FORMELSAMLING i GO040A STATISTIKK

Grunnleggende formler i sannsynlighetsregningen

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

Betinget sannsynlighet $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$

Gen. multiplikasjonsregel $P(A \cap B) = P(B) \cdot P(A|B)$ eller $P(A \cap B) = P(A) \cdot P(B|A)$

Total sannsynlighet $P(A) = \sum_{i=1}^{r} P(B_i) \cdot P(A|B_i)$ **Bayes lov** $P(B|A) = \dfrac{P(B) \cdot P(A|B)}{P(A)}$

A og B uavhengige $P(A \cap B) = P(A) \cdot P(B)$ $P(A|B) = P(A)$ $P(B|A) = P(B)$

Kombinatorikk

Antall forskjellige utvalg når s enheter trekkes fra en populasjon på N enheter:

Ordnet utvalg, med tilbakelegging N^s

Ordnet utvalg, uten tilbakelegging $(N)_s = N(N-1) \cdots (N-s+1) = \dfrac{N!}{(N-s)!}$

Uordnet utvalg, uten tilbakelegging $\binom{N}{s} = \dfrac{(N)_s}{s!} = \dfrac{N!}{s!(N-s)!}$

Generelt om sannsynlighetsfordelinger for 1 variabel

Fordelingsfunksjon $F(x) = P(X \le x)$ $F(x) = \int_{-\infty}^{x} f(u)du$

$f(x) = F'(x)$ $P(a < X \le b) = F(b) - F(a)$

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

$E[g(X)] = \sum_x g(x) \cdot P(X = x)$ $E[g(X)] = \int_{-\infty}^{\infty} g(x) \cdot f(x)dx$

Varians $\sigma^2 = Var(X) = E[(X-\mu)^2] = E(X^2) - \mu^2$

Standardavvik $\sigma = SD(X) = \sqrt{Var(X)}$

Generelt om sannsynlighetsfordelinger for 2 variable

Simultanfordeling for X og Y $P[(X = x) \cap (Y = y)]$

Forventning $E[g(X,Y)] = \sum_x \sum_y g(x,y) \cdot P[(X = x) \cap (Y = y)]$

Kovarians $Cov(X,Y) = E[(X - \mu_1)(Y - \mu_2)] = E(X \cdot Y) - \mu_1 \cdot \mu_2$

Korrelasjonskoeffisient $\rho(X,Y) = \dfrac{Cov(X,Y)}{\sigma_1 \cdot \sigma_2}$

Spesielle diskrete sannsynlighetsfordelinger

Binomisk fordeling $X \sim \text{bin}(n, p)$:

$$P(X = x) = \binom{n}{x} \cdot p^x \cdot (1-p)^{n-x}$$

$E(X) = np$ $\quad$ $Var(X) = np \cdot (1-p)$

Hypergeometrisk fordeling $X \sim \text{hypergeom}(N, M, n)$:

$$P(X = x) = \frac{\binom{M}{x} \cdot \binom{N-M}{n-x}}{\binom{N}{n}}$$

$E(X) = n\theta$ $\quad$ $Var(X) = \frac{N-n}{N-1} \cdot n\theta(1-\theta)$, der $\theta = \frac{M}{N}$

Poissonfordeling $X \sim Po(\lambda)$:

$$P(X = x) = \frac{\lambda^x}{x!} \cdot e^{-\lambda}$$

$E(X) = \lambda$ $\quad$ $Var(X) = \lambda$

Spesielle kontinuerlige sannsynlighetsfordelinger

Eksponensialfordeling $T \sim \text{eksp}(\lambda)$:

$f(t) = \lambda \cdot e^{-\lambda t}$ for $t > 0$ $\quad$ $F(t) = 1 - e^{-\lambda t}$ for $t > 0$

$E(T) = \frac{1}{\lambda}$ $\quad$ $Var(T) = \frac{1}{\lambda^2}$

Rektangulær fordeling $X \sim R(\alpha, \beta)$:

$f(x) = \frac{1}{\beta - \alpha}$ for $\alpha < x < \beta$

$E(X) = \frac{\alpha + \beta}{2}$ $\quad$ $Var(X) = \frac{1}{12}(\beta - \alpha)^2$

Standard normalfordeling $U \sim N(0,1)$:

$P(U \leq u) = G(u)$ $\quad$ $G(-u) = 1 - G(u)$

Generell normalfordeling $X \sim N(\mu, \sigma^2)$:

$$F(x) = P(X \leq x) = G\left(\frac{x - \mu}{\sigma}\right)$$

Regler for forventning og varians

$$E(aX + b) = a\,E(X) + b$$

$$E(X_1 + X_2) = E(X_1) + E(X_2)$$

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

$$Var(X_1 + X_2) = Var(X_1) + Var(X_2)\,, \quad \text{når } X_1 \text{ og } X_2 \text{ er uavhengige.}$$

Punktestimering av forventning μ og varians σ^2 i målemodellen

Punktestimator for forventning: $\hat{\mu} = \overline{X} = \frac{1}{n} \cdot \sum_{i=1}^{n} X_i$

$$E(\hat{\mu}) = \mu \qquad\qquad Var(\hat{\mu}) = \frac{\sigma^2}{n}$$

Punktestimator for varians: $S^2 = \frac{1}{n-1} \cdot \sum_{i=1}^{n} (X_i - \overline{X})^2 = \frac{1}{n-1} \cdot (\sum_{i=1}^{n} X_i^{\,2} - n \cdot \overline{X}^2)$

$$E(S^2) = \sigma^2$$

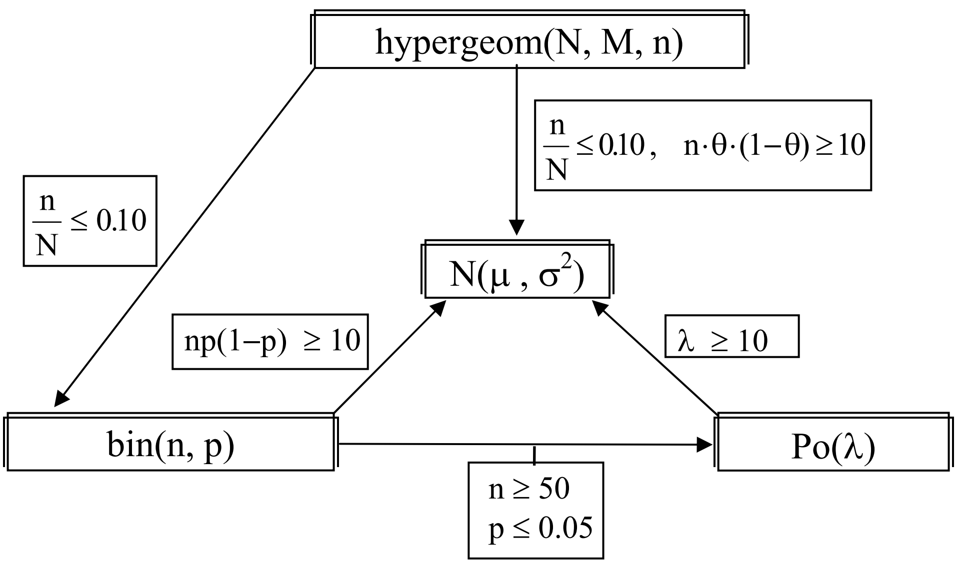
Tilnærminger

Sentralgrensesetningen (gjelder alle fordelinger):

Dersom $X_1, X_2, \ldots, X_n$ er uavhengige og identisk fordelte stokastiske variable med forventning μ og varians σ^2, så er for *store* verdier av n $(n \geq 30)$

$$X_1 + X_2 + \ldots + X_n \approx N(n\mu, n\sigma^2) \quad \text{og} \quad \overline{X} = \frac{1}{n}(X_1 + X_2 + \ldots + X_n) \approx N(\mu, \frac{\sigma^2}{n})$$

Sammenheng mellom spesielle fordelinger:



Intervallestimering og hypotesetesting i målemodellen

σ **kjent**, normalfordelte observasjoner eller stort antall observasjoner ($n \geq 30$):

$(1-\alpha)\cdot 100\%$ konfidensintervall for forventningen μ: $\overline{X} \pm u_{\alpha/2} \cdot \dfrac{\sigma}{\sqrt{n}}$

Utvalgsstørrelse: $n = \left(\dfrac{u_{\alpha/2} \cdot \sigma}{d} \right)^2$, der d er konfidensintervallets feilmargin.

Testobservator for H_0: $\mu = \mu_0$: $U_0 = \dfrac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N\left(\dfrac{\mu - \mu_0}{\frac{\sigma}{\sqrt{n}}}, 1\right)$

σ **ukjent**, normalfordelte observasjoner:

$(1-\alpha)\cdot 100\%$ konfidensintervall for forventningen μ: $\overline{X} \pm t_{\alpha/2,\, n-1} \cdot \dfrac{S}{\sqrt{n}}$

Testobservator for H_0: $\mu = \mu_0$: $T_0 = \dfrac{\overline{X} - \mu_0}{S/\sqrt{n}}$

Når H_0 er sann, er T_0 t-fordelt med $(n-1)$ frihetsgrader.

Intervallestimering og hypotesetesting i Poissonmodell, vha. normaltilnærming

Tilnærmet $100(1-\alpha)\%$ konfidensintervall for λ: $\hat{\lambda} \pm u_{\alpha/2}\sqrt{\hat{\lambda}}$, der $\hat{\lambda} = X$

Testobservator for H_0: $\lambda = \lambda_0$: $U_0 = \dfrac{\hat{\lambda} - \lambda_0}{\sqrt{\lambda_0}}$

Intervallestimering og hypotesetesting i binomisk modell, vha. normaltilnærming

Tilnærmet $100(1-\alpha)\%$ konfidensintervall for p: $\hat{p} \pm u_{\alpha/2}\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$, der $\hat{p} = \dfrac{X}{n}$

Testobservator for H_0: $p = p_0$: $U_0 = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$

Intervallestimering og hypotesetesting i hypergeometrisk modell, vha. normaltilnærming

Tilnærmet $100(1-\alpha)\%$ konfidensintervall for $\theta = \dfrac{M}{N}$: Analogt med binomisk modell.

Testobservator for H_0: $\theta = \theta_0$: Analogt med binomisk modell.

Korrelasjon

$$S_X{}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n}\sum_{i=1}^{n}X_i{}^2 - \overline{X}^2 \; ; \qquad S_Y{}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \frac{1}{n}\sum_{i=1}^{n}Y_i{}^2 - \overline{Y}^2$$

$$S_{XY} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^{n}X_iY_i - \overline{X}\cdot\overline{Y}$$

Empirisk korrelasjonskoeffisient: $R = \dfrac{S_{XY}}{S_X S_Y}$

Regresjonsmodellen

n par observasjoner av x og Y: $(x_1, Y_1), (x_2, Y_2), \ldots (x_n, Y_n)$

$Y_1, Y_2, \ldots, Y_n$ er uavhengige og normalfordelte stokastiske variable.

$E(Y_i) = \beta_0 + \beta_1 x_i$, $i = 1, 2, \ldots, n$; der $x_1, x_2, \ldots, x_n$ er kjente tall.

$\mathrm{Var}(Y_i) = \sigma^2$, $i = 1, 2, \ldots, n$.

Minste kvadraters estimatorer:

$$\hat{\beta}_1 = \frac{1}{M}\sum_{i=1}^{n}(x_i - \overline{x})Y_i \;, \quad \text{der} \quad M = \sum_{i=1}^{n}(x_i - \overline{x})^2$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{x}$$

$$\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{M}) \;, \qquad \hat{\beta}_0 \sim N(\beta_0, \frac{\sigma^2}{nM}\sum_{i=1}^{n}x_i{}^2)$$

Estimert regresjonslinje: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Intervallestimering og hypotesetesting i regresjonsmodellen

σ kjent:

Konfidensintervall for β_1: $\hat{\beta}_1 \pm u_{\alpha/2}\dfrac{\sigma}{\sqrt{M}}$

Testobservator for β_1: $U_0 = \dfrac{\hat{\beta}_1 - \beta_1{}^0}{\sigma/\sqrt{M}} \sim N(\dfrac{\beta_1 - \beta_1{}^0}{\sigma/\sqrt{M}}, 1)$

σ ukjent: Ikke pensum