Formelblad – Sannolikhetsteori 1

Bayes formel: Låt A och D vara två händelser. Då gäller

$$P(A|D) = \frac{P(D|A)P(A)}{P(D)}.$$

Chebyshevs olikhet: Låt X vara en stokastisk variabel med väntevärde μ och varians σ^2. Då gäller för alla $c > 0$ att

$$P(|X - \mu| \geq c) \leq \frac{\sigma^2}{c^2}.$$

Geometrisk serie: För en geometrisk serie gäller då $|x| < 1$

$$\sum_{i=0}^{\infty} ax^i = \frac{a}{1-x},$$

och delsumman S_n (summan av de n första termerna) gäller då $x \neq 1$

$$S_n = a\frac{1-x^n}{1-x}.$$

Diskreta fördelningar

Binomialfördelning: $X \sim Bin(n, p)$

- frekvensfunktionen $P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, ..., n$
- väntevärdet np
- variansen $np(1-p)$
- momentgenererande funktionen $M_X(t) = (pe^t + 1 - p)^n$

Poissonfördelning: $X \sim Pois(\lambda)$

- frekvensfunktionen $P(X = x) = e^{-\lambda} \cdot \frac{\lambda^x}{x!}, \quad x = 0, 1, 2, ...$
- väntevärdet λ
- variansen λ
- momentgenererande funktionen $M_X(t) = \exp(\lambda(e^t - 1))$

Geometrisk fördelning: $X \sim Geom(p)$

- frekvensfunktionen $P(X = x) = (1-p)^{x-1}p, \quad x = 1, 2, ...$
- väntevärdet $\frac{1}{p}$
- variansen $\frac{1-p}{p^2}$

Negativ binomialfördelning: $X \sim NegBin(r, p)$

- frekvensfunktionen $P(X = x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$, $x = r, r+1, ...$
- väntevärdet $\frac{r}{p}$
- variansen $\frac{r(1-p)}{p^2}$

Hypergeometrisk fördelning: $X \sim HypGeom(n, N, m)$

- frekvensfunktionen $P(X=x) = \dfrac{\begin{pmatrix} m \\ x \end{pmatrix}\begin{pmatrix} N-m \\ n-x \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}}, \quad x = 0, 1, ..., n$
- väntevärdet $\frac{nm}{N}$
- variansen $\frac{nm}{N}\left(\frac{(n-1)(m-1)}{N-1} + 1 - \frac{nm}{N}\right)$

Kontinuerliga fördelningar

Likformig fördelning: $X \sim Lik(a, b)$

- täthetsfunktionen $f(x) = \frac{1}{b-a}, \quad a < x < b$
- väntevärdet $\frac{1}{2}(a+b)$
- variansen $\frac{1}{12}(b-a)^2$
- momentgenererande funktionen $M_X(t) = \frac{e^{tb}-e^{ta}}{t(b-a)}$

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

- täthetsfunktionen $f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \quad -\infty < x < \infty$
- väntevärdet μ
- variansen σ^2
- momentgenererande funktionen $M_X(t) = \exp(\mu t + \sigma^2 t^2/2)$

Standardiserad normalfördelning: $Z \sim N(0, 1)$

- täthetsfunktionen $f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty$
- momentgenererande funktionen $M_Z(t) = \exp(t^2/2)$

Exponentialfördelning: $X \sim Exp(\lambda)$

- täthetsfunktionen $f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$
- väntevärdet $\frac{1}{\lambda}$
- variansen $\frac{1}{\lambda^2}$
- momentgenererande funktionen $M_X(t) = \frac{\lambda}{\lambda - t}$

Gammafördelning: $X \sim Gamma(s, \lambda)$

- täthetsfunktionen $f(x) = \frac{\lambda e^{-\lambda x}(\lambda x)^{s-1}}{\Gamma(s)}, \quad x \geq 0$, där $\Gamma(s) = \int_0^\infty e^{-y} y^{s-1}\, dy$
- väntevärdet $\frac{s}{\lambda}$
- variansen $\frac{s}{\lambda^2}$
- momentgenererande funktionen $M_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^s$

Student t-fördelning: $T \sim t_r$, $r = 1, 2, \ldots$

- täthetsfunktionen $f(t) = \frac{\Gamma((r+1)/2)}{\sqrt{r\pi}\Gamma(r/2)}\left(1 + \frac{t^2}{r}\right)^{-(r+1)/2}, \quad t \in \mathbb{R}$, där $\Gamma(s) = \int_0^\infty e^{-y} y^{s-1}\, dy$
- väntevärdet 0
- variansen $\frac{r}{r-2}$, för $r \geq 3$

χ^2-fördelning: $X \sim \chi^2_r$, $r = 1, 2, \ldots$

- täthetsfunktionen $f(x) = \frac{1}{r^{r/2}\Gamma(r/2)} x^{r/2-1} e^{-x/2}$, $\quad x \in \mathbb{R}$, där $\Gamma(s) = \int_0^\infty e^{-y} y^{s-1}\, dy$
- väntevärdet r
- variansen $2r$

Bivariat normalfördelning: En två-dimensionell stokastisk variabel (X, Y) har en bivariat normalfördelning med parametrar μ_1, μ_2, σ_1^2, σ_2^2 och ρ, om dess täthetsfunktion är

$$f(x, y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot$$
$$\exp\left(-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right)\right).$$

Den betingade fördelningen av Y givet $X = x$ är normalfördelning med väntevärde $\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1)$ och varians $\sigma_2^2(1-\rho^2)$. Om (X, Y) har en bivariat normalfördelning med parametrar $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2$ och ρ, då har $\left(\frac{X-\mu_1}{\sigma_1}, \frac{Y-\mu_2}{\sigma_2}\right)$ bivariat normalfördelning med parametrar 0, 0, 1, 1 och ρ.

Exempel om hur olika fördelningar kan tillämpas

- **Binomial:** Antal succéer i n oberoende försök, där succésannolikheten är p, är binomialfördelat.
- **Poisson:** Kan användas som fördelning för antal punkter i en stokastisk punktprocess (under några enkla antaganden). Approximerar binomialfördelningen, när n är stor och p liten, $\lambda = np$.
- **Geometrisk:** Antal försök, som behövs tills en händelse med sannolikhet p inträffar, har geometrisk fördelning.
- **Negativ binomial:** Antal försök, som behövs tills en händelse med sannolikhet p inträffar för r-te gången, har negativ binomialfördelning.
- **Hypergeometrisk:** Man använder fördelningen om man drar utan återläggning från en ändlig population som har två olika slags individer.
- **Likformig:** Används som fördelning för väntetider och avrundning av mätningsfel.
- **Normal:** Under generella antaganden är en summa av ett stort antal stokastiska variabler approximativt normalfördelat (centrala gränsvärdessatsen).
- **Standardiserad normal:** Om $X \sim N(\mu, \sigma)$, då har $\frac{X-\mu}{\sigma}$ standardiserad normalfördelning.
- **Exponential:** Fördelning för livslängd (utan åldrande).
- **Gamma:** Fördelning för summan av n oberoende stokastiska variabler, som är exponentialfördelade med parameter λ.

Summor av oberoende stokastiska variabler

- $X_1 \sim Bin(n_1, p)$ och $X_2 \sim Bin(n_2, p)$: $X_1 + X_2 \sim Bin(n_1 + n_2, p)$
- $X_1 \sim Pois(\lambda_1)$ och $X_2 \sim Pois(\lambda_2)$: $X_1 + X_2 \sim Pois(\lambda_1 + \lambda_2)$
- $X_1 \sim Exp(\lambda)$ och $X_2 \sim Exp(\lambda)$: $X_1 + X_2 \sim Gamma(2, \lambda)$

- $X_1 \sim Gamma(n_1, \lambda)$ och $X_2 \sim Gamma(n_2, \lambda)$:
  $X_1 + X_2 \sim Gamma(n_1 + n_2, \lambda)$
- $X_1 \sim N(\mu_1, \sigma_1^2)$ och $X_2 \sim N(\mu_2, \sigma_2^2)$: $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$
- $X_1, X_2, \ldots, X_n$ är $N(0,1)$: $\sum_{k=1}^n X_k^2 \sim \chi_n^2$
- $X_1 \sim \chi_n^2$ och $X_2 \sim \chi_m^2$: $X_1 + X_2 \sim \chi_{n+m}^2$
- $X_1 \sim N(0,1)$ och $X_2 \sim \chi_n^2$: $\frac{X_1}{\sqrt{X_2/n}} \sim t_n$

Statistikor

- **Ett stickprov**: Låt $X_1, X_2, \ldots, X_n$ vara oberoende normalfördelade stokastiska variabler med väntevärde μ och varians σ^2.
  - *Stickprovsväntevärde*: $\bar{X} = \frac{1}{n}\sum_{k=1}^n X_k$ är en väntevärdesriktig skattare för μ, och
  $$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$
  - *Stickprovsvarians*: $s^2 = \frac{1}{n-1}\sum_{k=1}^n (X_k - \bar{X})^2 = \frac{1}{n-1}\left(\sum_{k=1}^n X_k^2 - n\bar{X}\right)$ är en väntevärdesriktig skattare för σ^2, och
  $$\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2.$$
  - *Ett stickprov t-statistika*:
  $$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$
- **Två stickprov, lika varians**: Låt $X_1, X_2, \ldots, X_n$ och $Y_1, Y_2, \ldots, Y_m$ vara två oberoende stickprov från en $N(\mu_1, \sigma^2)$- respektive $N(\mu_2, \sigma^2)$-fördelning. Låt s_1^2 respektive s_2^2 beteckna stickprovsvariansen för vardera stickprov.
  - *Sammanvägd stickprovsvarians*: $s_p^2 = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2}$ är en väntevärdesriktig skattare för σ^2, och
  $$\frac{(n+m-2)s_p^2}{\sigma^2} \sim \chi_{n+m-2}^2.$$
  - *Två stickprov t-statistika*:
  $$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{s_p\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-2}$$
- **Två stickprov, parvis observationer**: Låt $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ vara ett stickprov av parvis observationer sådana att (X_k, Y_k) är bivariat normalfördelade med parametrar μ_k, $\mu_k + \Delta$, σ_1^2, σ_2^2 och ρ. Sätt $D_k = X_k - Y_k$ och låt s_D^2 vara stickprovsvariansen för $D_1, D_2, \ldots, D_n$. Då är
  $$D_k \sim N(\Delta, \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2).$$
  - *t-statistika*:
  $$T = \frac{\bar{D} - (\mu_1 - \mu_2)}{s_D\sqrt{n}} \sim t_{n-1}$$

- *Stickprovskorrelationscoefficient*:

$$R = \frac{\sum_{k=1}^{n}(X_k - \bar{X})(Y_k - \bar{Y})}{\sqrt{\sum_{k=1}^{n}(X_k - \bar{X})^2 \sum_{k=1}^{n}(Y_k - \bar{Y})^2}} = \frac{\sum_{k=1}^{n} X_k Y_k - n\bar{X}\bar{Y}}{\sqrt{\left(\sum_{k=1}^{n} X_k^2 - n\bar{X}^2\right)\left(\sum_{k=1}^{n} Y_k^2 - n\bar{Y}^2\right)}}$$

- Om $\rho = 0$, så är $T = \frac{(n-2)R}{\sqrt{1-R^2}} \sim t_{n-2}$.

• **Okända sannolikheter/populationsandelar**: Låt $X \sim Bin(n,p)$ och $Y \sim Bin(m,q)$ vara oberoende.

  - $\hat{p} = \frac{X}{n}$ och $\hat{q} = \frac{Y}{m}$ är väntevärdesriktiga skattare för p respektive q.
  - För stora n gäller

$$\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} \overset{d}{\approx} N(0,1),$$

  enligt centrala gränsvärdessatsen. Eftersom $\hat{p} \to p$ i sannolikhet då $n \to \infty$, så gäller vidare att

$$\frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})/n}} \overset{d}{\approx} N(0,1),$$

  för stora n.
  - För $\min(n,m)$ stort gäller

$$\hat{p} - \hat{q} \overset{d}{\approx} N\left(p - q, \frac{p(1-p)}{n} + \frac{q(1-q)}{m}\right),$$

  enligt centrala gränsvärdessatsen. Eftersom $\hat{p} - \hat{q} \to p - q$ i sannolikhet då $\min(n,m) \to \infty$, så gäller vidare att

$$\frac{\hat{p} - \hat{q} - (p-q)}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{\hat{q}(1-\hat{q})}{m}}} \overset{d}{\approx} N(0,1),$$

  då $\min(n,m)$ är stort.

• **Jämförelse av sannolikheter/populationsandelar**: Antag att utfallet av ett försök kan hamna i r olika kategorier, vardera med sannolikhet $p_1, p_2, \ldots, p_r$. Låt X_k beteckna antalet utfall i kategori k vid n oberoende försök.

  - $\hat{p}_k = \frac{X_k}{n}$ är en väntevärdesriktig skattare för p_k, för $k = 1, 2, \ldots, r$.
  - *Parvis jämförelse av okända sannolikheter/populationsandelar*: För stora n gäller

$$\hat{p}_k - \hat{p}_l \overset{d}{\approx} N\left(p_k - p_l, \frac{p_k + p_l - (p_k - p_l)^2}{n}\right),$$

  enligt centrala gränsvärdessatsen.
  - *Jämförelse av uppmätta och teoretiska frekvenser*:

$$\chi^2 = \sum_{k=1}^{r} \frac{(X_k - np_k)^2}{np_k} \overset{d}{\approx} \chi^2_{r-1}$$

  Approximationen är ej pålitlig om $np_k < 5$ för något $k = 1, 2, \ldots, r$.

- **Icke-parametriska metoder**: Låt $X_1, X_2, \ldots, X_n$ vara oberoende stokastiska variabler från en kontinuerlig fördelning med väntevärde μ och median m. Låt R_k vara rangen av X_k då $|X_k - \mu|$ ordnas i stigande ordning. Låt $I_k = I_{\{X_k > \mu\}}$ indikera om $X_k > \mu$.
  - *Teckenstatistika*: $N_+ = \#\{k : X_k > m\} \sim Bin(n, \frac{1}{2})$.
  - *Wilcoxon teckenrang-statistika*: $W = \sum_{k=1}^n R_k I_k$
  - Om stickprovsfördelningen är symmetrisk, så är $m = \mu$ och W symmetriskt fördelad över $0, 1, \ldots, \frac{n(n+1)}{2}$.
  - Om stickprovsfördelningen är symmetrisk, så gäller för stora n att

$$W \overset{d}{\approx} N\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right).$$

Linjär regression

Låt $(x_1, Y_1), (x_2, Y_2), \ldots, (x_n, Y_n)$ vara oberoende observationer sådana att $Y_k \sim N(\alpha + \beta x_k, \sigma^2)$ för några konstanter $\alpha, \beta \in \mathbb{R}$. Låt

$$S_x = \sum_{k=1}^n x_k, \qquad S_Y = \sum_{k=1}^n Y_k,$$
$$S_{xx} = \sum_{k=1}^n x_k^2, \qquad S_{xY} = \sum_{k=1}^n x_k Y_k.$$

- **Skattning av koeficienter**: Maximum Likelihood-skattarna för β respektive α ges av

$$\hat{\beta} = \frac{\sum_{k=1}^n (x_k - \bar{x})(Y_k - \bar{Y})}{\sum_{k=1}^n (x_k - \bar{x})^2} = \frac{nS_{xY} - S_x S_Y}{nS_{xx} - S_x^2},$$
$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{x}.$$

  - $\hat{\alpha} \sim N\big(\alpha, \mathrm{Var}(\hat{\alpha})\big)$ och $\hat{\beta} \sim N\big(\beta, \mathrm{Var}(\hat{\beta})\big)$, där

$$\mathrm{Var}(\hat{\alpha}) = \frac{\sigma^2 \sum_{k=1}^n x_k^2}{n \sum_{k=1}^n (x_k - \bar{x})^2} = \frac{\sigma^2 S_{xx}}{nS_{xx} - S_x^2},$$
$$\mathrm{Var}(\hat{\beta}) = \frac{\sigma^2}{\sum_{k=1}^n (x_k - \bar{x})^2} = \frac{\sigma^2}{S_{xx} - S_x^2/n}.$$

  - $s^2 = \frac{1}{n-2} \sum_{k=1}^n (Y_k - \hat{\alpha} - \hat{\beta}x_k)^2$ är väntevärdesriktig skattare för σ^2, och

$$\frac{(n-2)s^2}{\sigma^2} \sim \chi^2_{n-2}.$$

  - $T_\alpha = \frac{\hat{\alpha}-\alpha}{s}\sqrt{n - \frac{S_x^2}{S_{xx}}} \sim t_{n-2}$ och $T_b = \frac{\hat{\beta}-\beta}{s}\sqrt{S_{xx} - \frac{S_x^2}{n}} \sim t_{n-2}$.

- **Prediktion av ny observation**: Låt (x, Y) vara en observation sådan att $Y \sim N(\alpha + \beta x, \sigma^2)$. Givet x, så är $\hat{\alpha} + \hat{\beta}x$ är en väntevärdesriktig skattare för Y. Om (x, Y) är oberoende av tidigare observationer så är

$$T = \frac{Y - (\hat{\alpha} + \hat{\beta}x)}{s\sqrt{1 + \frac{1}{n} + \frac{n(x-\bar{x})^2}{nS_{xx} - S_x^2}}} \sim t_{n-2}.$$