MS-A0509 Grundkurs i sannolikhetskalkyl och statistik Sammanfattning, del II

G. Gripenberg

Aalto-universitetet

14 februari 2014

Mätskalor

- *Nominalskala: Olika grupper utan naturlig ordning.*
- *Ordinalskala: Olika grupper med en naturlig ordning.*
- *Intervallskala: Numeriska värden, skillnader meningsfulla, nollan godtycklig.*
- *Kvotskala: Numeriska värden, naturligt nollvärde.*

Stickprov

- *Målsättningen är att få information om slumpvariabeln X som inte behöver vara reell.*
- *För att få information gör man tex. n mätningar som ger resultaten $x_1, x_2, \ldots, x_n$ och man tänker att x_j är värdet av en slumpvariabel X_j.*
- *Slumpvariablerna $X_1, \ldots, X_n$ är ett stickprov av storleken n och $x_1, x_2, \ldots, x_n$ är ett observerat stickprov av storleken n.*
- *Vi antar (vanligen och utan att säga det explicit) att $X_1, X_2, \ldots, X_n$ är oberoende och har samma fördelning, som är fördelningen av den slumpvariabel vi är intresserade av.*

Obs!

Antagandet att slumpvariablerna X_j i ett stickprov förutsätter att vi använder "dragning med återläggning", men detta villkor uppfylls sällan! Det finns dessutom många andra större svårigheter när man i praktiken skall ta ett stickprov och detta är ett viktigt problem!

Aritmetiskt medelvärde

Om X_j, $j = 1, \ldots, n$ är ett stickprov av slumpvariabeln X så är dess (aritmetiska) medelvärde

$$\overline{X} = \frac{1}{n}\sum_{j=1}^{n} X_j$$

och

$$\mathrm{E}(\overline{X}) = \mathrm{E}(X) \quad \textit{och} \quad \mathrm{Var}(\overline{X}) = \frac{1}{n}\mathrm{Var}(X).$$

Stickprovsvarians

Om X_j, $j = 1, \ldots, n$ är ett stickprov av slumpvariabeln X så är dess stickprovsvarians

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

och *(vilket är motiveringen för valet av $n-1$ istället för n i nämnaren)*

$$\mathrm{E}(S^2) = \mathrm{Var}(X).$$

χ^2-fördelningen

Ifall $X_j \sim \mathrm{N}(0,1)$ är oberoende och

$$Y = \sum_{i=1}^{n} X_i^2$$

så säger vi att Y är χ^2-fördelad med n frihetsgrader eller $Y \sim \chi^2(n)$. Då är

$$\mathrm{E}(Y) = n \quad \textit{och} \quad \mathrm{Var}(Y) = 2n$$

och Y har täthetsfunktionen

$$f(x) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} \mathrm{e}^{-\frac{x}{2}}, \quad x \geq 0.$$

och $f(x) = 0$ då $x < 0$.

Stickprovsvarians för normalfördelningen

Om X_j, $j = 1, 2, \ldots, n$ är ett stickprov av en $\mathrm{N}(\mu, \sigma^2)$ fördelad slumpvariabel så gäller för stickprovsvariansen

$$\frac{n-1}{\sigma^2} S^2 \sim \chi^2(n-1).$$

t-fördelningen

Ifall $Z \sim \mathrm{N}(0,1)$ *och* $Y \sim \chi^2(m)$ *är oberoende och*

$$W = \frac{Z}{\sqrt{\frac{1}{m}Y}}$$

så säger vi att W *är* t*-fördelad med* m *frihetsgrader eller* $W \sim t(m)$.
Då är $\mathrm{E}(W) = 0$ *om* $m > 1$ *och* $\mathrm{Var}(W) = \frac{m}{m-2}$ *om* $m > 2$ *och täthetsfunktionen för* W *är*

$$f(x) = \frac{1}{\sqrt{m\pi}} \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m}{2}\right)} \left(1 + \frac{x^2}{m}\right)^{-\frac{m+1}{2}}, \quad x \in \mathbb{R}.$$

Stickprov av normalfördelningen

Om X_j, $j = 1, 2, \ldots, n$ *är ett stickprov av en* $\mathrm{N}(\mu, \sigma^2)$ *fördelad slumpvariabel så är*

$$\frac{\overline{X} - \mu}{\sqrt{\frac{1}{n}S^2}} \sim t(n-1).$$

Punktestimat och estimator

Antag att vi vet (eller tror) att X *är en slumpvariabel med frekvens- eller täthetsfunktion* $f(x, \theta)$ *där parametern* θ *(som också kan vara en vektor) är okänd. Vad kan man göra för att estimera eller skatta* θ*?*

- *Ta ett observerat stickprov* x_j, $j = 1, \ldots, n$
- *Räkna ut ett estimat* $\hat{\theta} = g(x_1, x_2, \ldots, x_n)$ *där* g *är någon funktion.*
- *Observera att* $\hat{\theta}$ *är ett tal eller en vektor medan* $\hat{\Theta} = g(X_1, X_2, \ldots, X_n)$ *är en slumpvariabel.*
- *Ibland är det funktionen* g *som avses med ordet estimator och ibland slumpvariabeln* $\hat{\Theta}$.

Intervallestimat

Istället för att bara räkna ut ett tal (eller en vektor) som estimat för en parameter kan man också räkna ut ett intervall.

Momentmetoden

Om frekvens- eller täthetsfunktionen $f(x,\theta)$ för en sannolikhetsfördelning är sådan att θ kan skrivas som en funktion av $\mathrm{E}(X)$, dvs. $\theta = h(\mathrm{E}(X))$ där X har täthetsfunktionen $f(x,\theta)$ så är momentestimatorn av θ

$$\hat{\Theta} = h\left(\frac{1}{n}\sum_{j=1}^{n} X_j\right).$$

Om parametern, eller parametrarna kan skrivas som en funktion $h(\mathrm{E}(X), \mathrm{E}(X^2))$ blir estimatorn på motsvarande sätt $\hat{\Theta} = h\left(\frac{1}{n}\sum_{j=1}^{n} X_j, \frac{1}{n}\sum_{j=1}^{n} X_j^2\right)$.

"Maximum likelihood"- metoden

Om $f(x,\theta)$ är en frekvens- eller täthetsfunktion för en sannolikhetsfördelning så är "Maximum likelihood"-estimatet av θ talet $\widehat{\theta}$ sådant att

$$L(\widehat{\theta}, x_1, x_2, \ldots, x_n) = \max_{\theta} L(\theta, x_1, x_2, x_n),$$

där

$$L(\theta, x_1, x_2, x_n) = f(x_1,\theta)\cdot f(x_2,\theta)\cdot \ldots \cdot f(x_n,\theta)$$

är den sk. "likelihood"-funktionen och x_j, $j = 1,\ldots,n$ är ett observerat stickprov av en slumpvariabel med frekvens- eller täthetsfunktionen $f(x,\theta)$. I det diskreta fallet är $L(\theta, x_1, x_2, x_n)$ sannolikheten för att man då parametern är θ får det observerade stickprovet x_j, $j = 1,\ldots,n$. I fallet med täthetsfunktion är $(2h)^n L(\theta, x_1, \ldots, x_n)$ för små positiva h ungefär sannolikheten att få ett observerat stickprov y_j, $j = 1,\ldots,n$ så att $|y_j - x_j| < h$ för alla j.

Konfidensintervall

Ett konfidensintervall med konfidensgraden α för en parameter θ i en sannolikhetsfördelning är en intervallestimator
$I(X_1, X_2, \ldots, X_n) = [a(X_1, X_2, \ldots, X_n), b(X_1, X_2, \ldots, X_n)]$ *så att*

$$\Pr\big(\theta \in I(X_1, X_2, \ldots, X_n)\big) = \alpha.$$

Oftast används också ordet konfidensintervall för intervallet $I(x_1, x_2, \ldots, x_n)$, *dvs. värdet av slumpvariabeln när man fått ett observerat stickprov* x_j, $j = 1, \ldots, n$.

Obs!

Vanligen väljer man konfidensintervallet symmetriskt så att

$$\Pr(\theta < a(X_1, X_2, \ldots, X_n)) = \Pr(\theta > b(X_1, X_2, \ldots, X_n)) = \frac{1}{2}(1 - \alpha).$$

Oftast får man nöja sig med att villkoren för konfidensintervallet gäller endast approximativt.

Konfidensintervall för väntevärdet då $X \sim \mathrm{N}(\mu, \sigma^2)$

Om $X_1, X_2, \ldots, X_n$ *är ett stickprov med medelvärde* $\overline{X}$ *och stickprovsvarians* S^2 *av en* $\mathrm{N}(\mu, \sigma^2)$ *fördelad slumpvariabel så är*

$$\left[\overline{X} - \sqrt{\frac{S^2}{n}} F^{-1}_{\mathrm{t}(n-1)}\left(\frac{1+\alpha}{2}\right), \overline{X} + \sqrt{\frac{S^2}{n}} F^{-1}_{\mathrm{t}(n-1)}\left(\frac{1+\alpha}{2}\right)\right],$$

ett konfidensintervall för μ *med konfidensgraden* α.

Konfidensintervall för p då $X \sim$ Bernoulli(p)

Om $X_1, X_2, \ldots, X_n$ *är ett stickprov med medelvärde* $\overline{X}$ *av en* Bernoulli(p)*-fördelad slumpvariabel så är*

$$\left[\overline{X} - \sqrt{\frac{\overline{X}(1-\overline{X})}{n}} F^{-1}_{\mathrm{N}(0,1)}\left(\frac{1+\alpha}{2}\right), \overline{X} + \sqrt{\frac{\overline{X}(1-\overline{X})}{n}} F^{-1}_{\mathrm{N}(0,1)}\left(\frac{1+\alpha}{2}\right)\right]$$

ett **approximativt** *konfidensintervall för* μ *med konfidensgraden* α.

Obs!

Ofta används beteckningen

$$t_\alpha = t_\alpha(m) = -F^{-1}_{\mathrm{t}(m)}(\alpha),$$

vilket alltså betyder att om X *är en* $\mathrm{t}(m)$ *fördelad slumpvariabel så är*

$$\Pr(X \leq -t_\alpha) = \Pr(X \geq t_\alpha) = \alpha \quad \textit{och} \quad \Pr(|X| \geq t_\alpha) = 2\alpha.$$

Motsvarande beteckning för normalfördelningen $\mathrm{N}(0,1)$ *är* z_α.

Konfidensintervall för σ^2 då $X \sim \mathrm{N}(\mu, \sigma^2)$

Om $X_1, X_2, \ldots, X_n$ *är ett stickprov med stickprovsvarians* S^2 *av en* $\mathrm{N}(\mu, \sigma^2)$ *fördelad slumpvariabel så är*

$$\left[\frac{(n-1)S^2}{F^{-1}_{\chi^2(n-1)}\left(\frac{1+\alpha}{2}\right)}, \frac{(n-1)S^2}{F^{-1}_{\chi^2(n-1)}\left(\frac{1-\alpha}{2}\right)} \right]$$

ett konfidensintervall för σ^2 *med konfidensgraden* α.

Hypotesprövning

- *Man prövar om det finns skäl att förksasta en hypotes* H_0, *nollhypotesen, och i stället acceptera dess alternativ* H_1.
- *För att kunna göra några beräkningar måste man som nollhypotes välja ett tillräckligt entydigt påstående, tex.* $\theta = \theta_0$ *och inte* $\theta \neq \theta_0$ *som är för diffust.* *Oftast räcker det om nollhypotesen har ett entydigt extremfall, tex.* $\theta \leq \theta_0$.
- *I nollhypotesen ingår oftast många andra antaganden om fördelningar, oberoende osv. som kan ha stor betydelse för resultatet.*
- *När man tagit ett stickprov räknar man ut en testvariabel vars fördelning man åtminstone approximativt känner till.*
- *Med stöd av nollhypotesen räknar man ut sannolikheten, det sk.* **p-värdet**, *för att testvariabeln får ett minst lika "extremt" värde i förhållande till nollhypotesen som det observerade stickprovet gav.*
- *Om p-värdet är mindre än en given* **signifikansnivå** *förkastar man nollhypotesen och accepterar den alternativa hypostesen* H_1.
- *Signifikansnivån är alltså sannolikheten för att man förkastar nollhypotesen trots att den gäller.*

Normalfördelning, H_0: Väntevärde $= \mu_0$

- X_j, $j = 1, 2, \ldots, n$ *antas vara ett stickprov av* X *där* $X \sim N(\mu, \sigma^2)$.
- $H_0 : \mu = \mu_0$.
- *Testvariabel:* $T = \dfrac{\overline{X} - \mu_0}{\sqrt{\frac{S^2}{n}}} \sim t(n-1)$.
- *p-värde:* $2F_{t(n-1)}\left(-\dfrac{|\overline{x} - \mu_0|}{\sqrt{\frac{s^2}{n}}}\right)$.
- *Hypotesen förkastas på signifikansnivån* α *om* $p < \alpha$ *dvs. om testvariabeln får ett värde i mängden* $(-\infty, -t_{\frac{\alpha}{2}}) \cup (t_{\frac{\alpha}{2}}, \infty)$ *där* $t_{\frac{\alpha}{2}} = -F^{-1}_{t(n-1)}(\frac{\alpha}{2}) = F^{-1}_{t(n-1)}(1 - \frac{\alpha}{2})$.

Normalfördelning, H_0: Väntevärde $\leq \mu_0$ (eller $\geq \mu_0$)

- X_j, $j = 1, 2, \ldots, n$ *antas vara ett stickprov av* X *där* $X \sim N(\mu, \sigma^2)$.
- $H_0 : \mu \leq \mu_0$ *(eller* $\mu \geq \mu_0$*).*
- *Testvariabel:* $T = \dfrac{\overline{X} - \mu_0}{\sqrt{\frac{S^2}{n}}} \sim t(n-1)$.
- *p-värde:* $F_{t(n-1)}\left(-\dfrac{\overline{x} - \mu_0}{\sqrt{\frac{s^2}{n}}}\right)$ $\left(\text{ eller } F_{t(n-1)}\left(\dfrac{\overline{x} - \mu_0}{\sqrt{\frac{s^2}{n}}}\right)\right)$.
- *Hypotesen förkastas på signifikansnivån* α *om* $p < \alpha$ *dvs. om testvariabeln får ett värde i mängden* (t_α, ∞) *(eller* $(-\infty, t_\alpha)$*) där* $t_\alpha = -F^{-1}_{t(n-1)}(\alpha) = F^{-1}_{t(n-1)}(1 - \alpha)$.

Andel eller sannolikhet, H_0: $p = p_0$

- X_j, $j = 1, 2, \ldots, n$ *antas vara ett stickprov av* X *där* $X \sim \text{Bernoulli}(p)$.
- $H_0 : p = p_0$.
- *Testvariabel:* $Z = \dfrac{\overline{X} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim_a N(0,1)$.
- *Approximativt p-värde:* $2F_{N(0,1)}\left(-\dfrac{|\overline{x} - p_0|}{\sqrt{\frac{p_0(1-p_0)}{n}}}\right)$
- *Hypotesen förkastas på signifikansnivån* α *om* $p < \alpha$ *dvs. om testvariabeln får ett värde i mängden* $(-\infty, -z_{\frac{\alpha}{2}}) \cup (z_{\frac{\alpha}{2}}, \infty)$ *där* $z_{\frac{\alpha}{2}} = -F^{-1}_{N(0,1)}(\frac{\alpha}{2}) = F^{-1}_{N(0,1)}(1 - \frac{\alpha}{2})$.
- *Alternativt exakt med p-värdet* $1 - F_{\text{Bin}(n,p)}(n(p_0 + |\overline{x} - p_0|) - 1) + F_{\text{Bin}(n,p)}(n(p_0 - |\overline{x} - p_0|))$.
- *Likadana modifikationer för ensidiga hypoteser som för väntevärdet av en normalfördelad slumpvariabel.*

Normalapproximation

- X_j, $j = 1, 2, \ldots, n$ *antas vara ett stickprov av* X *med en fördelning så att* $f(X_1, \ldots, X_n) \sim_a N(\theta, \sigma^2)$.
- $H_0 : E(f(X_1, \ldots, X_n)) = \theta_0$.
- *Testvariabel:* $Z = \dfrac{f(X_1, \ldots, X_n) - \mu_0}{\widehat{\sigma}} \sim_a N(0,1)$ *där* $\widehat{\sigma}$ *är något estimat av* σ *när* H_0 *gäller.*
- *Approximativt p-värde:* $2F_{N(0,1)}\left(-\frac{|f(X_1,\ldots,X_n) - \theta_0|}{\widehat{\sigma}}\right)$.
- *Hypotesen förkastas på signifikansnivån* α *om* $p < \alpha$ *dvs. om testvariabeln får ett värde i mängden* $(-\infty, -z_{\frac{\alpha}{2}}) \cup (z_{\frac{\alpha}{2}}, \infty)$ *där* $z_{\frac{\alpha}{2}} = -F^{-1}_{N(0,1)}(\frac{\alpha}{2}) = F^{-1}_{N(0,1)}(1 - \frac{\alpha}{2})$.

Normalfördelning, två stickprov, samma varians, H_0: Samma väntevärde

- X_j, $j = 1, 2, \ldots, n_x$ *och* Y_j, $j = 1, 2, \ldots, n_y$ *antas vara stickprov av* $X \sim N(\mu_x, \sigma^2)$ *och* $Y \sim N(\mu_y, \sigma^2)$ *och alla slumpvariabler är oberoende.*
- $H_0 : \mu_x = \mu_y$.
- *Testvariabel:* $T = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\frac{(n_x-1)S_x^2+(n_y-1)S_y^2}{n_x+n_y-2} \cdot (\frac{1}{n_x} + \frac{1}{n_y})}} \sim t(n_x + n_y - 2)$.
- *p-värde:* $2F_{t(n_x+n_y-2)}\left(- \dfrac{|\overline{x} - \overline{y}|}{\sqrt{\frac{(n_x-1)s_x^2+(n_y-1)s_y^2}{n_x+n_y-2} \cdot (\frac{1}{n_x} + \frac{1}{n_y})}} \right)$
- *Hypotesen förkastas på signifikansnivån* α *om* $p < \alpha$ *dvs. om testvariabeln får ett värde i mängden* $(-\infty, -t_{\frac{\alpha}{2}}) \cup (t_{\frac{\alpha}{2}}, \infty)$ *där* $t_{\frac{\alpha}{2}} = -F^{-1}_{t(n_x+n_y-2)}(\frac{\alpha}{2}) = F^{-1}_{t(n_x+n_y-2)}(1 - \frac{\alpha}{2})$.

Två andelar eller sannolikheter

- X_j, $j = 1, 2, \ldots, n_x$ *och* Y_j, $j = 1, \ldots, n - y$ *antas vara ett stickprov av* X *och* Y *där* $X \sim \text{Bernoulli}(p_x)$ *och* $Y \sim \text{Bernoulli}(p_y)$.
- $H_0 : p_x = p_y$ *(eller* $p_x \le p_y$*).*
- *Testvariabel:* $Z = \dfrac{\overline{X} - \overline{Y}}{\sqrt{\hat{P}(1 - \hat{P})(\frac{1}{n_x} + \frac{1}{n_y})}} \sim_a N(0, 1)$ *där*

  $\hat{P} = \dfrac{n_x\overline{X} + n_y\overline{Y}}{n_x + n_y}$.
- *Approximativt p-värde:* $2F_{N(0,1)}\left(- \dfrac{|\overline{x} - \overline{y}|}{\sqrt{\hat{p}(1 - \hat{p})(\frac{1}{n_x} + \frac{1}{n_y})}} \right)$

  $\left(\text{eller } F_{N(0,1)}\left(- \dfrac{\overline{x}-\overline{y}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_x}+\frac{1}{n_y})}} \right) \right)$.
- *Hypotesen förkastas på signifikansnivån* α *om* $p < \alpha$ *dvs. om testvariabeln får ett värde i mängden* $(-\infty, -z_{\frac{\alpha}{2}}) \cup (z_{\frac{\alpha}{2}}, \infty)$ *där* $z_{\frac{\alpha}{2}} = -F^{-1}_{N(0,1)}(\frac{\alpha}{2}) = F^{-1}_{N(0,1)}(1 - \frac{\alpha}{2})$ *(eller i mängden* (z_α, ∞)*).*

Anpassning eller "Goodness-of-fit"

- X_j, $j = 1, \ldots, n$ *antas vara ett stickprov av en slumpvariabel med värdemängd* $\cup_{k=1}^{m} A_k$ *där mängderna* A_k *är disjunkta.*
- $H_0 : \Pr(X \in A_k) = p_k$, $k = 1, \ldots, m$.
- *Testvariabel:* $C = \sum_{k=1}^{m} \frac{(O_k - E_k)^2}{E_k} \sim_a \chi^2(m-1)$ *där* $E_k = np_k$ *och* O_k *är antalet element i* $\{ j : X_j \in A_k \}$.
- *p-värde:* $1 - F_{\chi^2(m-1)}(c)$
- *Hypotesen förkastas på signifikansnivån* α *om* $p < \alpha$ *dvs. om testvariabeln får ett värde i mängden* $(F^{-1}_{\chi^2(m-1)}(1-\alpha), \infty)$.
- *Om* $p_k = p_k(\theta_1, \ldots, \theta_j)$ *och dessa parametrar estimeras med stickprovet så gäller* $C \sim_a \chi^2(m-j-1)$.

Obs!

Om X_j, $j = 1, \ldots, n$ *är ett stickprov av en* Bernoulli(p)*-fördelad slumpvariabel så är* $\left(\frac{\overline{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \right)^2 = \sum_{k=0}^{1} \frac{(O_k - E_k)^2}{E_k}$ *där* $E_0 = (1-p)n$, $E_1 = pn$, $O_0 = n - \sum_{j=1}^{n} X_j$ *och* $O_1 = \sum_{j=1}^{n} X_j$, *dvs. de förväntade och observerade antalen nollor och ettor.*

Oberoende

- X_j, $j = 1, \ldots, n$ *antas vara ett stickprov av en slumpvariabel med värdemängd* $\cup_{i=1}^{r} \cup_{k=1}^{c} A_{i,k}$ *där mängderna* $A_{i,k}$ *är disjunkta.*
- $H_0 : \Pr(X \in A_{i,k}) = p_i p_k$, $i = 1, \ldots, r$ *och* $k = 1, \ldots, c$.
- *Testvariabel:* $C = \sum_{i=1}^{r} \sum_{k=1}^{c} \frac{(O_{i,k} - E_{i,k})^2}{E_{i,k}} \sim_a \chi^2((r-1)(c-1))$ *där* $O_{i,k}$ *är antalet element i* $\{ j : X_j \in A_{i,k} \}$ *och* $E_{i,k} = \frac{1}{n} \left(\sum_{m=1}^{c} O_{i,m} \right) \left(\sum_{m=1}^{r} O_{m,k} \right)$.
- *p-värde:* $1 - F_{\chi^2((r-1)(c-1))}(c)$
- *Hypotesen förkastas på signifikansnivån* α *om* $p < \alpha$ *dvs. om testvariabeln får ett värde i mängden* $(F^{-1}_{\chi^2((r-1)(c-1))}(1-\alpha), \infty)$.

F-fördelningen

Ifall $Y \sim \chi^2(m)$ *och* $X \sim \chi^2(n)$ *så är* $\dfrac{\frac{1}{m}Y}{\frac{1}{n}X} \sim \mathrm{F}(m,n)$, *dvs. följer*

F-*fördelningen med parametrarna (eller frihetsgraderna)* m *och* n.

Täthetsfunktionen för denna fördelning är $f_{\mathrm{F}(m,n)}(x) = \frac{1}{x\,\mathrm{B}\left(\frac{m}{2},\frac{n}{2}\right)}\sqrt{\frac{(mx)^m\, n^n}{(mx+n)^{m+n}}}$ *där* $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$.

Varianser

- X_j, $j = 1, 2, \ldots, n_x$ *och* Y_j, $j = 1, 2, \ldots, n_y$ *antas vara stickprov av* $X \sim \mathrm{N}(\mu_x, \sigma_x^2)$ *och* $Y \sim \mathrm{N}(\mu_y, \sigma_y^2)$ *och alla slumpvariabler är oberoende.*
- $\mathrm{H}_0 : \sigma_x^2 = \sigma_y^2$.
- *Testvariabel:* $F = \dfrac{S_y^2}{S_x^2} \sim \mathrm{F}(n_y - 1, n_x - 1)$.
- *p-värde:* $2(1 - F_{\mathrm{F}(n_y-1,n_x-1)}\left(\dfrac{s_y^2}{s_x^2}\right)$ *om* $\frac{s_y^2}{s_x^2} \geq 1$ *och*

  $2F_{\mathrm{F}(n_y-1,n_x-1)}\left(\dfrac{s_y^2}{s_x^2}\right)$ *om* $\frac{s_y^2}{s_x^2} \leq 1$.
- *Hypotesen förkastas på signifikansnivån* α *om* $p < \alpha$ *dvs. om testvariabeln får ett värde i mängden* $(-\infty, F^{-1}_{\mathrm{F}(n_y-1,n_x-1)}(\frac{\alpha}{2})) \cup (F^{-1}_{\mathrm{F}(n_y-1,n_x-1)}(1 - \frac{\alpha}{2}), \infty)$.
- *I detta fall är antagandet beträffande normalfördelning viktigt!*

Korrelation

Korrelationskoefficienten mellan slumpvariablerna X och Y är

$$\rho_{XY} = \mathrm{Cor}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}},$$

och om (X_j, Y_j), $j = 1, \ldots, n$ är ett stickprov av slumpvariabeln (X, Y) så är **stickprovskorrelationskoefficienten**

$$R_{XY} = \frac{\sum_{j=1}^{n}(X_j - \overline{X})(Y_j - \overline{Y})}{\sqrt{\sum_{j=1}^{n}(X_j - \overline{X})^2 \sum_{j=1}^{n}(Y_j - \overline{Y})^2}} = \frac{S_{xy}}{\sqrt{S_x^2 S_y^2}},$$

där

$$S_{xy} = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})(Y_j - \overline{Y}),$$

och

$$S_x^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X})^2, \quad S_y^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \overline{Y})^2.$$

Stickprovskorrelationskoefficientens fördelning

- *Ifall (X_j, Y_j), $i = 1, \ldots, n$ är ett stickprov av en normalfördelad slumpvariabel (X, Y) med korrelationskoefficient $\rho_{xy} = 0$ (och $\sigma_x^2 > 0$ och $\sigma_y^2 > 0$) så gäller*

$$\frac{R_{XY}\sqrt{n-2}}{\sqrt{1 - R_{XY}^2}} \sim \mathrm{t}(n-2).$$

- *Ifall (X_j, Y_j), $i = 1, \ldots, n$ är ett stickprov av en normalfördelad slumpvariabel (X, Y) med $-1 < \rho_{xy} < 1$ (och $\sigma_x^2 > 0$ och $\sigma_y^2 > 0$) så gäller*

$$\tfrac{1}{2}\ln\left(\frac{1 + R_{XY}}{1 - R_{XY}}\right) \sim_a \mathrm{N}\left(\tfrac{1}{2}\ln\left(\frac{1 + \rho_{XY}}{1 - \rho_{XY}}\right), \frac{1}{n-3}\right)$$

Minsta-kvadrat-metoden då $y \approx b_0 + b_1 x$

Om man antar att sambandet mellan x och y är $y \approx b_0 + b_1 x$, punkterna (x_j, y_j), $j = 1, \ldots, n$ är givna och man vill bestämma a och b_1 så att $\sum_{j=1}^{n} (y_j - b_0 - b_1 x_j)^2$ minimeras så kan det av många skäl vara bra att först räkna ut $\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$ och $\overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$ och sedan istället minimera

$$f(\tilde{b}_0, b_1) = \sum_{j=1}^{n} \Big((y_j - \overline{y}) - \tilde{b}_0 - b_1 (x_j - \overline{x}) \Big)^2 .$$

Eftersom $\sum_{j=1}^{n}(x_j - \overline{x}) = \sum_{j=1}^{n}(y_j - \overline{y}) = 0$ är $f_{\tilde{b}_0}(\tilde{b}_0, b_1) = 2n\tilde{b}_0$ så optimeringsvillkoret $f_{\tilde{b}_0}(\tilde{b}_0, b_1) = 0$ ger $\tilde{b}_0 = 0$. Nu är $f_{b_1}(0, b_1) = 2\sum_{j=1}^{n} \big(b_1 (x_j - \overline{x}) - (y_j - \overline{y}) \big)(x_j - \overline{x})$ så att ekvationen $f_{b_1}(0, b_1) = 0$ har lösningen

$$b_1 = \frac{\sum_{j=1}^{n}(x_j - \overline{x})(y_j - \overline{y})}{\sum_{j=1}^{n}(x_j - \overline{x})^2}.$$

Minsta-kvadrat-metoden, forts.

Koefficienten b_0 i uttrycket $y = b_0 + b_1 x$ blir då

$$b_0 = \overline{y} - b_1 \overline{x}.$$

Ett annat sätt att formulera samma räkning är att definiera matrisen M med $M(j,1) = 1$ och $M(j,2) = x_j$, dvs. $M = \begin{bmatrix} 1 & x_! \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$, vektorn Y med $Y(j,1) = y_j$, dvs. $Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$ och vektorn $C = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$. Funktionen som skall minimeras då kan skrivas som $\sum_{j=1}^{n} \big(Y(j,1) - \sum_{k=1}^{2} M(j,k) C(k,1) \big)^2 = \| Y - MC \|^2$. Minimipunkten uppnås därför då

$$\begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = C = (M^{\mathsf{T}} M)^{-1} M^{\mathsf{T}} Y$$

Regression

- *Slumpvariabeln Y antas förutom slumpen bero på variabeln x så att*

$$Y = \beta_0 + \beta_1 x + \varepsilon$$

där ε är en slumpvariabel som antas vara oberoende av x.

- *Ett stickprov av Y är därför av typen (x_j, Y_j), $j = 1, \ldots, n$ där $\varepsilon_j = Y_j - \beta_0 - \beta_1 x_j$ är oberoende slumpvariabler med samma fördelning, som vanligen antas vara $\mathrm{N}(0, \sigma^2)$.*
- *Med minsta kvadratmetoden (som är förnuftig precis då $\varepsilon \sim \mathrm{N}(0, \sigma^2)$) får vi följande estimatorer för β_1, β_0 och σ^2:*

$$B_1 = \frac{S_{xy}}{s_x^2},$$
$$B_0 = \overline{Y} - B_1 \overline{x},$$
$$S^2 = \frac{1}{n-2} \sum_{j=1}^{n} (Y_j - B_0 - B_1 x_j)^2,$$

där $S_{xy} = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \overline{x})(Y_j - \overline{Y})$.

Regression, testvariabler

- *Antag att $\varepsilon_j \sim \mathrm{N}(0, \sigma^2)$, $j = 1, \ldots, n$ är oberoende och $Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j$, $j = 1, \ldots, n$. Då är*

$$B_0 \sim \mathrm{N}\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}\right)\right),$$
$$B_1 \sim \mathrm{N}\left(\beta_1, \frac{\sigma^2}{(n-1)s_x^2}\right),$$
$$\frac{(n-2)S^2}{\sigma^2} \sim \chi^2(n-2).$$

- *Som testvariabler kan man använda*

$$T_0 = \frac{B_0 - \beta_0}{\sqrt{S^2 \left(\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}\right)}} \sim \mathrm{t}(n-2),$$
$$T_1 = \frac{B_1 - \beta_1}{\sqrt{\frac{S^2}{(n-1)s_x^2}}} \sim \mathrm{t}(n-2).$$

Samband mellan estimatorerna

Av definitionerna ovan följer också att

$$S^2 = \frac{n-1}{n-2} S_y^2 (1 - R_{xy}^2),$$

$$R_{xy} = B_1 \sqrt{\frac{S_x^2}{S_y^2}},$$

och

$$\frac{B_1}{\sqrt{\frac{S^2}{(n-1)S_x^2}}} = \frac{R_{xy}\sqrt{n-2}}{\sqrt{1-R_{xy}^2}}.$$

Det senare resultatet visar att test av nollhypoteserna $\beta_1 = 0$ och $\rho_{xy} = 0$ ger samma resulat (då man antar normalfördelning).
Talet r_{xy}^2, dvs. värdet av slumpvariabeln R_{xy}^2 sägs vara reggressionsmodellens förklaringsgrad.

Extrapolering

Om man har gjort mätningar av något slag och fått resultaten (x_j, y_j), $j = 1, \ldots, n$ så vill man ofta veta vilket värde y skulle få om $x = x_0$. Ett sätt att räkna ut ett rimligt svar är att anta att $y \approx b_0 + b_1 x$, bestämma b_0 och b_1 och sedan räkna ut $b_0 + b_1 x_0$. Ett enkelt sätt att förutom att göra denna räkning också få en uppfattning om hur stort felet kan bli är att ersätta värdena x_j, $j = 1, \ldots, n$ med $x_j - x_0$ och sedan i normal ordning räkna ut estimat och göra hypotesprövningar för β_0 i regressionsmodellen $Y = \beta_0 + \beta_1 x + \varepsilon$.