Sannsynlighetsfordelinger:

Binomisk Fordeling:

1. n uavhengige delforsøk
2. Suksess eller ikke
3. P(A)=p i alle forsøk

- X = Antall ganger A intreffer på n forsøk.
- $X \sim binom(n,p)$
- $f(x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x},\ x = 0, 1, 2, \ldots, n$
- $P(X \leq x) = \sum_{k=0}^{x} P(X = k)$
- $E(X) = np \quad Var(X) = np(1-p)$

Hypergeometrisk:

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

- X, antallet suksesser.
- $f(x) = \frac{\binom{k}{x} \cdot \binom{N-k}{n-x}}{\binom{N}{n}}$
- $E(X) = np \quad Var(X) = np(1-p)\frac{N-n}{N-1},\ p = k/N$

Negativ-Binomisk:

X er antall forsøk en må gjøre for at en hendelse A skal intreffe k ganger

- $f(x) = \binom{x-1}{k-1} \cdot p^x (1-p)^{x-k},\ x = k, k+1, k+2, \ldots$
- $E(X) = k/p \quad Var(x) = k \cdot \frac{1-p}{p^2}$

Geometrisk:

X er antall forsøk en må gjøre for at hendelsen A intreffer første gang.

- $g(x) = P(X = x) = p(1-p)^{X-1}$
- $E(X) = 1/p \quad Var(X) = \frac{1-p}{p^2}$

Geometrisk fordeling er minneløs!

Poisson:

Antall forekomster av hendelsen A er Poisson-fordelt hvis:

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

- X = antall forekomster av A i et tidsrom t
- $f(x) = \frac{(\lambda t)^x \cdot e^{-\lambda t}}{x!},\ x = 0, 1, 2, \ldots$
- $E(X) = \lambda t \quad Var(X) = \lambda t$
- $P(X \leq x) = \sum_{k=0}^{x} P(X = k)$
- Ventetida til hendelse k er gammafordelt med $\alpha = k$ og $\beta = 1/\lambda$
- Ventetida til første hendelse,og mellom etterfølgende hendelser, er eksponensialfordelt

Uniform fordeling:

En kontinuerlig uniformt fordelt variabel, har samme sannsynlighet for alle verdier innen et intervall. Generelt har vi tetthetsfunksjonen: $f(x) = \begin{cases} \frac{1}{B-A}, & A \leq x \leq B \\ 0, & \text{ellers} \end{cases}$

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

Gammafordeling:

En kontinuerlig variabel X er gammfordelt med parameter $\alpha > 0$ og $\beta > 0$ dersom tetthetsfunksjonen er gitt ved: $f(x; \alpha, \beta) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}, & x > 0 \\ 0, & \text{ellers} \end{cases}$

- $E(X) = \alpha\beta \quad Var(X) = \alpha\beta^2$
- $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-\alpha} dt$
- $\Gamma(\alpha) = (\alpha - 1)!$

Eksponensialfordeling:

- $f(x; \beta) = \begin{cases} \frac{1}{\beta} e^{-\frac{x}{\beta}}, & x > 0 \\ 0, & ellers \end{cases}$
- $E(X) = \beta \quad Var(X) = \beta^2$

Eksponensialfordelinga er minneløs!

Normalfordeling:

- $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- $P(a \leq X \leq b) = \int_a^b f(x)dx$

Standard normalfordeling:

- Alle normalfordelinger kan skrives som Standard normalfordeling
- $Z = \frac{X-\mu}{\sigma}$
- $F(x) = F(X \leq x) = P\left(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}\right) = P\left(Z \leq \frac{x-\mu}{\sigma}\right) = \Phi\left(\frac{x-\mu}{\sigma}\right)$

Anta at $X_1, X_2, \ldots, X_n$ er uavhengige og normalfordelt. Da er: $Y = \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_n X_n$ Være normalfordelt med:

- $E(Y) = \sum_{i=1}^{n} \alpha_i \mu_i \quad Var(Y) = \sum_{i=1}^{n} \alpha_i^2 \sigma_i^2$

Inferens:

QQ-Plot:

- Plotter observasjoner mot teoretiske("ideelle") observasjoner fra en aktuell fordeling.
- Teoretiske observasjoner er gitt ved invers kumulativ fordeling av jevnt spredte Sannsynlighetsfordelinger mellom 0 og 1.
- Om antatt fordeling stemmer skal plottet gi tilnermet rett linje.

Estimering:

Viktige estimatoregenskaper:

- En punktestimator Θ for en paramaeter θ er forventningsrett hvis $E(\Theta) = \theta$

- Variansen $Var(\Theta)$ burde synke med økende antall observasjoner
- Om en har to ulike estimatorer, så er den estimatoren med minst varians den mest effektive estimatoren.

Vanlige estimatorer:

Alle estimatorene vist til her er forventningsrett.

- μ: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad E(\overline{X}) = \mu \quad Var(\overline{X}) = \frac{\sigma^2}{n}$
- σ^2: $S^2 = \frac{1}{1-n}\sum_{i=1}^{n}(X_i - \overline{X})^2 \quad E(S^2) = \sigma^2 \quad Var(S^2) = \frac{2\sigma^4}{n-1}$
- p: $\hat{p} = \frac{X}{n} \quad E(\hat{p}) = p \quad Var(\hat{p}) = \frac{p(1-p)}{p}$ Binomisk
- $\mu_1 - \mu_2$: $\overline{X_1} - \overline{X_2} \quad Var(\overline{X_1} - \overline{X_2}) = \frac{\sigma_1^2}{n_1} - \frac{\sigma_2^2}{n_2}$
- $\frac{\sigma_1^2}{\sigma_2^2}$: $\frac{S_1^2}{S_2^2}$
- $p_1 - p_2$: $\hat{p_1} - \hat{p_2}$, Binomisk
- $\mu_D : \overline{D}$

Utvalgsfordelinger:

$\overline{\mathbf{X}}$, Z:

$\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \quad Z = \frac{\overline{X} - E(\overline{X})}{\sqrt{Var\overline{X}}} = \frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$ Selv om populasjonen ikke er normalfordelt gjelder dette når $n \to \infty$.Regner vanligvis tilnærmina for god når $n > 30$

Sentralgrenseteoremet:

Når utvalgsstørrelsen $N \to \infty$ så vil $\overline{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ for uansett fordeling av X. Godkjener dette for $N \geq 30$

T

Hvis ukjent varians: $T = \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$ Dette gjelder tilnærmet andre fordelinger som har klokkeliknende form.

$\mathbf{S^2}$:

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

$\hat{\mathbf{p}}$:

Binomisk forsøk med sannsynlighet p, gitt at n er stor nok: $Z = \frac{\hat{p}-E(\hat{p})}{\sqrt{Var(\hat{p})}} = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$

$\hat{\mathbf{X_1}} - \hat{\mathbf{X_2}}$:

Kjent varians:

- $\hat{X_1} - \hat{X_2} \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$
- $Z = \frac{(\overline{X_1}-\overline{X_2})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$

Ukjent varians:

- $\sigma_1^2 = \sigma_2^2$: $T = \frac{(\overline{X_1}-\overline{X_2})-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \sim t_{n_1+n_2-2}$
  $S_p^2 = \frac{S_1^2(n_1-1)+S_2^2(n_2-1)}{n_1+n_2-2}$
- $\sigma_1^2 \neq \sigma_2^2$: $T' = \frac{(\overline{X_1}-\overline{X_2})-(\mu_1-\mu_2)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \sim t_v$,
  $v = \frac{(S_1^2/n_1+s_2^2/n_2)^2}{\frac{(S_1^2/n_1)^2}{n_1-1}+\frac{(S_2^2/n_2)^2}{n_2-1}}$

F

Fra to uavhengige NF utvalg: $F = \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} \sim F_{n_1-1,n_2-1}$

$\hat{\mathbf{p_1}} - \hat{\mathbf{p_2}}$

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

$\overline{\mathbf{D}}$ Differanse av parvis utvalg

Gitt normalfordeling: $T = \frac{\overline{D}-\mu_D}{S_D/\sqrt{n}} \sim t_{n-1}$
Fra utvalgsfordelinger kan en utlede testobservatorer og konfidensintervaller!

Utlede Konfidensintervall:

Anta at vi har $X_1, X_2, \ldots, X_n$ stokastiske variabler, hvor sannsynlighetsfordelingen til disse inneholder en ukjent parameter θ. Anta også at vi har observasjoner $x_1, x_2, \ldots, x_n$. Har lyst å bruke disse for å finne et $100(1-\alpha)\%$ konfidensintervall:

1. Bestem en stokastiske variabel
   $Z = h(X_1, X_2, \ldots, X_n, \theta)$ som følger en kjent fordeling. Altså finn utvalgsfordelingen for paramaeteren θ
2. Finn kvantilene $Z_{\frac{\alpha}{2}}$ og $Z_{1-\frac{\alpha}{2}}$. Da har en at: $P(Z_{1-\frac{\alpha}{2}} \leq h(X_1, X_2, \ldots, X_n, \theta) \leq Z_{\frac{\alpha}{2}})$
3. Da er løsningen på ulikhetene
   $Z_{1-\frac{\alpha}{2}} \leq h(X_1, X_2, \ldots, X_n, \theta)$ og
   $Z_{1-\frac{\alpha}{2}} \geq h(X_1, X_2, \ldots, X_n, \theta)$ konfidensintervallet.

Prediksjonsintervall:

μ, σ kjent:

$P\left(-z_{\frac{\alpha}{2}} \leq \frac{X_0-\mu}{\sigma} \leq z_{\frac{\alpha}{2}}\right) = 1-\alpha$

μ kjent og σ ukjent:

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

μ, σ ukjent:

$P\left(-t_{\frac{\alpha}{2}} \leq \frac{X_0-\overline{X}}{S\sqrt{1+\frac{1}{n}}} \leq t_{\frac{\alpha}{2}}\right) = 1-\alpha$

Hypotesetesting:

Velger testobservator med kjent fordeling(Velger utvalgsfordelingen til parameteren) når nullhypotesen er sann. Dersom utregna testobservator gir en verdi som er veldig usansynlig hvis nullhypotesen er sann forkastes nullhypotesen.

Forkastningsområde:

Forkastnings område velges slik at det skal være en sannsynlighet α for å få en så ekstrem verdi, dersom H_0 er sant. Kritisk verdi blir da z_α, og vi forkaster H_0 om $Z > z_\alpha$

Kritisk region:

Hypotesetest for parameter θ og fordeling z:

$\theta < \theta_0$	$z < -z_\alpha$
$\theta > \theta_0$	$z < z_\alpha$
$\theta \neq \theta_0$	$z < z_{\frac{\alpha}{2}}$ eller $z > z_{\frac{\alpha}{2}}$

Type 1 og type 2 feil:

- Type 1: Forkaste h_0 når h_0 er sann: $\alpha = P(\text{Forkaste } h_0 \mid h_0 \text{ sann})$
- Type 2: Forkaster ikke h_0 når h_1 er sann: $\beta = P(\text{Beholder } h_0 \mid h_1 \text{ sann})$

P-verdi:

$P(\text{minst like ekstremt resultat som vi fikk}|H_0 \text{ sann})$

Styrken til hypotesetest:

Styrken $= 1 - P(\text{type II-feil}) = 1 - \beta$

Enkel lineær regresjon:

Regresjonsmodell:

- $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,\ i = 1, \dots, n$
- Forutsatt: $E(\epsilon_i) = 0,\ Var(\epsilon_i) = \sigma^2$
- $E(Y_i) = \mu_{Y|x_i} = \beta_0 + \beta_1 x_i,\ Var(Y_i) = \sigma^2_{Y|x_i} = \sigma^2$

Minste kvadraters metode:

- Brukes for å finne β_0, β_1 fra data
- Vil minimere $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - b_0 - b_1)$
- Minste verdier av b_0, b_1 kan finnes ved partiellderivasjon.
- $b_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^n (x_i - \overline{x})^2},\ b_0 = \overline{Y} - b_1 \overline{x}$
- b_0, b_1 forventningsrette estimatorer for β_0, β_1
- $Var(b_1) = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}$
- $Var(b_0) = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \overline{x})^2}$
- $S^2 = \frac{SSE}{n-2}$

Inferens av parametere:

- $T = \frac{b_1 - E(b_1)}{S\hat{E}(b_1)} = \frac{b_1 - \beta_1}{\frac{S}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}}}$, t-fordelt, $n-2$ frihetsgrader
- $T = \frac{b_0 - E(b_0)}{S\hat{E}(b_0)} = \frac{b_0 - \beta_0}{S\sqrt{\frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \overline{x})^2}}}$, t-fordelt, $n-2$ frihetsgrader

Inferens av σ^2 og $\mu_{Y|x_0}$:

- σ^2: $V = \frac{(n-2)S^2}{\sigma^2}$, χ^2-fordelt, $n-2$ frihetsgrader
- $\mu_{Y|x_0}$: $T = \frac{\hat{Y}_0 - \mu_{Y|x_0}}{S\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}}}$, t-fordelt, $n-2$ frihetsgrader

Generell sannsynlighet og statistikkregler:

- Addisjonsregel $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Betinget sannsynlighet $P(A|B) = \frac{P(A \cap B)}{P(B)}$
- Multiplikasjonsregelen: $P(A \cap B) = P(B)P(A|B) = P(A)P(B|A)$
- Kumulativ fordelingsfunksjon: $F(X) = P(X \leq x) = \begin{cases} \sum_{t \leq x} P(X = t), \text{ diskret} \\ \int_{-\infty}^x f(t)dt, \text{ Kont.} \end{cases}$
- Forventningsverdi: $E(X) = \begin{cases} \sum_x X f(x), \text{ Diskret} \\ \int_{-\infty}^{\infty}, X f(x) \text{ Kont.} \end{cases}$
- Varians: $Var(X) = \begin{cases} \sum_x (x-\mu)^2 f(x) = E(X^2) - E(X)^2 \text{ Disk} \\ \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx \cdot \mu^2 \end{cases}$
- $E(X+Y) = E(X) + E(Y) \quad E(cX) = cE(X)$
- $Var(X \pm Y) = Var(X) + Var(Y)$
  $Var(aX + b) = a^2 Var(X)$ Ved uavhengighet.

Simultanfordeling for to variabler:

- $P((x,y) \in A) = \begin{cases} \int \int_A f(x,y)dxdy \text{ Kont.} \\ \sum \sum_A F(x,y) \text{ Diskret} \end{cases}$
- $g(x) = \int_{-\infty}^{\infty} f(x,y)dy$ eller $\sum_y f(x,y)$ Samme for $h(y)$
- $f(y|x) = \frac{f(x,y)}{g(x)}$
- $\sigma_{XY} = Cov(X,y) = E[(X - \mu_X)(Y - \mu_Y)] = \begin{cases} \sum_x \sum_y (x - \mu_X)(y - \mu_Y) f(x,y), \text{ Diskret} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f(x,y)dxdy \text{ Kont.} \end{cases}$
- $\rho_{XY} = Cor(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$
- X, Y uavhengig $\Rightarrow Cov(x,y) = 0$

Sannsynlighetsmaksimeringsestimator:

Brukes hvis en ikke har/vet en naturlig estimator for en parameter θ

1. $L(\theta) = f(x_1, x_2, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i; \theta)$
2. Bruk $\ln(L)$ og deriver med hensyn på theta.
3. Løs $\frac{\partial L}{\partial \theta}(\theta) = 0$ for θ