fungerad i x timmar. Vad är då sannolikheten att den fungerar ytterligare y timmar?

$$P(\xi > x + y \mid \xi > x) = \frac{P(\xi > x + y \text{ och } \xi > x)}{P(\xi > x)}$$
$$= \frac{P(\xi > x + y)}{P(\xi > x)} = \frac{e^{-\lambda(x+y)}}{e^{-\lambda x}} = e^{-\lambda y} = P(\xi > y).$$

Livslängden är oberoende av hur länge komponenten redan har fungerad!

4 Oberoendemått, summor av stokastiska variabler och centrala gränsvärdessatsen

4.1 Funktioner av en stokastisk variabel

Ett inledande exempel: Observera att den s.v. X från Exempel 1 i Kapitel 2 kan uppfattas som en funktion av den s.v. X_0 som modellerar en tärningskast (d.v.s. som är likformig fördelad över $\{1,2,3,4,5,6\}$)!

Låt X vara en stokastisk variabel med känd fördelningsfunktion F_X och $g : \mathbb{R} \to \mathbb{R}$ en funktion. Betrakta den s.v. $Y = g(X)$. Vad kan vi säger on Y:s fördelning?

Exempel 1: $Y = aX + b$ med $a > 0$. Vi får

$$F_Y(y) = P(Y \le y) = P(aX + b \le y) = P(X \le \frac{y-b}{a}) = F_X(\frac{y-b}{a})$$

Täthetsfunktionen erhålls genom derivering: $f_Y(y) = -\frac{1}{a} f_X(\frac{y-b}{a})$. I punkterna där fördelningsfunktionen ej är deriverbar kan man sätta täthetsfunktionen t.ex. lika med 0.

Exempel 2: Låt $X \in \mathrm{U}(0,1)$ vara en s.v. som antar bara positiva värden. Betrakta den logaritmiska transformationen $Y = -\frac{1}{\lambda} \ln X$ med $\lambda > 0$.

$$\begin{aligned} F_Y(y) &= P(-\frac{1}{\lambda} \ln X \le y) = P(\ln X \ge -\lambda y) = P(X \ge e^{-\lambda y}) \\ &= 1 - P(X < e^{-\lambda y}) = 1 - P(X \le e^{-\lambda y}) = \begin{cases} 0, & y < 0, \\ 1 - e^{-\lambda y}, & y \ge 0 \end{cases} \implies Y \in \mathrm{Exp}(\lambda) \end{aligned}$$

Observation: Låt X vara en kontinuerlig s.v. Om g är strängt växande resp. avtagande, så finns den inversa funktionen g^{-1} till g och vi får

a) $F_Y(y) = P(g(X) \le y) = P(X \le g^{-1}(y)) = F_X(g^{-1}(y))$ om g är strängt växande,

b) $F_Y(y) = P(g(X) \le y) = P(X \ge g^{-1}(y)) = 1 - F_X(g^{-1}(y))$ om g är strängt avtagande.

Väntevärdet av den s.v. $Y = g(X)$ beräknas genom

$$E(Y) = \begin{cases} \sum\limits_j g(x_j) P(X = x_j), & X \text{ diskret}, \\ \int\limits_{-\infty}^{\infty} g(t) f_X(t)\, dt, & X \text{ kontinuerlig}. \end{cases}$$

Exempel 3: Den s.v. $X \in \mathrm{U}(-\frac{1}{2}, \frac{1}{2})$ modellerar avrundningsfel. Låt $Y = X^2$. Som kvadratiskt avrundningsfel kan man vänta sig

$$E(Y) = \int_{-\infty}^{\infty} t^2 f_X(t)\, dt = \int_{-\frac{1}{2}}^{\frac{1}{2}} t^2\, dt = \frac{1}{12}$$

Exempel 4: St. Petersburgparadoxon.

Man kastar ett mynt tills en krona erhålls. Om detta inträffar i kast j får man 2^j kronor. Vad är spelets väntevärde?

Låt X betecknar kastet där krona erhålls. Vi är intresserade i väntevärdet av $Y = 2^X$.

$$E(Y) = \sum_j 2^j P(X = j) = \sum_j 2^j \left(\frac{1}{2}\right)^j = \sum_j 1 = \infty.$$

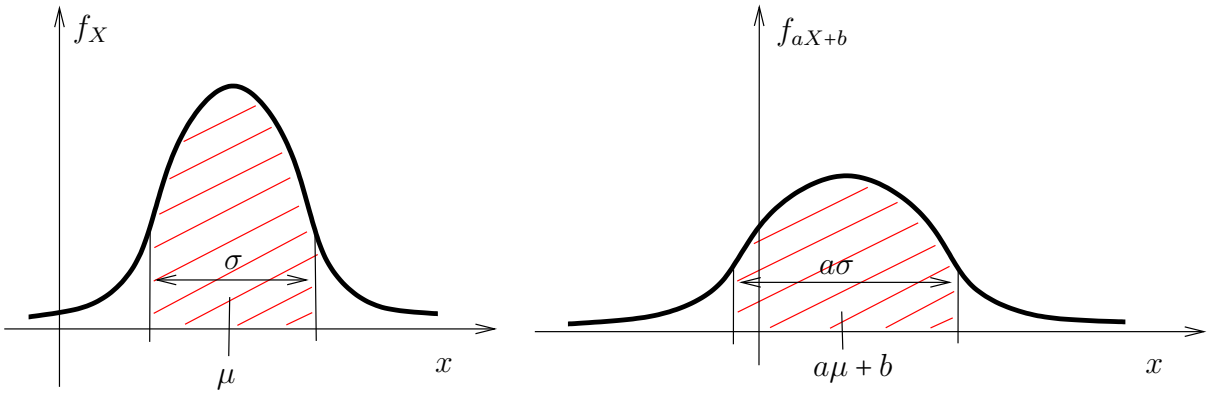
Man kan få stokastiska variabler med överaskande väntevärden genom att välja den stokastiska variablens värde mycket stor för en händelse med mycket liten sannolikhet.

Sats 1: Låt X vara en stokastisk variabel och a, b konstanter.
Då gäller:

a) $E(aX + b) = aE(X) + b$,
b) $V(aX + b) = a^2 V(X)$.

Resultatet i **a)** är rimligt eftersom förskutningen av sannolikhetsmassan medför motsvarande förskutningen av väntevärdet. Vidare medför förstorningen av värdena motsvarande förstorning av väntevärdet.

Resultatet i **b)** är rimligt eftersom förskutningen inte påverkar spridningen. Vidare medför förstorningen av värdena motsvarande förstorning i kvadrat av variansen (som mäter kvadratisk avvikelse).



Bevis:

$$\begin{aligned}
E(aX+b) &= \int_{-\infty}^{\infty}(at+b)f_X(t)\,dt = a\underbrace{\int_{-\infty}^{\infty}tf_X(t)\,dt}_{=E(X)}+b\underbrace{\int_{-\infty}^{\infty}f_X(t)\,dt}_{=1}\\
&= aE(X)+b\\
E\big((aX+b)^2\big) &= \int_{-\infty}^{\infty}(at+b)^2f_X(t)\,dt\\
&= a^2\underbrace{\int_{-\infty}^{\infty}t^2f_X(t)\,dt}_{=E(X^2)}+2ab\underbrace{\int_{-\infty}^{\infty}tf_X(t)\,dt}_{=E(X)}+b^2\underbrace{\int_{-\infty}^{\infty}f_X(t)\,dt}_{=1}\\
&= a^2E(X^2)+2abE(X)+b^2\\
\Longrightarrow V(aX+b) &= E\big((aX+b)^2\big)-\big(E(aX+b)\big)^2\\
&= a^2\Big(E(X^2)-\big(E(X)\big)^2\Big) = a^2V(X)
\end{aligned}$$

Följdsats: Låt X vara en stokastisk variabel med $E(X)=\mu$ och $V(X)=\sigma^2$. Då gäller för den standardiserade stokastiska variabeln

$$Y := \frac{X-\mu}{\sigma}$$

att $E(Y)=0$ och $V(Y)=1$.

4.2 Funktioner av flera stokastiska variabler

4.2.1 Kort om flerdimensionella stokastiska variabler

Mål: Att studera två (eller flera) slumpmässigt varierande storleker, t.ex. koordinaterna vid en pilkastning.

Även om vi senare kommer att betrakta summor av n stokastiska variabler beskränkar vi oss i den följande teorin på det tvådimensionella fallet.

Definition: En tvådimensionell s.v. (X,Y) är en funktion definierad på ett utfallsrum Ω med värden i $\mathbb{R}^2$.

Definition: Funktionen $F_{(X,Y)}:\mathbb{R}^2\to[0,1]$ given genom $F_{X,Y}(x,y)=P(X\le x, Y\le y)$ kallas för fördelningsfunktionen för (X,Y). Observera att $P(X\le x, Y\le y)$ betyder $P(X\le x \text{ och } Y\le y)$.

Om X, Y är kontinuerliga stokastiska variabler kallas en funktion $f_{X,Y}$ med

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

täthetsfunktionen för (X,Y). För att förstå sambandet mellan täthetsfunktionen $f_{X,Y}$ för (X,Y) och täthetsfunktionen f_X för X tittar vi på fördelningsfunktionerna

$$\begin{aligned} F_X(x) &= P(X \le x, Y \text{ godtycklig}) = \lim_{y\to\infty} F_{X,Y}(x,y) \\ &= \lim_{y\to\infty} \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(x,y)\ dy\ dx \\ &= \int_{-\infty}^{x} \underbrace{\int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dy}_{=f_X(x)}\ dx \end{aligned}$$

Som resultatet får vi att täthetsfunktionen f_X för X kan erhållas från den gemensamma täthetsfunktionen genom

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

Begreppet oberoende kan utvidgas till stokastiska variabler.

Definition: Två s.v. X, Y kallas oberoende om $P(X \le x, Y \le y) = P(X \le x)P(Y \le y)$.

Om X, Y är kontinuerliga stokastiska variabler är en ekvivalent definition att

$$f_{X,Y}(x,y) = f_X(x)f_Y(y).$$

4.2.2 Funktioner av två stokastiska variabler

Betrakta den s.v. $Z = g(X,Y)$. Vad kan vi säger om Z:s fördelning?

Exempel 5: Låt X och Y vara stokastiska variabler som är definierade på samma utfallsrum och som är oberoende. **a)** Betrakta $Z = \max(X,Y)$. Eftersom $Z \le z$ om och endast om både $X \le z$ och $Y \le z$ får man

$$F_Z(z) = P(Z \le z) = P(X \le z, Y \le z) = F_{X,Y}(z,z) = F_X(z)F_Y(z).$$

b) För $Z = \min(X,Y)$ använder vi att $Z > z$ omm både $X > z$ och $Y > z$.

$$\begin{aligned} F_Z(z) &= 1 - P(Z > z) = 1 - P(X > z)P(Y > z) = 1 - (1 - P(X \le z))(1 - P(Y \le z)) \\ &= 1 - (1 - F_X(z))(1 - F_Y(z)). \end{aligned}$$

Väntevärdet av den s.v. $Z = g(X,Y)$ beräknas genom

$$E(Z) = \begin{cases} \sum\limits_{j,k} g(x_j, y_k) P(X = x_j, Y = y_k), & X, Y \text{ diskreta,} \\ \int\limits_{-\infty}^{\infty} \int\limits_{-\infty}^{\infty} g(s,t) f_{X,Y}(,st)\ ds\ dt, & X, Y \text{ kontinuerliga.} \end{cases}$$

4.3 Summor av stokastiska variabler

Sats 2: För stokastiska variabler X, Y gäller
a) $E(X+Y) = E(X) + E(Y)$,
b) $V(X+Y) = V(X) + V(Y)$, om X och Y är oberoende.

Bevis för a)

$$\begin{aligned} E(X+Y) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x+y) f_{X,Y}(x,y)\ dx\ dy \\ &= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dy\ dx + \int_{-\infty}^{\infty} y \int_{-\infty}^{\infty} f_{X,Y}(x,y)\ dx\ dy \\ &= \int_{-\infty}^{\infty} x f_X(x)\ dx + \int_{-\infty}^{\infty} y f_Y(y)\ dy \\ &= E(X) + E(Y) \end{aligned}$$

Anmärkning: b) är en följd av Sats 6 och 7 (se avsnitt 4.5 för beviset)

Denna sats kan utvidgas till ändligt många stokastiska variabler $X_1, X_2, \ldots, X_n$. Som en följd får man följande resultat för det aritmetiska medelvärdet $\overline{X}$ som ges av

$$\overline{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n) = \frac{1}{n} \sum_{j=1}^{n} X_j$$

Sats 3: Låt $X_1, X_2 \ldots, X_n$ vara oberoende stokastiska variabler, där alla har väntevärde $E(X_j) = \mu$ och varians $V(X_j) = \sigma^2$. Då gäller

$$E(\overline{X}) = \mu \quad \text{och} \quad V(\overline{X}) = \frac{\sigma^2}{n}.$$

Stora talens lag: Låt $X_1, X_2, \ldots, X_n, \ldots$ vara en följd av oberoende och likafördelade s.v. med väntevärdet μ och låt $\epsilon > 0$. Då gäller att

$$P\Big(\mu - \epsilon < \frac{1}{n} \sum_{j=1}^{n} X_j < \mu + \epsilon\Big) \to 1 \quad \text{då} \quad n \to \infty.$$

I det fall att alla inblandade stokastiska variabler är normalfördelade kan man rentav bestämma fördelningen till deras summa.

Sats 4: Låt $a_1, a_2, \ldots, a_n$ vara givna konstanter. Om $X_1, X_2 \ldots, X_n$ är oberoende och $X_j \in \mathrm{N}(\mu_j, \sigma_j)$ för $j = 1, \ldots, n$, så gäller

$$\sum_{j=1}^{n} a_j X_j \in \mathrm{N}\Big(\sum_{j=1}^{n} a_j \mu_j, \sqrt{\sum_{j=1}^{n} a_j^2 \sigma_j^2} \Big)$$

Följdsats: Om $X_1, X_2 \ldots, X_n$ är oberoende och alla $X_j \in \mathrm{N}(\mu, \sigma)$, då gäller

$$\overline{X} = \mathrm{N}\Big(\mu, \frac{\sigma}{\sqrt{n}}\Big) \quad \text{för} \quad \overline{X} = \frac{1}{n} \sum_{j=1}^{n} X_j$$

Följdsats: Om alla $X_1, X_2 \ldots, X_{n_1}$ är $\mathrm{N}(\mu_1, \sigma_1)$ och alla $Y_1, Y_2 \ldots, Y_{n_2}$ är $\mathrm{N}(\mu_2, \sigma_2)$ och alla variabler är oberoende, så gäller att

$$\overline{X} - \overline{Y} \in \mathrm{N}\Big(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\Big) \quad \text{för} \quad \overline{X} = \frac{1}{n_1} \sum_{j=1}^{n_1} X_j, \quad \overline{Y} = \frac{1}{n_2} \sum_{j=1}^{n_2} Y_j$$

Exempel: En hiss tålar högst 10 personer eller 800 kg. Personvikten kan anses som s.v. X med väntevärdet 70 och standardavvikelsen 10.

a) Hur stor är vikten i genomsnitt där 10 personer är i hissen?
b) Antag att X är normalfördelad. Hur stor är sannolikhetenh för 10 personer at överlasta hissen?

Lösning: a) Låt X_j vara vikten av den personen nummer j.

$$E\Big(\sum_{j=1}^{10} X_j \Big) = \sum_{j=1}^{10} E(X_j) = 10 \cdot 70 = 700(\mathrm{kg}).$$

b) För $Y = \sum_{j=1}^{10} X_j$ gäller $Y \in \mathrm{N}(10 \cdot 70, \sqrt{10 \cdot 10^2})$, alltså $Z = \dfrac{Y - 700}{10\sqrt{10}} \in \mathrm{N}(0, 1)$.

$$P\Big(\sum_{j=1}^{10} X_j > 800 \Big) = 1 - P\Big(\frac{Y - 700}{10\sqrt{10}} \leq \frac{800 - 700}{10\sqrt{10}} \Big) = 1 - \Phi(\underbrace{\sqrt{10}}_{\approx 3{,}2}) = 0,0008$$

Det är 0,08%.

4.4 Centrala gränsvärdessatsen

Centrala gränsvärdessatsen: Låt $X_1, X_2, X_3 \ldots$ vara en oändlig följd av oberoende och likafördelade stokastiska variabler med väntevärdet μ och standardavvikelsen σ. Då gäller det att

$$P\left(\frac{\sum_{j=1}^{n} X_j - n\mu}{\sigma\sqrt{n}} \leq x \right) \longrightarrow \Phi(x) \text{ då } n \to \infty.$$

Följdsats: $\sum_{j=1}^{n} X_j$ är approximativt $\mathrm{N}(n\mu, \sigma\sqrt{n})$-fördelad då n är stort.

Resultatet gäller oavsett av den ursprungliga fördelningen! Det är detta resultat som förklarar varför många fenomen i naturen beteer sig approximativt "normal".

Approximativa egenskaper för binomial- och Poissonfördelning:

(1) Låt $X \in \mathrm{Bin}(N, p)$. Då kan X tolkas att ange hur ofta en viss händelse inträffar under N oberoende upprepningar av ett försök. Vi skriver

$$X = \sum_{j=1}^{N} X_j$$

där X_j anger om händelsen inträffar i försök nummer j.
Observera $X_j \in \mathrm{Bin}(1, p)$.

Centrala gränsvärdessatsen säger att X är approximativt normalfördelad med

$$\begin{aligned} E(X) &= N \cdot E(X_j) = N(1 \cdot p), \\ V(X) &= N \cdot V(X_j) = N(1 \cdot p(1-p)). \end{aligned}$$

Vi har visat följande

Sats: Om $X \in \mathrm{Bin}(N, p)$ så gäller det att X är approximativt $\mathrm{N}(Np, \sqrt{Np(1-p)})$ då N är tillräckligt stort.

(2) Man kan visa:

Sats: Om $X \in \mathrm{Po}(\lambda)$ så gäller att X är approximativt $\mathrm{N}(\lambda, \sqrt{\lambda})$ då λ är tillräckligt stort.

4.5 Oberoendemått

Låt X och Y vara stokastiska variabler som är definierade på samma utfallsrum. Kovariansen mellan X och Y är

$$C(X,Y) = E\big((X - E(X))(Y - E(Y))\big).$$

Observera att $C(X,X) = V(X)$.

Eftersom väntevärdet är linjär (se Sats 1) får man

Sats 5: $C(X,Y) = E(XY) - E(X)E(Y)$

Om $C(X,Y) = 0$ sägs X och Y kallas okorrelerade. Följande sats säger att om X och Y är oberoende så är de också okorrelerade.

Sats 6: X, Y oberoende $\Longrightarrow E(XY) = E(X)E(Y)$

Bevis:

$$E(XY) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} xy \underbrace{f_{X,Y}(x,y)}_{=f_X(x)f_Y(y)}\, dx\, dy = \int_{-\infty}^{\infty} x f_X(x)\, dx \int_{-\infty}^{\infty} y f_Y(y)\, dy = E(X)E(Y)$$

Följande exempel visar att okorrelerade variabler kan vara beroende.

Exempel: Låt $X \in \mathrm{U}(-1,1)$ och $Y = X^2$. Uppenbarligen är X, Y inte oberoende. Eftersom X:s fördelning är symmetrisk är $E(X) = 0$ och $E(XY) = E(X^3) = 0$. Enligt Sats 2 är $C(X,Y) = 0$, d.v.s. X, Y är okorrelerade.

Faktiskt mäter kovariansen graden av *linjärt* beroende.

Sats 7: $V(X \pm Y) = V(X) + V(Y) \pm 2C(X,Y)$

Bevis:

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

Korrelationskoefficienten för X och Y definieras av $\rho(X,Y) = \dfrac{C(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}}$.

Observera att $-1 \le \rho(X,Y) \le 1$.

Eftersom $V(X \pm Y) \ge 0$ följer $2\,|C(X,Y)| \le V(X) + V(Y)$ med Sats 4. Genom att ersätta X med aX och Y med bY för $a = \frac{1}{\sqrt{V(X)}}$, $b = \frac{1}{\sqrt{V(Y)}} > 0$ får man

$$2ab\,|C(X,Y)| \;=\; 2\,|C(aX,bY)| \;\le\; V(aX) + V(bY) \;=\; a^2V(X) + b^2V(Y) \;=\; 2,$$

alltså $|C(X,Y)| \le (ab)^{-1} = \sqrt{V(X)}\sqrt{V(Y)}$.

5 Beskrivande statistik

Mål: Beskriva ett siffermaterial på ett överskådligt sätt.

Exempel 1: Vid ett universitet får studenterna som går programmen A och B absolvera en matematikkurs. Kursen examineras genom en tenta med 24 poäng. Sista året deltog 20 studenter, 10 från varje program. De fick följande resultat:

$$\begin{array}{lllllllllll} A: & 11, & 22, & 7, & 5, & 11, & 16, & 10, & 19, & 15, & 14, \\ B: & 10, & 13, & 10, & 11, & 13, & 10, & 12, & 13, & 13, & 15. \end{array}$$

5.1 Storheter som är karakteristiska för materialet

Givet n observationer $x_1, \ldots x_n$. Beteckna samma observationer sedan de har storleksordnats med

$$\widehat{x}_1 \leq \ldots \leq \widehat{x}_n.$$

Största observationen betecknas också med $x_{\max}$ och minsta observationen med $x_{\min}$, det vill säger $x_{\min} = \widehat{x}_1$, $x_{\max} = \widehat{x}_n$.

Lägesmått:

- <u>medelvärde</u> $\quad \overline{x} = \dfrac{1}{n}\sum_{j=1}^{n} x_j$
- <u>median</u> $\quad \widetilde{x}_{0.50} = \begin{cases} \widehat{x}_{m+1}, & \text{om } n = 2m+1, \\ \dfrac{\widehat{x}_m + \widehat{x}_{m+1}}{2}, & \text{om } n = 2m \end{cases}$

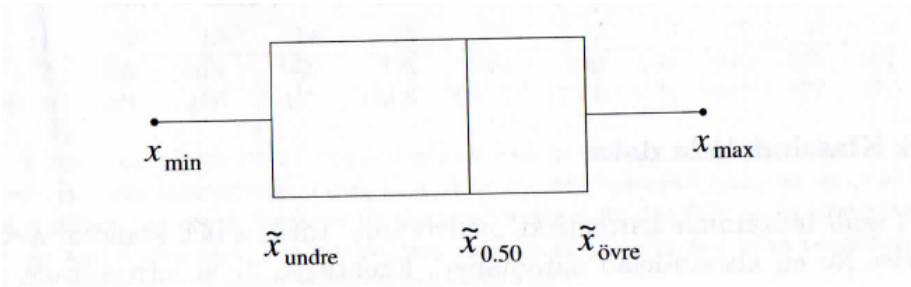
Observera att medianen påverkas inte av extrema observationer eftersom den inte tar hänsyn till de enskilda observationerna.

Vi inför också <u>övre kvartilen</u> $\widetilde{x}_{\text{övre}}$ och <u>undre kvartilen</u> $\widetilde{x}_{\text{undre}}$: Låt medianen dela upp observationerna i två delar, en undre del och en övre del. Om n är udda så räknas medianen in i bägge delarna och om n är jämnt inte i någon av delarna. Nu beräknas $\widetilde{x}_{\text{undre}}$ som medianen i undre delen och $\widetilde{x}_{\text{övre}}$ som medianen i övre delen.

Spridningsmått:

- varians $s^2 = \dfrac{1}{n-1}\sum_{j=1}^{n}(x_j - \overline{x})^2$ och standardavvikelse s
- variationsbredd $R = x_{\max} - x_{\min}$
- kvartilsavstånd $\widetilde{x}_{\text{övre}} - \widetilde{x}_{\text{undre}}$

Medianen, kvartilerna och variationsbredden hos ett material kan illustreras med hjälp av en boxplot eller ett lådagram:



Sats: $s^2 = \dfrac{1}{n-1}\Big(\sum_{j=1}^{n} x_j^2 - n\overline{x}^2\Big)$

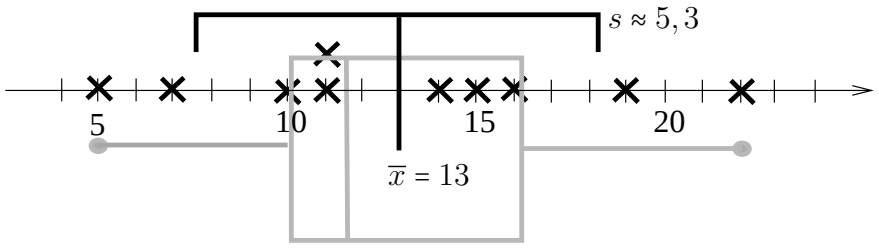
Bevis:

$$\sum_{j=1}^{n}(x_j - \overline{x})^2 = \sum_{j=1}^{n}(x_j^2 - 2\overline{x}x_j + \overline{x}^2) = \sum_{j=1}^{n} x_j^2 - 2\overline{x}\underbrace{\sum_{j=1}^{n} x_j}_{=n\overline{x}} + \overline{x}^2\underbrace{\sum_{j=1}^{n} 1}_{=n} = \sum_{j=1}^{n} x_j^2 - n\overline{x}^2$$

Exempel 1 (forts.):

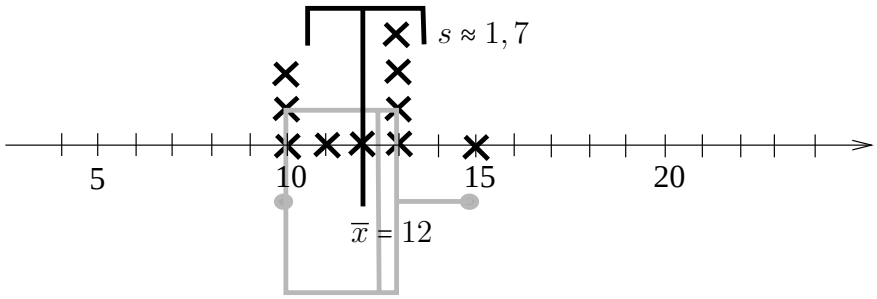
Grupp A: $x_{\min} = 5$, $x_{\max} = 22$ ger variationsbredden $R = 17$.

Medianen är $\widetilde{x}_{0.05} = \frac{11+14}{2} = 12,5$ och för kvartilerna får vi $\widetilde{x}_{undre} = 10$, $\widetilde{x}_{\text{övre}} = 16$ som ger kvartilsavståndet $16 - 10 = 6$.



Grupp B: $x_{\min} = 10$, $x_{\max} = 15$ ger variationsbredden $R = 5$.

Medianen är $\widetilde{x}_{0.05} = \frac{12+13}{2} = 12,5$ och för kvartilerna får vi $\widetilde{x}_{undre} = 10$, $\widetilde{x}_{\text{övre}} = 13$ som ger kvartilsavståndet $13 - 10 = 3$.



5.2 Grupperade data

Exempel 2: Man undersökte 35 tändsticksaskar och noterade för varje ask hur många tändstickor den innehöll. Följande värden erhölls:

51	52	49	51	52	51	53
52	48	52	50	53	49	50
51	53	51	52	50	51	53
53	55	50	49	53	50	51
51	52	48	53	50	49	51

För illustrationen av materialet behöver vi några beteckningar: Låt n vara antalet av observationer och $y_1 \leq \ldots \leq y_m$ vara de möjliga värden som observationerna kan anta.

- frekvens f_j = antalet förekomster av y_j
- relativ frekvens $p_j = \dfrac{f_j}{n}$

Anmärkning: **a)** $\sum_j f_j = n, \quad \sum_j p_j = 1$

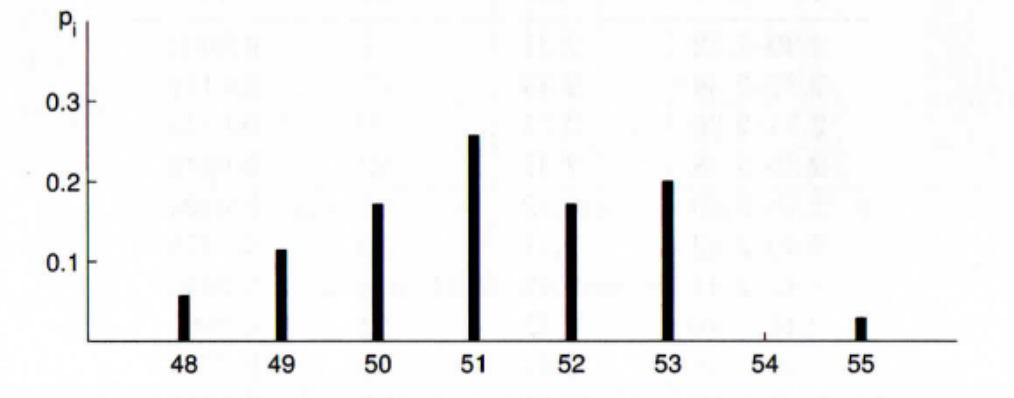
b) En relativ frekvens översätts i procent genom att multiplicera med 100. Till exempel motsvarar 1 100% och 0,01 motsvarar 1 %,

Resultaten kan sammanfattas i en frekvenstabell. Grafiskt kan de presenteras i ett stolpdiagram där man ritar en stolpe för varje variabelvärde sådan att stolpens längd motsvarar dess relativa frekvens.

Exempel 2 (forts.):

Frekvenstabell för antal tändstickor i tändsticksaskar.

Klass y_i	Absolut frekvens f_i	Relativ frekvens (%) $100\,p_i$
48	2	5.7
49	4	11.4
50	6	17.1
51	9	25.7
52	6	17.1
53	7	20.0
54	0	0.0
55	1	2.9
S:a	35	100.0



Det gäller:

(1) $\overline{x} = \dfrac{1}{n}\sum_{j=1}^{m} f_j y_j$

(2) $s^2 = \dfrac{1}{n-1}\sum_{j=1}^{m} f_j(y_j - \overline{x})^2 = \dfrac{1}{n-1}\Big(\sum_{j=1}^{m} f_j y_j^2 - n\overline{x}^2\Big) = \dfrac{1}{n-1}\Bigg(\sum_{j=1}^{m} f_j y_j^2 - \dfrac{1}{n}\Big(\sum_{j=1}^{m} f_j y_j\Big)^2\Bigg)$

Exempel 2 (forts.):

$$\sum_j f_j y_j = 2\cdot 48 + 4\cdot 49 + 6\cdot 50 + 9\cdot 51 + 6\cdot 52 + 7\cdot 53 + 0\cdot 54 + 1\cdot 55 = 1789$$

$$\sum_j f_j y_j^2 = 2\cdot 48^2 + 4\cdot 49^2 + 6\cdot 50^2 + 9\cdot 51^2 + 6\cdot 52^2 + 7\cdot 53^2 + 0\cdot 54^2 + 1\cdot 55^2 = 91533$$

$$\Longrightarrow \overline{x} = \frac{1789}{35} \approx 51,1, \qquad s^2 = \frac{1}{35-1}\Big(91533 - \frac{1789^2}{35}\Big) = \frac{1567}{595} \approx 2,63, \qquad s \approx 1,62$$

$$\widetilde{x}_{0.50} = \widehat{x}_{18} = 51, \qquad \widetilde{x}_{undre} = \frac{\widehat{x}_9 + \widehat{x}_{10}}{2} = 50, \qquad \widetilde{x}_{\text{övre}} = \frac{\widehat{x}_{26} + \widehat{x}_{27}}{2} = 52$$

5.3 Klassindelade data

I praktiken är materialet ofta så stort (och antar så många värden) att man måste förenkla det för att få det överskådligt.

Exempel 3: Man mätte kapacitansen hos 630 kondensatorer.

Tabell 10.2 *Kapacitans hos kondensatorer (utdrag ur större datamängd).*

2.504	2.616	2.627	2.541	2.618	2.476	2.328	2.404	2.506	2.413
2.578	2.638	2.596	2.362	2.519	2.520	2.372	2.483	2.501	2.395
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
2.547	2.574	2.528	2.460	2.467	2.418	2.427	2.451	2.602	2.546
2.424	2.446	2.491	2.475	2.601	2.352	2.621	2.558	2.322	2.459

Därtill klassindelar man materialet i ett antal klasser (intervall) $I_1, \ldots, I_m$ där $I_j \cap I_k = \varnothing$, $j \neq k$, och där $\bigcup_{j=1}^{m} I_j$ teckar alla möjliga värden.

Med y_j betecknar man nu $I_j : s$ klassmitten (medelpunkt av intervallet).

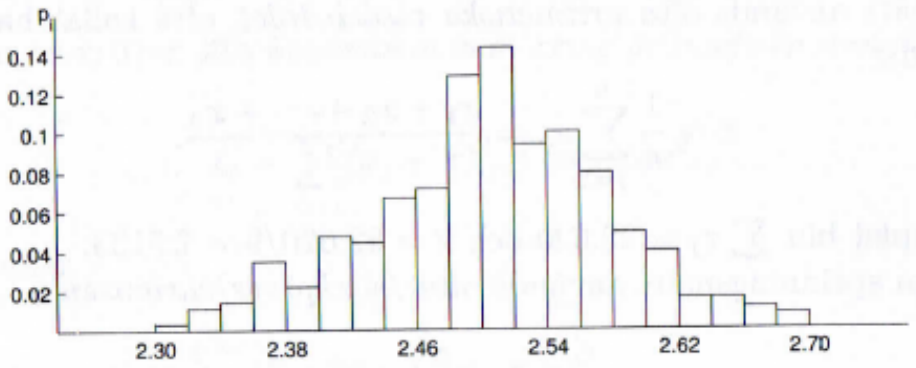
Exempel 3 (forts.):

Tabell 10.3 *Kapacitans hos 630 kondensatorer.*

Klassgränser		Klassmitt	Absolut frekvens	Relativ frekvens
g_i	g_{i+1}	y_i	f_i	p_i
2.30	2.32	2.31	2	0.0032
2.32	2.34	2.33	7	0.0111
2.34	2.36	2.35	9	0.0143
2.36	2.38	2.37	22	0.0349
2.38	2.40	2.39	14	0.0222
2.40	2.42	2.41	30	0.0476
2.42	2.44	2.43	28	0.0444
2.44	2.46	2.45	42	0.0667
2.46	2.48	2.47	45	0.0714
2.48	2.50	2.49	81	0.1286
2.50	2.52	2.51	90	0.1429
2.52	2.54	2.53	59	0.0937
2.54	2.56	2.55	63	0.1000
2.56	2.58	2.57	50	0.0794
2.58	2.60	2.59	31	0.0492
2.60	2.62	2.61	25	0.0397
2.62	2.64	2.63	10	0.0159
2.64	2.66	2.65	10	0.0159
2.66	2.68	2.67	7	0.0111
2.68	2.70	2.69	5	0.0080

Observera att den övre gränsen till en klass inte är inkluderad i klassen! T. ex. första klassen är intervallet $[2,30; 2,32)$.

Grafiskt presenteras materialet i ett histogram. Notera att arean i histogrammet är lika med 1 om klasslängerna nomeras till 1.



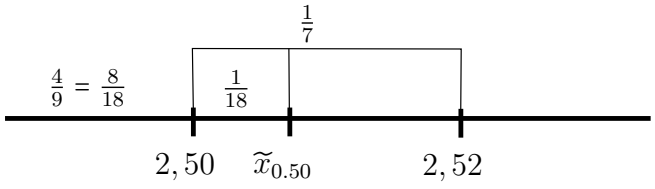
Figur 10.2 *Histogram.*

För klassindelade data beräknas de karakteristiska storheter med avseende på den valda klassindelningen. T.ex.:

- Medelvärdet och variansen beräknas som för grupperade data. Observera dock att y_j betecknar nu klassmitterna av intervallen.
- Medianen definieras som det värde för vilket arean i histogrammet till vänster av värdet och till höger av värdet är likandana
- R definieras som avståndet mellan de mest extrema klassgränserna

Exempel 3 (forts.): $R = 2,70 - 2,30 = 0,40$

Enligt tabellen innehåller klasserna 1 - 9 totalt 280 observationer. Arean i histogrammet över dessa klasser motsvarar därför $\frac{280}{630} = \frac{4}{9}$ av hela arean. Klass 10 innehåller 90 observationer motsvarande $\frac{90}{630} = \frac{1}{7}$ av hela arean.



$$\Longrightarrow \frac{\widetilde{x}_{0.50} - 2,5}{2,52 - 2,5} = \frac{1/18}{1/7}$$

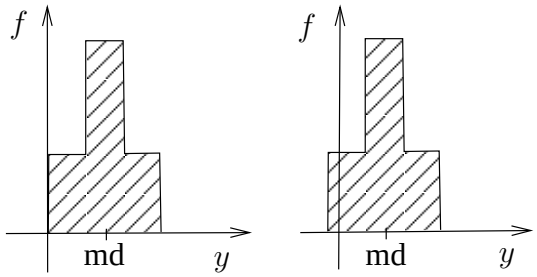
$$\Longrightarrow \widetilde{x}_{0.50} = 2,5 + \frac{7}{18} \cdot (2,52 - 2,50) \approx 2,5078$$

Anmärkning: a) Klassindelning förenklar materialet: De karakteristikor man beräknar med det ursprungliga materialet är inte nödvändigtvis lika med de man får efter klassindelning.

b) Olika klassindelningar kan ge olika karakteristikor!

Exempel: 0,5; 0,7; 1,1; 1,1; 1,2; 1,3; 1,4; 1,4; 2,1; 2,1

- klassindelning 1: [0,1), [1,2), [2,3) ger $\widetilde{x}_{0.50} = 1,5$
- klassindelning 2: [-0.2;0,8), [0,8;1,8), [1,8;2,8) ger $\widetilde{x}_{0.50} = 1,3$



5.4 Korrelation

Låt datamängden består av parade observationer $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Beroendemått:

- kovarians mellan x- och y-värdena $\quad c_{xy} = \dfrac{1}{n-1}\sum_{j=1}^{n}(x_j - \overline{x})(y_j - \overline{y})$
- korrelationskoefficient $\quad r = \dfrac{c_{xy}}{s_x s_y}$

  där s_x och s_y är standardavvikelsen för x- respektice y-värdena

Om data variera så att till stora x-värdena tillhör oftast stora y-värden förväntas positiv kovarians.

Det gäller: $\quad c_{xy} = \dfrac{1}{n-1}\Big(\sum_{j=1}^{n} x_j y_i - n\,\overline{x}\,\overline{y}\Big)$

Den bakomliggande modellen

Vi betraktar ett visst experiment som beskrivs av en stokastisk variabel X med fördelningsfunktion F. Antag att fördelningen beror på en okänd parameter θ. Ett typsikt exempel är att $\theta = E(X)$.

Vi vill, på grund av experimentella data $x_1, \ldots, x_n$, få information om θ. I de följande kapitlen kommer vi att

6. **Punktskattning:** skatta θ
7. **Intervallskattning:** konstruera ett intervall som täcker θ med föreskriven sannolikhet
8. **Hypotesprövning:** testa en hypotes om θ

Låt oss anta att vi har fått datamaterialet $x_1, \ldots, x_n$ genom n oberoende upprepningar av experimentet. Stokastiskt kan vi beskriva experiment nummer j genom en stokastisk variabel X_j som har samma fördelning som X. Datamaterialet kan alltså tolkas som en observation av

$$n \text{ oberoende likafördelade stokastiska variabler } X_1, \ldots, X_n.$$

I denna situation kallas $x_1, \ldots, x_n$ för ett stickprov av storlek n.

6 Punktskattning

Mål: För att skatta θ är idén att ta en lämplig vald funktion $\theta^\star$ av stickprovet.

Funktionen $\theta^\star = \theta^\star(X_1, \ldots, X_n)$ kallas stickprovsvariabel. Observera att den är en stokastisk variabel. Den konkreta observationen $\theta^\star_{\text{obs}} = \theta^\star(x_1, \ldots, x_n)$ kallas en punktskattning för θ.

Punktskattningen $\theta^\star_{\text{obs}}$ är alltså ett utfall av den stickprovsvariabeln $\theta^\star$.

6.1 Punktskattningens grundläggande egenskaper

Vilka funktioner kan anses lämplig för att skatta till exempel väntevärde och varians?

a) väntevärdet $\mu = E(X)$:

$$\mu_1^\star =: \overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j,$$

$$\mu_2^\star = \frac{1}{2}\big(\max_j X_j + \min_j X_j\big).$$

Observera att $\overline{X}$ är stickprovsvariabeln, alltså en s.v.! Motsvarande punktskattningen är medelvärdet $\overline{x}$ av det konkreta datamaterialet.

b) variansen $\sigma^2 = V(X)$:

$$(\sigma_1^2)^\star = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2,$$

$$(\sigma_2^2)^\star = \frac{1}{2}\Big(\max_j X_j - \min_j X_j\Big).$$

Definition: En punktskattning $\theta^\star_{\text{obs}}$ av parametern θ är väntevärdesriktig om $E(\theta^\star) = \theta$ gäller för den tillhörande stickprovsvariabeln $\theta^\star$.

Sats: **a)** $\mu^\star_{1,\text{obs}}$ är en väntevärdesriktig punktskattning av μ.
b) $(\sigma_1^2)^\star_{\text{obs}}$ är en väntevärdesriktig punktskattning av σ^2

Observera att satsen inte beror på X:s fördelning!

Bevis: b)

$$\begin{aligned}
\sum_{j=1}^{n}(X_j - \overline{X})^2 &= \sum_{j=1}^{n}\Big((X_j - \mu) - (\overline{X} - \mu)\Big)^2 \\
&= \sum_{j=1}^{n}\Big((X_j - \mu)^2 - 2(X_j - \mu)(\overline{X} - \mu) + (\overline{X} - \mu)^2\Big) \\
&= \sum_{j=1}^{n}(X_j - \mu)^2 - 2(\overline{X} - \mu)\underbrace{\sum_{j=1}^{n}(X_j - \mu)}_{=n(\overline{X}-\mu)} + (\overline{X} - \mu)^2\underbrace{\sum_{j=1}^{n}1}_{=n} \\
&= \Big(\sum_{j=1}^{n}(X_j - \mu)^2\Big) - n(\overline{X} - \mu)^2 \\
\Longrightarrow E\Big(\sum_{j=1}^{n}(X_j - X)^2\Big) &= \sum_{j=1}^{n}\underbrace{E\Big((X_j - \mu)^2\Big)}_{=V(X_j)=\sigma^2} - n\,\underbrace{E\Big((\overline{X} - \mu)^2\Big)}_{=V(\overline{X})=\frac{\sigma^2}{n}} \\
&= n\sigma^2 - n\frac{\sigma^2}{n} \\
&= (n-1)\sigma^2
\end{aligned}$$

I allmänheten är $\mu^\star_{2\text{obs}}$ inte väntevärdesriktig. Ett fall där $\mu^\star_{2,\text{obs}}$ dock är det är om X har en symmetrisk frekvensfunktion.

Definition: Låt $\theta^\star_{1,\text{obs}}$, $\theta^\star_{2,\text{obs}}$ vara väntevärdesriktiga punktskattningar av θ. Vi säger att $\theta^\star_{1,\text{obs}}$ är effektivare än $\theta^\star_{2,\text{obs}}$ om motsvarande stickprovsvariabler uppfyller $V(\theta^\star_1) < V(\theta^\star_2)$.

Man kan visa att $\mu^\star_{1,\text{obs}}$, $(\sigma^2_1)^\star_{\text{obs}}$ är de effektivaste skattningar för μ, σ^2 förutsatt att stickprovet kommer från en normalfördelning.

6.2 Maximum-likelihood-metoden (ML-metoden)

Antag att vi vet att sannolikhetsfunktionen p (resp. täthetsfunktionen f) av den s.v. X tillhör en viss familj $p(k,\theta)$ (resp. $f(x;\theta)$) som beror på θ.

ML-metoden: För det givna stickprovet $x_1,\ldots,x_n$ definierar vi likelihood-funktionen L genom

$$L(\theta) = \begin{cases} p(x_1;\theta)\ldots p(x_n;\theta), & \text{(diskreta fallet)},\\ f(x_1;\theta)\ldots f(x_n;\theta), & \text{(kontinuerliga fallet)}. \end{cases}$$

Om det finns ett unikt värde θ^*_{obs} där $L(\theta)$ antar sitt största värde så kallas θ^*_{obs} för ML-skattningen av θ.

Exempel 1: Låt X = antalet telefonsamtal som invånarna i en by börjar inom ett givet tidsintervall. Vi antar att X är Poissonfördelad med en positiv parameter θ. Fem oberoende mätningar ger antalen 10, 12, 7, 10, 4. Bestäm ML-skattningen av θ.

Lösning: Eftersom X har sannolikhetsfunktionen

$$p(k;\theta) = P(X=k;\theta) = \frac{\theta^k}{k!}e^{-\theta}, \qquad k = 0,1,2,\ldots,$$

blir likelihood-funktionen

$$L(\theta) \;=\; p(10;\theta)\; p(12;\theta)\; p(7;\theta)\; p(10;\theta)\; p(4;\theta) \;=\; \frac{\theta^{43}e^{-5\theta}}{10!\;12!\;7!\;10!\;4!}.$$

För att maximera täljaren $\theta^{43}e^{-5\theta}$ undersöker vi derivatan $(\theta^{43}e^{-5\theta})' = (43-5\theta)\theta^{42}e^{-5\theta}$och ser att funktionen är växande för $0<\theta<43/5$, avtagande för $\theta>43/5$ och har ett globalt maximum i $\theta^*_{\text{obs}} = 43/5 = 8,6$.

Anmärkning: Likadant visar man att, för ett givet stickprov $x_1,\ldots,x_n$, ML-skattningen för θ är $\overline{x} = \frac{1}{n}\sum_{j=1}^n x_j$ om $X \in \text{Po}(\theta)$.

Anmärkning: Ofta är det lättare att maximera $\ln L(\theta)$ istället av $L(\theta)$.

Exempel 2: För n glödlampor mätar man livstider $x_1, \ldots, x_n$. Antag att livstiden X av en glödlampa är exponentialfördelad med parameter θ. Bestäm ML-skattningen av θ.

Lösning: Eftersom X har täthetsfunktionen $f(x) = \frac{1}{\theta} e^{-x/\theta}$ får vi

$$\begin{aligned} L(\theta) &= \prod_{j=1}^{n} \frac{e^{-x_j/\theta}}{\theta} = \frac{e^{-(\sum_{j=1}^{n} x_j)/\theta}}{\theta^n} \\ \Longrightarrow \ln L(\theta) &= -\frac{\sum_{j=1}^{n} x_j}{\theta} - n \ln \theta \\ \Longrightarrow \frac{d}{d\theta} \ln L(\theta) &= \frac{1}{\theta} \left(\frac{\sum_{j=1}^{n} x_j}{\theta} - n \right). \end{aligned}$$

Det visar att likelihood-funktionen är växande för $0 < \theta < \frac{1}{n} \sum_{j=1}^{n} x_j = \overline{x}$, avtagande för $\theta > \overline{x}$ och har ett globalt maximum i $\theta^{\star}_{\text{obs}} = \overline{x}$.

6.3 Tillämpning till normalfördelning

6.3.1 Ett stickprov

En okänd parameter. Låt $x_1, \ldots, x_n$ vara ett stickprov från $\mathrm{N}(\mu, \sigma^2)$ där en av parametrar μ, σ är okänd. Som i sista avsnittet kan man bestämma ML-skattningen för denna parameter:

- μ **okänt,** σ **känd:** ML-skattningen för μ är $\mu^{\star}_{\text{obs}} = \overline{x}$.
- μ **känt,** σ **okänd:** ML-skattningen för σ^2 är $(\sigma^2)^{\star}_{\text{obs}} = \dfrac{1}{n} \displaystyle\sum_{j=1}^{n} (x_j - \mu)^2$.

I båda fall kan man visa att ML-skattningen är väntevärdesriktig.

Två okända parametrar. Den vanligare situationen är dock att både parametrar är okända. Observera att likelihood-funktionen i detta fall beror på två variabler. Eftersom

$$\begin{aligned} L(\mu,\sigma^2) &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\sum_{j=1}^n (x_j-\mu)^2/(2\sigma^2)} \\ \Longrightarrow \quad \frac{d}{d\mu}\ln L(\mu,\sigma^2) &= \frac{1}{\sigma^2}\sum_{j=1}^n (x_j-\mu), \\ \frac{d}{d\sigma^2}\ln L(\mu,\sigma^2) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{j=1}^n (x_j-\mu)^2. \end{aligned}$$

blir resultatet

- μ**,** σ **okända:** ML-skattningar för μ, σ^2 är

$$\mu^\star_{\text{obs}} = \overline{x}, \qquad (\sigma^2)^\star_{\text{obs}} = \frac{1}{n}\sum_{j=1}^n (x_j-\overline{x})^2.$$

Anmärkning: ML-skattningen för μ är väntevärdesriktig men ML-skattningen för σ^2 är det inte! En väntevärdesriktig skattning ger följande adjustering

$$s^2 = \frac{n}{n-1}(\sigma^2)^\star_{\text{obs}}.$$

6.3.2 Två stickprov

Till slut tar vi upp fallet att man har två oberoende stickprov $x_1,\ldots,x_{n_1}$ från $\mathrm{N}(\mu_1,\sigma_1^2)$ och $y_1,\ldots,y_{n_2}$ från $\mathrm{N}(\mu_2,\sigma_2^2)$ där $\sigma_1 = \sigma_2 =: \sigma$. Här är likelihood-funktionen

$$\begin{aligned} L(\mu_1,\mu_2,\sigma^2) &= L_1(\mu_1,\sigma^2)\cdot L_2(\mu_2,\sigma^2) \text{ där} \\ L_1(\mu_1,\sigma^2) &= \frac{1}{(2\pi\sigma^2)^{n_1/2}} e^{-\sum_{j=1}^{n_1}(x_j-\mu_1)^2/(2\sigma^2)}, \\ L_2(\mu_2,\sigma^2) &= \frac{1}{(2\pi\sigma^2)^{n_2/2}} e^{-\sum_{j=1}^{n_2}(y_j-\mu_2)^2/(2\sigma^2)}. \end{aligned}$$

ML-skattningarna blir

$$(\mu_1)^\star_{\text{obs}} = \overline{x}, \qquad (\mu_2)^\star_{\text{obs}} = \overline{y}, \qquad (\sigma^2)^\star_{\text{obs}} = \frac{1}{n_1+n_2}(Q_1+Q_2)$$

med $Q_1 = \sum_{j=1}^{n_1}(x_j-\overline{x})^2$, $Q_2 = \sum_{j=1}^{n_2}(y_j-\overline{y})^2$. De första två skattningarna är väntevärdesriktiga. För den sista skattningen blir följande adjustering väntevärdesriktig

$$s^2 = \frac{n_1+n_2}{(n_1-1)+(n_2-1)}(\sigma^2)^\star_{\text{obs}}.$$

7 Intervallskattning

Mål: Låt $0 \le \alpha \le 1$. Vi vill ange ett intervall $I^\star$ sådant att

$$P(\theta \in I^\star) = 1 - \alpha.$$

Observera att $I^\star$ måste beteckna en stokastisk variabel! För ett givet intervall $I \subset \mathbb{R}$ gäller nämligen antingen $P(\theta \in I) = 0$ (om $\theta \notin I$) eller $P(\theta \in I) = 1$ (om $\theta \in I$).

En konkret observation $I^\star_{\text{obs}}$ av den stokastiska variablen $I^\star$ kallas för ett konfidensintervall för θ med konfidensgrad $1 - \alpha$.

7.1 Teckentest

(se föreläsning).

7.2 Tillämpning till Normalfördelning

Från och med nu antar vi att stickprovet kommer från en normalfördelning. Observera att centrala gränsvärdessatsen säger att det är approximativt fallet för n tillräckligt stort.

7.2.1 Konfidensintervall för μ i $\mathrm{N}(\mu, \sigma)$ där σ är känd

Idé: Betrakta $I^\star = [\overline{X} - a, \overline{X} + b]$ och bestäm lämpliga konstanterna a, $b > 0$ sådana att $P(\overline{X} - a \le \mu \le \overline{X} + b) = 1 - \alpha$.

Vi vet att $\overline{X} \in \mathrm{N}(\mu, \dfrac{\sigma}{\sqrt{n}})$. Eftersom

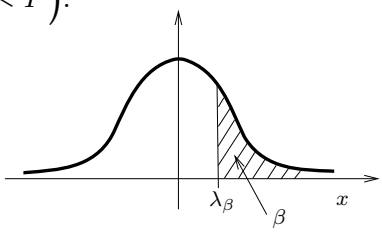
$$\overline{X} - a \le \mu \le \overline{X} + b \quad \Longleftrightarrow \quad -b \le \overline{X} - \mu \le a$$

och $\overline{X}$:s fördelning är symmetrisk kring μ är det rimligt att välja $b = a$. Vi vill välja a sådan att

$$\begin{aligned} \alpha &= P(\mu < \overline{X} - a) + P(\overline{X} + a < \mu) = P(a < \overline{X} - \mu) + P(\overline{X} - \mu < -a) \\ &= 2P(a < \overline{X} - \mu) = 2P\Big(\frac{a}{(\sigma/\sqrt{n})} < \frac{\overline{X} - \mu}{(\sigma/\sqrt{n})}\Big). \end{aligned}$$

Sätt $Y := (\overline{X} - \mu)/(\sigma/\sqrt{n})$ och observera att $Y \in \mathrm{N}(0,1)$. Enligt ovanstående resonemanget måste vi välja a sådan att

$$\frac{\alpha}{2} = P\Big(\frac{\sqrt{n}}{\sigma}a < Y\Big). \tag{1}$$



Låt λ_β beteckna det värde för vilket $P(\lambda_\beta < Y) = \beta$ där $Y \in \mathrm{N}(0,1)$. För de viktigaste värdena av β kan λ_β hittas i tabellen för standardnormalfördelningen.

För att få (1) väljer vi nu konstanten a sådan att

$$\frac{\sqrt{n}}{\sigma}a = \lambda_{\frac{\alpha}{2}} \quad \Longrightarrow \quad a = \frac{\sigma}{\sqrt{n}}\lambda_{\frac{\alpha}{2}}.$$

Resultat: Låt $x_1, \ldots, x_n$ vara ett stickprov från $\mathrm{N}(\mu, \sigma)$ där σ är känd. Då är ett konfidensintervall för μ med konfidensgrad $1 - \alpha$ givet genom

$$I^\star_{\text{obs}} = \Big[\ \overline{x} - \frac{\sigma}{\sqrt{n}}\lambda_{\frac{\alpha}{2}}, \overline{x} + \frac{\sigma}{\sqrt{n}}\lambda_{\frac{\alpha}{2}}\ \Big].$$

Anmärkning: Intervallets längd är $2\sigma\lambda_{\frac{\alpha}{2}}/\sqrt{n}$. Det medför att

- intervallet blir större om man ökar konfidensgraden $1 - \alpha$,
- intervallet blir mindre om man ökar n.

Anmärkning: Eftersom man vill kontrollera felsannolikheten, d.v.s. man vill garantera att $P(\mu \notin I^\star) \leq \alpha$, bör avrundningen av konfidensintervallet göras utåt!

Anmärkning: Ovan har vi bestämt ett tvåsidigt konfidensintervall.

Ibland är man också intresserad av ett ensidigt konfidensintervall, t.ex. $I^\star = (-\infty, \overline{X} + a]$ med ett lämpligt tal a. Ett likadant resonemang som ovan ger konfidensintervallet

$$I^\star_{\text{obs}} = \Big(-\infty,\ \overline{x} + \frac{\sigma}{\sqrt{n}}\lambda_\alpha\Big]$$

för konfidensgraden $1 - \alpha$.

Exempel 1: Vi har gjort 9 mätningar för livslängder av en viss sorts maskindelar och fått värdena

$$115,3;\ 106,5;\ 110,6;\ 106,9;\ 95,4;\ 115,1;\ 112,9;\ 107,7;\ 109,7$$

(i timmar). Dessa ser vi som ett utfall av oberoende stokastiska variabler $X_1, \ldots, X_9$ som alla är $\mathrm{N}(\mu, 15)$. Vi vill bestämma ett 95% konfidensintervall för μ.

Lösning: Vi beräknar $\overline{x} = 108,9$. Eftersom $\alpha = 0,05$ får vi från tabellen för standardnormalfördelningen att $\lambda_{0,025} = 1,9600$. Ett (två-sidigt) konfidensintervall för μ med konfidensgraden $0,95$ är alltså

$$I^{\star}_{\text{obs}} = \Big[108,9 + \frac{15}{\sqrt{9}}1,96;\ 108,9 + \frac{15}{\sqrt{9}}1,96\Big] \approx \Big[99,1;\ 118,7\Big].$$

7.2.2 Konfidensintervall för μ i $\mathrm{N}(\mu, \sigma)$ där σ är okänd

Idé: Skatta σ genom punktskattningen $s = \sigma^{\star}_{\text{obs}}$. Den bakomliggande stokastiska variablen är

$$\sigma^{\star} = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2}.$$

I detta fall är

$$Y = \frac{\overline{X} - \mu}{\sigma^{\star}/\sqrt{n}}$$

inte längre normalfördelad utan t-fördelad med $n-1$ frihetsgrader. Denna fördelning är också symmetrisk.

För några f, β är värdena $t_{\beta}(f)$ sådana att $P(t_{\beta}(f) < Y) = \beta$, där Y är t-fördelad med f frihetsgrader, tabellerade.

Likadant som i förra avsnittet får vi nu följande

Resultat: $I^{\star}_{\text{obs}} = \Big[\ \overline{x} - \frac{s}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1);\ \overline{x} + \frac{s}{\sqrt{n}}t_{\frac{\alpha}{2}}(n-1)\ \Big]$ är ett konfidensintervall för μ med konfidensgrad $1-\alpha$.

Observera att intervallängden nu beror på stickprovet!

Exempel 1 (forts.): Om σ är okänd skattar vi den med s. Vi beräknar $s = 6,05$. För $\alpha = 0,05$ får vi nu från tabellen för t-fördelningen att $t_{0,025}(9-1) = 2,306$. Ett (två-sidigt) konfidensintervall för μ med konfidensgraden $0,95$ är alltså

$$I^{\star}_{\text{obs}} = \Big[108,9 + \frac{6,05}{\sqrt{9}}2,306;\ 108,9 + \frac{6,05}{\sqrt{9}}2,306\Big] \approx \Big[104,2;\ 113,6\Big].$$

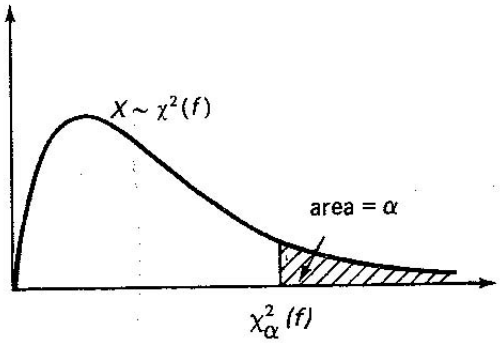
7.2.3 Konfidensintervall för σ^2 i $\mathrm{N}(\mu,\sigma)$ där μ är okänt

Den stokastiska variablen

$$\frac{1}{\sigma^2}\sum^{n}(X_j - \overline{X})^2$$

är χ^2-fördelad med $n-1$ frihetsgrader.

Enligt definitionen av $\chi^2_{\alpha}(f)$, se figur, är sannolikhetsmassan mellan $\chi^2_{1-\alpha/2}(f)$ och $\chi^2_{\alpha/2}(f)$ lika med $1-\alpha$.



Alltså gäller med sannolikhet $1-\alpha$ att

$$\frac{1}{\chi^2_{\alpha/2}(n-1)}\sum_{j=1}^{n}(X_j - \overline{X})^2 \ < \ \sigma^2 \ < \ \frac{1}{\chi^2_{1-\alpha/2}(n-1)}\sum_{j=1}^{n}(X_j - \overline{X})^2$$

och vi får

Resultat: $I^{\star}_{\text{obs}} = \Big[\ \frac{1}{\chi^2_{\alpha/2}(n-1)}\sum_{j=1}^{n}(x_j - \overline{x})^2;\ \frac{1}{\chi^2_{1-\alpha/2}(n-1)}\sum_{j=1}^{n}(x_j - \overline{x})^2\ \Big]$ är ett konfidensintervall för σ^2 med konfidensgrad $1-\alpha$.

7.2.4 Två stickprov. Konfidensintervall för differens mellan väntevärden

Antag att vi har två oberoende stickprov $x_1,\ldots,x_{n_1}$ från $\mathrm{N}(\mu_1,\sigma_1)$ och $y_1,\ldots,y_{n_2}$ från $\mathrm{N}(\mu_2,\sigma_2)$ där $\sigma_1 = \sigma_2 =: \sigma$. Antag att σ också är okänd. Vi vill bestämma ett konfidensintervall för differensen $\mu_1 - \mu_2$.

Differensen $\overline{X} - \overline{Y}$ är normalfördelad med väntevärde $\mu_1 - \mu_2$ och varians $\sigma_1^2/n_1 + \sigma_2^2/n_2 = \sigma^2(1/n_1 + 1/n_2)$. Enligt Avsnitt 6.3.2 ledar

$$(\sigma^2)^\star = \Big(\sum_{j=1}^{n_1} (X_j - \overline{X})^2 + \sum_{j=1}^{n_2} (Y_j - \overline{Y})^2 \Big) / (n_1 + n_2 - 2)$$

till en väntevärdesriktig punktskatting av σ^2. Som i Avsnitt 7.2.2 får man att den stokastiska variabeln

$$\frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sigma^\star \sqrt{1/n_1 + 1/n_2}}$$

är t-fördelad med $n_1 + n_2 - 2$ frihetsgrader.

Resultat: Ett konfidensintervall för $\mu_1 - \mu_2$ med konfidensgrad $1 - \alpha$ är

$$I^\star_{\text{obs}} = \Big[\ (\overline{x} - \overline{y}) - \sigma^*_{\text{obs}} \sqrt{\tfrac{1}{n_1} + \tfrac{1}{n_2}} \ t_{\frac{\alpha}{2}}(n_1 + n_2 - 2); \ (\overline{x} - \overline{y}) + \sigma^*_{\text{obs}} \sqrt{\tfrac{1}{n_1} + \tfrac{1}{n_2}} \ t_{\frac{\alpha}{2}}(n_1 + n_2 - 2) \ \Big]$$

där $\sigma^*_{\text{obs}} = \sqrt{(Q_1 + Q_2)/(n_1 + n_2 - 2)}$ med $Q_1 = \sum_{j=1}^{n_1} (x_j - \overline{x})^2$, $Q_2 = \sum_{j=1}^{n_2} (y_j - \overline{y})^2$.

7.2.5 Stickprov i par

Inledande exempel: Två tekniker A och B både mäter n olika objekt. A får mätvärdena $x_1, \ldots, x_n$ däremot B får mätvärdena $y_1, \ldots, y_n$. Observera att i denna situation kan finnas skillander mellan de olika objekt.

Uppenbarligen är det rimligt att gruppera mätvärdena i par.

	Objekt			
	1	2	...	n
A	x_1	x_2	...	x_n
B	y_1	y_2	...	y_n

Antag att x_j kommer från $\mathrm{N}(\mu_j, \sigma_1)$ och y_j kommer från $\mathrm{N}(\mu_j + \Delta, \sigma_2)$.

Idén är att skillnader mellan μ_j motsvarar skillnader mellan de olika objekt. Δ står för den systematiska skillnaden mellan teknikernas mätmetod och σ_1, σ_2 för deras (inte nödvändigtvis lika) noggrannhet.

Trick: Betrakta $z_j = y_j - x_j$. Då kan $z_1, \ldots, z_n$ anses som ett stickprov från $\mathrm{N}(\Delta, \sigma_1^2 + \sigma_2^2)$ och vi kan använda våra resultat från Avsnitt 7.2.1 och 7.2.2.

Exempel 2: Vid en undersökning av alkohols inverkan på slumpvis utvalda personer fick man följande resultat (tid i sekunder):

person	1	2	3	4	5	6
före alkohol	0,15	0,10	0,10	0,25	0,25	0,05
efter alkohol	0,55	0,60	1,00	0,55	0,55	0,35

Antag att mätningar före och efter alkohol på person nr. j är en observation från $\mathrm{N}(\mu_j, \sigma_1)$ resp. $\mathrm{N}(\mu_j + \Delta, \sigma_2)$. Bestäm ett 95% konfidensintervall för Δ.

Lösning: Betraktar observationerna

$$z_j = \text{ mätning efter - mätning före alkohol på person nr. j.}$$

Dessa är observationer från $\mathrm{N}(\Delta, \sigma)$ där $\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$ är okänd. Vi beräknar $\overline{z} = 0,45$ och $s \approx 0,2345$. Enligt Avsnitt 7.2.2 får vi konfidensintervallet

$$I^*_{\text{obs}} = \Big[0,45 - \frac{0,2345}{\sqrt{6}} t_{0,025}(6-1);\ 0,45 + \frac{0,2345}{\sqrt{6}} t_{0,025}(6-1)\Big] \approx \Big[0,20; 0,70\Big].$$

Exempel 3: Vid en undersökning av hur hållfastheten hos cement beror på härdningstiden bestämdes hållfastheten hos provkroppar som härdats under 2 respektive 7 dagar. Man fick följande resultat:

Härdningstid	Hållfasthet (i kp/m^2)						
2 dagar	21,8	21,7	20,0	22,5	22,0	22,1	21,9
7 dagar	32,4	31,8	34,5	33,9	34,4	34,2	34,3

Hållfastheten vid båda härdningstiderna kan antas vara normalfördelad med samma standardavvikelse σ. Bestäm ett 95% konfidensintervall för skillnaden på den genomsnittliga hållfastheten.

Lösning: Vi beräknar för första stickprovet (2 dagars härdningstid) $\overline{x} = 21,714$ och $s_1^2 = 0,6381$. För andra stickprovet (7 dagars härdningstid) får vi $\overline{y} = 33,643$ och $s_2^2 = 1,1762$. Att "väga ihop" (spoola) s_1^2 och s_2^2 ger

$$(\sigma^2)^*_{\text{obs}} = \frac{6s_1^2 + 6s_2^2}{12} = \frac{s_1^2 + s_2^2}{2} \approx 0,90715 \implies \sigma^*_{\text{obs}} \approx 0,9524.$$

Eftersom $\overline{x} - \overline{y} = -11,929$, $\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = \sqrt{\frac{2}{7}} \approx 0,5345$ och $t_{\alpha/2}(n_1 + n_2 - 2) = t_{0,025}(12) = 2,179$ blir konfidensintervallet

$$I^*_{\text{obs}} = \Big[-11,929 - 0,5345 \cdot 0,9524 \cdot 2,179;\ -11,929 + 0,5345 \cdot 0,9524 \cdot 2,179\Big] \approx \Big[-13,1; -10,8\ \Big].$$

8 Hypotesprövning

Vi konfroterar en hypotes H_0 om en stokastisk fördelning med resultatet av ett experiment. Vi förkastar hypotesen H_0 om den förklarar resultatet "dåligt", d.v.s. att resultatet är osannolikt förutsatt att hypotesen H_0 är sann. Annars förkastar vi H_0 inte.

Anmärkning: Den typsika situtationen är att man litar på hypotesen H_0 så att man kräver starka argument innan man förkastar den.

Exempel 1 (ESP, extrasensory perception): En person påstår att han kan avgöra med förbundna ögon om krona eller klave har kommit upp vid kastet av ett mynt. Vi misstänker att personen bara gissar, han får alltså övertyga oss att han har rätt.

Vi vill pröva personens påstående p.g.a. följande experiment: Vi kastar ett mynt 12 gånger och noterar

$$X = \text{antalet rätta svar.}$$

Observera att $X \in \text{Bin}(12, p)$ där p är sannolikheten att personen ger rätt svar.

Vår hypotes är $H_0 : p = \frac{1}{2}$.

Vi är beredda att förkasta vår hypotes om personen ger tillräckligt många rätta svar. Hur många rätta svar a bör vi kräva? Vi vill kontrollera risken att personen övertyger oss trots att den bara gissar. Denna risk är

$$P(X \geq a \,|\, H_0 \text{ sann}).$$

Om H_0 är sann är $X \in \text{Bin}(12; \frac{1}{2})$ och

$$P(X \geq a) = \sum_{i=a}^{12} \binom{12}{i} (\tfrac{1}{2})^{12}$$

a	$P(X \geq a)$
12	0,00024
11	0,00317
10	0,01929
9	0,07300

Om vi är beredda att ta risken $\alpha = 0,05$ måste vi alltså kräva $a = 10$. Med andra ord förkastar vi H_0 om personen har minst 10 rätta svar utav 12. Annars förkastar vi H_0 inte, d.v.s. personen kunde inte övertyga oss.

Risken $\alpha = 0,05$ vi är beredda att ta kallas för testets signifikansnivå. Den observerade signifikansnivån $P(X \geq a \,|\, H_0 \text{ sann}) = 0,01929$ kallas för p-värdet.

8.1 Styrkefunktionen

Låt oss anta att vi testar hypotesen $H_0 : \theta = \theta_0$ mot $H_1 : \theta \in \Theta$. Styrkefunktionen för testet är

$$h(\theta) = P(\text{förkasta } H_0 \,|\, \theta \text{ sann})$$

där $\theta \in \Theta \cup \{\theta_0\}$.

Ett test är bra om H_0 förkastas med stor sannolikhet om den inte är sann, d.v.s att $h(\theta)$ är stor för $\theta \in \Theta$.

Observera att $h(\theta_0) = \alpha$.

Exempel 1 (forts.): Personen påstår att han kan avgöra resultatet av ett myntkast med sannolikhet $0,9$. Vi tar samma test som förut med

$$H_0 : p = \frac{1}{2}, \qquad H_1 : p = \frac{9}{10}.$$

Testets styrka är

$$h(\tfrac{9}{10}) = \sum_{i=10}^{12} \binom{12}{i} (\tfrac{9}{10})^i (1 - \tfrac{9}{10})^{12-i} \approx 0,89.$$

Sannolikheten är alltså ganska stor att man med testets hjälp kommer att upptäcka personens ESP.

8.2 Tillämpning till normalfördelning

Vi fokuserar på följande

Problem: **1.** Den bakomliggande fördelningen är en normalfördelning med väntevärdet μ och känd standardavvikelse σ, **2.** Vi testar hypotesen H_0: $\mu = \mu_0$.

Idé: Låt $0 \le \alpha \le 1$. För att testa H_0 tar vi fram ett kritiskt område C och förkastar hypotesen om teststorheten $\overline{X}$ hamnar i C. Vi väljer C sådant att

$$P(\ \overline{X} \in C \mid H_0 \text{ sann }) = \alpha.$$

α kallas för signifikansnivån av testet. Observera att $P(\ \overline{X} \in C \,|\, H_0 \text{ sann }\,)$ är sannolikheten att förkasta H_0 trots att H_0 är sann!

Under förutsättningen att H_0 är sann gäller $X_j \in \mathrm{N}(\mu_0, \sigma)$ och därför

$$\overline{X} \;=\; \frac{1}{n}\sum_{j=1}^{n} X_j \;\in\; \mathrm{N}\Big(\mu_0, \frac{\sigma}{\sqrt{n}}\Big).$$

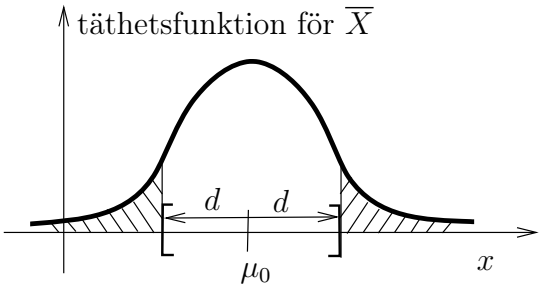
Eftersom denna normalfördelning är symmetrisk kring μ_0 tar vi det kritiska området av formen

$$C = \mathbb{R}\backslash[\mu_0 - d, \mu_0 + d]$$

med ett reelt tal d.

Vi väljer d så att

$$P(\overline{X} \in C) = \alpha.$$



$$\alpha \;\overset{!}{=}\; P(\overline{X} \in C) \;=\; P\big(\overline{X} \notin [\mu_0 - d, \mu_0 + d]\big) \;=\; P\big(\overline{X} < \mu_0 - d\big) + P\big(\mu_0 + d < \overline{X}\big)$$
$$= \;2P\big(\mu_0 + d < \overline{X}\big) \;=\; 2P\left(\frac{d}{(\sigma/\sqrt{n})} < \frac{\overline{X} - \mu_0}{(\sigma/\sqrt{n})}\right) \;=\; 2P\left(\frac{\sqrt{n}}{\sigma} d < Y\right)$$

där Y är standardnormalfördelad. Värdet $\lambda_{\frac{\alpha}{2}}$ där $P(\lambda_{\frac{\alpha}{2}} < Y) = \frac{\alpha}{2}$ kan tas från standardnormalfördelningen:s tabell. Vi får

$$d = \frac{\sigma}{\sqrt{n}} \lambda_{\frac{\alpha}{2}}.$$

Kom ihåg att $\overline{x}$ betecknar medelvärdet av experimentets mätvärden. Resonemanget ger då följande

Test: Förkasta H_0 om $\overline{x} \notin \Big[\mu_0 - \frac{\sigma}{\sqrt{n}}\lambda_{\frac{\alpha}{2}}, \mu_0 + \frac{\sigma}{\sqrt{n}}\lambda_{\frac{\alpha}{2}}\Big]$, bibehåll H_0 annars.

Anmärkning: Ovanstående metod kallas för tvåsidigt test.

Exempel 2: Låt X beskriva livslängden av en viss sorts maskindelar. Av någon anledning vet vi att $X \in \mathrm{N}(\mu, 15)$. Vår hypotes är

$$H_0: \quad \mu = 100.$$

Uppgiften är att testa H_0 med 5% signifikansnivå.

Vi genomför ett experiment där vi mäter livslängder hos 9 maskindelar och får följande mätvärden:

$$115,3;\ 106,5;\ 110,6;\ 106,9;\ 95,4;\ 115,1;\ 112,9;\ 107,7;\ 109,7.$$

Ger mätvärdena upphov för att förkasta H_0?

Lösning: För $\alpha = 0,05$ får vi $\lambda_{0,025} = 1,96$ från tabellen. Detta ger oss följande intervall:

$$\begin{aligned} I &= \Big[\ 100 - \frac{15}{\sqrt{9}}1,96\ ,\ 100 + \frac{15}{\sqrt{9}}1,96\ \Big] \\ &= [\ 100 - 9,8\ ;\ 100 + 9,8\ ] \\ &= [\ 90,2\ ;\ 109,8\ ]. \end{aligned}$$

Eftersom $\overline{x} = 108,9$ ligger i detta intervall kan vi **inte** förkasta H_0.

Ibland är man bara intresserad av ett ensidigt test. Här vill vi förkasta hypotesen H_0: $\mu = \mu_0$ endast i det fall att $\overline{x}$ ligger "långt ifrån" μ_0 på en viss sida om μ_0.

Problem: Testa hypotesen H_0: $\mu = \mu_0$ under bivillkoret att man förkastar H_0 endast i fallet $\overline{x} > \mu_0$.

Här söker vi d så att $P(\overline{X} > d) = \alpha$ förutsatt att H_0 är sann.

$$\alpha \overset{!}{=} P(d < \overline{X}) = P\Big(\frac{d - \mu_0}{(\sigma/\sqrt{n})} < \underbrace{\frac{\overline{X} - \mu_0}{(\sigma/\sqrt{n})}}_{= \eta \in \mathrm{N}(0,1)}\Big) \implies \frac{d - \mu_0}{(\sigma/\sqrt{n})} = \lambda_\alpha.$$

Vi förkastar alltså H_0 om

$$d = \mu_0 + \frac{\sigma}{\sqrt{n}}\lambda_\alpha < \overline{x}.$$

Exempel 2 (forts.): Testa ovanstående hypotes H_0: $\mu = 100$ under bivillkoret att man förkastar H_0 endast i fallet $\overline{x} > 100$.

Lösning: Enligt ovanstående resongement beräknar vi (observera att tabellen ger $\lambda_\alpha = 1,6449$ för $\alpha = 0,05$)

$$d \;=\; 100 + \frac{15}{\sqrt{9}} 1,6449 \;=\; 108,2245.$$

I experimentet har vi fått $\overline{x} = 108,9 > d$. Alltså förkastar vi H_0 den här gången.

8.3 Samband mellan signifikanstest och konfidensintervall

Vi antar fortfarande att stickprovet kommer från $\mathrm{N}(\mu, \sigma)$ där σ är känd. I det två-sidiga testet för hypotesen $H_0 : \mu = \mu_0$ med signifikansnivå α förkastade vi H_0 inte om

$$\begin{aligned} &\mu_0 - \frac{\sigma}{\sqrt{n}}\lambda_{\alpha/2} \;\le\; \overline{x} \;\le\; \mu_0 + \frac{\sigma}{\sqrt{n}}\lambda_{\alpha/2} \\ \iff\quad &\overline{x} - \frac{\sigma}{\sqrt{n}}\lambda_{\alpha/2} \;\le\; \mu_0 \;\le\; \overline{x} + \frac{\sigma}{\sqrt{n}}\lambda_{\alpha/2}. \end{aligned}$$

Testet kan alltså utföras på följande sätt: Man bestämmer (det två-sidiga) konfidensintervallet för μ_0 med konfidensgrad $1 - \alpha$ och ser efter om det hypotetiska värdet μ_0 ligger i intervallet. Man förkastar H_0 om det inte är fallet.

Vi säger att testet är utfört enligt konfidensmetoden.

Anmärkning: Vi kan använda alla våra resultat om konfidensintervall (en-sidigt konfidensintervall, konfidensintervall för μ om man har en normalfördelning där σ är okänd o.s.v.) för att konstruera signifikanstest.

8.4 χ^2-metoden

8.4.1 Test av fördelning

Betrakta ett experiment med r olika utfall $A_1, \ldots, A_r$. Sannolikheterna $P(A_1), \ldots, P(A_r)$ är okända men vi gör hypotesen

$$H_0 : P(A_1) = p_1, \;\ldots\; , P(A_r) = p_r.$$

Vi gör n oberoende upprepningar av experimentet och noterar hur ofta varje utfall A_j har inträffat. Resultatet sammanfattas i en tabell

A_1	A_2	$\ldots$	A_r	antal försök
x_1	x_2	$\ldots$	x_r	n

där x_j betecknar den absoluta frekvensen för A_j.

Mål: Vi vill p.g.a. detta experimentella data pröva H_0!

Exempel 3: Vi kastar en tärning 96 gånger och får 15 ettor, 7 tvåor, 9 treor, 20 fyror, 26 femmor och 19 sexor. Vi vill pröva hypotesen att tärningen är symmetrisk.

Låt alltså A_j, $j = 1, 2, \ldots, 6$, beteckna utfallet att man får j ögon. Hypotesen vi vill pröva är att $P(A_j) = 1/6$ gäller för alla j.

Låt X_j vara den s.v. som anger hur ofta händelsen A_j har inträffats. Under förutsättningen att H_0 är sann kan man visa att

$$Q^\star = \sum_{j=1}^{r} \frac{(X_j - np_j)^2}{np_j}$$

är approximativt χ^2-fördelad med $r-1$ frihetsgrader. Denna observation ledar till

χ^2-metoden: : Vi beräknar

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

För en signifikansnivå α förkastar vi H_0 om $Q^\star_{\text{obs}} > \chi^2_\alpha(r-1)$ och bekräftar H_0 annars.

Anmärkning: Tumregeln är att n är tillräckligt stort om $np_j \geq 5$ gäller för alla j.

Exempel 3 (forts.): $p_j = 1/6$, alltså $np_j = 96/6 = 16$. Med $x_1 = 15, x_2 = 7, x_3 = 9, x_4 = 20, x_5 = 26, x_6 = 19$ får vi

$$\begin{aligned} Q^\star_{obs} &= \frac{(15-16)^2}{16} + \frac{(7-16)^2}{16} + \frac{(9-16)^2}{16} + \frac{(20-16)^2}{16} + \frac{(26-16)^2}{16} + \frac{(19-16)^2}{16} \\ &= \frac{1}{16}(1^2 + 9^2 + 7^2 + 4^2 + 10^2 + 3^2) \;=\; 16 \end{aligned}$$

För $\alpha = 0,01$ överstiger detta resultat $\chi^2_{0,01}(6-1) = 15,1$ och vi förkastar hypotesen att tärningen är symmetrisk.

8.4.2 Homogenitetstest med χ^2-metoden

χ^2-metoden kan man också använda om man har genomfört två oberoende serier av experiment och vill pröva hypotesen att serierna är homogena, d.v.s att sannolikheterna $P(A_1), \ldots, P(A_r)$ är samma i varje serie. Observera att vi inte anger $P(A_1), \ldots, P(A_r)$ konkret.

Låt oss antar att vi har fått följande resultat:

serie	A_1	A_2	$\ldots$	A_r	antal försök
1	x_1	x_2	$\ldots$	x_r	n_1
2	y_1	y_2	$\ldots$	y_r	n_2
summa	S_1	S_2	$\ldots$	S_r	n

Homogenitetstest: För att skatta den gemensamma sannolikheten $p_j = P(A_j)$ från vårt datamaterial tar vi $p^\star_{j,\text{obs}} = S_j/n$. Låt

$$Q^\star_{\text{obs}} = \sum_{j=1}^{r} \frac{(x_j - n_1 p^\star_{j,\text{obs}})^2}{n_1 p^\star_{j,\text{obs}}} + \sum_{j=1}^{r} \frac{(y_j - n_2 p^\star_{j,\text{obs}})^2}{n_2 p^\star_{j,\text{obs}}}.$$

För en signifikansnivå α förkastar vi homogenitetshypotesen om $Q^\star_{\text{obs}} > \chi^2_\alpha(r-1)$.

Tumregel för god approximation: $n_i p^*_{j,\text{obs}} \geq 5$ för alla i, j.

Exempel 4: Slumpmässigt väljer man 556 par med barn och 260 utan barn som alla söker bostad. Av den första gruppen får 324 och av den andra 98 en lägenhet inom ett år. De övriga (232 i grupp 1 och 162 i grupp 2) får vänta längre.

Med beteckningarna

$$A_1 = \text{"väntetid högst ett år"}, \quad A_2 = \text{"väntetid längre än ett år"},$$

kan datamaterialet sammanfattas:

	A_1	A_2	antal par
med barn	324	232	556
utan barn	98	162	260
summa	422	394	816

Vi vill testa hypotesen att A_1, A_2 händer med samma sannolikhet inom båda grupper.

Enligt tabellen är

$$x_1 = 324,\ x_2 = 232,\ y_1 = 98,\ y_2 = 162 \text{ och } n_1 = 556,\ n_2 = 260.$$

Dessutom gäller $p^\star_{1,\text{obs}} = 422/816 \approx 0,517$ och $p^\star_{2,\text{obs}} = 394/816 \approx 0,483$ varmed vi kan beräkna

$$\begin{aligned} Q^\star_{\text{obs}} &= \frac{(324 - 556 \cdot 0,517)^2}{556 \cdot 0,517} + \frac{(232 - 556 \cdot 0,483)^2}{556 \cdot 0,483} \\ &\qquad + \frac{(98 - 260 \cdot 0,517)^2}{260 \cdot 0,517} + \frac{(162 - 260 \cdot 0,483)^2}{260 \cdot 0,483} \\ &\approx 30,0. \end{aligned}$$

Eftersom $Q^\star_{\text{obs}} > \chi^2_{0,001}(2-1) = 10,8$ kan vi förkasta homogenitetshypotesen med felrisk mindre än $0,001$.

Anmärkning: Testet kan utvidgas till fler serier av experiment, se [Blom et al.] för detaljer.

8.4.3 Oberoendetest med χ^2-metoden

I ett experiment mätar vi samtidigt två egenskaper A, B. Det finns r möjliga utfall för A nämligen $A_1, \ldots, A_r$ och s möjliga utfall för B nämligen $B_1, \ldots, B_s$, alltså totalt rs möjliga utfall

$$B_i \cap A_j, \quad i = 1, \ldots, s,\ j = 1, \ldots, r.$$

Man vill pröva hypotesen att egenskaperna A och B är oberoende.

Observera att vi inte antar att sannolikheterna $p_{ij} = P(B_i \cap A_j)$ är kända. Sannolikheterna för de enskilda egenskaperna kan beräknas genom

$$p_{\bullet j} = P(A_j) = \sum_{i=1}^{s} P(B_i \cap A_j) = \sum_{i=1}^{s} p_{ij}, \quad p_{i\bullet} = P(B_i) = \sum_{j=1}^{r} p_{ij}.$$

Att A och B är oberoende betyder

$$p_{ij} = p_{i\bullet} p_{\bullet j}$$

för alla i, j.

Efter n oberoende experimentet har vi fått de absoluta frekvenserna x_{ij} för utfallen $B_i \cap A_j$. Resultatet kan sammanfattas i en kontingenstabell:

	A_1	A_2	$\dots$	A_r
B_1	x_{11}	x_{12}	$\dots$	x_{1r}
B_2	x_{21}	x_{22}	$\dots$	x_{2r}
$\vdots$				
B_s	x_{s1}	x_{s2}	$\dots$	x_{sr}

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

Oberoendetest: De exakta sannolikheterna $p_{\bullet j}$, $p_{i\bullet}$ för de enskilda egenskaperna kan skattas med

$$p^*_{\bullet j,\text{obs}} = \sum_{i=1}^{s} x_{ij}/n, \quad p^*_{i\bullet,\text{obs}} = \sum_{j=1}^{r} x_{ij}/n.$$

Låt

$$Q^\star_{\text{obs}} = \sum_{i=1}^{s} \sum_{j=1}^{r} \frac{\big(x_{ij} - n\, p^*_{i\bullet,\text{obs}}\; p^*_{\bullet j,\text{obs}} \big)^2}{n\, p^*_{i\bullet,\text{obs}}\; p^*_{\bullet j,\text{obs}}}.$$

För en signifikansnivå α förkastar vi oberoendehypotesen om $Q^\star_{\text{obs}} > \chi^2_\alpha((r-1)(s-1))$.

Tumregel för god approximation: $n_i p^*_{j,\text{obs}} \geq 5$ för alla i, j.

Exempel 4 (forts.): Med betäckningarna

$$B_1 = \text{"med barn"}, \quad B_2 = \text{"utan barn"},$$

får vi kontingenstabellen

	A_1	A_2	summa
B_1	324	232	556
B_2	98	162	260
summa	422	394	$n = 816$

Eftersom $Q^\star_{\text{obs}} \approx 30,0 > 10,8 = \chi^2_{0,001}\big((2-1)(2-1)\big)$ kan vi förkasta med felrisk mindre än $0,001$ hypotesen att väntetid och förekomst av barn är oberoende.

Anmärkning: Numeriskt är oberoendetest och homogenitetstest (med s serier) identiska. För oberoendetestet antar man dock att datamaterialet är ett enda stickprov, däremot antar man för homogenitetstestet att materialet kommer från s försöksserier, d.v.s. man har s oberoende stickprov.

9 Linjär regression

9.1 Minsta-kvadrat-metoden (MK-metoden)

Exempel 1: Låt en kula falla n gånger och bestäm den tillryggalagda vägen efter en given tid t_0. Vi får ett stickprov $x_1, \ldots, x_n$ för

$$X = \text{fallsträckan av kulan.}$$

Vi är intresserade av X:s väntevärde (som uppenbarligen också beror på t_0). Enligt kända fysikaliska lagar antar vi att

$$E(X) = \theta \frac{t_0^2}{2}.$$

för ett reellt tal θ. Vi vill nu skatta θ p.g.a. vårt experimentella datamaterial.

Problem: För den stokastiska variablen X är väntevärdet $E(X)$ okänt men vi vet att det är av formen $m(\theta)$ där m är en känd funktion.

Mål: Skatta θ på grund av ett stickprov $x_1, \ldots, x_n$.

MK-metoden: Kvadratsumman

$$Q(\theta) = \sum_{j=1}^{n} \big(x_j - m(\theta)\big)^2$$

är en icke-negativ function av θ. Om det finns ett unikt värde θ^*_{obs} där $Q(\theta)$ antar sitt minsta värde så kallas θ^*_{obs} för MK-skattningen av θ.

Exempel 1 (forts.): Grafen till

$$\begin{aligned} Q(\theta) &= \sum_{j=1}^{n} \Big(x_j - \theta \frac{t_0^2}{2}\Big)^2 \\ &= \sum_{j=1}^{n} \Big(x_j^2 - 2\, x_j \theta \frac{t_0^2}{2} + \theta^2 \frac{t_0^4}{4}\Big) \\ &= \frac{n t_0^4}{4} \theta^2 - t_0^2 \Big(\sum_{j=1}^{n} x_j\Big) \theta + \Big(\sum_{j=1}^{n} x_j^2\Big) \end{aligned}$$

är en parabel med öppningen uppåt. Minimipunkten är nollstället av derivatan

$$\frac{d}{d\theta}Q(\theta) = \frac{nt_0^4}{2}\,\theta - t_0^2\left(\sum_{j=1}^{n} x_j\right).$$

MK-skattningen för θ är alltså

$$\theta^*_{\text{obs}} = \frac{\sum_{j=1}^{n} x_j}{nt_0^2/2} = \frac{2\overline{x}}{t_0^2}.$$

Anmärkning: Här har vi varken betraktat det allmänare fall att $x_1, \ldots, x_n$ är mätningar för olika variabler $X_1, \ldots, X_n$ eller diskuterat väntevärdesriktighet. Båda synpunkter tas upp i avsnittet om linjär regression.

9.2 Enkel linjär regression

Betrakta n par av värden

$$(x_1, y_1), \ldots, (x_n, y_n),$$

där $x_1, \ldots, x_n$ är givna storheter och $y_1, \ldots, y_n$ är mätningar till oberoende s. v. $Y_1, \ldots, Y_n$. Vi antar att $Y_j \in N(\mu_j, \sigma)$ (med samma σ) och att det linjära sambandet

$$\mu_j = \alpha + \beta x_j$$

gäller.

Mål: Skatta α och β!

Exempel: För att kalibrera en våg lägger vi successivt n kända vikter $x_1, \ldots, x_n$ på vågen och avläser resultaten $y_1, \ldots, y_n$ från vågens skala. Kunskap om vågens konstruktion berättigar oss att anta att

$$Y_j = \text{mätningsresultat för vikten } x_j$$

är normalfördelad enligt $N(\mu_j, \sigma)$ och att $\mu_j = \alpha + \beta x_j$.

För att skatta α, β med MK-metoden ska vi minimera

$$Q(\alpha, \beta) = \sum_{j=1}^{n} (y_j - \mu_j)^2 = \sum_{j=1}^{n} (y_j - \alpha - \beta x_j)^2.$$

Eftersom $Q(\alpha,\beta) \geq 0$ är grafen en paraboloid med unik minimipunkt som hittas genom att lösa det linjära systemet

$$0 = \frac{\partial Q}{\partial \alpha} = -2\sum_{j=1}^{n}(y_j - \alpha - \beta x_j) = 2\left(n\alpha + \beta\sum_{j=1}^{n} x_j - \sum_{j=1}^{n} y_j\right)$$
$$0 = \frac{\partial Q}{\partial \beta} = -2\sum_{j=1}^{n}(y_j - \alpha - \beta x_j)x_j = 2\left(\alpha\sum_{j=1}^{n} x_j + \beta\sum_{j=1}^{n} x_j^2 - \sum_{j=1}^{n} x_j y_j\right).$$

Lösningen blir

$$\boxed{\beta^*_{\text{obs}} = \frac{S_{xy}}{S_{xx}},\ \alpha^*_{\text{obs}} = \overline{y} - \beta^*_{\text{obs}}\overline{x}}$$

med

$$S_{xy} = \Big(\sum_{j=1}^{n} x_j y_j\Big) - n\overline{xy} = \sum_{j=1}^{n}(x_j y_j - x_j\overline{y} - \overline{x}y_j + \overline{xy}) = \sum_{j=1}^{n}(x_j - \overline{x})(y_j - \overline{y}),$$
$$S_{xx} = \Big(\sum_{j=1}^{n} x_j^2\Big) - n\overline{x}^2 = \sum_{j=1}^{n}(x_j^2 - 2x_j\overline{x} + \overline{x}^2) = \sum_{j=1}^{n}(x_j - \overline{x})^2.$$

Linjen

$$y = \alpha^*_{\text{obs}} + \beta^*_{\text{obs}} x$$

kallas för den skattade regressionslinjen. För varje givet x_0 kan linjen användas för att skatta tillhörande väntevärde $\mu_0 = \alpha + \beta x_0$ genom

$$\mu^\star_{0,\text{obs}} = \alpha^*_{\text{obs}} + \beta^*_{\text{obs}} x_0.$$

(denna procedur kallas skattning av en punkt på den teoretiska regressionslinjen). Notera att $\mu^\star_{0,\text{obs}} = \alpha^*_{\text{obs}}$ för $x_0 = 0$.

Anmärkning: a) Observera att de motsvarande stickprovsvariablerna kan skrivas

$$\beta^\star = \sum_{j=1}^{n} \frac{x_j - \overline{x}}{S_{xx}} Y_j,$$
$$\mu_0^\star = \sum_{j=1}^{n} \Big(\frac{1}{n} + \frac{(x_0 - \overline{x})(x_j - \overline{x})}{S_{xx}}\Big) Y_j$$

d.v.s. att de är linjära uttryck i Y_j. Eftersom Y_j är normalfördelade är alltså också $\beta^\star$, $\mu_0^\star$ normalfördelade.

b) Man kan utnyttja uttrycken i **a)** för att visa att skattningarna β^*_{obs}, $\mu^*_{0,\text{obs}}$ är väntevärdesriktiga, d.v.s.

$$E(\beta^\star) = \beta, \qquad E(\mu_0^\star) = \mu_0.$$

Behandlingen av variansen ger

$$V(\beta^\star) = \sigma^2/S_{xx}, \qquad V(\mu_0^\star) = \sigma^2\Big(\frac{1}{n} + \frac{(x_0-\overline{x})^2}{S_{xx}}\Big).$$

I grova drag betyder det att den blir mindre (och därmed skattningen bättre) om S_{xx} blir större.

c) Enligt **a)** och **b)** gäller

$$\beta^\star \in \mathrm{N}\Big(\beta, \sigma^2/S_{xx}\Big), \qquad \mu_0^\star \in \mathrm{N}\Big(\mu_0, \sigma^2\big(1/n + (x_0-\overline{x})^2/S_{xx}\big)\Big).$$

Som en konsekvens kan man använda våra resultat i Avsnitt 7.2.1 och 7.2.2 för att bestämma konfidensintervall för β, μ_0 och utföra hypotestest enligt konfidensmetoden.

Anmärkning: Metoden ger också nyttiga resultat om y bara beror approximativt linjärt på x.