1 Beskrivende statistikk

For n observasjoner $x_1, \ldots, x_n$, og n parvise observasjoner $(x_1, y_1), \ldots, (x_n, y_n)$ finner vi:

Gjennomsnitt

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Empirisk varians

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Empirisk standardavvik

$$s = \sqrt{s^2}$$

Empirisk korrelasjon

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Minste kvadratsums rette linje $y = a + bx$

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$a = \bar{y} - b\bar{x}.$$

2 Hendelser og sannsynlighet

Addisjonsregelen

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

Betinget sannsynlighet

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

Total sannsynlighet La hendelsene A_1, A_2, ..., A_n danne en partisjon (oppdeling) av utfallsrommet. Da er

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

Bayes regel

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

Uavhengige hendelser

$$P(A \cap B) = P(A) \cdot P(B)$$
$$P(A|B) = P(A),\ P(B|A) = P(B)$$

2.1 Urnemodeller

En urne inneholder n kuler. Antall mulige trekninger av r kuler er:

1. Ordnet utvalg, trekning med tilbakelegging
$$n^r$$
2. Ordnet utvalg, trekning uten tilbakelegging
$${}_nP_r = \frac{n!}{(n-r)!}$$
3. Ikke-ordnet utvalg, trekning uten tilbakelegging
$${}_nC_r = \binom{n}{r} = \frac{n!}{r!(n-r)!}$$

3 Stokastiske variabler

3.1 Diskret stokastisk variabel X

Kumulativ fordeling

$$F(x) = P(X \leq x) = \sum_{t \leq x} P(X = t)$$

Forventningsverdi

$$\mathrm{E}(X) = \mu = \sum_{x} x \cdot P(X = x)$$

Varians

$$\mathrm{Var}(X) = \sigma^2 = \mathrm{E}((X-\mu)^2) = \sum_{x}(x-\mu)^2 \cdot P(X = x)$$
$$= \sum_{x} x^2 \cdot P(X = x) - \mu^2$$

Standardavvik

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

3.2 Kontinuerlig stokastisk variabel X

Kumulativ fordeling

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\,dt$$

Forventningsverdi

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

Varians

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

Standardavvik

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

3.3 Kovarians og korrelasjon

$$\mathrm{Cov}(X,Y) = \mathrm{E}\left((X-\mu_X)\cdot(Y-\mu_Y)\right) = \mathrm{E}(XY) - \mu_X\mu_Y$$
$\mathrm{Cov}(X,Y) = 0$ dersom X og Y er uavhengige.
$$\mathrm{Corr}(X,Y) = \rho_{XY} = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \cdot \sigma_Y}.$$

3.4 Regneregler

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

$$E(aX + b) = aE(X) + b$$
$$\mathrm{Var}(aX + b) = a^2\mathrm{Var}(X)$$

$$E(aX + bY) = aE(X) + bE(Y)$$
$$\mathrm{Var}(aX + bY) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) + 2ab\mathrm{Cov}(X, Y)$$

$$E(a_1X_1 + a_2X_2 + \dots a_nX_n) = a_1E(X_1) + a_2E(X_2) + \dots + a_nE(X_n)$$

For uavhengige stokastiske variabler gjelder

$$\mathrm{Var}(a_1X_1 + a_2X_2 + \dots a_nX_n) = a_1^2\mathrm{Var}(X_1) + a_2^2\mathrm{Var}(X_2) + \dots + a_n^2\mathrm{Var}(X_n)$$

4 Sannsynlighetsfordelinger

Diskret uniformfordeling

$$X \sim \mathrm{Uniform}(k)$$
$$P(X = x) = \frac{1}{k}, \text{ for } x = x_1, x_2, \dots, x_k$$
$$E(X) = \frac{1}{k}\sum_{i=1}^{k} x_i, \quad \mathrm{Var}(X) = \frac{1}{k}\sum_{i=1}^{k}(x_i - E(X))^2$$

Kontinuerlig uniformfordeling

$$X \sim \mathrm{Uniform}(a, b)$$
$$f(x) = \frac{1}{b-a}, \text{ for } a \leq x \leq b$$
$$E(X) = \frac{a+b}{2}, \quad \mathrm{Var}(X) = \frac{(b-a)^2}{12}$$

Binomisk fordeling

$$X \sim \mathrm{Binom}(n,\, p)$$
$$P(X = x) = \binom{n}{x} p^x(1-p)^{n-x}, \quad \text{for } x = 0, 1, 2, \dots, n$$
$$E(X) = np, \quad \mathrm{Var}(X) = np(1-p)$$

Geometrisk fordeling

$$X \sim \mathrm{Geom}(p)$$
$$P(X = x) = p(1-p)^{x-1}, \quad \text{for } x = 1, 2, \dots$$
$$F(x) = P(X \leq x) = 1 - (1-p)^x$$
$$E(X) = \frac{1}{p}, \quad \mathrm{Var}(X) = \frac{1-p}{p^2}$$

Poissonfordeling

$$X \sim \mathrm{Poisson}(\lambda t)$$
$$P(X = x) = \frac{(\lambda t)^x}{x!}e^{-\lambda t}, \quad \text{for } x = 0, 1, 2, \dots$$
$$E(X) = \lambda t, \quad \mathrm{Var}(X) = \lambda t$$

Eksponentialfordeling

$$T \sim \mathrm{Eksponential}(\lambda)$$
$$f(t) = \lambda e^{-\lambda t}, \quad \text{for } t > 0$$
$$F(t) = P(T \leq t) = 1 - e^{-\lambda t}$$
$$E(T) = \frac{1}{\lambda}, \quad \mathrm{Var}(T) = \frac{1}{\lambda^2}$$

Weibullfordeling

$$T \sim \mathrm{Weibull}(\alpha, \lambda)$$
$$f(t) = \alpha\lambda(\lambda t)^{\alpha-1}e^{-(\lambda t)^\alpha}, \quad \text{for } t > 0$$
$$F(t) = P(T \leq t) = 1 - e^{-(\lambda t)^\alpha}$$
$$E(T) = \frac{1}{\lambda}\Gamma(1 + \tfrac{1}{\alpha})$$
$$\mathrm{Var}(T) = \frac{1}{\lambda^2}\left(\Gamma(1 + \tfrac{2}{\alpha}) - \Gamma(1 + \tfrac{1}{\alpha})^2\right)$$

Normalfordeling

$$X \sim \mathrm{N}(\mu, \sigma)$$
$$f(x) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad \text{for } -\infty < x < \infty$$
$$F(x) = P(X \leq x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$
$$E(X) = \mu, \quad \mathrm{Var}(X) = \sigma^2$$

Standard normalfordeling

$$Z \sim \mathrm{N}(0, 1)$$
$$f(z) = \phi(z) = \frac{1}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}, \quad \text{for } -\infty < z < \infty$$
$$F(z) = P(Z \leq z) = \Phi(z)$$
$$E(Z) = 0, \quad \mathrm{Var}(Z) = 1$$

4.1 Regneregler normalfordeling

Vi ser på n uavhengige stokastiske variabler $X_1, X_2, \dots, X_n$ slik at $X_i \sim N(\mu, \sigma)$, for $i = 1, \dots, n$. Da er

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} \sim \mathrm{N}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

og

$$X_1 + X_2 + \dots + X_n \sim \mathrm{N}\left(n\mu, \sqrt{n}\cdot\sigma\right)$$

4.1.1 Normaltilnærminger

$\mathrm{Binom}(n,\, p) \approx \mathrm{N}(np, \sqrt{np(1-p)})$ hvis $np(1-p) \geq 5$,
$\mathrm{Poisson}(\lambda t) \approx \mathrm{N}(\lambda t, \sqrt{\lambda t})$ hvis $\lambda t > 10$

4.1.2 Sentralgrenseteoremet

Dersom $X_1, X_2, \ldots, X_n$ er uavhengige stokastiske variabler fra samme sannsynlighetsfordeling med forventning $\mathrm{E}(X_i) = \mu$ og varians $\mathrm{Var}(X_i) = \sigma^2$, for $i = 1, \ldots, n$, og dersom $n > 30$, så er

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} \text{ tilnærmet N} \left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

og

$$X_1 + X_2 + \cdots + X_n \text{ tilnærmet N} \left(n\mu, \sqrt{n} \cdot \sigma\right)$$

4.2 Begreper fra levetidsanalyse

Pålitelighetsfunksjon

$$R(t) = 1 - F(t) = P(X > t)$$

Sviktrate

$$h(t) = \frac{f(t)}{R(t)}$$

Systempålitelighet seriekobling

$$R(t) = R_1(t) \cdot R_2(t) \cdots R_n(t)$$

Systempålitelighet parallellkobling

$$R(t) = 1 - F_1(t) \cdot F_2(t) \cdots F_n(t)$$

5 Punktestimering

5.1 Forventningsverdi og varians

For et tilfeldig utvalg $X_1, X_2, \ldots, X_n$ der $\mathrm{E}(X_i) = \mu$ og $\mathrm{Var}(X_i) = \sigma^2$, for $i = 1, \ldots, n$, så er

Estimator for forventningsverdien (μ):

$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Estimator for varians (σ^2):

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Estimator for standardavviket (σ):

$$S = \sqrt{S^2}$$

5.2 Sannsynligheten p i binomisk fordeling

Dersom X teller antall suksesser i en binomisk forsøksrekke av n forsøk så er estimator for sannsynligheten for suksess (p):

$$\hat{p} = \frac{X}{n}$$

5.3 Raten λ i poissonfordelingen

Dersom X teller antall hendelser i en poissonprosess med rate λ over et intervall/område av lengde/størrelse t, så er estimator for raten (λ):

$$\hat{\lambda} = \frac{X}{t}$$

6 Konfidensintervall

6.1 Forventningsverdi μ

For et tilfeldig utvalg $X_1, X_2, \ldots, X_n$, $X_i \sim N(\mu, \sigma)$, $i = 1, \ldots, n$, der standardavviket σ er *kjent*, så er

$$\left[\bar{X} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right]$$

et $100(1-\alpha)\%$ konfidensintervall for μ.

Dersom standardavviket σ er *ukjent*, så er

$$\left[\bar{X} - t_{\alpha/2, n-1} \cdot \frac{S}{\sqrt{n}},\ \bar{X} + t_{\alpha/2, n-1} \cdot \frac{S}{\sqrt{n}}\right]$$

et $100(1-\alpha)\%$ konfidensintervall for μ.

6.2 Sannsynligheten p i binomisk fordeling

Under forutsetningen om at $n\hat{p}(1-\hat{p}) \geq 5$, så er

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

et tilnærmet $100(1-\alpha)\%$ konfidensintervall for p.

6.3 Raten λ i poissonfordeling

Under forutsetningen om at $\hat{\lambda}t > 10$, så er

$$\left[\hat{\lambda} - z_{\alpha/2} \cdot \sqrt{\frac{\hat{\lambda}}{t}},\ \hat{\lambda} + z_{\alpha/2} \cdot \sqrt{\frac{\hat{\lambda}}{t}}\right]$$

et tilnærmet $100(1-\alpha)\%$ konfidensintervall for λ.

7 Hypotesetesting

Noen begreper:

- *Type I-feil:* Forkaste nullhypotesen H_0 selv om H_0 er sann.
- *Type II-feil (eller type 2-feil):* Ikke forkaste nullhypotesen H_0 selv om den alternative hypotesen H_1 er sann.
- *Teststyrke:* Teststyrken til en test er sannsynligheten for å forkaste nullhypotesen H_0 til fordel for den alternative hypotesen H_1 når den alternative hypotesen er sann og vi kjenner den riktige parameterverdien.

- *P-verdi:* P-verdien er sannsynligheten for det vi har observert, eller noe mer ekstremt i retning den alternative hypotesen H_1, når vi antar at nullhypotesen H_0 er sann.

7.1 Forventningsverdi μ i normalfordelingen

Testobservator for $H_0 : \mu = \mu_0$ mot

1. $H_1 : \mu \neq \mu_0$, eller
2. $H_1 : \mu > \mu_0$, eller
3. $H_1 : \mu < \mu_0$,

dersom standardavviket σ er *kjent* (Z-test):

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \overset{H_0}{\sim} N(0,1)$$

og dersom standardavviket er *ukjent* (T-test):

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \overset{H_0}{\sim} t_{n-1}$$

Forkast H_0 dersom

1. $|z| > z_{\alpha/2}$ eller $|t| > t_{\alpha/2,n-1}$
2. $z > z_\alpha$ eller $t > t_{\alpha,n-1}$
3. $z < -z_\alpha$ eller $t < -t_{\alpha,n-1}$

7.2 Sannsynligheten p i binomisk fordeling

Testobservator for $H_0 : p = p_0$ (Z-test):

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \overset{H_0}{\approx} N(0,1)$$

8 Kovarians og korrelasjon

Estimator for **kovarians** $\mathrm{Cov}(X,Y)$ og **korrelasjon**:

$$\widehat{\mathrm{Cov}}(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}),$$

$$R = \frac{\widehat{\mathrm{Cov}}(X,Y)}{S_x \cdot S_y},$$

der S_x og S_y er estimatorer for standardavvikene til X og Y.

9 Enkel lineær regresjon

La $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ være n uavhengige par der x-ene er kjente tall, og Y-ene er stokastiske variabler slik at

$$Y_i | X_i = x_i \sim \mathrm{N}(\beta_0 + \beta_1 x_i, \sigma) \text{ for } i = 1, \ldots, n$$

En annen måte å formulere modellen på er:

$$Y_i = \beta_0 + \beta_1 x_i + e_i, \quad e_i \sim \mathrm{N}(0, \sigma) \quad \text{for } i = 1, \ldots, n$$

Minste kvadratsums estimatorer for β_0 og β_1:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \text{og}$$

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

Estimert regresjonslinje:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Estimator for varians σ^2 er:

$$S^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

og estimator for standardavviket σ, er $S = \sqrt{S^2}$.

Godhetsmål for regresjonsmodellen:

$$R^2 = \frac{\mathrm{SST} - \mathrm{SSE}}{\mathrm{SST}}, \text{ der}$$

$$\mathrm{SST} = \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \text{ og,}$$

$$\mathrm{SSE} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

Konfidensintervall for β_1:

$$\left[\hat{\beta}_1 - t_{\alpha/2,\,n-2} \cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + t_{\alpha/2,\,n-2} \cdot \mathrm{SE}(\hat{\beta}_1)\right],$$

der

$$\mathrm{SE}(\hat{\beta}_1) = \frac{S}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

Testobservator for $H_0 : \beta_1 = 0$ mot $H_1 : \beta_1 \neq 0$:

$$T = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)} \overset{H_0}{\sim} t_{n-2}$$

10 Noen Python-kommandoer

Eks: Punktsannsynligheter og kumulative sannsynligheter i binomisk fordeling:

```
from scipy import stats
stats.binom.pmf(x,n,p)
stats.binom.cdf(x,n,p)
```

Eks: Sannsynlighetstetthet og kumulative sannsynligheter i normalfordeling:

```
from scipy import stats
stats.norm.pdf(x, mu, sigma)
stats.norm.cdf(x, mu, sigma)
```

Eks: Tilpasse en lineær regresjonsmodell:

```
import statsmodels.formula.api as smf
modell = smf.ols('y ~ x',data).fit()
modell.summary()
```