Estimering

Viktige estimatoregenskaper:

Estimatoren $\hat{\Theta}$ bør være **forventningsrett**, DVS $E(\hat{\Theta}) = \Theta$, hvor Θ er et parameter du prøver å estiemre. **Variansen** til $\hat{\Theta}$ bør være synkende med økende antall observasjoner.

Om du har to estimatorer $\hat{\Theta}_1, \hat{\Theta}_2$ er estimatoren med minst varians den mest effektive estimatoren for Θ.

Noen vanlige estimatorer, standardsituasjoner:

μ: For et tilfeldig utvalg av størrelse *n* fra en populasjon med forventning μ og varians σ^2 er en estimator for μ gitt ved:
$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad E(\bar{X}) = \mu \quad Var(\bar{X}) = \frac{\sigma^2}{n}$

σ^2: For et tilfeldig utvalg av størrelse *n* fra en populasjon med forventning μ og varians σ^2 er en estimator for σ^2 gitt ved:
$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \; E(S^2) = \sigma^2 \; Var(S^2) = \frac{2\sigma^4}{n-1}$

p: For et tilfeldig utvalg av størrelse *n* fra et binomisk forsøk (Bernulli-forsøksrekke) med sannsynlighet *p*. En estimator for *p* er gitt ved
$\hat{p} = \frac{X}{n} \qquad E(\hat{p}) = p \qquad Var(\hat{p}) = \frac{p(1-p)}{n}$

$\mu_1 - \mu_2$: For to uavhengige utvalg av størrelser n_1, n_2 fra populasjoner med forventning μ_1, μ_2 og varianser σ_1^2, σ_2^2 er en estimator for $\mu_1 - \mu_2$ gitt ved
$\bar{X}_1 - \bar{X}_2 \; E(\bar{X}_1 - \bar{X}_2) = \mu_1 - \mu_2 \; Var(\bar{X}_1 - \bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$

$\frac{\sigma_1^2}{\sigma_2^2}$: For to tilfeldig utvalg av størrelser n_1, n_2 fra normalfordelte populasjoner med forventninger μ_1, μ_2 of varianser σ_1^2, σ_2^2 er en estimator for $\frac{\sigma_1^2}{\sigma_2^2}$ gitt ved: $\frac{S_1^2}{S_2^2}$

$p_1 - p_2$: For to uavhengige utvalg fra binomiske forsøk med sannsynligheter p_1, p_2 er en estimator for $p_1 - p_2$ gitt ved:
$\hat{p}_1 - \hat{p}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2} \; E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$
$Var(\hat{p}_1 - \hat{p}_2) = \frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}$

μ_D: For to parvise tilfeldig utvalg av størrlese *n* der differansene fra populasjonene med forventning μ_D og varians σ_D^2 er en estimator for μ_D gitt ved:

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

Utvalgsfordelinger

En utvalgsfordeling er fordelinga for en observator (funksjon av de stokastiske variablene i utvalget) for et (tilfeldig) utvalg data. Vi er gjerne interresert i fordelinga til (de spesielle observatorene som er) estimatorer for paramerere i populasjonen, ofte på standarisert form.

Standardsituasjonene (Utvalgsfordelinger)

$\bar{X}, Z$: For et tilfeldig utvalg av størrelse *n* fra en normalfordelt populasjon med forventning μ og varians σ^2 vil:
$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \qquad Z = \frac{\bar{X} - E(\bar{X})}{\sqrt{Var(\bar{X})}} = \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$

SGT: Selv om populasjonen ikke er normalfordelt vil resultatet over gjelde når $n \to \infty$. Regner vanlighvis tilnærminga for god når $n \geq 30$

T: For et tilfeldig utvalg av størrelser *n* fra en normalfordelt populasjon med forventning μ og varians σ^2, der variansen estimeres ved S^2 fra utvalget har vi at:
$T = \frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} = \frac{\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{S^2}{\sigma^2}}} \sim t_{n-1}$

S^2: For et tilfeldig utvalg av sørrelse *n* fra en normalfordelt populasjon med forventning μ og varians σ^2:
$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2_{n-1}$

$\hat{p}$: For et tilfeldig utvalg av størrelse *n* fra et binomisk forsøk med sannsynlighet *p* har vi tilnærmet at:
$Z = \frac{\hat{p} - E(\hat{p})}{\sqrt{Var(\hat{p})}} = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0, 1)$

$\bar{X}_1 - \bar{X}_2$: For to uavhengige tilfeldig utvalg n_1, n_2 (normalfordelt) med forventning μ_1, μ_2 og varianser σ_1^2, σ_2^2 vil:
$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$
Eller
$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$
Gjelder også for tilnærma uten å forutsette normalfordelt (stor n_1 og n_2)

For ukjent men lik varians $\sigma_1^2 = \sigma_2^2 = \sigma_p^2$ estimeres med S_p^2:
$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim T_{n_1+n_2-2}$

For ukjent og ulik varians $\sigma_1^2 \neq \sigma_2^2$, estimeres med S_1^2, S_2^2
$T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \sim T_\nu \quad \nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2 - 1}}$

F: For to uavhengige tilfeldige utvalg n_1, n_2 (normalfordelt) med forventning μ_1, μ_2 og varianser σ_1^2, σ_2^2 har vi at:
$F = \frac{\frac{S_1^2}{n_1}}{\frac{S_2^2}{n_2}} = \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} \sim F_{n_1-1, n_2-1}$

$\hat{p}_1 - \hat{p}_2$: med to tilfeldig utvalg n_1, n_2 (binomisk) med sannsynligheter p_1, p_2 har vi tilnærma at (om n_1, n_2 stor nok):
$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0, 1)$

$\bar{D}$: 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:

$$T = \frac{\bar{D} - \mu_D}{\frac{S_D}{\sqrt{n}}} \sim t_{n-1}$$

Intervallestimering

Konfidensintervall for μ (KI)

KI for μ med kjent σ^2:
$P\left(-z_{\frac{\alpha}{2}} < \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < z_{\frac{\alpha}{2}}\right) = 1 - \alpha$
$\Rightarrow \mu = \bar{X} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

KI for μ med S^2 som estimator for ukjent σ^2 med $n - 1$ frihetsgrader: $P\left(-t_{\frac{\alpha}{2}} < \frac{\bar{X}-\mu}{\frac{s}{\sqrt{n}}} < t_{\frac{\alpha}{2}}\right) = 1 - \alpha$
$\Rightarrow \mu = \bar{X} \pm t_{\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$

Konfidensintervall for σ^2 med $n - 1$ (KI)

$P\left(\mathcal{X}^2_{1-\frac{\alpha}{2}} < \frac{(n-1)S^2}{\sigma^2} < \mathcal{X}^2_{\frac{\alpha}{2}}\right)$
$$\Rightarrow \sigma^2 \in \left[\frac{(n-1)S^2}{\mathcal{X}^2_{\frac{\alpha}{2}}}, \frac{(n-1)S^2}{\mathcal{X}^2_{1-\frac{\alpha}{2}}}\right]$$

Konfidensintervall for *p*

KI for *p* med utvalg av størrelse *n* fra et binomisk forsøk med sannsynlighet *p* (OBS: $n > 30$ /stor nok)
$P\left(-z_{\frac{\alpha}{2}} < \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\frac{\alpha}{2}}\right) = 1 - \alpha$
$\Rightarrow p = \hat{p} \pm z_{\frac{\alpha}{2}}\sqrt{\frac{p(1-p)}{n}}$ Erstatt *p* i utrykket med $\hat{p}$ for å kunne regne ut

Konfidensintervall for $\mu_1 - \mu_2$

For to uavhengige tilfelig utvalgte utvalg av størrelser n_1 og n_2 fra normalfordelte populasjoner med forventning μ_1, μ_2 og vareanser σ_1^2, σ_2^2:
For **kjente** vareanser σ_1^2, σ_2^2:
$\mu_1 - \mu_2 = (\bar{x}_1 - \bar{x}_2) \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = 1 - \alpha$
For **ukjente men like** vareanser $\sigma_1^2 = \sigma_2^2$ med antall frihetsgrader $\nu = n_1 + n_2 - 2$:
$\mu_1 - \mu_2 = (\bar{x}_1 - \bar{x}_2) \pm s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 1 - \alpha$
For **ukjente men ulike** vareanser $\sigma_1^2 \neq \sigma_2^2$ der antall frihetsgrader ν er gitt fra utvalgsfordelinga:
$\mu_1 - \mu_2 = (\bar{x}_1 - \bar{x}_2) \pm t_{\frac{\alpha}{2}}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = 1 - \alpha$

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

Normalfordelt, med μ_1, μ_2 og σ_1^2, σ_2^2:
$P\left(f_{1-\frac{\alpha}{2},\nu_1,\nu_2} < \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} < f_{\frac{\alpha}{2},\nu_1,\nu_2}\right)$
$$\Rightarrow \frac{\sigma_1^2}{\sigma_2^2} \in \left[\frac{S_1^2}{S_2^2}\frac{1}{f_{\frac{\alpha}{2},\nu_1,\nu_2}}, \frac{S_1^2}{S_2^2}f_{\frac{\alpha}{2},\nu_1,\nu_2}\right]$$

Konfidensintervall for $p_1 - p_2$

Uavhengig tilfeldig utvalg n_1, n_2 fra binomisk forsøk med sannsynlighet p_1, p_2 har vi (om n_1, n_2 store nok):
$$(\hat{p}_1 - \hat{p}_2) \pm z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = 1 - \alpha$$

Konfidensintervall for μ_D

For to parvise tilfeldige utvalg, størrelse *n* med normalfordelte differanse med forventning μ_D og varians σ_D^2 (Variansen estimeres med S_D^2)
$\bar{d} \pm t_{\frac{\alpha}{2}}\frac{S_D}{\sqrt{n}} = 1 - \alpha$

Lengden av KI

For en normalfordelt estimator med kjent varians σ^2 kan vi regne ut hva lengden vil bli for et gitt valg av *n* og α
$L = 2 \cdot z_{\frac{\alpha}{2}} \cdot SE(\hat{\Theta})$, to typiske tilfeller:

1. Konstruere KI for μ basert på estimatoren: $\bar{X}$ blir lengda:
   $L = 2 \cdot z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} \Rightarrow n = \left(\frac{2z_{\frac{\alpha}{2}}\sigma}{L}\right)^2$
   Når σ^2 må estiemres kan vi finne approx forventa lengde
2. Når vi konstruerer KI for *p* basert på andelsestimatoren $\hat{p}$ ($l \approx L$):
   $L = 2 \cdot z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{l}} \leq \frac{z_{\frac{\alpha}{2}}}{\sqrt{n}} \Rightarrow n = \left(\frac{2z_{\frac{\alpha}{2}}\sqrt{\hat{p}(1-\hat{p})}}{l}\right)^2$

Estimeringsfeil

Når fordelinga til estimatoren er kjent (utvalgsfordelinga) kan vi regne ut hvor stor feil vi 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)
$\bar{X}$ (normalfordelt):
$P(|\bar{X} - \mu| < e = 1 - \alpha) \Rightarrow n = \left(\frac{z_{\frac{\alpha}{2}}\sigma}{e}\right)^2 \Rightarrow e = z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}$ $\hat{p}$ (binomisk):
$P(|\hat{p} - p| < e = 1 - \alpha) \Rightarrow n \approx \left(\frac{z_{\frac{\alpha}{2}}\sqrt{\hat{p}1-\hat{p}}}{e}\right)^2 \Rightarrow$
$$e \approx z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Noen regneregler

$(u \cdot v)' = u'v + u \cdot v'$
$\left(\frac{u}{v}\right) = \frac{u'v - u \cdot v'}{v^2}$
$\int \ln(x)\,dx = x\ln(x) - x + C$
u er et utrykk med *x*
$\int a^u\,dx = \frac{a^u}{u' \cdot \ln(a)} + C$
$\int \frac{1}{u} = \frac{\ln(x)}{u'} + C$
$\int u \cdot v'\,dx = u \cdot v - \int u'v\,dx$

$E(X)$ og $Var(X)$

$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y)$
$E(\chi^2_{n-1}) = n - 1$
$E(T_\nu) = 0$
$Var(T_\nu) = \frac{\nu}{\nu - 2}, \qquad \nu \geq 3$

Prediksjonintervall

Et intervall som sier noe om neste verdi X_0 i en normalfordelt populasjon: $X \sim N(\mu, \sigma)$
Kjent μ og σ (spredningsintervall)
$P\left(\mu - z_{\frac{\alpha}{2}} \cdot \sigma < X_0 < \mu + z_{\frac{\alpha}{2}} \cdot \sigma\right) = 1 - \alpha$
$\Rightarrow X_0 = \mu \pm z_{\frac{\alpha}{2}} \cdot \sigma$
Ukjent μ, kjent σ
$P\left(\bar{X} - z_{\frac{\alpha}{2}} \cdot \sigma\sqrt{1 + \frac{1}{n}} < X_0 < \bar{X} + z_{\frac{\alpha}{2}} \cdot \sigma\sqrt{1 + \frac{1}{n}}\right) = 1 - \alpha$
$\Rightarrow X_0 = \bar{x} \pm z_{\frac{\alpha}{2}} \cdot \sigma\sqrt{1 + \frac{1}{n}}$
Ukjent μ, σ frihetsgrad $\nu = n - 1$
$X_0 = \bar{x} \pm z_{\frac{\alpha}{2}} \cdot S\sqrt{1 + \frac{1}{n}}$
Formel for S_P^2 (Gjelder for alt på arket)
$$S_P^2 = \frac{S_1(n_1 - 1) + S_2^2(n_2 - 1)}{n_1 + n_2 - 2}$$

Fordelinger

Binomisk fordeling
1. *n* uavhengige delforsøk
2. To utfall: suksess/ikke suksess
3. Samme sannsynlighet for $p = P(a)$ (suksess) i alle delforsøk

Hypergeometrisk fordeling
1. Populasjon med *N* elementer
2. *k* av disse regnes som "suksess", $N - k$ som "fiasko"
3. Trekker *n* elementer uten tilbakelegging

Sannsynligheten *p* endrer seg mellom hvert delforsøk.

Negativ binomisk fordeling
Antall forsøk du må gjøre for at hendelsen *A* (suksess) skal intreffe *k* ganger

Geometrisk fordeling
Antall forsøk du må gjøre for at hendelsen *A* (suksess) skal intreffeførste gang

Poisson-fordeling
$\mu = \lambda t$, $\sigma^2 = Var(X) = \lambda t$
$f(x) = \frac{\mu^x}{x!}e^{-x}$
1. Antallet av *A* disjunkte tidsintervall er uavhengige
2. Forventa antall av *A* er konstant list λ (raten) per tidsenhet
3. Kan ikke få to forekomster samtidig

Gammafordeling
En kontinuerlig vareabel *X* er gammafordelt med parameter $\alpha > 0$ og $\beta > 0$ dersom tetthetsfunksjonen er gitt ved (se blå tabell). 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 handelser) i en Poisson-prosess følger en eksponensialfordeling. En kontinuerlig variabel *X* har eksponentialfordeling med parameter $\beta > 0$ dersom tetthetsfunksjonen er gitt ved (se tabellbok: $f(x)$). Eksponensialfordelinga er en variant av gammafordelinga med $\alpha = 1$

Stokastisk variabel

En variabel
$$\mu = E(X) = \begin{cases} \sum_x x \cdot f(x), & \text{deskret} \\ \int_{-\infty}^{\infty} x \cdot f(x)dx, & \text{kontinuerlig} \end{cases}$$
$$\sigma^2 = Var(X) = \begin{cases} \sum_x (x - \mu)^2 f(x), & \text{deskret} \\ \int_{-\infty}^{\infty} (x - \mu)^2 \cdot f(x)dx, & \text{kontinuerlig} \end{cases}$$
$F(x) = \int_0^x f(t)\,dt$

Funksjoner av stokastiske variabler
$$\mu_{g(X)} = E(g(X)) = \begin{cases} \sum_x g(x) \cdot f(x), & \text{deskret} \\ \int_{-\infty}^{\infty} g(x) \cdot f(x)dx, & \text{kontinuerlig} \end{cases}$$
$$\sigma^2_{g(X)} = Var(g(X)) = \begin{cases} \sum_x (g(x) - \mu_{g(X)})^2 \cdot f(x), & \text{deskret} \\ \int_{-\infty}^{\infty} (g(x) - \mu_{g(X)})^2 \cdot f(x)dx, & \text{kontinuerlig} \end{cases}$$
$$\mu_{g(X,Y)} = E(g(X, Y)) = \begin{cases} \sum_x \sum_x g(x, y)f(x, y), & \text{deskret} \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x, y)f(x, y)\,dx\,dy & \text{kontinuerlig} \end{cases}$$

Simultanfordeling for to variabler
$$P(X, Y) = \begin{cases} \sum\sum f(x, y), & \text{deskret} \\ \int\int f(x, y)\,dx\,dy, & \text{kontinuerlig} \end{cases}$$

Marginale fordelinger (to variabler)
$$g(x) = \begin{cases} \sum_y f(x, y) \\ \int_{-\infty}^{\infty} f(x, y)\,dy \end{cases}$$
$$h(y) = \begin{cases} \sum_x f(x, y) \\ \int_{-\infty}^{\infty} f(x, y)\,dx \end{cases}$$

Betinga fordeling (for Y gitt X)
$f(y|x) = \frac{f(x,y)}{g(x)}, \quad g(x) > 0$

Kovarians
$$\sigma_{XY} = Cov(X, Y) = \begin{cases} \sum_x \sum_y (x - \mu_X)(y - \mu_Y)f(x, y) \\ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y)f(x, y)\,dx\,dy \end{cases}$$
Korrelasjon:
$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{Var(x) \cdot Var(y)}}, \qquad -1 < \rho_{XY} < 1$

Sannsynlighetsregler

Addisjonsregelen:
$P(A \cup B) = P(A) + P(B) - P(A \cap B)$
Multiplikasjonsregelen:
$P(A \cap B) = P(A) \cdot P(B|A) = P(B) \cdot P(A|B)$
Betinga sannsynlighet: $P(A|B) = \frac{P(A \cap B)}{P(B)}$
Total sannsynlighet:
$P(A) = \sum_{i=1}^{n} P(B_i \cap A) = \sum_{i=1}^{k} P(B_i) \cdot P(A|B_i)$
Bayes setning: $P(B|A) = \frac{P(B) \cdot P(A|B)}{P(A)}$
Tilfeller ved uavhengighet:
$P(A \cap B) = P(A) \cdot P(B)$
$P(A|B) = A(A)$

Testing

Lager tester som tar stilling til påstander. Har da: nullhypotesen (H_0) og en utfordrende hypotese (H_1). Vi bruker testobservatorer for å kunne si nø om H_0 er sann eller ikke.

Normalfordelt estimator:
Dersom vi har en normalfordelt forventningsrett estimator $\hat{\Theta}$ for parameter Θ: $\hat{\Theta} \sim N(\Theta, SE(\hat{\Theta}))$
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.
Forkastningsområdet velges slik at med sannsynlighet α for å få en så stor verdi, dersom H_0 er sann. Kritisk verdi blir da z_α om $Z > z_\alpha$

P-verdi:
Som et alternativ fil å finne et forkastningsområde for *Z* kan man finne en $p - verdi$ som er det minste signifikansnivået som ville forkastet nullhypotesen.

Feil av type I/II:
P(Feil av type I) = P(Forkaste en rett H_0)=α
P(Feil av type II) = P(ikke forkaste gal H_0)=β
Det ideelle er å få α og β minst mulig.
$1 - \beta$ kalles for styrken til en test og sier hvor sannsynlig det er å forkaste H_0 som funskjon av den sanne parameterverdien. Ved å finne denne for ulike verdier av parameteren akn vi skissere en styrkefunksjon.

Ulike hypoteseoppsett:
Om vi har $H_0 \Rightarrow \hat{\Theta} = \Theta_0$ har vi tre tester. Venstresidig, tosidig og høyresidig. H_1 blir da respektivt:
$\hat{\Theta} < \Theta_0 \qquad \hat{\Theta} \neq \Theta_0 \qquad \hat{\Theta} > \Theta_0$
Hvis ikke en retning er indikert er en tosidig test å foretrekke.

Noen vanlige tester, standardsituasjoner
μ: fra normalfordelt utvalg av størrelse *n*, kjent σ^2.
Testobservator blir da (for alle testene): $Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$
Som forkaster når (VS Tosidig HS)
$z < -z_\alpha \qquad |z| > z_{\frac{\alpha}{2}} \qquad z > z_\alpha$

μ: fra normalfordelt utvalg av størrelse *n*, ukjent σ^2, bruker S^2. Testobservator vil bli $t - fordelt$ med $n - 1$ firhetsgrader
Testobservator blir da (for alle testene): $t = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}}$ Som forkaster når (VS Tosidig HS)
$t < -t_\alpha \qquad |t| > t_{\frac{\alpha}{2}} \qquad t > t_\alpha$

σ^2: fra normalfordelt utvalg av størrelse *n*. Testobservator vil bli *kjikvadratfordelt* (χ^2) med $n - 1$ frihetsgrader.
Testobservator for alle testene: $\chi^2 = \frac{(n-1)S^2}{\sigma_0^2}$
Som forkastes når: (VS Tosidig HS)
$\chi^2 < -\chi^2_{1-\alpha} \qquad \chi^2 < \chi^2_{1-\frac{\alpha}{2}} \vee \chi^2 > \chi^2_{\frac{\alpha}{2}}$
$\chi^2 > \chi^2_\alpha$

p for et tilfeldig utvalg av størrelse *n* fra et binomisk forsøk med sannsynlighet *p*. Hvis *n* stor nok ($n \geq 30$) kan tilnærmes normalfordelt.
Testobservator blir da: $Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$, grenser = grensene for μ
Hvis $n < 30$ må man bruke binomisk fordeling, vanlig testobservator er da $X =$ antall suksesser.
Vanlig å bruke **p-verdi** for dette tilfellet.

$\mu_1 - \mu_2$: For to uavhengige utvalg av størrelser n_1, n_2 (normalfordelt) med forventninger μ_1, μ_2 og varianser σ_1^2, σ_2^2. Dette gir H_0: $\mu_1 - \mu_2 = d_0$ (en vanlig test er $d_0 = 0$)
Kjent varians σ_1^2, σ_2^2: Testobservatoren blir da
$Z = \frac{(\bar{X}_1 - \bar{X}_2) - d_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$ forkasningsgrensene blir som en vanlig Z test (μ med ukjent σ^2)
ukjent og lik varians $\sigma_1^2 = \sigma_2^2 = \sigma^2$:
Testobservatoren blir $t - fordelt$ med frihetsgrad $n_1 + n_2 - 2$: $T = \frac{(\bar{X}_1 - \bar{X}_2) - d_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$
Forkastningsgrensene blir som en vanlig *T* test (μ med ukjent σ^2)
ukjent og ulik varians $\sigma_1^2 \neq \sigma_2^2$:
Testobservatoren blir *t-fordelt* med frihetsgrad gitt fra utvalgsfordelinga. Testobservator blir da: $T = \frac{(\bar{X}_1 - \bar{X}_2) - d_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$

$\frac{\sigma_1^2}{\sigma_2^2}$: for to uavhengige utvalg av størrelser n_1, n_2 (normalfordelt) med forventninger μ_1, μ_2 og varianser σ_1^2, σ_2^2. Testobservatoren blir da *f-fordelt* med firhetsgrader $n_1 - 1$ og $n_2 - 1$. Testobservatoren blir da (for alle testene): $F = \frac{S_1^2}{S_2^2}$ som forkastes når
(VS Tosidig HS)
$F < \frac{1}{F_{\alpha,\nu_1,\nu_2}} \qquad F_{\frac{\alpha}{2},\nu_1,\nu_2} < F < \frac{1}{F_{\frac{\alpha}{2},\nu_1,\nu_2}}$
$F > F_{\alpha,\nu_1,\nu_2}$

$p_1 - p_2$: For to uavhengige tilfeldige utvalg av størrelse n_1, n_2 (binomisk) med sannsynligheter p_1, p_2. Får $H_0 : p_1 - p_2 = d_0$. Testobservator blir
$\frac{(\hat{p}_1 - \hat{p}_2) - d_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$
Hvis $d_0 = 0$: $\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{p(1-p)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ Standardnormalfordelt ved stor nok n_1, n_2. Forkastningsgrense blir som en vanlig Z test (μ med kjent σ^2)

μ_D: Differansene er normalfordelt, variansen σ_D^2 må estimeres S_D^2. Får hypotesen $H_0 : \mu_D = d_0$
Testobservatoren er *t-fordelt* med $n - 1$ firhetsgrader.
Testobservatoren blir da: $T = \frac{\bar{D} - d_0}{\frac{S_D}{\sqrt{n}}}$.
Forkastningsområdet blir deretter som en vanlig T-test.

Styrkefunksjon

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

$$1 - \beta(\Theta) = P(\text{Forkaste } H_0, \text{sann verdi} = \Theta)$$
$$1 - \beta(\mu) = 1 - \Phi\left(z_\alpha - \frac{\mu - \mu_0}{\frac{\alpha}{\sqrt{n}}}\right) \qquad n = \frac{(z_\alpha - z_\beta)^2\sigma^2}{(\mu - \mu_0)}$$

Løningsforslag

Kombinatorikk

Multiplikasjonsregelen
Forsøk i k etapper, med $m_1, m_2 \cdots, m_k$ mulige utfall i etappene. Totalt antall utfall: $\prod_{i=1}^{k} m_i$
Potensregelen
n merka enheter, velger k med tilbakelegging. Antall ordna utfall: n^k
Permutasjonsregelen (nPr)
n merka enheter, velger k uten tilbakelegging, antall ordna utfall:
${}_nP_r = \frac{n!}{(n-k)!}$
Kombinasjonsregelen

n merka enheter, velger k uten tilbakelegging, antall ikke-ordna utfall:

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

Forenklinger for lineærkombinasjoner

$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2ab \cdot Cov(X, Y)$
Merk: hvis X og Y er uavhengige er $Cov(X, Y) = 0$
$E(aX + b) = aE(X) + b$
Tsjebysjeffs teorem: Sannsynligheten for at verdien på en variabel X ligger innafor k avstander fra forventningsverdien er minst $1 - \frac{1}{k^2}$
Sannsynlighetsmaksimeringsfunksjonen:
$L(\Theta) = \prod_{i=1}^{n} f(x_i, \Theta)$
Løses ved å:
1. Ta algoritmen av funksjonen
$\ln\left(\frac{a}{b}\right) = \ln(a) - \ln(b), \qquad \ln(ab) = \ln(a) + \ln(b)$
2. Deriver utrykker med hensyn på variabelen Θ oftest β
3. Sett utrykket lik 0 og løs med hensyn på variabelen.

Tilfeldig utvalg:
Når et utvalg er tilfeldig, kan vi se på det som en mengde identiske og uavhengige observasjoner.
QQ-plot:
- Plotter utvalgskvantiler (Observasjonene ordna etter størrelse) mot teoretiske kvantiler ("Ideelle observasjoner") fra aktuell fordeling
- Teoretiske kvantiler er gitt ver invers kumulativ fordeling "jevnt spredte" sannsynligheter mellom 0 og 1
- Om antatt fordeling stemmer skal plotter gi en tilnerma rett linje $x = y$

Enkel linær regresjon

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

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

Forutsetningene er at

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

og $\epsilon_1, \ldots, \epsilon_n$ er inbyrdes uavhengige. Dette medfører at
$E(Y_i) = \mu_{Y|x_i} = \beta_0 + \beta_1 x_i, \quad Var(Y_i) = \sigma^2_{Y|x_i} = \sigma^2, \quad i = 1, \ldots, n$
der $Y_1, \ldots, Y_n$ også er innbyrdes uavhengige.
MKM
Bruker miste 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

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

der

$$E(B_1) = \beta_1, \qquad Var(B_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

$$E(B_0) = \beta_0, \qquad Var(B_0) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

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 anntar at feilledda er normalfordelte

Oppgave

Ola samler spillekort. Ønsker et spesifikt spillekort (A). X er anntall kort han kjøper før han får dette. Det er 10 forskjellige spillekort i serien. Han kjøper pakker av kort med $n = 20$ kort i hver pakke. Hver pakke er uavhengig.
Hva er fordelinga for antall kort i pakka som er med spiller A?
Hva er sannsynligheten for at Ola får minst ett kort med spiller A i pakka?
Hva er forventa antall ulike spillekort i pakka?
Ola gjør $n = 20$ kjøp. Resultatet av hvert kjøp er uavhengige, og sannsynligheten for suksess (kort A) er lik $\frac{1}{10}$ for hvert kjøp. Da blir Y = tall på kort A, binomisk fordelt med $n = 20$ delforsøk og sannsynlighet $p = \frac{1}{10}$:
$Y \sim \text{binom}\left(n = 20, p = \frac{1}{10}\right)$
$P(Y \geq 1) = 1 - P(X = 0) = 1 - \binom{20}{0}\left(\frac{1}{10}\right)^0\left(1 - \frac{1}{10}\right)^{20} =$
$1 - 0.122 = 0.878$
Definerer korttypene $i \in \{A, B, \cdots, J\}$:
$I_i = \begin{cases} 1, & \text{minst ett kort av typen } i. \\ 0, & \text{ingen kort av typen } i \end{cases}$
Da blir
$E(I_i) = 1 \cdot P(I_i = 1) + 0 \cdot P(I_i = 0) = P(I_i = 1) = P(Y \geq 1) = 0.878$
La tallet på ulike kort være $N = \sum_i I_i$. Da blir
$E(N) = E(\sum_i I_i) = \sum_i E(I_i) = \sum_i 0.878 = 10 \cdot 0.878 = 8.78$

Oppgave

Fordelinga til en kontinuerlig stokastisk variabel X er gitt ved følgende sannsynlighetstetthetsfunksjon:

$$f(x) = \begin{cases} \frac{2}{9}x, & 0 < x < 3 \\ 0, & \textit{ellers} \end{cases}$$

Vi er interresert i følgende hendelser:

$$X = \{X > 1\} \qquad \textit{og} \qquad B = \{X > 2\}$$

er disse hendelsene disjunkte?
Hva er sannsynligheten for B?
Hva er sannsynligheten for B gitt A?
Er disse hendelsene uavhengige?
Nei, A og B er ikke disjunkte, det er mulig at $X > 1$ sammtidig som $X > 2$ $(X = 2.5)$

$$P(B) = P(X > 2) = \int_2^3 f(x)\,dx = \int_2^3 \frac{2}{9}x\,dx \left[\frac{1}{9}x^2\right]_2^3 = 1 - \frac{4}{9} = \frac{5}{9}$$

$$P(B|A) = \frac{P(B \cap A)}{P(A)} = \frac{P(A|B)P(B)}{P(A)} = \frac{1 \cdot P(B)}{P(A)} = \frac{\frac{5}{9}}{\int_1^3 \frac{1}{9}x\,dx} = \frac{5}{8}$$

Vi ser at $P(B|A) \neq P(B)$, så A og B er ikke uavhengige

Oppgave

Vi skal se på rekkevidde for en elektrisk bilmodell (kjørelengde fra betteriet er fullt til tomt). Det blir påstått at bilprodusenter oppgir urealistisk lang rekkevidde, derfor lar vi $n = 15$ tilfeldig valgte sjåfører bruke bilen til normal kjøring til batteriet er tomt, for hver sjåfør er kjørelengda (km) registrert. Vi går ut ifra at disse kjørelengdene er et tilfeldig utvalg fra en populasjon med forventning μ ($\hat{\mu} = 201.0667$) og varians σ^2 ($\hat{\sigma}^2 = 111.0667$).
Produsenten av bilmodellen påstår er forventa rekkevidde er under tilsvarende forhold 210 (km). Vi vil undersøke om det er grunnlag for å påstå at en reell rekkevidde er lavere enn dette.
Sett opp hypoteser og testobservator, og utfør en test for problemstillinga over med signifikansnivå 5%.
Finn både forkastningsområde og (tilnærma) p-verdi. Hva blir konklusjonen?
Vi går ut ifra at observatorene i et tilfeldig utvalg $X_1, \ldots, X_{15}$ fra en normalfordelt populasjon med forventning μ og standardavvik σ. Vi vil teste hypotesene:

$$H_0 : \mu = 210 \qquad H_1 : \mu < 210$$

Ettersom σ er ukjent og maa estimeres fra data ved S blir dette en T-test, testobservator:

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

som er *t-fordelt* med $n - 1 = 14$ firhetsgrader når H_0 er sann. Insett data:

$$t = \frac{201.067 - 210}{\frac{10.539}{\sqrt{n}}} = -3.282$$

Forkaster H_0 om $t < t_{0.02,14} = -1.761$. Så H_0 blir forkasta, og vi kan påstå at forventnia rekkevidde μ er under 210km.
Kan finne *p-verdien* fra:

$$\text{p-verdi} = P(T_{14} < -3.282) = \begin{cases} < 0.005 \\ > 0.001 \end{cases}$$

Så fra tabellen kan vi se at p-verdien er mellom 0.1% og 0.5%. Da dette er lavere enn signifikansnivået på 0.5% blir H_0 forkasta.

Deloppgave

Vi er og interresert i hvor mye rekkevidda kan variere.
Utled et 90%-konfidensintervall for variansen σ^2.
Finn intervallestimatet fra observatorene. Hva kan du konkludere fra intervallet?
Utleder et 90%-konfidensintervall fra variansen, σ^2, bruker at vi kjenner fordelinga for den standariserte observatoren

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$$

$$P\left(\chi^2_{0.95} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{0.05}\right) = 0.90$$

$$P\left(\frac{\chi^2_{0.95}}{(n-1)S^2} < \frac{1}{\sigma^2} < \frac{\chi^2_{0.05}}{(n-1)S^2}\right) = 0.90$$

$$P\left(\frac{(n-1)S^2}{\chi^2_{0.05}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{0.95}}\right) = 0.90$$

Intervallestimat fra observasjonene og $n - 1 = 14$ frihetsgrader

$$\left(\frac{(n-1)\cdot s^2}{\chi^2_{0.05}}, \frac{(n-1)\cdot s^2}{\chi^2_{0.05}}\right) = \left(\frac{(15-1)\cdot 111.067}{23.685}, \frac{(15-1)\cdot 111.067}{6.571}\right)$$

$$= (65.65, 236.65)$$

Med 90% sannsynlighet for at intervallet inneholder den sanne verdien for variansen σ^2. Merk at fra dette kan vi og lage et 90%-konfidensintervall for standardavviket σ:

$$(\sqrt{65.65}, \sqrt{236.65}) = (8.10, 15.38)$$

Deloppgave

Vi ønsker og å sammenligne modellem med en annen bilmodemm. Vi lar ei anna grupe $n_2 = 15$ sjåfører bruke denne andre modellen. Resultater: $\hat{\mu} = 184.467$, $\hat{\sigma}^2 = 115.124$. Vi går ut ifra at variansen for rekkevidde er lik for de to ulike populasjonene.
Gør de nødvendige forutsetningene for de to utvalga, og regn ut et estimert 95%-konfidensintervall for forskjell i forventa rekkevidde mellom de to modellene. Hva kan du konkludere fra dette intervallet?
Vi går ut ifra at vi har to tilfeldige utvalg $X_{1,1}, \ldots, X_{1,15}$ fra en normalfordelt populasjon med forventning μ_1, og $X_{2,1}, \ldots, X_{2,15}$ fra en normalfordelt populasjon med forventning μ_2. Vi går og utifra at de to utvalga er uavhengige. I teksten er det gjort en anntakelse om at variansen σ^2 er den samme for de to populasjonene.
Et 95%-konfidensintervall for $\mu_1 - \mu_2$ er da gitt ved

$$\left(\overline{X}_1 - \overline{X}_2 - t_{0.025} \cdot S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, \overline{X}_1 - \overline{X}_2 + t_{0.025} \cdot S_p \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

Estimatet for den felles variansen blir

$$s_p^2 = \frac{s_1^2 \cdot (n_1 - 1) + s_2^2 \cdot (n_2 - 1)}{n_1 + n_2 - 2} = \frac{111.067 \cdot 14 + 115.124 \cdot 14}{15 + 15 - 2}$$

Dermed blir intervallestimatet med insatte observasjoner:

$$201.067 - 194.467 \pm 2.048 \cdot \sqrt{113.095} \cdot \sqrt{\frac{1}{15} + \frac{1}{15}} = (-1.353, 14.553)$$

Her er det brukt $n_1 + n_2 - 2 = 28$ frihetsgradet i t-fordelinga.
Vi kan med 95% sannsynlighet konkludre at den sanne forventa rekkeviddedifferansen ligger innenfor intervallgrensene. Da 0 er innafor intervallet vet vi og at en tosidig test (5% signifikansnivå) ikke ville forkasta H_0 om ingen forskjell, kan ikke konkludere at forventa rekkevidde er ulike.