1 Ett stickprov

1.1 Ett normalfördelat stickprov

Antag att $X_1, \ldots, X_n$ är oberoende $N(\mu, \sigma^2)$-fördelade variabler. Låt

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

beteckna stickprovsmedelvärdet och

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

stickprovsvariansen.

1.1.1 Ett normalfördelat stickprov. Känd varians

Då σ^2 är känd är

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

1.1.2 Ett normalfördelat stickprov. Okänd varians

Om σ^2 är okänd använder vi istället att S^2 är en väntevärdesriktig skattare av σ^2 och att

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

1.2 Skattning av proportioner

Antag att n oberoende försök görs och låt X vara frekvensen för en händelse med sannolikhet p. Då är $X \sim Bin(n,p)$. Om $np > 10$ och $n(1-p) > 10$ gäller att

$$\frac{p^* - p}{\sqrt{p^*(1-p^*)/n}} \approx \frac{p^* - p}{\sqrt{p(1-p)/n}} \overset{appr}{\sim} N(0,1).$$

där $p^* = X/n$.

2 Två stickprov

Not: Om vi har parade normalfördelade variabler $(X_1, Y_1), \ldots, (X_n, Y_n)$ där $X_i \sim N(\mu_X, \sigma_X^2)$, $Y_i \sim N(\mu_Y, \sigma_Y^2)$, $i = 1, \ldots, n$ och X_i, Y_i är mäts på samma individ/enhet i, kan vi bilda differenser $Z_i = X_i - Y_i$ och behandla de n nya variablerna som i enstickprovsfallet.

2.1 Två oberoende normalfördelade stickprov

Låt $X_1, \ldots, X_n$ vara oberoende, alla $N(\mu_X, \sigma_X^2)$-fördelade med stickprovsmedelvärde $\overline{X}$ samt stickprovsvarians S_X^2. Låt på samma sätt $Y_1, \ldots, Y_m$ vara oberoende, alla $N(\mu_Y, \sigma_Y^2)$-fördelade med stickprovsmedelvärde $\overline{Y}$ samt stickprovsvarians S_Y^2.

2.1.1 Två oberoende stickprov. Känd varians.

Då σ_X^2 och σ_Y^2 är kända är

$$\frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} \sim N(0, 1).$$

2.1.2 Två oberoende stickprov. Okänd, men lika varians.

Om $\sigma_X = \sigma_Y = \sigma$, men σ okänd är

$$S_P^2 = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n + m - 2}$$

(P:et står för *pooled*) en väntevärdesriktig skattare av σ^2 och

$$\frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{S_P\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t(n + m - 2).$$

2.1.3 Två oberoende stickprov. Varianser olika.

Då σ_X och σ_Y inte kan antas lika använder vi istället att

$$\frac{\overline{X} - \overline{Y} - (\mu_X - \mu_Y)}{\sqrt{S_X^2/n + S_Y^2/m}} \overset{appr}{\sim} t(v).$$

där frihetsgradsantalet skattas från våra observationer som

$$v = \left\lfloor \frac{\left(s_X^2/n + s_Y^2/m\right)^2}{\frac{(s_X^2/n)^2}{n-1} + \frac{(s_Y^2/m)^2}{m-1}} \right\rfloor .$$

Not: Antag att $X_1, \ldots, X_n$ oberoende, alla med samma fördelning som X med $E[X] = \mu_X$ och $Var(X) = \sigma_X^2$ och på samma sätt att $Y_1, \ldots, Y_m$ oberoende, alla med samma fördelning som Y och $E[Y] = \mu_Y$ och $Var(Y) = \sigma_Y^2$. Enligt centrala gränsvärdessatsen är påståendet ovan approximativt giltigt för stora stickprovsstorlekar n och m även då X och Y inte är normalfördelade.

2.2 Jämförelse av proportioner

Antag att $X \sim Bin(n, p_X)$ och $Y \sim Bin(m, p_Y)$ är oberoende. För stora n, m gäller

$$\frac{p_X^* - p_Y^* - (p_X - p_Y)}{\sqrt{p_X^*(1-p_X^*)/n + p_Y^*(1-p_Y^*)/m}} \overset{appr}{\sim} N(0,1)$$

där $p_X^* = X/n$, $p_Y^* = Y/m$. För specialfallet $p_X = p_Y = p$ gäller att

$$\frac{p_X^* - p_Y^*}{\sqrt{p^*(1-p^*)\,(1/n + 1/m)}} \overset{appr}{\sim} N(0,1)$$

där $p^* = (X+Y)/(n+m)$.

3 Goodness-of-fit test

3.1 Jämförelse mellan uppmätta och teoretiska frekvenser. Teoretiska frekvenser kända.

Låt $N_1, \ldots, N_k$ vara frekvenser för kategorierna $A_1, \ldots, A_k$ i totalt $n = \sum_{i=1}^k n_i$ oberoende försök. Låt $p_i = P(A_i),\ i = 1, \ldots, k$. Då är

$$Q = \sum_{i=1}^{k} \frac{(N_i - np_i)^2}{np_i} \overset{appr}{\sim} \chi^2(k-1)$$

då $np_i > 5,\ i = 1, \ldots, k$.

3.2 Jämförelse mellan uppmätta och teoretiska frekvenser. Teoretiska frekvenser skattade.

Låt $N_1, \ldots, N_k$ vara frekvenser för kategorierna $A_1, \ldots, A_k$ i totalt $n = \sum_{i=1}^{k} n_i$ oberoende försök. Låt $p_i(\theta) = P(A_i),\ i = 1, \ldots, k$ bero av en okänd parametervektor av längd m. Om θ skattas med maximum likelihoodskattaren $\hat{\theta}$ är

$$Q = \sum_{i=1}^{k} \frac{(N_i - np_i(\hat{\theta}))^2}{np_i(\hat{\theta})} \overset{appr}{\sim} \chi^2(k-1-m)$$

då $np_i > 5,\ i = 1, \ldots, k$.

4 Enkel linjär regression

Modell: $Y_1, \ldots, Y_n$ är oberoende och sådana att

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

där $\varepsilon_i \sim N(0, \sigma^2)$. Låt

$$S_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2,\ S_{xY} = \sum_{i=1}^{n} (x_i - \overline{x})(Y_i - \overline{Y}).$$

Minsta kvadratskattarna av β_0 och β_1 är

$$\beta_1^* = \frac{S_{xY}}{S_{xx}},\ \beta_0^* = \overline{Y} - \beta_1^* \overline{x}.$$

En väntevärdesriktig skattare av σ^2 är

$$S^2 = \frac{SSE}{n-2} = \frac{\sum_{i=1}^{n} (Y_i - (\beta_0^* + \beta_1^* x_i))^2}{n-2}.$$

Det gäller att

$$\frac{\beta_1^* - \beta_1}{S/\sqrt{S_{xx}}} \sim t(n-2).$$