Kapitel 10: "Comparing Two Means and Two Variances"

$X_1, \dots, X_m$ är stickprov på $N(\mu_x, \sigma_x^2)$
$Y_1, \dots, Y_n$ är stickprov på $N(\mu_y, \sigma_y^2)$ } oberoende

Sats $\bar{X} - \bar{Y}$ är $N(\mu_x - \mu_y, \sigma_x^2/m + \sigma_y^2/n)$

(Bevisas mha MGF som i kapitel 7, se även kapitel 9.)

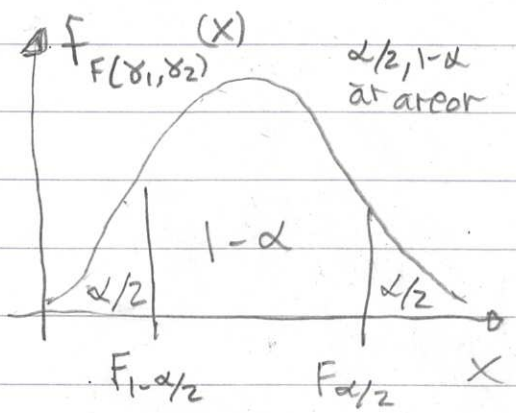
Definition Om $X^2_{\gamma_1}$ och $Y^2_{\gamma_2}$ är oberoende $\chi^2(\gamma_1)$- respektive $\chi^2(\gamma_2)$-fördelade är

$$\frac{X^2_{\gamma_1}/\gamma_1}{Y^2_{\gamma_2}/\gamma_2} \quad F(\gamma_1, \gamma_2)\text{-fördelad.}$$

$F(\gamma_1, \gamma_2)$ variabler är positiva.

I bokens tabell IX finner man talet F_α sådant att

$P(F(\gamma_1, \gamma_2) > F_\alpha) = \alpha$ för α nära 0 och 1



Sats $\dfrac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2}$ är $F(m-1, n-1)$-fördelad.

(Här är förstås S_x^2 och S_y^2 X- och Y-stickprovens stickprovsvarianser.)

Bevis Följer direkt av föregående definition med $\gamma_1 = m-1$ och $\gamma_2 = n-1$ eftersom $(m-1)S_x^2/\sigma_x^2$ och $(n-1)S_y^2/\sigma_y^2$ är $\chi^2(m-1)$ resp $\chi^2(n-1)$-fördelade. □

Sats $\dfrac{\sigma_x^2}{\sigma_y^2} \in \left[\dfrac{S_x^2}{S_y^2}\dfrac{1}{F_{\alpha/2}}, \dfrac{S_x^2}{S_y^2}\dfrac{1}{F_{1-\alpha/2}}\right]$ med konfidensgrad $1-\alpha$. ($m-1, n-1$ frihetsgrader)

Bevis Händelsen ovan är samma som $\dfrac{S_x^2/\sigma_x^2}{S_y^2/\sigma_y^2} \in [F_{1-\alpha/2}, F_{\alpha/2}]$. □

* Vi önskar testa $H_0: \sigma_x^2 = \sigma_y^2$ mot en av

$H_1: \sigma_x^2 \neq \sigma_y^2$
$H_1: \sigma_x^2 > \sigma_y^2$
$H_1: \sigma_x^2 < \sigma_y^2$

Sats Förkasta H_0 om $S_x^2/S_y^2 \notin \begin{cases} [F_{1-\alpha/2}, F_{\alpha/2}] \\ (0, F_\alpha] \\ [F_{1-\alpha}, \infty) \end{cases}$ med α-fel, α. ($m-1, n-1$ frihetsgrader)

(Om H_0 ej förkastas enligt satsen så godtages H_0.)

(Notera att de tre varianterna av testen är designade upptäcka motsvarande tre H_1-avvikelser från H_0.)

Sats Om $\sigma_x^2 = \sigma_y^2 = \sigma^2$ är $\dfrac{\bar{X}-\bar{Y}-(\mu_x-\mu_y)}{\sqrt{S_p^2(1/m+1/n)}}$ $T(m+n-2)$-fördelad

där $S_p^2 = \dfrac{(m-1)S_x^2+(n-1)S_y^2}{m+n-2}$.

Bevis. Överkurs. $\bar{X}-\bar{Y}$ är $N(\mu_x-\mu_y, \sigma^2/m+\sigma^2/n)$ och $((m-1)S_x^2+(n-1)S_y^2)/\sigma^2$ är $\chi^2(m+n-2)$ vilket ger

$$\frac{(\bar{X}-\bar{Y})-(\mu_x-\mu_y)}{\sqrt{\sigma^2(1/m+1/n)}} \Big/ \sqrt{S_p^2/\sigma^2} = \frac{N(0,1)}{\sqrt{\chi^2(m+n-2)/(m+n-2)}} = T(m+n-2). \square$$

Sats Om $\sigma_x^2 = \sigma_y^2$ är ($m+n-2$ frihetsgrader)

$$\mu_x-\mu_y \in \bar{X}-\bar{Y} \pm t_{\alpha/2}\sqrt{S_p^2(1/m+1/n)} \text{ med konfidensgrad } 1-\alpha.$$

Bevis. Följer direkt av föregående sats. $\square$

Sats Om $\sigma_x^2 = \sigma_y^2$ kan $H_0: \mu_x = \mu_y$ testas med α-fel α genom ($m+n-2$ frihetsgrader)

förkasta H_0 mot $\begin{cases} H_1: \mu_x \neq \mu_y \\ H_1: \mu_x > \mu_y \\ H_1: \mu_x < \mu_y \end{cases}$ om $\dfrac{\bar{X}-\bar{Y}}{\sqrt{S_p^2(1/m+1/n)}} \notin \begin{cases} [-t_{\alpha/2}, t_{\alpha/2}] \\ (-\infty, t_\alpha] \\ [-t_\alpha, \infty) \end{cases}$.

Bevis. Följer också direkt av ovan. $\square$

Man kan lätt modifiera testen ovan till test av $H_0: \mu_x - \mu_y = \mu_0$ mot något av alternativen $H_1: \mu_x - \mu_y \neq \mu_0$, $H_1: \mu_x - \mu_y > \mu_0$ eller $H_1: \mu_x - \mu_y < \mu_0$.

* Om man ej vet att $\sigma_x^2 = \sigma_y^2$ (= ofta realistiskt!) kan man ersätta

$$\sqrt{s_p^2(1/m + 1/n)} \text{ med } \sqrt{s_x^2/m + s_y^2/n}$$

i alla tre föregående resultat varefter de blir approximativ (istf exakt) sanna med

$T(\gamma)$-fördelning med $\gamma = \dfrac{(s_x^2/m + s_y^2/n)^2}{\dfrac{(s_x^2/m)^2}{m-1} + \dfrac{(s_y^2/n)^2}{n-1}}$:)

* Om $m = n$ behöver man ej veta att $\sigma_x^2 = \sigma_y^2$ utan man kan studera sk parade data

$Z_i = X_i - Y_i$, $i = 1, \dots, n$ som är $N(\underbrace{\mu_x - \mu_y}_{=\mu}, \underbrace{\sigma_x^2 + \sigma_y^2}_{=\sigma^2})$

och sedan m ha dem testa $H_0: \mu = \mu_x - \mu_y = \mu_0$ mot $H_1: \mu_x - \mu_y \gtrless \mu_0$ enligt metodiken från kapitel 8.

Wilcoxon rank summe test

(Liksom i slutet av kapitel 8 tittar vi nu på icke-parametriska tester där vi ej behöver antaga att våra stickprov är N-fördelade.)

* $X_1, \dots, X_m$ har teoretisk median M_x
  $Y_1, \dots, Y_n$ har teoretisk median M_y $\quad m \leq n$

* Vi vill testa $H_0: M_x = M_y$ mot en av

$H_1: M_x \neq M_y$
$H_1: M_x > M_y$
$H_1: M_x < M_y$

1) Ordna alla data $X_1, \dots, X_m, Y_1, \dots, Y_n$ i växande ordning $Z_1 < Z_2 < \dots < Z_{m+n}$ ~ samma siffror ordnade

2) Till dela $Z_1, \dots, Z_{m+n}$ rankerna $R_1 = 1, \dots, R_{m+n} = m+n$ och låt W_x och W_y vara summan av alla X- respektive Y-ranker så att

$$W_x + W_y = \frac{(m+n)(m+n+1)}{2}$$

3) Om H_0 är sann bör $W_x \approx \frac{m(m+n+1)}{2}$. Vad som är tillräcklig avvikelse från detta för kunna förkasta H_0 framgår av tabell X i boken.

4) Tabell X visar onormalt små värde och sedan är

$$\underset{\text{stort värde}}{\text{onormalt}} = \frac{m(m+n+1)}{2} + \left(\frac{m(m+n+1)}{2} - \underset{\text{litet värde}}{\text{onormalt}}\right) = m(m+n+1) - \underset{\text{litet värde}}{\text{onormalt}}$$

Exempel (Example 10.6.1 i M&A) Jag har två datamaterial med $m=12$ och $n=15$

Brand B		Brand A	
69.3	52.6	28.6	30.6
56.0	34.4	25.1	31.8
22.1	60.2	26.4	41.6
47.6	43.8	34.9	21.1
53.2		29.8	36.0
48.1		28.4	37.9
23.2		38.5	13.9
13.8		30.2	

Jag utför steg 1-2 i Wilcoxon ranksumme test:

Observation	13.8	13.9	21.1	22.1	23.2	25.1	26.4	28.4	28.6
Brand	*B*	*A*	*A*	*B*	*B*	*A*	*A*	*A*	*A*
Rank	1	2	3	4	5	6	7	8	9

Observation	29.8	30.2	30.6	31.8	34.4	34.9	36.0	37.9	38.5
Brand	*A*	*A*	*A*	*A*	*B*	*A*	*A*	*A*	*A*
Rank	10	11	12	13	14	15	16	17	18

Observation	41.6	43.8	47.6	48.1	52.6	53.2	56.0	60.2	69.3
Brand	*A*	*B*	*B*	*B*	*B*	*B*	*B*	*B*	*B*
Rank	19	20	21	22	23	24	25	26	27

$W_B = 1+4+5+14+20+21+22+23+24+25+26+27 = 212$

och önskar testa $H_0: M_B = M_A$ mot $H_1: M_B > M_A$ ensidigt.
Enligt tabell X är ett $\alpha = 0.05$ onormalt stort värde

$m(m+n+1) - \underset{\text{litet värde}}{\text{onormalt}} = 12(12+15+1) - 134.202 = 201.8$

så H_0 kan förkastas och H_1 godtas.

Om $\underline{m=n}$ kan man studera $\underline{\text{parade data}}$

$Z_i = X_i - Y_i$ för $i = 1, \dots, m$

så kan $H_0: M_X = M_Y$ testas mot $H_1: M_X \overset{\neq}{\gtrless} M_Y$ genom använda Wilcoxon signed rank test från slutet kapitel 8 för Z-data och testa $H_0: M_Z = 0$ mot $H_1: M_Z \overset{\neq}{\gtrless} 0$ mha. denna/dessa

Kapitel 11: "Simple Linear Regression and Correlation"

* **Linjär regression.** Jag har ett datamaterial $(x_i, Y_i)_{i=1}^n$ där $Y_i = \beta_0 + \beta_1 x_i + E_i$ med

$\begin{cases} x_1, \dots, x_n \text{ kända deterministiska data} \\ \beta_0 \text{ och } \beta_1 \text{ parametrar med okända värden} \\ E_1, \dots, E_n \text{ oberoende } N(0, \sigma^2) \text{ variabler med okänt värde} \\ Y_1, \dots, Y \text{ kända observerade stokastiska data} \end{cases}$

* Uppgiften är att skatta β_0 och β_1.

* Lösning: Minsta kvadratmetoden - minimera

$$SSE = \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 x_i))^2 \text{ map } \beta_0 \text{ och } \beta_1$$

$$\begin{cases} \frac{\partial SSE}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - (\beta_0 + \beta_1 x_i)) = 0 \\ \frac{\partial SSE}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (Y_i - (\beta_0 + \beta_1 x_i)) = 0 \end{cases} \text{ ger}$$

Satz.
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x} \text{ och } \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i Y_i - n \bar{x} \bar{Y}}{\sum_{i=1}^{n} x_i^2 - n \bar{x}^2}$$

(**Bevis** Bara att flytta om i ekvationerna ovan. □)

Sats
$\hat{\beta}_0$ är $N\left(\beta_0, \frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i-\bar{x})^2}\sigma^2\right)$, $\hat{\beta}_1$ är $N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right)$

(Boken sid 388 står en del fakta som kan göra enklare beräkning av summor men de är lite utdaterade - det är länge sedan datorer blev billigare använda än utbildade människor.)

(Bevis. Övekurs. Detta går göra själv med noga användning av det jag lärt er men är "risigt". □)

Sats $\hat{\sigma}^2 = \frac{SSE}{n-2}$ med $(\beta_0, \beta_1) = (\hat{\beta}_0, \hat{\beta}_1)$ är uvr skattning av σ^2

Sats SSE/σ^2 är $\chi^2_{(n-2)}$ så att följande är $T(n-2)$:

$$(\hat{\beta}_0 - \beta_0) \Big/ \sqrt{\sum_{i=1}^{n} x_i^2 \frac{SSE}{n-2} \Big/ n\sum_{i=1}^{n}(x_i-\bar{x})^2} \quad \text{o} \quad (\hat{\beta}_1 - \beta_1) \Big/ \sqrt{\frac{SSE}{n-2} \Big/ \sum_{i=1}^{n}(x_i-\bar{x})^2}$$

(Mha satsen ovan går göra konfidensintervall för β_0 och β_1 i analogi med kapitel 8 samt testa hypoteser angående värdet för β_0 och β_1.)

Sats ($n-2$ frihetsgrader)

$$\mu_{Y|x} = E(Y) = \beta_0 + \beta_1 x \in \hat{\beta}_0 + \hat{\beta}_1 x \pm t_{\alpha/2}\sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\sqrt{\frac{SSE}{n-2}}$$

$$Y \in \hat{\beta}_0 + \hat{\beta}_1 x \pm t_{\alpha/2}\sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}\sqrt{\frac{SSE}{n-2}}$$

med konfidensgrad $1-\alpha$.

(Konfidensintervallet för Y gäller för en "ny" observation av Y variabeln, inte $Y_1, \dots, Y_n$.)

Korrelation. Nu ändras grundförutsättningarna:

$(X_i, Y_i)_{i=1}^n$ är stickprov på (X, Y)

(Här kan X och Y (liksom då X_i och Y_i) bero av varandra medan $(X_1,Y_1),\dots,(X_n,Y_n)$ är oberoende.)

Sats $\rho = \dfrac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$ skattas med $\hat{\rho} = \dfrac{\frac{1}{n}\sum_{i=1}^n (x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^n (x_i-\bar{x})^2 \frac{1}{n}\sum_{i=1}^n (y_i-\bar{y})^2}}$

Sats

$H_0: \rho = 0$ förkastas mot $\begin{cases} H_1: \rho \neq 0 \\ H_1: \rho > 0 \\ H_1: \rho < 0 \end{cases}$ om $\dfrac{\hat{\rho}\sqrt{n-2}}{\sqrt{1-\hat{\rho}^2}} \notin \begin{cases} [-t_{\alpha/2}, t_{\alpha/2}] \\ (-\infty, t_\alpha] \\ [-t_\alpha, \infty) \end{cases}$

med α-fel α — $n-2$ frihetsgrader

Sats $\frac{1}{2}\ln\left(\frac{1+\hat{\rho}}{1-\hat{\rho}}\right) \approx N\left(\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right), \frac{1}{n-3}\right)$ så att

$\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) \in \frac{1}{2}\ln\left(\frac{1+\hat{\rho}}{1-\hat{\rho}}\right) \pm \lambda_{\alpha/2}\sqrt{\frac{1}{n-3}}$ approximativ, med konfidensgrad $1-\alpha$.

Genom "lösa ut" ρ fås konfidensintervall för ρ.

Man kan testa tex $H_0: \rho = \rho_0$ mot $H_1: \rho \neq \rho_0$ med α-fel α genom förkasta H_0 om $\frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right) \notin$ intervallet.