SAMMENDRAG

Vil med stor sannsynlighet innholde skrivefeil.

Utgangspunkt: Har en "prosess" som gir resultater med iboende variasjon.

Eks:

- Levetid til Dataskjerm.
- Avlingsmengde.
- Terningkast.
- Utdeling av pokerhand.

Med andre ord: resultata vil variere.

Statistikk: Vil kunne besvare spørsmål som:

- Sannsynligheter for utfall terningkast kan gi oss?
- Hvilke pokerutfall skal vi satse på?
- Konklusjoner om gjennomsnittlig skjermlevetid?
- Beslutning om ny medisin gir lavere blodtrykk enn gammel medisin.
- Beslutning om sammenheng mellom avlingsmengde og gjødningsmengde.

Konklusjonene vi kan trekke vil ofte sjøl være usikre, av typen:

- 50% sannsynlighet for ett par.
- 95% sikre på at forventa skjermlevetid er innafor gitt intervall.
- Under 5% sannylighet for å få et tilsvarende resultat om den forventa skjermlevetida følger nullhypotesen.

Trenger:

- Kunne regne med sannsynligheter (**sannsynlighetsregning**).
- Kunne lage modeller som beskriver fenomenet (**stokastiske fordelinger, regresjonsmodell**).
- Kunne trekke konklusjoner fra modell og data (**statistisk inferens**).

GRUNNLEGGENDE SANNSYNLIGHETSREKNING

Observasjon av tilfeldig fenomen: **stokastisk forsøk**. Da gjelder:

- Kjenner mulige utfall.
- Resultat: kun ett av utfalla kan inntreffe.
- Resultatet er ukjent på forhand.

Vi definerer også:

Utfallsrommet (S) er mengden av alle mulige utfall (resultat) av det stokastiske (statistiske) forsøket.

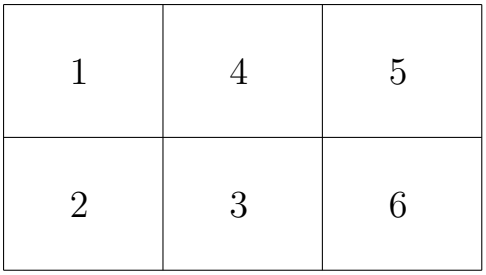
En hendelse (A) er en delmengde av utfallsrommet (ett eller fleire utfall) som oppfyller visse karakteristika.

Eks: Terningkast

$$S = \{1, 2, 3, 4, 5, 6\}$$

$$A = \{\text{Får minst } 4\} = \{4, 5, 6\}$$

Venn-diagram:



Komplementet til en hendelse A er alle utfall som ikke er med i A. Skriver A' (eller A^c, $\overline{A}$)

Snittet av to hendelser A og B er alle utfall som både er i A og i B. $A \cap B$.

Disjunkte hendelser: To hendelser A og B er disjunkte om snittet er tomt: $A \cap B = \emptyset$. De har altså ingen felles utfall.

Unionen av to hendelser A og B er alle utfall som enten er i A eller B eller i begge. $A \cup B$.

Kombinatorikkk:

Uniforme sannsynlighetsmodeller: Finner sannsynlighet ved antall gunstige delt på antall mulige. Kombinatorikk hjelper oss å telle disse.

Multiplikasjonsregelen: Forsøk i k etapper, med $m_1, m_2, \ldots, m_k$ mulige utfall i etappene. Totalt antall utfall:

$$m_1 \cdot m_2 \cdots m_k$$

Potensregelen: n merka enheter, velger k **med** tilbakelegg, antall **ordna** utfall:

$$n^k$$

Permutasjonsregelen: n merka enheter, velger k **uten** tilbakelegg, antall **ordna** utfall:

$$_nP_r = \frac{n!}{(n-k)!}$$

Kombinasjonsregelen: n merka enheter, velger k **uten** tilbakelegg, antall **ikke-ordna** utfall:

$$_nC_r = \binom{n}{k} = \frac{n!}{(n-k)! \cdot k!}$$

Generalisering: Kan dele ei samling med n merka enheter inn i r celler med n_1 enheter i første celle, n_2 i andre etc på

$$\binom{n}{n_1, n_2, \ldots, n_r} = \frac{n!}{n_1! \cdot n_2! \cdots n_r!}$$

mulige måter, der $n = n_1 + n_2 + \cdots n_r$.

(Tilsvarer antall permutasjoner av n enheter der n_1 er er av en type 1 (samme merking), n_2 av type 2, ..., og n_k av type k)

Sannsynlighet for hendelse:

Hovedsaklig 3 måter å sette sannsynlighet av en hendelse:

- Uniform sannsynlighetsmodell, d.v.s. alle utfall like sannsynlige. Har N mulige utfall, alle like sannsynlige:

$$P(A) = \frac{\#\text{ utfall som gir } A}{\#\text{ mulige utfall}} = \frac{\#\text{ gunstige}}{\#\text{ mulige}} = \frac{n}{N}$$

- Utfør mange (∞) forsøk, registrer relativ frekvens (andel):

$$P(A) \overset{n\to\infty}{=} \frac{\#\text{ forsøk som gir A}}{\text{totalt }\#\text{ forsøk, } n}$$

  (Store talls lov.)

- Subjektiv sasnnynlighet

Eks: Terningkast, $A = \{4, 5, 6\}$

- $$P(A) = \frac{\#g}{\#m} = \frac{3}{6} = \underline{0.50}$$

- 1000 forsøk, 520 A

$$P(A) \approx \frac{520}{1000} = \underline{0.52}$$

Regler for sannsynlighetsrekning:

3 **aksiomer** (grunnsetninger):

1. $0 \leq P(A) \leq 1$, $A \in S$
2. $P(S) = 1$ og $P(\emptyset) = 0$
3. Om $A_1, A_2, \ldots, A_n$ er parvis disjunkte (ingen felles elementer) så er

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n)$$

Addisjonsregelen for disjunkte hendelser.

Fra disse får vi:

- **Komplementregel**:

$$P(A') = 1 - P(A)$$

- Generell **addisjonsregel** (kan utvides til fleire hendelser):

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

Om A og B er **disjunkte**:

$$P(A \cup B) = P(A) + P(B)$$

Dersom $A_1, A_2, \ldots, A_n$ er en oppdeling av utfallsrommet S, så er

$$P(A_1 \cup A_2 \cup \cdots \cup A_n) = P(A_1) + P(A_2) + \cdots + P(A_n) = P(S) = 1$$

Eks: Terningkast

$$P(1) = P(2) = \cdots = P(6) = \frac{1}{6}$$

$$P(A) = P(4) + P(5) + P(6) = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \underline{\underline{\frac{3}{6}}}$$

Betinga sannsynlighet

$$P(A \mid B)$$

Sannsynligheten for hendelsen A forutsatt at hendelsen B inntreffer.

Eks: Terningkast, $A = \{4, 5, 6\}$ $\quad B = \{2, 4, 6\}$

$$P(A) = \frac{1}{2} \qquad P(A \mid B) = \frac{2}{3}$$

Definisjon:

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

Eks forts.

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{2}{6}}{\frac{1}{2}} = \frac{2}{3}$$

Regneregler for betinga sannsynlighet

- Multiplikasjonsregelen (kan utvides til fleire hendelser):

$$P(A \cap B) = P(A) \cdot P(B \mid A) = P(B) \cdot P(A \mid B)$$

- Når bare en av hendelsene $B_1, B_2, \ldots, B_k$ kan inntreffe:

$$P(A) = \sum_{i=1}^{k} P(B_i \cap A) = \sum_{i=1}^{k} P(B_i) \cdot P(A \mid B_i) \qquad \textbf{Total sannsylighet}$$

$$P(B_r \mid A) = \frac{P(B_r) \cdot P(A \mid B_r)}{P(A)}) = \frac{P(B_r) \cdot P(A \mid B_r)}{\sum_{i=1}^{k} P(B_i) \cdot P(A \mid B_i)}) \qquad \textbf{Bayesregel}$$

Uavhengighet: Om A, B **uavhengige**:

$$\begin{aligned} P(A \cap B) &= P(A) \cdot P(B) \qquad \text{(Kan utvides til fleire enn to hendelser)} \\ P(A \mid B) &= P(A) \\ P(B \mid A) &= P(B) \end{aligned}$$

STOKASTISK VARIABEL X:

En stokastisk variabel er en funksjon som knytter en bestemt tallverdi til hvert utfall i utfallsrommet. Utfallsrommet er **diskret** om X kan anta høgst et tellbart antall verdier, ellers **kontinuerlig**.

Punktsannsynligheten til en diskret stokastisk variabel X, $f(x)$, har følgende egenskaper:

$$\begin{array}{lllll} 1) & f(x) & \geq & 0, & \forall x \\ 2) & \sum\limits_{x} f(x) & = & 1 & \\ 3) & P(A) & = & \sum\limits_{x \in A} f(x) & \text{(summerer sannsynligheter)} \end{array}$$

Tilsvarende for en **sannsynlighetstetthet** (kontinuerlig stok. variabel):

$$\begin{array}{lllll} 1) & f(x) & \geq & 0, & \text{for enhver } x \in R. \\ 2) & \int\limits_{-\infty}^{\infty} f(x)\,dx & = & 1 & \\ 3) & P(a < X < b) & = & \int\limits_{a}^{b} f(x)dx & \text{(arealet under kurven fra } a \text{ til } b). \end{array}$$

Kumulativ fordelingsfunksjon:

$$F(x) = P(X \leq x) = \begin{cases} \sum\limits_{\substack{t \leq x \\ x}} P(X = t) & X \text{ diskret} \\ \int\limits_{-\infty} f(t)\,dt & X \text{ kontinuerlig} \end{cases}$$

Har alltid at

$$P(a < X \leq b) = F(b) - F(a)$$

Eks: Terningkast, $X = \#$ terningøyne

x	$f(x) = P(X = x)$	$F(x) = P(X \leq x)$
1	$\frac{1}{6}$	$\frac{1}{6}$
2	$\frac{1}{6}$	$\frac{2}{6}$
3	$\frac{1}{6}$	$\frac{3}{6}$
4	$\frac{1}{6}$	$\frac{4}{6}$
5	$\frac{1}{6}$	$\frac{5}{6}$
6	$\frac{1}{6}$	$\frac{6}{6}$

Simultanfordeling for to variabler

$$f(x,y) = P(X = x, Y = y) = P(X = x \cap Y = y)$$

Tilsvarende egenskaper for sannsynlighetstetthet og punktsannsynlighet som for univariate fordelinger, merk spesielt at:

$$P[(X,Y) \in A] = \begin{cases} \int\int_A f(x,y)\,dxdy & X \text{ kontinuerlig} \\ \sum\sum_A f(x,y) & X \text{ diskret} \end{cases}$$

Marginale fordelinger:

$$g(x) = \sum_y f(x,y) \qquad \text{og} \qquad h(y) = \sum_x f(x,y)$$

$$g(x) = \int_{-\infty}^{\infty} f(x,y)\,dy \qquad \text{og} \qquad h(y) = \int_{-\infty}^{\infty} f(x,y)\,dx$$

Betinga fordeling for Y gitt $X = x$:

$$f(y \mid x) = \frac{f(x,y)}{g(x)}, \quad g(x) > 0$$

Dersom X og Y er **uavhengige**:

$$\begin{aligned} f(x,y) &= g(x) \cdot h(y) \qquad \forall\ (x,y) \\ f(x \mid y) &= g(x) \\ f(y \mid x) &= h(y) \end{aligned}$$

Forventningsverdi:

- Sentralmål.
- Gjennomsnitt i det lange løp.
- Tyngdepunkt i fordelinga.

Diskret:

$$\mu = E(X) = \sum_x x \cdot f(x)$$

Kontinuerlig:

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

Eks: Terningkast, $X = \#$ terningøyne

$$E(X) = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \cdots + 6 \cdot \frac{1}{6} = \underline{3.5}$$

Funksjoner av stokastiske variable:

$$\mu_{g(X)} = E[g(X)] = \begin{cases} \sum_x g(x) \cdot f(x) & X \text{ diskret} \\ \int_{-\infty}^{\infty} g(x) \cdot f(x)\, dx & X \text{ kont.} \end{cases}$$

$$\mu_{g(X,Y)} = E[g(X,Y)] = \begin{cases} \sum_x \sum_y g(x,y) \cdot f(x,y) & X, Y \text{ diskrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) \cdot f(x,y)\, dxdy & X, Y \text{ kont.} \end{cases}$$

Eks: Terningkast, forventningsverdi til kvadratet av antall øyne:

$$E(X^2) = 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + \cdots + 6^2 \cdot \frac{1}{6} = \underline{15.167}$$

Varians:

- Spredningsmål
- $E[(X-\mu)^2]$, forventa kvadratavvik fra forventningsverdien.
- Kvadratet av standardavviket.

Diskret:

$$\sigma^2 = Var(X) = \sum_x (x-\mu)^2 f(x) = \sum_x x^2 \cdot f(x) - \mu^2 = E(X^2) - [E(X)]^2$$

Kontinuerlig:

$$\sigma^2 = Var(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)\,dx = \int_{-\infty}^{\infty} x^2 \cdot f(x)\,dx - \mu^2 = E(X^2) - [E(X)]^2$$

Funksjoner av stokastiske variable:

$$\begin{aligned} \sigma^2_{g(X)} &= Var[g(X)] = E\{[g(X) - \mu_{g(X)}]^2\} \\ &= \begin{cases} \sum_x [g(x) - \mu_{g(X)}]^2 \cdot f(x) & X \text{ diskret} \\ \int_{-\infty}^{\infty} [g(x) - \mu_{g(X)}]^2 \cdot f(x)\,dx & X \text{ kont.} \end{cases} \end{aligned}$$

Kovarians:

$$\begin{aligned} \sigma_{XY} &= Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] \\ &= \begin{cases} \sum_x \sum_y (x-\mu_X)(y-\mu_Y) \cdot f(x,y) & X \text{ diskr.} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y) \cdot f(x,y)\,dxdy & X \text{ kont.} \end{cases} \end{aligned}$$

Alternativ formel: $Cov(X,Y) = E(X \cdot Y) - \mu_X \cdot \mu_Y$

Dersom X og Y er to **uavhengige** stokastiske variabler er

$$E(X \cdot Y) = E(X) \cdot E(Y)$$

$$X \text{ og } Y \text{ uavhengige} \quad \Rightarrow \quad Cov(X,Y) = 0$$

Korrelasjon:

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(x) \cdot Var(Y)}}, \quad -1 \le \rho_{XY} \le 1$$

Forenklinger for lineærkombinasjoner:

For funksjoner av to stokastiske variabler X og Y:

$$E[g(X,Y) \pm h(X,Y)] = E[g(X,Y)] \pm E[h(X,Y)]$$

Dette medfører:

$$\begin{aligned} E[g(X) \pm h(Y)] &= E[g(X)] \pm E[h(Y)] \\ E[X \pm Y] &= E[X] \pm E[Y] \end{aligned}$$

Merk òg:

$$E(aX + b) = a\,E(X) + b$$
$$E(\sum_{i=1}^{n} a_i X_i + b) = \sum_{i=1}^{n} a_i\,E(X_i) + b$$

For to stokastiske variabler X og Y:

$$Var(a\,X + b\,Y) = a^2\,Var(X) + b^2\,Var(Y) + 2ab\,Cov(X,Y)$$

For to **uavhengige** variable X og Y har vi da:

$$Var(X \pm Y) = Var(X) + Var(Y)$$

Merk òg:

$$Var(a\,X + b) = a^2\,Var(X)$$

Om $X_1, X_2, \ldots, X_n$ er **uavhengige** stokastiske variabler får vi

$$Var(\sum_{i=1}^{n} a_i\,X_i + b) = \sum_{i=1}^{n} a_i^2\,Var(X_i)$$

Ikke-lineære funksjoner:

For ikke-lineære funksjoner kan tilnærma verdier for forventning og varians finnes ved **Taylor-utvikling**.

Tsjebysjeffs teorem:

Sannsynligheten for at en verdien på en variabel X ligger innafor k standardavvik fra forveningsverdien er minst $1 - \frac{1}{k^2}$:

$$P(\mu_X - k \cdot \sigma_X < X < \mu_X + k \cdot \sigma_X) \geq 1 - \frac{1}{k^2}$$

VANLIGE SANNSYNLIGHETSFORDELINGER:

Binomisk fordeling:

Krav:

1. n uavhengige delforsøk.
2. Registrerer suksess (A inntreffer) eller ikke i hvert delforsøk.
3. $P(A) = p$ i alle delforsøk.

En slik rekke med uavhengige identiske delforsøk med to utfall kalles gjerne **Bernoulli-forøksrekke** eller **binomisk forsøk/prosess**.

X, antall ganger A inntreffer på n forsøk, er **binomisk fordelt**. Skriver:

$$X \sim \text{binom}(n, p)$$

$$f(x) = b(x; n, p) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}, \quad x = 0, 1, \ldots, n$$

For utrekning av sannsynligheter brukes formelen over eller tabellene i formelsamlinga (for noen valg av n og p). Merk at tabellen gir kumulative sannsynligheter:

$$P(X \leq x) = \sum_{k=0}^{x} P(X = k)$$

Om n og p er kjente kan vi enkelt finne

$$E(X) = n \cdot p \qquad\qquad Var(X) = n \cdot p \cdot (1-p)$$

Når n er stor og p liten kan binomisk fordeling tilnærmes med Poisson-fordeling.

$$b(x; n, p) \approx p(x; \lambda t)$$

med $\mu = np = \lambda t$.

Når ***n* er stor** nok kan binomisk tilnærmes med normalfordeling.

$$P(X \leq x) \approx \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right)$$

Tommelfingerregel for når dette kan brukes er $np \geq 5$ og $n(1-p) \geq 5$. (Litt ulike varianter av denne, noen bruker $np(1-p) \geq 5$.)

Enda bedre tilnærming oppnås ved **kontinuitetskorreksjon**.

Dersom vi har k mulige utfall (ikke 2 som i binomisk), hver med sannsynlighet p_i, får vi **multinomisk** fordeling:

$$f(x_1, x_2, \ldots, x_k; p_1, p_2, \ldots, p_k, n) = \begin{pmatrix} n \\ x_1, x_2, \ldots, x_k \end{pmatrix} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

Hypergeometrisk fordeling:

Krav:

1. Populasjon med N element.
2. k av disse reknes som "suksess", $N - k$ som "fiasko".
3. Trekker n enheter **uten** tilbakelegging.

X, antallet "suksesser" vi trekker ut, blir da **hypergeometrisk** fordelt med punktsannsynlighet:

$$f(x) = h(x; N, n, k) = \frac{\begin{pmatrix} k \\ x \end{pmatrix} \cdot \begin{pmatrix} N-k \\ n-x \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}},$$
$$max(0, n - (N - k)) \leq x \leq min(n, k)$$

Forventning og varians kan finnes ved

$$E(X) = n \cdot p \qquad Var(X) = n \cdot p \cdot (1 - p) \cdot \frac{N - n}{N - 1}$$

der $p = k/N$.

Når populasjonen er stor i forhold til utvalget ($N >> n$), kan hypergeometrisk tilnærmes med binomisk fordeling:

$$h(x; N, n, k) \approx b(x; n, p)$$

Negativ-binomisk fordeling:

La den stokastiske variabelen X være antall forsøk du må gjøre for å oppnå at hendelsen A (suksess) skal inntreffe ***k* ganger** i ei Bernoulli-forsøksrekke. X har da **negativ binomisk** fordeling med punktsannsynlighet

$$f(x) = b^*(x; k, p) = \begin{pmatrix} x-1 \\ k-1 \end{pmatrix} \cdot p^k \cdot (1-p)^{x-k}, \quad x = k, k+1, k+2, \ldots$$

$$E(X) = \frac{k}{p} \qquad Var(X) = k \cdot \frac{1-p}{p^2}$$

Geometrisk fordeling:

La den stokastiske variabelen X være antall forsøk du må gjøre for å oppnå at hendelsen A (suksess) skal inntreffe **første** gang i ei Bernoulli-forsøksrekke. X har da **geometrisk** fordeling med punktsannsynlighet

$$f(x) = g(x; p) = p \cdot (1-p)^{x-1}, \quad x = 1, 2, 3, \ldots$$

$$E(X) = \frac{1}{p} \qquad Var(X) = \frac{1-p}{p^2}$$

Geometrisk fordeling er "minneløs".

Poisson-fordeling:

Antall forekomster av hendelsen A er **poissonfordelt** dersom

1. Antallet av A i disjunkte tidsintervall er uavhengige.
2. Forventa antall av A er konstant lik λ (raten) per tidsenhet.
3. Kan ikke få to forekomster samtidig.

("Tid" kan her representere tid, lengde, areal, volum etc.)
Om krava er oppfylt sier vi at vi har en **Poisson-prosess**. Da er X, antallet forekomster av A i et tidsrom t, poissonfordelt. Skriver:

$$X \sim \text{poisson}(\lambda t)$$

Punktsannsynlighet

$$f(x) = p(x; \lambda t) = \frac{(\lambda t)^x \cdot e^{-\lambda t}}{x!}, \qquad x = 0, 1, 2, \ldots$$

$$E(X) = \lambda t \qquad Var(X) = \lambda t$$

Skriver ofte punktsannsynligheten med $\mu = \lambda t$

$$f(x) = p(x; \mu) = \frac{\mu^x \cdot e^{-\mu}}{x!}, \qquad x = 0, 1, 2, \ldots$$

I tabellene er kumulative sannsynligheter for X

$$P(X \leq x) = \sum_{k=0}^{x} P(X = k)$$

oppgitt for noen verdier av $\mu = \lambda t$.

Normalfordelinga:

Har at X er en **normalfordelt** variabel med forventning μ og varians σ^2. Notasjon og tetthetsfunksjon:

$$X \sim \mathrm{N}(\mu, \sigma) \quad \Leftrightarrow \quad f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad -\infty < x < \infty$$

$$E(X) = \mu \qquad\qquad Var(X) = \sigma^2$$

Sannsynligheter gitt ved integral (areal):

$$P(a \leq X \leq b) = \int_a^b f(x)dx = F(b) - F(a)$$

Tabeller: Finnes tabeller for kumulativ fordeling til **standardnormalfordelinga**, andre normalfordelinger kan transformeres til denne:

$$X \sim \mathrm{N}(\mu, \sigma) \qquad \Leftrightarrow \qquad Z \sim \mathrm{N}(0, 1)$$

$$\text{der:} \qquad Z = \frac{X - \mu}{\sigma}$$

Z tilsvarer altså antall standardavvik X er fra μ.

$$\begin{aligned} F(x) &= P(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = P\left(Z \leq \frac{x-\mu}{\sigma}\right) \\ &= \Phi\left(\frac{x-\mu}{\sigma}\right) \qquad \text{(Kan slås opp i tabell.)} \end{aligned}$$

Kvantiler: Det finnes òg tabell spesielt for kvantiler z_α, disse er definert til å være:

$$P(Z > z_\alpha) = \alpha$$

Det er altså f.eks. 5% sannsynlighet for at Z er større enn $z_{0.05} = 1.645$.

Intervall: Om $X \sim \mathrm{N}(\mu, \sigma)$ er det $100(1-\alpha)\%$ sikkert at den får en verdi i intervallet:

$$\mu \pm z_{\frac{\alpha}{2}} \cdot \sigma$$

Dette er det vi lengre ut i kurset kalte **prediksjonsintervall** med kjente parametrer. Kalles òg spredningsintervall.

Lineærkombinasjoner: Anta at $X_1, X_2, \ldots, X_n$ er uavhengige og normalfordelte, $X_i \sim \mathrm{N}(\mu_i, \sigma_i)$.
Da vil

$$Y = a_1X_1 + a_2X_2 + \cdots + a_nX_n = \sum_{i=1}^{n} a_iX_i$$

være **normalfordelt** med

$$\begin{aligned} E(Y) = \mu_Y &= E\left(\sum_{i=1}^{n} a_iX_i\right) = \sum_{i=1}^{n} a_iE(X_i) = \sum_{i=1}^{n} a_i\mu_i \\ Var(Y) = \sigma_Y^2 &= Var\left(\sum_{i=1}^{n} a_iX_i\right) = \sum_{i=1}^{n} a_i^2Var(X_i) = \sum_{i=1}^{n} a_i^2\sigma_i^2 \end{aligned}$$

Uniform fordeling:

En (kontinuerlig) **uniformt** fordelt variabel har samme sannsynlighet for alle verdier innen et intervall. Generelt for en uniformt fordelt stokastisk variabel definert på intervallet $[A, B]$ har vi tetthetsfunksjonen

$$f(x) = \begin{cases} \frac{1}{B-A}, & A \le x \le B \\ 0, & \text{ellers} \end{cases}$$

$$E(X) = \frac{A+B}{2} \qquad Var(X) = \frac{(A-B)^2}{12}$$

Gammafordeling:

En kontinuerlig variabel X er **gammafordelt** med parametrer $\alpha > 0$ og $\beta > 0$ dersom tetthetsfunksjonen er gitt ved

$$f(x;\alpha,\beta) = \begin{cases} \frac{1}{\beta^\alpha\Gamma(\alpha)}x^{\alpha-1}e^{-\frac{x}{\beta}}, & x > 0 \\ 0, & \text{ellers} \end{cases}$$

$$E(X) = \alpha\beta \qquad Var(X) = \alpha\beta^2$$

Ventetida til hendelse nummer k i en Poisson-prosess vil være gammafordelt med $\alpha = k$ og $\beta = \frac{1}{\lambda}$.

Eksponensialfordeling:

Ventetida til første hendelse (og mellom etterfølgende hendelser) i en Poisson-prosess følger en eksponensialfordeling. En kontinuerlig variabel X har **eksponensialfordeling** med parameter $\beta > 0$ dersom tetthetsfunksjonen er gitt ved

$$f(x;\beta) = \begin{cases} \frac{1}{\beta}e^{-\frac{x}{\beta}}, & x > 0 \\ 0, & \text{ellers} \end{cases}$$

$$E(X) = \beta \qquad Var(X) = \beta^2$$

En viktig egenskap ved eksponensialfordelinga er at den er **minneløs**. Eksponensialfordelinga er gammafordeling med $\alpha = 1$.

De siste fordelingene (kikvadrat, t-tordeling og F-fordeling), brukes vanligvis ikke som fordeling for populasjonen, men er fordelinger for viktige observatorer når populasjonsfordelinga er normalfordelt:

Kikvadratfordeling:

En kontinuerlig stokastisk variabel X er **kikvadratfordelt** med ν frihetsgrader om tetthetsfunksjonen er gitt ved:

$$f(x;v) = \begin{cases} \frac{1}{2^{\frac{\nu}{2}}\Gamma(\frac{\nu}{2})}x^{\frac{\nu}{2}-1}e^{-\frac{x}{2}}, & x > 0 \\ 0, & \text{ellers} \end{cases}$$

Der ν er et heiltall større enn 0.

$$E(X) = \nu \qquad Var(X) = 2\nu$$

(Egentlig gammafordelinga med $\alpha = \frac{\nu}{2}$ og $\beta = 2$.)
Tabellene oppgir kvantiler χ^2_α for ulike valg av antall frihetgrader.

$$P(\chi^2 > \chi^2_\alpha) = \alpha$$

Fordelinga er tar bare positive verdier og er ikke symmetrisk ($\chi^2_{1-\alpha} \neq -\chi^2_\alpha$).

Sum: Anta at $X_1, X_2, \ldots, X_n$ er uavhengige og kikvadratfordelte med frihetsgrader $\nu_1, \nu_2, \ldots, \nu_n$.
Da vil

$$Y = X_1 + X_2 + \cdots + X_n = \sum_{i=1}^{n} X_i$$

være kikvadratfordelt med $\nu = \sum_{i=1}^{n} \nu_i$ frihetsgrader.

T-fordeling:

Har at X er **t-fordelt** med ν frihetsgrader om tetthetsfunksjonen er gitt ved

$$f(t;\nu) = \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})\sqrt{\pi\nu}}\left(1+\frac{t^2}{\nu}\right)^{-\frac{(\nu+1)}{2}}, \qquad -\infty < t < \infty$$

$$E(X) = 0 \qquad Var(X) = \frac{\nu}{\nu-2}, \quad \nu \geq 3$$

Om Z er en standardnormal-fordelt variabel og V er en kikvadratfordelt variabel med ν frihetsgrader vil, om Z er uavhengig av V,

$$T = \frac{Z}{\sqrt{V/\nu}}$$

bli t-fordelt med ν frihetsgrader.
T-fordelinga konvergerer mot standardnormalfordelinga når $\nu \to \infty$. Tommelfingerregel: $n \geq 30$.
Tabellene oppgir kvantiler t_α for ulike valg av antall frihetgrader.

$$P(T > t_\alpha) = \alpha$$

F-fordeling:

Har at X er **F-fordelt** med ν_1 og ν_2 frihetsgrader dersom tetthetsfunksjonen er gitt ved

$$f(x;\nu_1,\nu_2) = \begin{cases} \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)\left(\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1}{2}}}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \frac{x^{\frac{\nu_1}{2}-1}}{\left(1+x\frac{\nu_1}{\nu_2}\right)^{\frac{\nu_1+\nu_2}{2}}}, & f > 0 \\ 0, & f \leq 0 \end{cases}$$

Om U og V er to uavhengige kikvadratfordelte stokastiske variabler med henholdsvis ν_1 og ν_1 frihetsgrader vil

$$F = \frac{U/\nu_1}{V/\nu_2}$$

være **F-fordelt** med ν_1 og ν_2 frihetsgrader.
Tabellene oppgir kvantilene f_α for ulike valg av antall frihetsgrader.

$$P(F > f_{\alpha,\nu_1,\nu_2}) = \alpha$$

Fordelinga er, ulik standardnormal- og t-fordeling, ikke symmetrisk rundt 0 ($f_{1-\alpha,\nu_1,\nu_2} \neq -f_{\alpha,\nu_1,\nu_2}$). Men kan likevel finne $f_{1-\alpha,\nu_1,\nu_2}$ fra sammenhengen

$$f_{1-\alpha,\nu_1,\nu_2} = \frac{1}{f_{\alpha,\nu_2,\nu_1}}$$

INFERENS:

Statistisk inferens:

Statistisk inferens er å ved hjelp av data og modell (fordeling) trekke slutninger om (parametrer i) populasjonen. Som oftest tar vi utgangspunkt i et **tilfeldig utvalg** data fra populasjonen.

Tilfeldig utvalg:

La $X_1, X_2, \ldots, X_n$ være n uavhengige stokastiske variable som alle har sannsynlighetsfordeling $f(x)$. Vi sier da at $X_1, X_2, \ldots, X_n$ er et **tilfeldig utvalg** av størrelse n fra populasjonen med populasjonsfordelinga $f(x)$.
Et tilfeldig utvalg kan ses på som en mengde med uavhengige og identisk fordelte observasjoner.

Observatorer:

En observator er en funksjon av de stokastiske variablene i utvalget for et (tilfeldig) utvalg data. For eksempel $\overline{X}$ eller S^2 eller $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ eller $\frac{(n-1)S^2}{\sigma^2}$.

QQ-plott:

- Plotter utvalgskvantiler (observasjonene, ordna etter størrelse) mot teoretiske kvantiler ("ideelle observasjoner") fra aktuell fordeling.
- Teoretiske kvantiler er gitt ved invers kumulativ fordeling av "jevnt spredte" sannsynligheter mellom 0 og 1.
- Om antatt fordeling stemmer skal plottet gi tilnærma rett linje.

ESTIMERING:

Punktestimator

En **punktestimator** for en parameter θ er en observator (funksjon av utvalget), $\hat{\Theta}$. Denne skal gi oss gode anslag for (den ukjente) parameteren θ. Når vi setter inn verdier for observerte data får vi et punktestimat.

Viktige estimatoregenskaper:

En punktestimator $\hat{\Theta}$ er **forventningsrett** om

$$E(\hat{\Theta}) = \theta$$

Et rimelig krav til variansen $Var(\hat{\Theta})$ er at den bør synke med økende antall observasjoner i utvalget.

Om en har to ulike forventningsrette estimatorer $\hat{\Theta}_1$ og $\hat{\Theta}_2$ for en parameter, θ, er den med minst varians den **mest effektive estimatoren** for θ.

Noen vanlige estimatorer, standardsituasjoner:

$\boldsymbol{\mu}$

For et tilfeldig utvalg av størrelse n fra en populasjon med forventning μ og varians σ^2 er en estimator for μ gitt ved

$$\mathbf{\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i} \qquad E(\overline{X}) = \mu \qquad Var(\overline{X}) = \frac{\sigma^2}{n}$$

$\boldsymbol{\sigma^2}$

For et tilfeldig utvalg av størrelse n fra normalfordelt populasjon med forventning μ og varians σ^2 er en estimator for σ^2 gitt ved

$$\mathbf{S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2} \qquad E(S^2) = \sigma^2 \qquad Var(S^2) = \frac{2\sigma^4}{n-1}$$

$\boldsymbol{p}$

For et tilfeldig utvalg av størrelse n fra et binomisk forsøk (Bernoulli-forsøksrekke) med sannsynlighet p en estimator for p gitt ved

$$\mathbf{\hat{p} = \frac{X}{n}} \qquad E(\hat{p}) = p \qquad Var(\hat{p}) = \frac{p(1-p)}{n}$$

$\boldsymbol{\mu_1 - \mu_2}$

For to uavhengige tilfeldig utvalg av størrelser n_1 og n_2 fra populasjoner med forventninger μ_1, μ_2 og varianser σ_1^2, σ_2^2 er en estimator for $\mu_1 - \mu_2$ gitt ved

$$\mathbf{\overline{X}_1 - \overline{X}_2} \qquad E(\overline{X}_1 - \overline{X}_2) = \mu_1 - \mu_2 \qquad Var(\overline{X}_1 - \overline{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$\boldsymbol{\frac{\sigma_1^2}{\sigma_2^2}}$

For to tilfeldig utvalg av størrelser n_1 og n_2 fra normalfordelte populasjoner med forventninger μ_1, μ_2 og varianser σ_1^2, σ_2^2 er en estimator for $\frac{\sigma_1^2}{\sigma_2^2}$ gitt ved

$$\mathbf{\frac{S_1^2}{S_2^2}}$$

$p_1 - p_2$

For to uavhengige tilfeldig utvalg av størrelser n_1 og n_2 fra binomiske forsøk med sannsynligheter p_1 og p_2 er en estimator for $p_1 - p_2$ gitt ved

$$\hat{\mathbf{p}}_1 - \hat{\mathbf{p}}_2 = \frac{\mathbf{X_1}}{\mathbf{n_1}} - \frac{\mathbf{X_2}}{\mathbf{n_2}} \qquad E\left(\hat{p}_2 - \hat{p}_2\right) = p_1 - p_2$$

$$Var\left(\hat{p}_2 - \hat{p}_2\right) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

μ_D

For to parvise tilfeldige utvalg av størrelse n der differansene er fra populasjon med forventning μ_D og varians σ_D^2 er en estimator for μ_D gitt ved

$$\overline{\mathbf{D}} \qquad E(\overline{D}) = \mu_D \qquad Var(\overline{D}) = \frac{\sigma_D^2}{n}$$

Sannsynlighetsmaksimering:

Systematisk metode for å finne estimatorer for parametrer dersom fordelinga er kjent. Baserer seg på å velge de verdiene for parametrene som maksimerer sannsynligheten av observasjonene, denne sannsynligheten er gitt ved den såkalte likelihoodfunksjonen.
Gitt uavhengige observasjoner med sannsynlighetstetthet (eller punktsannsynlighet) $f(x;\theta)$ vil sannsynlighetsmaksimeringsestimatoren (SME), $\hat{\theta}$, være den verdi av parameteren θ som maksimerer

$$L(\theta) = L(x_1, x_2, \ldots, x_n; \theta) = f(x_1;\theta) \cdot f(x_2;\theta) \cdots f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta)$$

Vil da velge den parameterverdien som gjør at vi får størst sannsynlighet for å få de observasjonene vi har observert.

I praksis vil en som oftest finne verdien som maksimerer logaritmen til likelihoodfunksjonen, typisk ved derivasjon.

Utvalgsfordelinger:

En utvalgsfordeling er fordelinga for en observator (funksjon av de stokastiske variablene i utvalget) for et (tilfeldig) utvalg data. Vi er gjerne interessert i fordelinga til (de spesielle observatorerene som er) **estimatorer** for parametrer i populasjonen, ofte på standardisert form. Her er noen av **standardsituasjonene**:

$\overline{X}$, Z

For et tilfeldig utvalg av størrelse n fra en normalfordelt populasjon med forventning μ og varians σ^2 vil

$$\overline{X} \sim \mathrm{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad \text{og} \quad Z = \frac{\overline{X} - E(\overline{X})}{\sqrt{Var(\overline{X})}} = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim \mathrm{N}(0,1)$$

SGT: Sjøl om populasjonen ikke er normalfordelt vil resultatet over gjelde når $n \to \infty$. Rekner vanligvis tilnærminga for god når $n \geq 30$.

T

For et tilfeldig utvalg av størrelse n fra en normalfordelt populasjon med forventning μ og varians σ^2, der variansen estimeres ved S^2 fra utvalget har vi at

$$T = \frac{\overline{X} - \mu}{\frac{S}{\sqrt{n}}} = \frac{\frac{(\overline{X}-\mu)}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{S^2(n-1)}{\sigma^2(n-1)}}} \sim t_{n-1}$$

Altså t-fordelt med $n-1$ frihetgrader.
Dette vil også gjelde tilnærma for andre fordelinger med klokke-liknende form.
T vil bli tilnærma standardnormalfordelt for store n ($s \to \sigma$). Dette vil også gjelde uansett form på populasjonsfordelinga pga. SGT.

$\boldsymbol{S^2}$

For et tilfeldig utvalg av størrelse n fra en normalfordelt populasjon med forventning μ og varians σ^2 har vi at

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \sim \chi^2_{n-1}$$

Altså kikvadratfordelt med $n-1$ frihetsgrader. Forutsetninga om normalfordeling er her viktig.

$\boldsymbol{\hat{p}}$

For et tilfeldig utvalg av størrelse n fra et binomisk forsøk (Bernoulli-forsøksrekke) med sannsynlighet p har vi tilnærma at

$$Z = \frac{\hat{p} - E(\hat{p})}{\sqrt{Var(\hat{p})}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim \mathrm{N}(0,1)$$

forutsatt at n er stor nok.

$\boldsymbol{\overline{X}_1 - \overline{X}_2}$

For to uavhengige tilfeldig utvalg av størrelser n_1 og n_2 fra normalfordelte populasjoner med forventninger μ_1, μ_2 og varianser σ_1^2, σ_2^2 vil

$$\overline{X}_1 - \overline{X}_2 \sim \mathrm{N}\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

eller

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim \mathrm{N}(0,1)$$

Gjelder også tilnærma uten å forutsette normalfordeling om n_1, n_2 er store nok.

Om $\sigma_1^2 = \sigma_2^2$ og estimeres ved S_p^2 fra utvalga har vi at

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}$$

Om $\sigma_1^2 \neq \sigma_2^2$ som estimeres fra utvalga ved S_1^2 og S_2^2 har vi tilnærma at

$$T' = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim t_\nu$$

Altså tilnærma t-fordelt, der antall frihetsgrader er gitt ved

$$\nu = \frac{\left(s_1^2/n_1 + s_2^2/n_2\right)^2}{\frac{\left(s_1^2/n_1\right)^2}{n_1 - 1} + \frac{\left(s_2^2/n_2\right)^2}{n_2 - 1}}$$

F

For to uavhengige tilfeldig utvalg av størrelser n_1 og n_2 fra normalfordelte populasjoner med forventninger μ_1, μ_2 og varianser σ_1^2, σ_2^2 har vi at

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} \sim F_{n_1-1, n_2-1}$$

Altså F-fordelt med $n_1 - 1$ og $n_2 - 1$ frihetsgrader.

$\hat{p}_1 - \hat{p}_2$

For to uavhengige tilfeldig utvalg av størrelser n_1 og n_2 fra binomiske forsøk med sannsynligheter p_1 og p_2 har vi tilnærma at

$$Z = \frac{(\hat{p_1} - \hat{p_2}) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim \mathrm{N}(0, 1)$$

forutsatt at n_1 og n_2 er store nok.

$\overline{D}$

For to parvise tilfeldige utvalg av størrelse n der differansene er normalfordelte med forventning μ_D og varians σ_D^2, og der variansen estimeres ved S_D^2 fra utvalget, har vi at

$$T = \frac{\overline{D} - \mu_D}{\frac{s_D}{\sqrt{n}}} \sim t_{n-1}$$

OBS: Fra utvalgsfordelingene over kan vi utlede **konfidensintervall og testobservatorer** for parametrene.

Estimeringsfeil:

Når fordelinga til estimatoren er kjent (utvalgsfordelinga) kan vi rekne ut hvor stor feil vi (sannsynligvis) gjør i estimeringa. Vil ofte ha en viss sannsynlighet for at feilen ikke skal overskride en verdi e:

$$P(|\hat{\theta} - \theta| < e) = 1 - \alpha$$

Vanlige tilfeller:

$\overline{\boldsymbol{X}}$

For tilfeldig utvalg fra normalfordelt populasjon har vi at dersom vi skal være $100(1-\alpha)\%$ sikre på at estimeringsfeilen $|\overline{X} - \mu|$ ikke skal overskride e, må vi ha antall observasjoner n:

$$P(|\overline{X} - \mu| < e) = 1 - \alpha \qquad \Rightarrow \qquad n = \left(\frac{z_{\frac{\alpha}{2}}\sigma}{e}\right)^2$$

Av dette ser vi også at med n observasjoner kan vi være $100(1-\alpha)\%$ sikre på at estimeringsfeilen ikke vil overskride $e = z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$.

$\hat{\boldsymbol{p}}$

For et binomisk forsøk (Bernoulli-forsøksrekke) med sannsynlighet p har vi at dersom vi skal være $100(1-\alpha)\%$ sikre på at estimeringsfeilen $|\hat{p} - p|$ ikke skal overskride e, må vi ha antall delforsøk n:

$$P(|\hat{p} - p| < e) = 1 - \alpha \qquad \Rightarrow \qquad n \approx \left(\frac{z_{\frac{\alpha}{2}}\sqrt{\hat{p}(1-\hat{p})}}{e}\right)^2$$

Tilnærming da den ukjente p estimeres av $\hat{p}$ bestemt på basis av en forstudie. Har da også at med n observasjoner kan vi være $100(1-\alpha)\%$ sikre på at feilen ikke vil overskride $e \approx z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

Er også alltid **minst** $100(1-\alpha)\%$ sikkert at feilen ikke vil overskride e om

$$n = \left(\frac{z_{\frac{\alpha}{2}}}{2e}\right)^2$$

Da følger også at med n observasjoner er vi $100(1-\alpha)\%$ sikre på at feilen ikke vil overskride $e \leq \frac{z_{\frac{\alpha}{2}}}{2\sqrt{n}}$.

INTERVALLESTIMERING

Når fordelinga til estimatoren er kjent (utvalgsfordelinga) kan vi finne et intervall som vi med en viss sannsynlighet kan si inneholder verdien av en ukjent parameter.

Generelt KI for θ:

Vi har et tilfeldig utvalg $X_1, \ldots, X_n$ fra en populasjon med fordeling $f(x;\theta)$. Observerer verdier $x_1, \ldots, x_n$, og vil finne et $100(1-\alpha)\%$-KI for θ.

Finn observator $W(\theta)$ som har kjent fordeling og ikke avhenger av andre parametrer. (Her brukes vanligvis estimatoren for θ, $\hat{\theta}$ som utgangspunkt.) Generelt har vi at

$$P(w_{1-\frac{\alpha}{2}} < W(\theta) < w_{\frac{\alpha}{2}}) = 1-\alpha$$

der $w_{1-\frac{\alpha}{2}}$ og $w_{\frac{\alpha}{2}}$ er kvantiler. Denne løser vi mhp. θ slik at

$$P(\hat{\theta}_L < \theta < \hat{\theta}_U) = 1-\alpha$$

Setter inn for $x_1, \ldots, x_n$ og får at et $100(1-\alpha)\%$-KI er gitt ved

$$[\hat{\theta}_L\,,\,\hat{\theta}_U]$$

KI for θ ved normalfordelt estimator:

Dersom estimatoren $\hat{\Theta}$ er normalfordelt

$$\hat{\Theta} \sim \mathrm{N}\left(\theta, SE(\hat{\Theta})\right)$$

kan vi alltid finne $100(1-\alpha)\%$-KI for θ ved å bruke at

$$W(\theta) = \frac{\hat{\Theta}-\theta}{SE(\hat{\Theta})} \sim \mathrm{N}(0,1)$$

(OBS: Bruker ofte standardfeil (SE) for standardavviket til en estimator.) Har da at

$$P\left(-z_{\frac{\alpha}{2}} < \frac{\hat{\Theta}-\theta}{SE(\hat{\Theta})} < z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

som kan løses for θ:

$$P\left(\hat{\Theta} - z_{\frac{\alpha}{2}} \cdot SE(\hat{\Theta}) < \theta < \hat{\Theta} + z_{\frac{\alpha}{2}} \cdot SE(\hat{\Theta})\right) = 1-\alpha$$

(Om $SE(\hat{\Theta})$ er ukjent, og estimeres ved bruk av S, får en t-fordeling.)

Noen vanlige KI, standardsituasjoner:

Fra utvalgsfordelingene kan vi finne KI for endel aktuelle parametrer.
De samme tillempningene av forutsetningene som spesifisert under utvalgsfordelingene gjelder fremdeles.

$\boldsymbol{\mu}$

For et tilfeldig utvalg av størrelse n fra en normalfordelt populasjon med forventning μ og varians σ^2 har vi at

$$P\left(-z_{\frac{\alpha}{2}} < \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

Da blir et $100(1-\alpha)\%$-KI for μ gitt ved

$$\left[\overline{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\,,\, \overline{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

For et tilfeldig utvalg av størrelse n fra en normalfordelt populasjon med forventning μ og varians σ^2, der ukjent varians estimeres ved S^2 fra utvalget vil

$$P\left(-t_{\frac{\alpha}{2}} < \frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}} < t_{\frac{\alpha}{2}}\right) = 1-\alpha$$

med $n-1$ frihetsgrader. Da blir et $100(1-\alpha)\%$-KI for μ

$$\left[\overline{x} - t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}\,,\, \overline{x} + t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}\right]$$

$\boldsymbol{\sigma^2}$

For et tilfeldig utvalg av størrelse n fra en normalfordelt populasjon med forventning μ og varians σ^2 vil

$$P\left(\chi^2_{1-\frac{\alpha}{2}} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\frac{\alpha}{2}}\right) = 1-\alpha$$

med $n-1$ frihetsgrader. Da blir et $100(1-\alpha)\%$-KI for σ^2

$$\left[\frac{(n-1)s^2}{\chi^2_{\frac{\alpha}{2}}}\,,\, \frac{(n-1)s^2}{\chi^2_{1-\frac{\alpha}{2}}}\right]$$

$\underline{\boldsymbol{p}}$

For et tilfeldig utvalg av størrelse n fra et binomisk forsøk (Bernoulli-forsøksrekke) med sannsynlighet p har vi tilnærma (dersom n stor nok) at

$$P\left(-z_{\frac{\alpha}{2}} < \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\frac{\alpha}{2}}\right) = 1-\alpha$$

Dette gir et tilnærma $(1-\alpha)\cdot 100\%$-KI:

$$\left[\hat{p} - z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\,,\ \hat{p} + z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}\right]$$

Vanligvis erstattes p med $\hat{p}$ i uttrykka for grensene.

$\underline{\boldsymbol{\mu_1 - \mu_2}}$

For to uavhengige tilfeldig utvalg av størrelser n_1 og n_2 fra normalfordelte populasjoner med forventninger μ_1, μ_2 og varianser σ_1^2, σ_2^2 vil et $100(1-\alpha)\%$-KI for $\mu_1 - \mu_2$ bli gitt ved

$$\left[(\overline{x}_1 - \overline{x}_2) - z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\,,\ (\overline{x}_1 - \overline{x}_2) + z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right]$$

Om $\sigma_1^2 = \sigma_2^2$ (ukjente) estimeres fra utvalga ved S_p^2 har vi at et $100(1-\alpha)\%$-KI for $\mu_1 - \mu_2$ blir gitt ved

$$\left[(\overline{x}_1 - \overline{x}_2) - t_{\frac{\alpha}{2}}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\,,\ (\overline{x}_1 - \overline{x}_2) + t_{\frac{\alpha}{2}}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right]$$

Her har vi $n_1 + n_2 - 2$ frihetgrader.

Om $\sigma_1^2 \neq \sigma_2^2$ (ukjente) som estimeres fra utvalga ved S_1^2 og S_2^2 har vi at et $100(1-\alpha)\%$-KI for $\mu_1 - \mu_2$ blir gitt ved

$$\left[(\overline{x}_1 - \overline{x}_2) - t_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\,,\ (\overline{x}_1 - \overline{x}_2) + t_{\frac{\alpha}{2}}\sqrt{\frac{s_2^2}{n_1} + \frac{s_2^2}{n_2}}\right]$$

Der antall frihetsgrader ν er gitt fra utvalgsfordelinga.

$\underline{\boldsymbol{\frac{\sigma_1^2}{\sigma_2^2}}}$

For to uavhengige tilfeldig utvalg av størrelser n_1 og n_2 fra normalfordelte populasjoner med forventninger μ_1, μ_2 og varianser σ_1^2, σ_2^2 vil

$$P\left(f_{1-\frac{\alpha}{2},\nu_1,\nu_2} < \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} < f_{\frac{\alpha}{2},\nu_1,\nu_2}\right) = 1-\alpha$$

der antall frihetsgrader er $\nu_1 = n_1 - 1$ og $\nu_2 = n_2 - 1$.
Har da at et $100(1-\alpha)\%$-KI for $\frac{\sigma_1^2}{\sigma_2^2}$ er gitt ved

$$\left[\frac{s_1^2}{s_2^2}\frac{1}{f_{\frac{\alpha}{2},\nu_1,\nu_2}}\,,\ \frac{s_1^2}{s_2^2}f_{\frac{\alpha}{2},\nu_2,\nu_1}\right]$$

$\underline{\boldsymbol{p_1 - p_2}}$

For to uavhengige tilfeldig utvalg av størrelser n_1 og n_2 fra binomiske forsøk med sannsynligheter p_1 og p_2 har vi (om n_1 og n_2 store nok) at et tilnærma $100(1-\alpha)\%$-KI for $p_1 - p_2$ er gitt ved

$$\left[(\hat{p}_1-\hat{p}_2) - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\,,\ (\hat{p}_1-\hat{p}_2) + z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}+\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}\right]$$

$\underline{\boldsymbol{\mu_D}}$:

For to parvise tilfeldige utvalg av størrelse n der differansene er normalfordelte med forventning μ_D og varians σ_D^2, og der (ukjent) varians estimeres ved S_D^2, har vi at et $100(1-\alpha)\%$-KI for μ_D er gitt ved

$$\left[\overline{d} - t_{\frac{\alpha}{2}}\frac{s_D}{\sqrt{n}}\,,\ \overline{d} + t_{\frac{\alpha}{2}}\frac{s_D}{\sqrt{n}}\right]$$

Lengde av KI:

Lengda av et KI er øvre grense - nedre grense. For normalfordelt estimator med kjent varians kan vi rekne ut hva denne lengda vil bli for et gitt valg av n og α

$$L = \hat{\Theta} + z_{\frac{\alpha}{2}} \cdot SE(\hat{\Theta}) - [\hat{\Theta} - z_{\frac{\alpha}{2}} \cdot SE(\hat{\Theta})] = 2 \cdot z_{\frac{\alpha}{2}} \cdot SE(\hat{\Theta})$$

I forelesningene (og i boka) er det spesielt sett på to tilfelle:

1. Når vi nå konstruerer KI for μ basert på gjennomsnittsestimatoren $\overline{X}$ blir lengda

   $$L = 2 \cdot z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

   Vi kan bruke dette (løse for n) til å finne hvor mange data vi må ha for at KI-et skal bli av en viss lengde l:

   $$n = \left(\frac{2 z_{\frac{\alpha}{2}} \sigma}{l} \right)^2$$

   For tilfeller hvor σ^2 må estimeres kan vi ofte finne forventa lengde av intervallet (som funksjon av σ) sjøl om vi ikke kan finne den eksakte lengda. Se for eksempel t-intervall i oppgave 4 fra juni 2005 og χ^2-intervall i oppgave 1 fra september 2008.

2. Når vi nå konstruerer KI for p basert på andelsestimatoren $\hat{p}$ blir lengda

   $$L = 2 \cdot z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq \frac{z_{\frac{\alpha}{2}}}{\sqrt{n}}$$

   Om $\hat{p}$ er bestemt på basis av en forstudie oppnås omtrentlig lengde l om

   $$n = \left(\frac{2 z_{\frac{\alpha}{2}} \sqrt{\hat{p}(1-\hat{p})}}{l} \right)^2$$

   Oppnår alltid maksimal lengde l om

   $$n = \left(\frac{z_{\frac{\alpha}{2}}}{l} \right)^2$$

Prediksjonsintervall (PI):

Vil lage et intervall som er slik at det med en viss sannsynlighet inneholder verdien av en stokastisk variabel når vi gjør et nytt forsøk.

X_0:

Typisk vil vi ha et intervall som med en viss s.s. inneholder verdien av en ny observasjon X_0 når vi allerede har observert et tilfeldig utvalg $X_1, \ldots, X_n$ fra en normalfordelt populasjon:

$$X \sim \mathrm{N}(\mu, \sigma)$$

Dersom parametrene μ og σ er **kjente** blir dette rett og slett (kalles også spredningsintervall):

$$\begin{aligned} P\left(-z_{\frac{\alpha}{2}} < \frac{X_0 - \mu}{\sigma} < z_{\frac{\alpha}{2}} \cdot\right) &= 1 - \alpha \\ P\left(\mu - z_{\frac{\alpha}{2}} \cdot \sigma < X_0 < \mu + z_{\frac{\alpha}{2}} \cdot \sigma\right) &= 1 - \alpha \end{aligned}$$

Dersom **μ er ukjent**, og må estimeres ved $\overline{X}$ fra utvalget, må vi ta hensyn til variansen i $\overline{X}$. Tar utgangspunkt i differansen mellom en ny observasjon og gjennomsnittet. Veit at

$$X_0 - \overline{X} \sim \mathrm{N}\left(0, \sigma\sqrt{1 + \frac{1}{n}}\right)$$

Og dermed

$$P\left(-z_{\frac{\alpha}{2}} < \frac{(X_0 - \overline{X}) - 0}{\sigma\sqrt{1 + \frac{1}{n}}} < z_{\frac{\alpha}{2}}\right) = 1 - \alpha$$

Da blir et $100(1-\alpha)\%$-PI for X_0 gitt ved

$$\left[\overline{x} - z_{\frac{\alpha}{2}} \cdot \sigma\sqrt{1 + \frac{1}{n}}\,,\ \overline{x} + z_{\frac{\alpha}{2}} \cdot \sigma\sqrt{1 + \frac{1}{n}}\right]$$

Kan tilsvarende lage PI når også **σ^2 er ukjent** ved t-fordeling med $n - 1$ frihetsgrader:

$$\left[\overline{x} - t_{\frac{\alpha}{2}} \cdot s\sqrt{1 + \frac{1}{n}}\,,\ \overline{x} + t_{\frac{\alpha}{2}} \cdot s\sqrt{1 + \frac{1}{n}}\right]$$

TESTING

I statistisk hypotesetesting tar vi stilling til en påstand om egenskaper ved populasjonen. For eksempel **hypoteser** om parameteren θ

$$H_0 : \ \theta = \theta_0 \qquad\qquad H_1 : \ \theta > \theta_0$$

Om nullhypotesen forkastes eller ikke bestemmes ved hjelp av en **testobservator** utrekna fra utvalg data $X_1, X_2, \ldots, X_n$ fra populasjonen. Testobservatoren er valgt slik at den har kjent fordeling når nullhypotesen er sann. Dersom den utrekna testobservatoren gir en verdi som (i følge fordelinga) er veldig **usannsynlig** når nullhypotesen er sann, så forkastes H_0. Dette gir opphav til et **forkastingsområde** for testen, eller eventuelt en **p-verdi**.

Normalfordelt estimator:

Dersom vi har en normalfordelt forventningsrett estimator $\hat{\theta}$ for parameteren θ:

$$\hat{\Theta} \sim \mathrm{N}\left(\theta, SE(\hat{\Theta})\right)$$

kan vi konstruere testobservator

$$Z = \frac{\hat{\Theta} - \theta_0}{SE(\hat{\Theta})_0}$$

(Måler hvor mange standardavvik $\hat{\Theta}$ er fra θ_0.)
- Om H_0 er sann vil Z bli standardnormalfordelt.
- Om H_1 er sann vil vi forvente at $\hat{\Theta}$ blir større enn θ_0, og dermed at Z blir stor (høgresidig test).

Forkastingsområdet velges slik at det skal være en sannsynlighet α (**signifikansnivået**, ofte 5%) for å få en så stor verdi dersom H_0 er sann. Kritisk verdi blir da z_α og vi forkaster H_0 om $Z > z_\alpha$.

P-verdi:

Som et alternativ til å finne et forkastingsområde for Z kan vi finne en **p-verdi**. Dette er det minste signifikansnivået som akkurat ville gitt forkasting av nullhypotesen. Kan finnes som

$$\text{p-verdi} = P(\text{minst like ekstremt resultat som vi fikk} \mid H_0 \text{ korrekt})$$

Om denne blir liten nok forkastes H_0. Om vi velger grensa til å gå ved α vil dette gi akkurat samme resultat som tradisjonell testing.

Feil av type I/II:

$$\begin{aligned} P(\text{feil av type I}) &= P(\text{forkaste en korrekt nullhypotese}) = \alpha \\ P(\text{feil av type II}) &= P(\text{ikke forkaste en gal nullhypotese}) = \beta \end{aligned}$$

Vi ønsker gjerne å ha både α og β minst mulige. I praksis velges et nivå for α, mens β vil avhenge av den sanne verdien for parameteren.

$1 - \beta$ kalles for **styrken** til en test og sier hvor sannsynlig det er at vi forkaster nullhypotesen som en funksjon av den sanne parameterverdien. Ved å finne denne for ulike verdier av parameteren kan vi skissere en **styrkefunksjon**.

Ulike hypoteseoppsett:

Hypotesetester kan settes opp som høgresidige, venstresidige eller tosidige:

$$\begin{array}{ccc} H_0 & & H_1 \\ \theta = \theta_0 & \text{mot} & \theta > \theta_0 \\ \theta = \theta_0 & \text{mot} & \theta < \theta_0 \\ \theta = \theta_0 & \text{mot} & \theta \neq \theta_0 \end{array}$$

Hva slags type det er må avgjøres av spørsmålsstillinga. Spørmål som ikke indikerer en spesiell "retning", altså av typen "ulik", "forskellig fra" etc. gjør at tosidig test bør foretrekkes.

Noen vanlige tester, standardsituasjoner:

Fra utvalgsfordelingene kan vi finne fordelinga for testobservatorer for en-del aktuelle parametrer.
De samme tillempingene av forutsetningene som spesifisert under utvalgs-fordelingene gjelder fremdeles.

$\boldsymbol{\mu}$

For et tilfeldig utvalg av størrelse n fra en normalfordelt populasjon med forventning μ og (**kjent**) varians σ^2 får vi at

Testoppsett		Testobservator	Forkast når
$H_0: \mu = \mu_0$	$H_1: \mu > \mu_0$	$Z = \frac{\overline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$	$z > z_\alpha$
$H_0: \mu = \mu_0$	$H_1: \mu < \mu_0$	$Z = \frac{\overline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$	$z < -z_\alpha$
$H_0: \mu = \mu_0$	$H_1: \mu \neq \mu_0$	$Z = \frac{\overline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}$	$\lvert z\rvert > z_{\frac{\alpha}{2}}$

Dette fordi testobservatoren

$$Z = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

er standardnormalfordelt når H_0 er sann.

For et tilfeldig utvalg av størrelse n fra en normalfordelt populasjon med forventning μ og varians σ^2, der variansen (**ukjent**) må estimeres ved S^2 fra utvalget får vi

Testoppsett		Testobservator	Forkast når
$H_0: \mu = \mu_0$	$H_1: \mu > \mu_0$	$T = \frac{\overline{X}-\mu_0}{\frac{S}{\sqrt{n}}}$	$t > t_\alpha$
$H_0: \mu = \mu_0$	$H_1: \mu < \mu_0$	$T = \frac{\overline{X}-\mu_0}{\frac{S}{\sqrt{n}}}$	$t < -t_\alpha$
$H_0: \mu = \mu_0$	$H_1: \mu \neq \mu_0$	$T = \frac{\overline{X}-\mu_0}{\frac{S}{\sqrt{n}}}$	$\lvert t\rvert > t_{\frac{\alpha}{2}}$

der $t_{\frac{\alpha}{2}}$ er $\frac{\alpha}{2}$-kvantilen fra t-fordeling med $n-1$ frihetsgrader.
Dette fordi testobservatoren

$$T = \frac{\overline{X} - \mu_0}{\frac{S}{\sqrt{n}}}$$

er t-fordelt med $n-1$ frihetsgrader når H_0 er sann.

σ^2

For et tilfeldig utvalg av størrelse n fra en normalfordelt populasjon med forventning μ og varians σ^2 får vi at

Testoppsett		Testobservator	Forkast når
$H_0: \sigma^2 = \sigma_0^2$	$H_1: \sigma^2 > \sigma_0^2$	$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$	$\chi^2 > \chi^2_\alpha$
$H_0: \sigma^2 = \sigma_0^2$	$H_1: \sigma^2 < \sigma_0^2$	$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$	$\chi^2 < \chi^2_{1-\alpha}$
$H_0: \sigma^2 = \sigma_0^2$	$H_1: \sigma^2 \neq \sigma_0^2$	$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$	$\chi^2 < \chi^2_{1-\frac{\alpha}{2}}$ eller $\chi^2 > \chi^2_{\frac{\alpha}{a}}$

Dette fordi testobservatoren

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

er kikvadratfordelt med $n-1$ frihetsgrader når H_0 er sann.

p

For et tilfeldig utvalg av størrelse n fra et binomisk forsøk (Bernoulli-forsøksrekke) med sannsynlighet p får vi testoppsett

$$\begin{array}{ll} H_0: p = p_0 & H_1: p > p_0 \\ H_0: p = p_0 & H_1: p < p_0 \\ H_0: p = p_0 & H_1: p \neq p_0 \end{array}$$

Naturlig testobservator er her

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0 \cdot (1-p_0)}{n}}}$$

som er tilnærma standardnormalfordelt (for stor nok n) når H_0 er sann. Forkastingsgrensene blir som for vanlig Z-test.

For små utvalg må vi bruke binomisk fordeling. Vanlig testobservator er da

$$X = \text{antall suksesser} \sim \text{binom}(n, p_0)$$

når er nullhypotesen er sann.
For denne testen er det vanlig å basere konklusjonen på en p-verdi.

$\mu_1 - \mu_2$

For to uavhengige tilfeldig utvalg av størrelser n_1 og n_2 fra normalfordelte populasjoner med forventninger μ_1, μ_2 og varianser σ_1^2, σ_2^2 får vi testoppsett

$$H_0: \ \mu_1 - \mu_2 = d_0 \qquad H_1: \ \mu_1 - \mu_2 > d_0$$
$$H_0: \ \mu_1 - \mu_2 = d_0 \qquad H_1 :; \mu_1 - \mu_2 < d_0$$
$$H_0: \ \mu_1 - \mu_2 = d_0 \qquad H_1: \ \mu_1 - \mu_2 \neq d_0$$

Det vanligste er å teste med nullhypotesen "**ingen forskjell**",'dvs. $d_0 = 0$.

Dersom σ_1^2 og σ_2^2 er kjente vil testobservatoren

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

være standard-normalfordelt når H_0 er sann ($\mu_1 - \mu_2 = d_0$).
Forkastingsgrensene blir deretter som vanlig Z-test.

Dersom $\sigma_1^2 = \sigma_2^2$ (ukjent) må estimeres ved S_p^2 fra utvalga vil testobservatoren

$$T = \frac{(\overline{X}_1 - \overline{X}_2) - d_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

være t-fordelt med $n_1 + n_2 - 2$ frihetsgrader når H_0 er sann ($\mu_1 - \mu_2 = d_0$).
Forkastingsgrensene blir deretter som vanlig T-test.

Dersom $\sigma_1^2 \neq \sigma_2^2$ (ukjente) og må estimeres ved S_1^2 og S_2^2 fra utvalga vil testobservatoren

$$T' = \frac{(\overline{X}_1 - \overline{X}_2) - d_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

være tilnærma t-fordelt når H_0 er sann ($\mu_1 - \mu_2 = d_0$). Antall frihetsgrader er gitt fra utvalgsfordelinga.
Forkastingsgrensene blir deretter som vanlig T-test.

$\frac{\sigma_1^2}{\sigma_2^2}$

For to uavhengige tilfeldig utvalg av størrelser n_1 og n_2 fra normalfordelte populasjoner med forventninger μ_1, μ_2 og varianser σ_1^2, σ_2^2 får vi at

Testoppsett		Testobservator	Forkast når
$H_0: \sigma_1^2 = \sigma_2^2$	$H_1: \sigma_1^2 > \sigma_2^2$	$F = \frac{S_1^2}{S_2^2}$	$f > f_{\alpha,\nu_1,\nu_2}$
$H_0: \sigma_1^2 = \sigma_2^2$	$H_1: \sigma_1^2 < \sigma_2^2$	$F = \frac{S_1^2}{S_2^2}$	$f < f_{1-\alpha,\nu_1,\nu_2} = \frac{1}{f_{\alpha,\nu_2,\nu_1}}$
$H_0: \sigma_1^2 = \sigma_2^2$	$H_1: \sigma_1^2 \neq \sigma_2^2$	$F = \frac{S_1^2}{S_2^2}$	$f < \frac{1}{f_{\frac{\alpha}{2},\nu_2,\nu_1}}$ eller $f > f_{\frac{\alpha}{2},\nu_1,\nu_2}$

der $\nu_1 = n_1 - 1$ og $\nu_2 = n_2 - 1$. Dette fordi testobservatoren

$$F = \frac{S_1^2}{S_2^2}$$

er F-fordelt med $n_1 - 1$ og $n_2 - 1$ frihetsgrader når H_0 er sann.

$p_1 - p_2$

For to uavhengige tilfeldig utvalg av størrelser n_1 og n_2 fra binomiske forsøk med sannsynligheter p_1 og p_2 får vi testoppsett

$$\begin{aligned} H_0: &\quad p_1 - p_2 = d_0 \qquad & H_1: &\quad p_1 - p_2 > d_0 \\ H_0: &\quad p_1 - p_2 = d_0 \qquad & H_1: &\quad p_1 - p_2 < d_0 \\ H_0: &\quad p_1 - p_2 = d_0 \qquad & H_1: &\quad p_1 - p_2 \neq d_0 \end{aligned}$$

Vanligst å teste H_0 "**ingen forskjell**",'dvs. $d_0 = 0$.
Naturlig testobservator blir her

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - d_0}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \text{ om } \underline{d_0} = 0 \; \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

som er tilnærma standardnormalfordelt (for store nok n_1 og n_2) når H_0 er sann, og p, p_1 og p_2 er estimert fra data.
Forkastingsgrensene blir deretter som vanlig Z-test.

μ_D:

For to parvise tilfeldige utvalg av størelse n der differansene er normalfordelte med forventning μ_D og varians σ_D^2, og der variansen (ukjent) må estimeres ved S_D^2, setter vi opp

$$\begin{aligned} H_0: &\quad \mu_D = d_0 \qquad & H_1: &\quad \mu_D > d_0 \\ H_0: &\quad \mu_D = d_0 \qquad & H_1: &\quad \mu_D < d_0 \\ H_0: &\quad \mu_D = d_0 \qquad & H_1: &\quad \mu_D \neq d_0 \end{aligned}$$

Det vanligste er å teste nullhypotesen "**ingen forskjell**",'dvs. $d_0 = 0$.
Naturlig testobservator er her

$$T = \frac{\overline{D} - d_0}{\frac{S_D}{\sqrt{n}}}$$

som er t-fordelt med $n-1$ frihetgrader når H_0 er sann.
Forkastingsgrensene blir deretter som vanlig t-test.

Styrkefunksjon og utvalgsstørrelse:

Ved å finne sannsynligheten for å forkaste H_0 for ulike verdier av parameteren, θ, finner vi styrkefunksjonen:

$$\begin{aligned} 1-\beta(\theta) &= P(\text{forkaste } H_0 \,;\, \text{sann verdi} = \theta) \\ &= P(\text{Testobservator i forkastningsområdet} \,;\, \text{sann verdi} = \theta) \end{aligned}$$

For eksempel vil denne bli for vanlig høgresidig test for μ med kjent standardavvik:

$$\begin{aligned} 1-\beta(\mu) &= P(\text{forkaste } H_0 \,;\, \text{sann verdi} = \mu) \\ &= P\left(\frac{\overline{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}} > z_\alpha \,;\, \text{sann verdi} = \mu\right) \\ &= P\left(\overline{X} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} \,;\, \text{sann verdi} = \mu\right) \\ &= P\left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} > z_\alpha - \frac{\mu-\mu_0}{\frac{\sigma}{\sqrt{n}}} \,;\, \text{sann verdi} = \mu\right) \\ &= P\left(Z > z_\alpha - \frac{\mu-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right) = 1-\Phi\left(z_\alpha - \frac{\mu-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right) \end{aligned}$$

Ofte vil en gjerne ha en viss sannsynlighet, $1-\beta$, for at testen skal forkaste H_0 (en viss styrke), og kan da finne ut hvor mange observasjoner som trengs for å oppnå dette. Fra styrken over:

$$\begin{aligned} P\left(Z > z_\alpha - \frac{\mu-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right) &= 1-\beta \\ \Phi\left(z_\alpha - \frac{\mu-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right) &= \beta \\ \Rightarrow \quad n &= \frac{[z_\alpha - \Phi^{-1}(\beta)]^2\sigma^2}{(\mu-\mu_0)^2} = \frac{(z_\alpha + z_\beta)^2\sigma^2}{(\mu-\mu_0)^2} \end{aligned}$$

Tilsvarende utrekninger av styrke og utvalgsstørrelse kan gjøres for venstresidige og tosidige tester (se bok) når standardavviket er kjent.
Kan også finnes for andre tester/fordelinger, for eksempel binomisk for ulike verdier av p, kikvadrattest for ulike verdier av σ^2, etc.

ENKEL LINEÆR REGRESJON

Har observasjonspar $(x_1, y_1), \ldots, (x_n, y_n)$.

Regresjonsmodellen:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad i = 1, \ldots, n$$

Forutsetningene er at

$$E(\epsilon_i) = 0, \qquad Var(\epsilon_i) = \sigma^2, \quad i = 1, \ldots, n$$

og $\epsilon_1, \ldots, \epsilon_n$ er innbyrdes uavhengige.

Dette medfører at

$$E(Y_i) = \mu_{Y|x_i} = \beta_0 + \beta_1 x_i, \qquad Var(Y_i) = \sigma^2_{Y|x_i} = \sigma^2, \quad i = 1, \ldots, n$$

der $Y_1, \ldots, Y_n$ også er innbyrdes uavhengige.

MKM:

Bruker minste kvadrat-metoden til å estimere β_0 og β_1 fra data for å få ei estimert (tilpassa) regresjonslinje for forventningslinja $\beta_0 + \beta_1 x$:

$$\hat{y} = b_0 + b_1 x$$

MKM baserer seg på å minimere

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Gir estimatorer

$$B_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \frac{\sum_{i=1}^n (x_i - \overline{x}) Y_i}{\sum_{i=1}^n (x_i - \overline{x})^2} \qquad B_0 = \overline{Y} - B_1 \overline{x}$$

der

$$\begin{array}{lll} E(B_1) = \beta_1 & \qquad & Var(B_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2} \\ E(B_0) = \beta_0 & & Var(B_0) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \overline{x})^2} \end{array}$$

I tillegg har vi (forventningsrett) estimator for σ^2:

$$S^2 = \frac{SSE}{n-2} = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2} = \frac{\sum_{i=1}^n (Y_i - B_0 - B_1 x_i)^2}{n-2}$$

Parameterinferens:

Om vi antar at feilledda er normalfordelte

$$\epsilon_i \sim \mathrm{N}(0, \sigma)$$

vil

$$B_1 \sim \mathrm{N}\left(\beta_1, \frac{\sigma}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}}\right) \qquad B_0 \sim \mathrm{N}\left(\beta_0, \sigma \cdot \sqrt{\frac{\sum_{i=1}^n x_i^2}{n \cdot \sum_{i=1}^n (x_i - \overline{x})^2}}\right)$$

og KI og tester kan konstrueres ved å ta utgangspunkt i at

$$V = \frac{(n-2)S^2}{\sigma^2}$$

er kikvadratfordelt med **n − 2** frihetsgrader, og dermed at

$$T = \frac{B_1 - \beta_1}{\widehat{SE(B_1)}} = \frac{B_1 - \beta}{\frac{S}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}}} \qquad T = \frac{B_0 - \beta_0}{\widehat{SE(B_0)}} = \frac{B_0 - \beta_0}{S\sqrt{\frac{\sum_{i=1}^n x_i^2}{n \cdot \sum_{i=1}^n (x_i - \overline{x})^2}}}$$

vil være t-fordelte med **n − 2** frihetsgrader. Dette gir oss da KI og tester for β_1, β_0 og σ.

Inferens for $\mu_{Y|x_0}$:

Dersom x er en gitt verdi, x_0, vil en estimator for forventa verdi av responsen bli

$$\hat{Y}_0 = B_0 + B_1 x_0$$

Fordelinga for $\hat{Y}_0$ blir:

$$\hat{Y}_0 \sim \mathrm{N}\left(\mu_{Y|x_0}, \sigma\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}\right)$$

Og da kan vi ta utgangspunkt i at

$$T = \frac{\hat{Y}_0 - \mu_{Y|x_0}}{S\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}}$$

er t-fordelt med **n − 2** frihetsgrader til for eksempel å lage $100(1-\alpha)\%$-**konfidensintervall** for $\mu_{Y|x_0}$ (forventa verdi av responsen Y nå $x = x_0$):

$$\left[\hat{y}_0 - t_{\frac{\alpha}{2}} s\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}\ ,\ \hat{y}_0 + t_{\frac{\alpha}{2}} s\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}\right]$$

PI for Y_0:

Ønsker ofte intervall som med sannsynlighet $100(1-\alpha)\%$ vil inneholde verdien av en ny observasjon Y_0 (når $x = x_0$), altså et **prediksjonsintervall**. Tar utgangspunkt i forskjellen mellom estimatoren for forventa respons $\hat{Y}_0$ og verdien av en ny observasjon, Y_0

$$\hat{Y}_0 - Y_0 \sim \mathrm{N}\left(0, \sigma\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}\right)$$

Da kan vi bruke at

$$T = \frac{\hat{Y}_0 - Y_0}{S\sqrt{1 + \frac{1}{n} + \frac{(x_0-\overline{x})^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}}}$$

er t-fordelt med **n − 2** frihetsgrader.

Et $100(1-\alpha)\%$-PI for responsen Y_0 vil da finnes ved

$$\left[\hat{y}_0 - t_{\frac{\alpha}{2}}s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}\,,\ \hat{y}_0 + t_{\frac{\alpha}{2}}s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}}\right]$$

Betydning av lineær modell?

Har kvadratsumsoppsplittinga:

Total variasjon	=	Forklart variasjon	+	Restvariasjon
$\sum_{i=1}^{n}(y_i - \overline{y})^2$	=	$\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$	+	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$
SST	=	SSR	+	SSE

Forklaringsgraden er definert ved

$$R^2 = \frac{\text{Forklart variasjon}}{\text{Total variasjon}} = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}, \qquad 0 \le R^2 \le 1$$

Dette er et vanlig mål på hvor stor andel av variasjonen i responsen som kan tilskrives den lineære sammenghengen $\mu_{Y|x} = \beta_0 + \beta_1 x$. (Resten tilskrives tilfeldige feil, ϵ.)