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

G. Gripenberg

Aalto-universitetet

13 februari 2015

Stickprov

- *Målsättningen är att få information om slumpvariabeln X.*
- *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 $X_1, X_2, \ldots, X_n$ är ett stickprov med storleken n och $x_1, x_2, \ldots, x_n$ är ett observerat stickprov med 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.*

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.*

Obs!

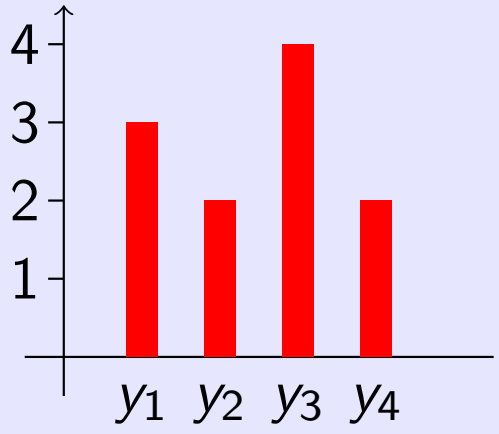
Antagandet att slumpvariablerna X_j i ett stickprov är oberoende förutsätter att vi använder "dragning med återläggning", men detta villkor uppfylls sällan! Det finns dessutom många andra större svårigheter när man i praktiken skall ta ett stickprov och detta är ett viktigt problem som inte behandlas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

Fördelningen av de observerade värdena och hur den beskrivs

Av de observerade värdena $x_1, x_2, \ldots, x_n$ i ett stickprov kan man bilda en diskret sannolikhetsfördelning, en sk. empirisk fördelning så att $\Pr(H = x) = \frac{1}{n}|\{\, j : x_j = x \,\}|$ *(som alltså är en jämn diskret fördelning om värdena är olika). Man kan beskriva denhär fördelningen med väntevärdet, variansen, medianen, andra kvantiler mm. men också med stapeldiagram eller histogram beroende på situationen.*

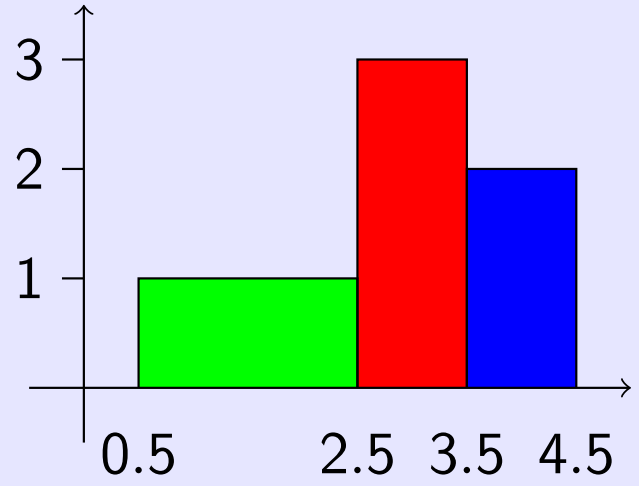
Stapeldiagram

Om mätskalan är en nominal- eller ordninalskala och/eller den ursprungliga slumpvariabeln är diskret så kan det observerade stickprovet $x_1, x_2, \ldots, x_n$ *beskrivas med ett stapeldiagram där höjden av varje stapel* y_k *är den observerade frekvensen* $f_k = |\{ j : x_j = y_k \}|$ *och alla staplar har samma bredd.*



Histogram

Om slumpvariabeln är kontinuerlig och mätskalan är en intervall- eller kvotskala så kan det observerade stickprovet $x_1, x_2, \ldots, x_n$ *beskrivas med ett histogram, dvs. klassindelade frekvenser så att man väljer klassgränser* $a_0 < a_1 < \ldots < a_m$*, räknar frekvenserna* $f_k = |\{ j : a_{k-1} < x_j \leq a_k \}|$ *och ritar dessa som rektanglar vars ytor är proportionella mot frekvenserna.*



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

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

eftersom väntevärdet är linjärt, variansen av en summa av oberoende slumpvariabler är summan av varianserna och $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$.

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

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

så att stickprovsvariansen är en väntevärdesriktig estimator av variansen vilket är motiveringen för valet av $n-1$ istället för n i nämnaren.

Obs

Om $x_1, x_2, \ldots, x_n$ är observerade värden i ett stickprov av slumpvariabeln X så är deras medeltal $\overline{x} = \frac{1}{n} \sum_{j=1}^{n} x_j$ och (observerade) stickprovsvarians $s^2 = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \overline{x})^2$.
Om man i Matlab/Octave har observationerna i vektorn `x` *så räknar man medelvärdet med kommandot* `mean(x)` *och stickprovsvariansen med kommandot* `var(x)`.

χ^2-fördelningen

Ifall $X_j \sim \mathrm{N}(0,1)$, $j = 1, 2, \ldots, m$, *är oberoende och*

$$C = \sum_{i=1}^{m} X_i^2$$

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

$$\mathrm{E}(C) = m \quad \textit{och} \quad \mathrm{Var}(C) = 2m,$$

och C *har täthetsfunktionen*

$$f_C(t) = \frac{1}{2^{\frac{m}{2}}\Gamma\left(\frac{m}{2}\right)} t^{\frac{m}{2}-1} \mathrm{e}^{-\frac{t}{2}}, \quad t \geq 0.$$

och $f_C(t) = 0$ *då* $t < 0$.

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* $C \sim \chi^2(m)$ *är oberoende och*

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

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* W *har täthetsfunktionen*

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

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 vi göra för att estimera eller skatta θ?

- *Vi tar ett observerat stickprov x_j, $j = 1, \ldots, n$ av X.*
- *Vi räknar 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 men om vi byter ut talen x_j mot motsvarande slumpvariabler X_j så får vi slumpvariablen $\hat{\Theta} = g(X_1, X_2, \ldots, X_n)$.*
- *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.

Exempel: Momentmetoden

Av slumpvariablen X har vi fått följande observationer 0.46, 0.20, 0.19, 0.09, 0.46 *och* 0.16. *Vi har skäl att tro att X är* $\mathrm{Exp}(\lambda)$*-fördelad men vi känner inte till parametern λ. Hur kan vi uppskatta, dvs. estimera λ? Eftersom vi vet att* $\mathrm{E}(X) = \frac{1}{\lambda}$ *så är det naturligt att räkna medelvärdet av de observerade värdena och vi får*

$$\overline{x} = \frac{1}{6}\sum_{j=1}^{6} = \frac{1}{6}(0.46 + 0.20 + 0.19 + 0.09 + 0.46 + 0.16) = 0.26,$$

och sedan använda detta tal istället för $\mathrm{E}(X)$ *i formeln* $\mathrm{E}(X) = \frac{1}{\lambda}$ *så att vi får estimatet*

$$\hat{\lambda} = \frac{1}{0.26} \approx 3.8.$$

För exponentialfördelningen kan vi alltså som estimator för parametern använda $\frac{1}{\overline{X}}$.

Denhär estimatorn är inte väntevärdesriktig eftersom $\mathrm{E}(\frac{1}{\overline{X}}) > \lambda$ *men då n växer närmar den sig det riktiga värdet, dvs.*

$\lim_{n\to\infty} \Pr\left(\left|\lambda - \left(\frac{1}{n}\sum_{j=1}^{n} X_j\right)^{-1}\right| > \epsilon\right) = 0$ *för alla* $\epsilon > 0$.

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. $\theta = h(\mathrm{E}(X))$ 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.

Exempel: Maximum-likelihood metoden mm

Du anländer till en främmande stad och på flygfältet ser du tre taxibilar med numrorna 57, 113 *och* 758. *Hur många taxibilar finns det i denhär staden?*

Vi antar att att det finns N *taxibilar med numrorna* $1, 2, \ldots, N$ *och att sannolikheten att en taxibil på flygfältet har nummer* j *är* $\frac{1}{N}$ *för alla* $j = 1, 2, \ldots, N$.

Om vi använder momentmetoden så skall vi räkna väntevärdet av en slumpvariabel X *som är jämnt fördelad i mängden* $\{1, \ldots, N\}$ *och det är* $\mathrm{E}(X) = \sum_{i=1}^{N} i \cdot \frac{1}{N} = \frac{N(N+1)}{2N} = \frac{N+1}{2}$, *så att* $N = 2\mathrm{E}(X) - 1$. *Sedan räknar vi medelvärdet av observationerna* $\overline{x} = \frac{1}{3}(57 + 113 + 758) = 309.33$ *och som estimat får vi* $\hat{N} = 2 \cdot 309.33 - 1 \approx 618$ *vilket är ett för litet antal.*

En annan möjlighet är att använda maximum-likelihood metoden: Om antalet taxibilar är N *så är sannolikheten* $\frac{1}{N}$ *att vi ser bilen med nummer* 57. *Samma sannolikhet gäller för bilarna med nummer* 113 *och* 758, *förutsatt att* $N \geq 758$ *för annars är sannolikheten* 0 *att vi ser en bil med nummer* 758.

Exempel: Maximum-likelihood metoden mm, forts.

Dethär betyder att

$$\mathcal{L}(N) = \Pr(\text{"}Du\ ser\ numrorna\ 57,\ 113\ och\ 758\text{"}) = \begin{cases} \frac{1}{N^3}, & N \geq 758, \\ 0, & N < 758. \end{cases}$$

I enlighet med maximum-likelihood metoden väljer vi estimatet $\hat{N}$ *så att likelihoodfunktionen* $\mathcal{L}(N)$ *får ett så stort värde som möjligt, dvs. i detta fall* $\hat{N} = 758$.
Motsvarande resultat gäller också mera allmänt, dvs. om $X_1, X_2, \ldots, X_k$ *är ett stickprov av en slumpvariabel som är jämnt fördelad i mängden* $\{1, 2, \ldots, N\}$ *(eller i det kontinuerliga fallet i intervallet* $[0, N]$*) så är maximum-likelihood estimatet av* N

$$\hat{N} = \max(X_1, X_2, \ldots, X_k).$$

Detta är inte ett väntevärdesriktigt estimat för det är klart att $\mathrm{E}(\hat{N}) < N$ *men vad är* $\mathrm{E}(\max(X_1, X_2, \ldots, X_k))$*?*

Exempel: Maximum-likelihood metoden mm, forts.

Nu är $\Pr(\max(X_1, X_2, \ldots, X_k) \leq m) = \Pr(X_j \leq m, j = 1, \ldots, k) = \left(\frac{m}{N}\right)^k$
av vilket följer att $\Pr(\max(X_1, X_2, \ldots, X_k) = m) = \left(\frac{m}{N}\right)^k - \left(\frac{m-1}{N}\right)^k$ *och väntevärdet blir*

$$\mathrm{E}\big(\max(X_1, X_2, \ldots, X_k)\big) = \sum_{m=1}^{N} m \left(\left(\frac{m}{N}\right)^k - \left(\frac{m-1}{N}\right)^k \right).$$

En följd av detta är att

$$\frac{k}{k+1} N < \mathrm{E}(\max(X_1, X_2, \ldots, X_k)) < \frac{k}{k+1} N + 1.$$

Dethär betyder att en bättre estimator för N kunde vara

$$\frac{k+1}{k} \max(X_1, X_2, \ldots, X_k),$$

som är väntevärdesriktigt i det kontinuerliga fallet
Ett bättre estimat för antalet taxibilar är alltså $\frac{4}{3} \cdot 758 \approx 1011$.

Exempel: Konfidensintervall för parametern i exponentialfördelningen

Vi antar att vi har ett stickprov av en $\mathrm{Exp}(\lambda)$*-fördelad slumpvariabel så att stickprovets storlek är* 50 *och medelvärdet är* 0.8*. Med momentmetoden får vi då estimatet* $\hat{\lambda} = \frac{1}{0.8} = 1.25$ *för parametern* λ *men här gäller det att bestämma ett intervall så att om vi med många olika stickprov med samma metod bestämmer ett intervall så kommer i stort sett tex.* 95% *av intervallen att vara sådana att parametern hör till det intervall vi räknat ut med hjälp av de observerade värdena i det fallet.*

För detta behöver vi en slumpvariabel vars fördelning vi åtminstone approximativt känner till, dvs. den innehåller inga okända parametrar. Med stöd av den centrala gränsvärdessatsen använder man för dethär ofta normalfördelningen $\mathrm{N}(0,1)$ *och det gör vi nu också.*

Vi struntar för en stund i de numeriska värdena och antar att vi har ett stickprov $X_1, X_2, \ldots, X_{50}$ *av en slumpvariabel* $X \sim \mathrm{Exp}(\lambda)$*. Väntevärdet av medelvärdet* $\overline{X} = \frac{1}{50}\sum_{j=1}^{n} X_j$ *är då* $\mathrm{E}(\overline{X}) = \mathrm{E}(X) = \frac{1}{\lambda}$ *och variansen* $\mathrm{Var}(\overline{X}) = \frac{1}{50}\mathrm{Var}(X) = \frac{1}{50} \cdot \frac{1}{\lambda^2}$.

Exempel: Konfidensintervall för parametern i exponentialfördelningen, forts.

Om vi tror att $n = 50$ *är tillräckligt stort så är*

$$\frac{\overline{X} - \frac{1}{\lambda}}{\sqrt{\frac{1}{50\lambda^2}}} \sim_a \mathrm{N}(0,1).$$

Ifall $Z \sim \mathrm{N}(0,1)$ *så gäller*

$$\Pr\left(F^{-1}_{\mathrm{N}(0,1)}(0.025) \leq Z \leq F^{-1}_{\mathrm{N}(0,1)}(0.975)\right) = \Pr(-1.96 \leq Z \leq 1.96) = 0.95,$$

så att

$$\Pr\left(-1.96 \leq \frac{\overline{X} - \frac{1}{\lambda}}{\sqrt{\frac{1}{50\lambda^2}}} \leq 1.96\right) \approx 0.95.$$

Nu är

$$-1.96 \leq \frac{\overline{X} - \frac{1}{\lambda}}{\sqrt{\frac{1}{50\lambda^2}}} \leq 1.96 \quad \Leftrightarrow \quad \frac{1 - \frac{1.96}{\sqrt{50}}}{\overline{X}} \leq \lambda \leq \frac{1 + \frac{1.96}{\sqrt{50}}}{\overline{X}},$$

Exempel: Konfidensintervall för parametern i exponentialfördelningen, forts.

så att sannolikheten att λ *ligger mellan slumpvariablerna* $\frac{0.72}{\overline{X}}$ *och* $\frac{1.28}{\overline{X}}$ *också är ungefär* 0.95. *Detta betyder att ett* 95% *approximativt konfidensintervall för parametern i exponentialfördelningen då stickprovets storlek är* 50 *är*

$$\left[\frac{0.72}{\overline{X}}, \frac{1.28}{\overline{X}}\right].$$

I dethär fallet blir konfidensintervallet $[0.9, 1.6]$.

För exponentialfördelningen är det inte speciellt svårt att få fram olikheter för parametern, men om detta inte skulle ha varit fallet (detta gäller tex. Bernoulli-fördelningen) så skulle vi i uttrycket $\frac{1}{\lambda^2}$ *för variansen ha kunnat använda estimatorn* $\overline{X}^{-1}$ *för* λ *och då skulle konfidensintervallet ha blivit*

$$\left[\frac{1}{\overline{X} + \frac{1.96}{\sqrt{50}}\overline{X}}, \frac{1}{\overline{X} - \frac{1.96}{\sqrt{50}}\overline{X}}\right] = \left[\frac{0.78}{\overline{X}}, \frac{1.38}{\overline{X}}\right],$$

och dethär konfidensintervallet blir $[0.97, 1.73]$ *om* $\overline{x} = 0.8$.

Konfidensintervall

Ett konfidensintervall med konfidensgraden $1-\alpha$ *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*

$$\Pr\big(\theta \in I(X_1, X_2, \ldots, X_n)\big) = 1 - \alpha.$$

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

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

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

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

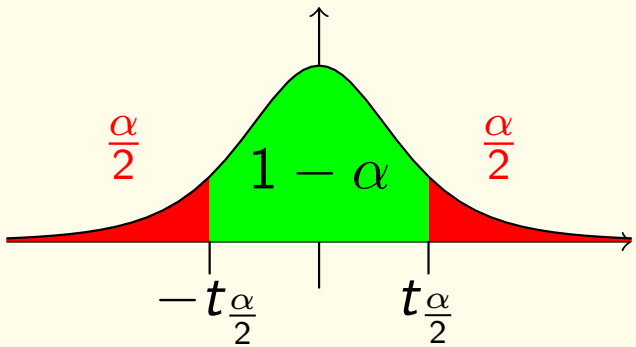
Om $X_1, X_2, \ldots, X_n$ är ett stickprov med medelvärde $\overline{X}$ och stickprovsvarians S^2 av en $\mathrm{N}(\mu, \sigma^2)$-fördelad slumpvariabel så är

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

ett konfidensintervall för μ med konfidensgraden $1-\alpha$.

Varför?

Ifall W är en $t(n-1)$-fördelad slumpvariabel och $t_{\frac{\alpha}{2}} = F^{-1}_{\mathrm{t}(n-1)}\left(1-\frac{\alpha}{2}\right) = -F^{-1}_{\mathrm{t}(n-1)}\left(\frac{\alpha}{2}\right)$ så är $\Pr(-t_{\frac{\alpha}{2}} \leq W \leq t_{\frac{\alpha}{2}}) = 1-\alpha$. Om nu $W = \frac{\overline{X}-\mu}{\sqrt{\frac{S^2}{n}}}$ så är $-t_{\frac{\alpha}{2}} \leq W \leq t_{\frac{\alpha}{2}}$ om och endast om $\overline{X} - t_{\frac{\alpha}{2}}\sqrt{\frac{S^2}{n}} \leq \mu \leq \overline{X} + t_{\frac{\alpha}{2}}\sqrt{\frac{S^2}{n}}$.



Konfidensintervall för p då $X \sim \text{Bernoulli}(p)$

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

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

ett **approximativt** *konfidensintervall för μ med konfidensgraden $1-\alpha$.*

Varför?

Om $\widetilde{Z}$ är approximativt $\mathrm{N}(0,1)$*-fördelad och $z_{\frac{\alpha}{2}} = F^{-1}_{\mathrm{N}(0,1)}\left(1-\frac{\alpha}{2}\right)$ så är* $\Pr(-z_{\frac{1}{2}\alpha} \leq \widetilde{Z} \leq z_{\frac{\alpha}{2}}) \approx 1-\alpha$. *Nu är $\frac{\overline{X}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim_a \mathrm{N}(0,1)$ men p ersätts i nämnaren med estimatorn $\overline{X}$ och om $\tilde{Z} = \frac{\overline{X}-p}{\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}}$ så är $-z_{\frac{1}{2}\alpha} \leq \widetilde{Z} \leq z_{\frac{\alpha}{2}}$ precis då $\overline{X} - z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}} \leq p \leq \overline{X} + z_{\frac{\alpha}{2}}\sqrt{\frac{\overline{X}(1-\overline{X})}{n}}$.*

Obs!

Ofta används beteckningen

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

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 *så att om* $Z \sim \mathrm{N}(0,1)$ *så är*

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

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

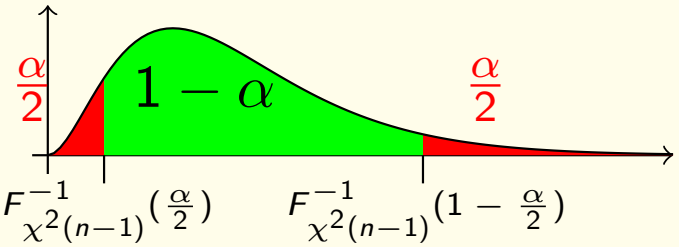
Om $X_1, X_2, \ldots, X_n$ är ett stickprov med stickprovsvarians S^2 av en $\mathrm{N}(\mu, \sigma^2)$ fördelad slumpvariabel så är

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

ett konfidensintervall för σ^2 med konfidensgraden $1-\alpha$.

Varför?

Om C är en $\chi^2(n-1)$ fördelad slumpvariabel så gäller $\Pr(C < F^{-1}_{\chi^2(n-1)}(\frac{\alpha}{2})) = \frac{\alpha}{2}$ *och* $\Pr(C > F^{-1}_{\chi^2(n-1)}(1-\frac{\alpha}{2})) = \frac{\alpha}{2}$. *Om nu* $C = \frac{(n-1)S^2}{\sigma^2}$ *så är* $\sigma^2 < \frac{(n-1)S^2}{F^{-1}_{\chi^2(n-1)}(1-\frac{\alpha}{2})}$ *då* $C > F^{-1}_{\chi^2(n-1)}(1-\frac{\alpha}{2})$ *och* $\sigma^2 > \frac{(n-1)S^2}{F^{-1}_{\chi^2(n-1)}(\frac{\alpha}{2})}$ *då* $C < F^{-1}_{\chi^2(n-1)}(\frac{\alpha}{2})$ *så att sannolikheten för båda händelserna är* $\frac{\alpha}{2}$.



Hypotesprövning

- *Vi undersöker om det finns skäl att förkasta en hypotes* H_0*, den sk.* **nollhypotesen**, *för att de resultat vi fått är mycket osannolika om nollhypotesen gäller eller om allt bara beror på slumpen.*
- *Nollhypotesen är vanligen ett motpåstående eller antites som vi behöver argument för att förkasta.*
- *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 men som man inte nödvändigtvis försöker förkasta.*

Hypotesprövning, forts.

- *När man tagit ett stickprov räknar vi ut värdet på en testvariabel som vi valt så att om nollhypotesen gäller så har testvariabeln (åtminstone approximativt) någon standardfördelning som vi känner väl 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.*
- *Signifikansnivån är alltså sannolikheten (ofta ett närmevärde och om nollhypotesen innehåller olikheter, en övre gräns) för att man förkastar nollhypotesen trots att den gäller.*
- *För att beräkna sannolikheten att man inte förkastar nollhypotesen fastän den inte gäller behövs specifika tilläggsantaganden vilket gör denna fråga svårare att behandla.*

Exempel: Hypotestestning

Till en poliklinik kommer i genomsnitt 9 *patienter i timmen. En dag då det varit halt väglag kommer det* 130 *patienter under* 12 *timmar.*
Kommer det mera patienter på grund av det dåliga väglaget eller är det frågan om slumpmässiga variationer?
Om det kommer i genomsnitt 9 *patienter i timmen så kan vi räkna med att väntevärdet av antalet patienter under* 12 *timmar är* $9 \cdot 12 = 108$ *och vi kan som nollhypotes ta antitesen till frågan om det kommit ovanligt många patienter att väntevärdet av antalet patienter är högst* 108.
Dessutom gör vi också antagandet att antalet patienter under 12 *timmar är* Poisson(λ)*-fördelat där alltså* $\lambda \leq 108$. *För räkningarna använder vi ändå extremfallet* $\lambda = 108$.
Det är ingen idé att räkna bara sannolikheten för att $\Pr(X = 130)$ *om* X *är antalet patienter, men däremot skall vi räkna sannolikheten* $\Pr(X \geq 130)$.
Om vi räknar med Poisson-fördelningens fördelningsfunktion får vi

$$p = \Pr(X \geq 130) = 1 - \Pr(X \leq 129) = 1 - F_{\text{Poisson}(108)}(129) = 0.021645.$$

Exempel: Hypotestestning, forts.

Om vi använder normalapproximation så får vi

$$p = \Pr(X \geq 130) = \Pr\left(\frac{X - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)}} \geq \frac{130 - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)}}\right)$$

$$= \Pr\left(\frac{X - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)}} \geq \frac{130 - 108}{\sqrt{108}}\right) = \Pr\left(\frac{X - \mathrm{E}(X)}{\sqrt{\mathrm{Var}(X)}} \geq 2.117\right) \approx 0.017132.$$

(Genom att räkna $1 - Pr(X \leq 129)$ *med normalapproximation kommar man närmare det exakta svaret.)*
Slutsatsen är i alla fall att nollhypotesen kan förkastas på signifikansnivån 0.05 *men inte på signifikansnivån* 0.01*.*
Om vi istället som nollhypotes tagit $\lambda = 108$*, vilket skulle ha varit förnuftigt om vi frågat om det varit en ovanlig dag på polikliniken, så borde vi också beakta möjligheten att det kommit väldigt få patienter och då skulle p-värdet ha blivit det dubbla (vilket inte exakt är* $\Pr(X \geq 130) + \Pr(X \leq 86)$*).*

Testa väntevärde, normalfördelning, exempel

Var mars 2014 en ovanlig månad beträffande nederbörden?
I mars 2014 var nederbördsmängderna på vissa mätstationer följande:

	1	2	3	4	5	6	7	8	9	10
Nederbörd	33	27	30	22	28	28	24	31	34	22

Motsarande medeltal för åren 1981–2010 var

	1	2	3	4	5	6	7	8	9	10
Medeltal	39	37	38	36	36	26	35	29	30	21

Nu är det förnuftigt att räkna hur mycket värdena för år 2014 avviker från medelvärdena och skillnaderna är följande:

	1	2	3	4	5	6	7	8	9	10
Skillnad	-6	-10	-8	-14	-8	2	-11	2	4	1

Testa väntevärde, normalfördelning, exempel, forts.

Eftersom frågan var om mars var en ovanlig månad så väljer vi som nollhypotes att den inte var det. Vi kan inte som nollhypotes använda antagandet att den var ovanlig för det ger ingenting som kan användas i räkningar och här sägs ingenting om på vilket sätt den eventuellt var ovanlig.

Nollhypotesen blir därför att skillnaden mellan nederbördsmängderna 2014 och medelvärdena från en längre tid är $\mathrm{N}(\mu, \sigma^2)$*-fördelade med* $\mu = 0$ *och att dehär skillnaderna på olika orter är oberoende.*

Medelvärdet av skillnaderna är -4.8 *och stickprovsvariansen är* 41.733*. Det betyder att testvariabeln* $W = \frac{\overline{X}-0}{\sqrt{\frac{S^2}{10}}}$ *får värdet* -2.3496*. Eftersom* W *enligt nollhypotesen har fördelningen* $t(10-1)$ *så blir p-värdet*

$$p = \Pr(|W-0| \geq |-2.3496-0|) = \Pr(W \leq -2.3496 \text{ eller } W \geq 2.3496)$$
$$= F_{\mathrm{t}(9)}(-2.3496) + 1 - F_{\mathrm{t}(9)}(2.3496) = 2F_{\mathrm{t}(9)}(-2.3496) = 0.043333,$$

så vi kan förkasta nollhypotesen på signifikansnivån 0.05*.*

Testa väntevärde, normalfördelning, exempel, forts.

Om frågan skulle ha varit om nederbördsmängden i mars 2014 var ovanligt liten skulle vi som nollhypotes ha valt påståendet att den inte var det, dvs. att fördelningen av skillnaderna är $\mathsf{N}(\mu, \sigma^2)$ *där* $\mu \geq 0$*. Testvariabeln skulle ha varit precis densamma men p-värdet skulle ha blivit*

$$p = \Pr(W \leq -2.3496) = F_{\mathrm{t}(9)}(-2.3496) = 0.021667.$$

Om frågan skulle ha varit om nederbördsmängden i mars 2014 var ovanligt stor skulle vi som nollhypotes ha valt påståendet att den inte var det, dvs. att fördelningen av skillnaderna är $\mathsf{N}(\mu, \sigma^2)$ *där* $\mu \leq 0$*. Eftersom medelvärdet är negativt är resultaten helt i enlighet med den här nollhypotesen så det finns inget skäl att förkasta den och vi behöver inte heller räkna ut stickprovsvariansen, det räcker att vi räknar medelvärdet.*

Normalfördelad slumpvariabel, testning av väntevärdet

Ifall X_j, $j = 1, 2, \ldots, n$ är ett stickprov av slumpvariabeln X som är $\mathrm{N}(\mu, \sigma^2)$-fördelad och nollhypotesen är $\mu = \mu_0$ (eller $\mu \leq \mu_0$ eller $\mu \geq \mu_0$) så väljer vi som testvariabel

$$\frac{\overline{X} - \mu_0}{\sqrt{\frac{S^2}{n}}} \sim \mathrm{t}(n-1),$$

där $\overline{X}$ är medelvärdet och S^2 stickprovsvariansen.

Obs!

Det är en följd av antagandet om normalfördelning att här inte används approximationer så det är inte nödvändigtvis ett problem om stickprovsstorleken n är liten.

Testning av andel eller sannolikhet med normalapproximation

Ifall X_j, $j = 1, 2, \ldots, n$ *är ett stickprov av en* Bernoulli(p)*-fördelad slumpvariabel och nollhypotesen är* $p = p_0$ *(eller* $p \leq p_0$ *eller* $p \geq p_0$*) så kan vi som testvariabel välja*

$$\frac{\overline{X} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \sim_a \mathrm{N}(0,1).$$

Vi kan lika väl räkna summan $Y = \sum_{j=1}^{n} X_j$ *av stickprovet som är* Bin(n, p)*-fördelad och testvariabeln (som inte ändras) kan vi skriva i formen*

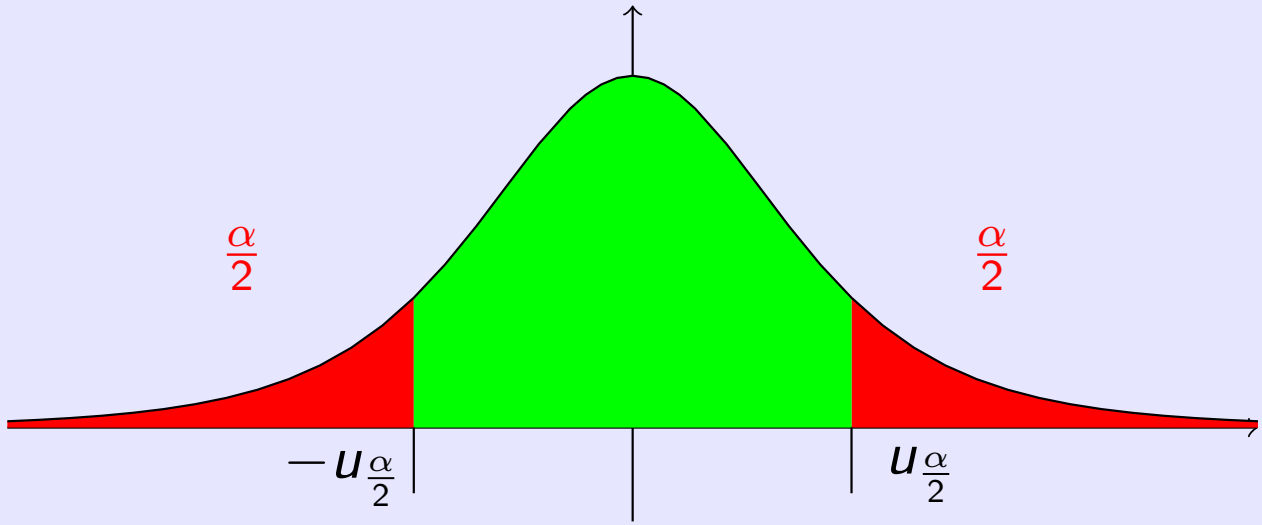
$$\frac{Y - np_0}{\sqrt{p_0(1-p_0)n}} \sim_a \mathrm{N}(0,1).$$

Obs!

I dethär fallet använder vi en approximativ fördelning och som en tumregel kan man använda att approximationen är tillräckligt bra om $\min(np_0, n(1-p_0)) \geq 10$.

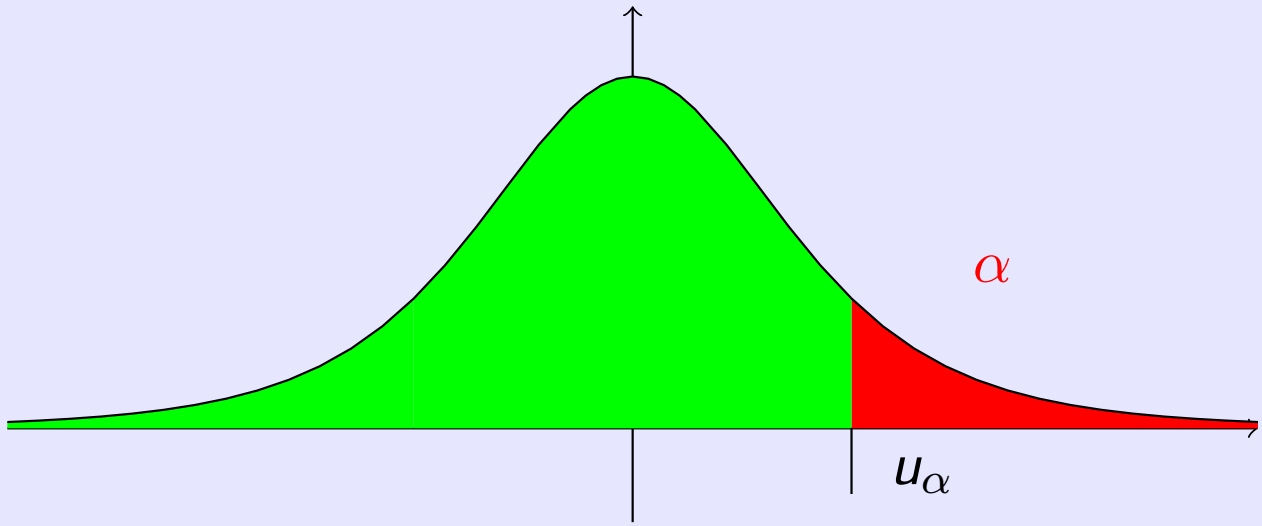
p-värde, kritiskt område, t(m)- eller N(0, 1)-testvariabel

- *Vi antar att testvariabeln U är $t(m)$- eller (approximativt) $N(0,1)$-fördelad så att dess fördelningsfunktion är F_U och att den i testet får värdet u_*.*
- *Om alternativet till nollhypotesen är tvåsidigt, dvs. nollhypotesen är $\mu = \mu_0$, $p = p_0$ osv., dvs. resultaten är helt i enlighet med nollhypotesen då testvariabeln är 0 så gäller:*
  - *p-värdet är $\Pr(U \leq -|u_*| \text{ eller } U \geq |u_*|) = 2F_U(-|u_*|)$.*
  - *Nollhypotesen förkastas på signifikansnivån α ifall $p < \alpha$ dvs. om testvariabelns värde ligger i det kritiska området $(-\infty, -u_{\frac{\alpha}{2}}) \cup (u_{\frac{\alpha}{2}}, \infty)$ där $u_{\frac{\alpha}{2}} = -F_U^{-1}(\frac{\alpha}{2}) = F_U^{-1}(1 - \frac{\alpha}{2})$.*



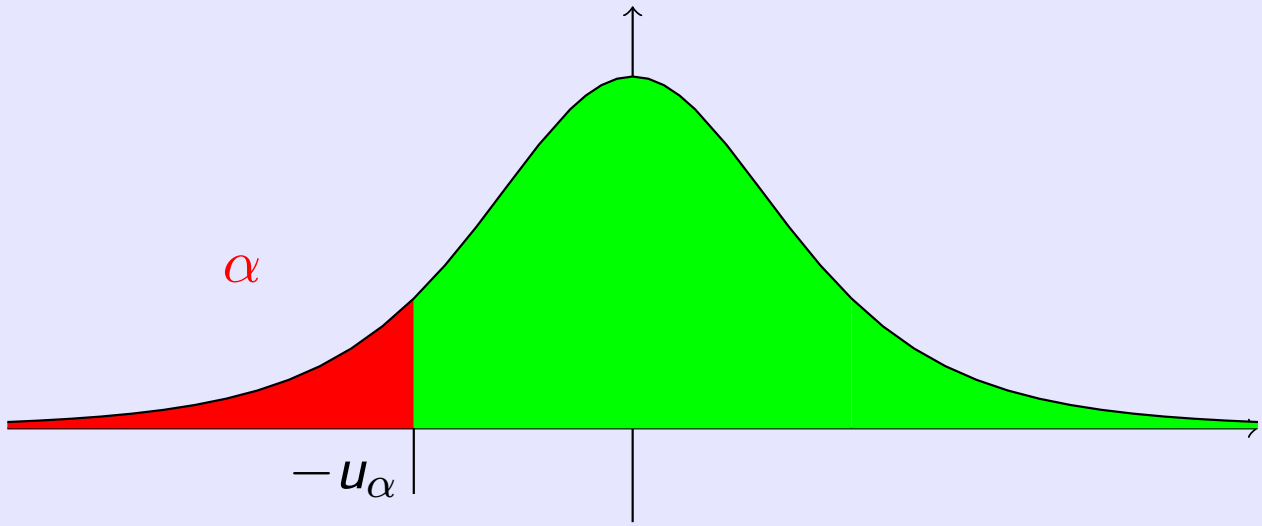
p-värde, kritiskt område, t(m)- eller N(0, 1)-testvariabel, forts.

- *Om alternativet till nollhypotesen är ensidigt, och nollhypotesen är $\mu \leq \mu_0$, $p \leq p_0$ osv., dvs. resultaten är helt i enlighet med nollhypotesen då testvariabeln är ≤ 0 så gäller*
  - *p-värdet är* $\Pr(U \geq u_*) = 1 - F_U(u_*)$.
  - *Nollhypotesen förkastas på signifikansnivån α om $p < \alpha$ dvs. om testvariablens värde ligger i det kritiska området (u_α, ∞) där* $u_\alpha = -F_U^{-1}(\alpha) = F_U^{-1}(1 - \alpha)$.



p-värde, kritiskt område, t(m)- eller N(0, 1)-testvariabel, forts.

- *Om alternativet till nollhypotesen är ensidigt, och nollhypotesen är $\mu \geq \mu_0$, $p \geq p_0$ osv., dvs. resultaten är helt i enlighet med nollhypotesen då testvariabeln är ≥ 0 så gäller*
  - *p-värdet är* $\Pr(U \leq u_*) = F_U(u_*)$.
  - *Nollhypotesen förkastas på signifikansnivån α om $p < \alpha$ dvs. om testvariabelns värde ligger i det kritiska området $(-\infty, -u_\alpha)$ där* $u_\alpha = -F_U^{-1}(\alpha) = F_U^{-1}(1-\alpha)$.



Testning av två väntevärden, normalfördelning, samma varians

Om X_j, $j = 1, 2, \ldots, n_x$ *och* Y_j, $j = 1, 2, \ldots, n_y$ *är (oberoende) stickprov av slumpvariablerna* X *och* Y *där* $X \sim \mathrm{N}(\mu_x, \sigma^2)$ *och* $Y \sim \mathrm{N}(\mu_y, \sigma^2)$ *och nollhypotesen är* $\mu_x = \mu_y$ *(eller* $\mu_x \leq \mu_y$ *eller* $\mu_x \geq \mu_y$*) så väljer vi som testvariabel*

$$\frac{\overline{X} - \overline{Y}}{\sqrt{\frac{(n_x-1)S_x^2+(n_y-1)S_y^2}{n_x+n_y-2} \cdot \left(\frac{1}{n_x} + \frac{1}{n_y}\right)}} \sim \mathrm{t}(n_x + n_y - 2).$$

Varför $\sqrt{\frac{(n_x-1)S_x^2+(n_y-1)S_y^2}{n_x+n_y-2} \cdot \left(\frac{1}{n_x} + \frac{1}{n_y}\right)}$

Eftersom $X_j \sim \mathrm{N}(\mu_x, \sigma^2)$ *och* $Y_j \sim \mathrm{N}(\mu_y, \sigma^2)$ *så gäller* $\frac{(n_x-1)S_x^2}{\sigma^2} \sim \chi^2(n_x - 1)$ *och* $\frac{(n_y-1)S_y^2}{\sigma^2} \sim \chi^2(n_y - 1)$ *och eftersom* X*- och* Y*-slumpvariablerna och därmed* S_x^2 *och* S_y^2 *är oberoende så är* $\frac{1}{\sigma^2}((n_x - 1)S_x^2 + (n_y - 1)S_y^2) \sim \chi^2(n_x - 1 + n_y - 1)$*. Testvariabeln kan alltså skrivas i formen* $\frac{Z}{\sqrt{\frac{1}{m}C}}$ *där* $Z = \frac{\overline{X}-\overline{Y}}{\sqrt{\sigma^2(\frac{1}{n_x}+\frac{1}{n_y})}} \sim \mathrm{N}(0, 1)$*,* $m = n_x + n_y - 2$ *och* $C = \frac{1}{\sigma^2}((n_x - 1)S_x^2 + (n_y - 1)S_y^2)$*.*

Exempel: Skillnaden mellan andelar

Under åren 1660–1740 föddes i Paris 377 649 *flickor och* 393 535 *pojkar och under samma tid föddes i London* 698 900 *flickor och* 737 687 *pojkar. Finns det skillnader i andelen flickor?*

Låt X_j *vara en slumpvariabel som får värdet* 1 *om barn nummer* j *i Paris är en flicka och* 0 *om det är en pojke och låt* Y_j *vara motsvarande slumpvariabel för barnen i London. Dessutom antar vi att alla dehär slumpvariablerna är oberoende och att* $\Pr(X_j = 1) = p_P$ *och* $\Pr(Y_j = 1) = p_L$. *Nollhypotesen är i detta fall* $H_o : p_P = p_L$. *Nollhypotesen säger inte vad* $p_P = p_L$ *är men vi kan räkna ett estimat* $\hat{p}$ *för denhär sannolikheten genom att konstatera att det föddes sammanlagt* 2 207 771 *barn och av dessa var* 1 076 549 *flickor så att* $\hat{p} = \frac{1\,076\,549}{2\,207\,771} \approx 0.48762$. *Vi kan också räkna medelvärdena av de observerade stickproven och de är* $\overline{x} = 0.4897$ *och* $\overline{y} = 0.4865$.

Slumpvariabelns $\overline{X}$ *varians är ungefär* $\frac{\hat{p}(1-\hat{p})}{n_P}$ *där* $n_P = 771184$ *är antalet barn födda i Paris.*

Exempel: Skillnaden mellan andelar, forts.

På samma sätt är variansen av $\overline{Y}$ *ungefär* $\frac{\hat{p}(1-\hat{p})}{n_L}$ *där* $n_L = 771184$ *är antalet barn födda i London.*
Det här betyder att slumpvariabelns $\overline{X} - \overline{Y}$ *varians är ungefär* $\frac{\hat{p}(1-\hat{p})}{n_P} + \frac{\hat{p}(1-\hat{p})}{n_L}$ *så att testvariabeln*

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_P} + \frac{1}{n_L})}}$$

är i stort sett $\mathrm{N}(0,1)$*-fördelad.*
I dethär fallet får testvariabeln värdet

$$z = \frac{0.48970 - 0.48650}{\sqrt{0.48762 \cdot (1 - 0.48762) \cdot \left(\frac{1}{771184} + \frac{1}{1436587}\right)}} = 4.5350.$$

p-värdet blir nu

$$p \approx \Pr(|Z| \geq 4.535) = 2 \cdot F_{\mathrm{N}(0,1)}(-4.5350) = 0.00000576,$$

vilket betyder att vi har goda skäl att förkasta nollhypotesen.

Testning av två andelar eller sannolikheter

Om X_j, $j = 1, 2, \ldots, n_x$ *och* Y_j, $j = 1, \ldots, n_y$ *är två (oberoende) stickprov av slumpvariablerna* X *och* Y *där* $X \sim \text{Bernoulli}(p_x)$ *och* $Y \sim \text{Bernoulli}(p_y)$ *och nollhypotesen är* $p_x = p_y$ *(eller* $p_x \leq p_y$ *eller* $p_x \geq p_y$*) så väljer vi som testvariabel*

$$\frac{\overline{X} - \overline{Y}}{\sqrt{\widehat{P}(1-\widehat{P})(\frac{1}{n_x} + \frac{1}{n_y})}} \sim_a \text{N}(0,1)$$

där

$$\widehat{P} = \frac{n_x \overline{X} + n_y \overline{Y}}{n_x + n_y}.$$

Exempel: Skillnaden mellan två väntevärden, allmänt fall

Från en viss process har vi samlat in data för att säkerställa produktkvaliteten och sedan gjorde vi ändringar i processen för att minska på variansen. Detta lyckades också men vi hoppas och också mätvärdena, dvs. kvaliteten också stigit. För att undersöka detta gjorde vi mätningar före och efter förändringarna:

	Stickprovsstorlek	*Medelvärde*	*Stickprovsvarians*
Före	*220*	*4.50*	*0.08*
Efter	*250*	*4.56*	*0.04*

Här har vi alltså stickprov $X_1, X_2, \ldots, X_{220}$ *(före) och* $Y_1, Y_2, \ldots, Y_{250}$ *(efter) och vi antar att alla dessa slumpvariabler är oberoende, slumpvariablerna* X_j *har samma fördelning och slumpvariablerna har samma fördelning. Däremot antar vi inte att de har samma varians eller är normalfördelade men nog att de är sådana att medelvärdena* $\overline{X}$ *och* $\overline{Y}$ *är ungefär normalfördelade på gund av den centrala gränsvärdessatsen.*

Exempel: Skillnaden mellan två väntevärden, allmänt fall, forts.

Då gäller också

$$\overline{X} - \overline{Y} \sim_a N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{220} + \frac{\sigma_Y^2}{250}\right).$$

I dethär fallet väljer vi som nollhypotes $\mu_X \geq \mu_Y$ *som motpåstående till vår förmodan att kvaliteten förbättrades, dvs.* $\mu_Y > \mu_X$. *Vi vet inte vad* σ_X^2 *och* σ_Y^2 *är men vi kan estimera dem med stickprovsvarianserna* S_X^2 *och* S_Y^2 *så att testvariabeln blir*

$$Z = \frac{\overline{X} - \overline{Y}}{\sqrt{\frac{S_x^2}{220} + \frac{S_Y^2}{250}}} \sim_a \mathrm{N}(0,1).$$

Värdet av testvariabeln är i detta fall -2.622 *och eftersom positiva värden på testvariabeln är i samklang med nollhypotesen så blir p-värdet*

$$p = \Pr(Z \leq -2.622) \approx F_{\mathrm{N}(0,1)}(-2.622) = 0.0044.$$

Det här betyder att vi kan förkasta nollhypotesen på signifikansnivån 0.01.

Anpassning eller "Goodness-of-fit"

Om X_j, $j = 1, \ldots, n$ är ett stickprov av slumpvariabeln X vars värdemängd är $\cup_{k=1}^{m} A_k$ där mängderna A_k är disjunkta och nollhypotesen är

$$\mathrm{H}_0 : \Pr(X \in A_k) = p_k, \quad k = 1, \ldots, m$$

så väljer vi som testvariabel

$$\sum_{k=1}^{m} \frac{(O_k - np_k)^2}{np_k} \sim_a \chi^2(m-1),$$

där O_k är antalet element i mängden $\{ j : X_j \in A_k \}$.

Exempel: Singla slant

Antag att vi singlar slant 400 *gånger och får* 170 *klavor och* 230 *kronor. Som nollhypotes tar vi* $H_0 : p = 0.5$ *där* $p = \Pr(T)$. *Om* Y *är antalet klavor så är* $Y \sim \text{Binom}(n, p)$ *med* $n = 400$ *och* $p = 0.5$. *Det betyder att* $\frac{Y-np}{\sqrt{np(1-p)}} \sim_a N(0,1)$ *så p-värdet blir, eftersom alternativet till nollhypotesen är tvåsidigt,*

$$
\begin{aligned}
p &= 2 \cdot \Pr(Y \leq 170) \\
&= 2 \cdot \Pr\left(\frac{Y - np}{\sqrt{np(1-p)}} \leq \frac{170 - 200}{\sqrt{400 \cdot 0.5 \cdot 0.5}} \right) \\
&= 2 \cdot \Pr\left(\frac{Y - np}{\sqrt{np(1-p)}} \leq -3 \right) \approx 0.0026998.
\end{aligned}
$$

Exempel: Singla slant, forts.

Ett annat sätt är att skriva de observerade talen i en tabell:

T	H
170	*230*

och räkna värdet av testvariabeln $C = \sum_{k=1}^{m} \frac{(O_k - np_k)^2}{np_k}$ *i* χ^2*-anpassningstestet och det blir*

$$c = \frac{(170 - 400 \cdot 0.5)^2}{400 \cdot 0.5} + \frac{(230 - 400 \cdot 0.5)^2}{400 \cdot 0.5} = \frac{30^2}{200} + \frac{30^2}{200} = 9.$$

Nu är C *ungefär* $\chi^2(2-1)$*-fördelad och det är bara stora värden på* C *som motsäger nollhypotesen så testets p-värde blir*

$$p = \Pr(C \geq 9) = 1 - F_{\chi^2(1)}(9) = 0.0026998.$$

Exempel: Singla slant, forts.

Hur kommer det sig att vi får exakt samma svar i båda fallen?
Om $Y \sim \text{Binom}(n, p)$ *är antalet klavor så är* $n - Y$ *antalet kronor och*

$$\frac{(Y-np)^2}{np} + \frac{((n-Y)-n(1-p))^2}{n(1-p)} = \frac{(Y-np)^2}{np} + \frac{(-Y+np)^2}{n(1-p)}$$

$$= \frac{(Y-np)^2}{n}\left(\frac{1}{p} + \frac{1}{1-p}\right) = \frac{(Y-np)^2}{np(1-p)} = \left(\frac{Y-np}{\sqrt{np(1-p)}}\right)^2,$$

så att testvariabeln i χ^2*-testet är kvadraten av testvariabeln i normalapproximationen av den binomialfördelade slumpvariabeln* Y *och en* $\chi^2(1)$*-fördelad slumpvariabel är enligt definitionen kvadraten av en* $\text{N}(0,1)$*-fördelad slumpvariabel.*
Ifall antalet klasser m *i* χ^2*-testet är större än 2 så är det betydligt besvärligare att visa att* $C \sim_a \chi^2(m-1)$.

Test av variansen, normalfördelning

Om X_j, $j = 1, 2, \ldots, n$ är ett stickprov av slumpvariabeln X som är $\mathrm{N}(\mu, \sigma^2)$-fördelad och nollhypotesen är $\sigma^2 = \sigma_0^2$ (eller $\sigma^2 \leq \sigma_0^2$ eller $\sigma^2 \geq \sigma_0^2$) så väljer vi som testvariabel

$$\frac{(n-1)S^2}{\sigma_0^2} \sim \chi^2(n-1),$$

där S^2 är stickprovsvariansen.

p-värde, kritiskt område, χ^2-testvariabel

- *Vi antar att testvariabeln C är (approximativt) $\chi^2(k)$-fördelad och att den i testet får värdet c_*.*
- *Om alternativet till nollhypotesen är ensidigt och små värden av testvariabeln är förenliga med nollhypotesen så gäller:*
  - *p-värdet är $\Pr(C \geq c_*) = 1 - F_{\chi^2(k)}(c_*)$.*
  - *Nollhypotesen förkastas på signifikansnivån α om $p < \alpha$ dvs. om testvariabeln får sitt värde i det kritiska området $(F^{-1}_{\chi^2(k)}(1-\alpha), \infty)$.*
- *Om alternativet till nollhypotesen är ensidigt och stora värden av testvariabeln är föreneliga med nollhypotesen så gäller*
  - *p-värdet är $\Pr(C \leq c_*) = F_{\chi^2(k)}(c_*)$.*
  - *Nollhypotesen förkastas på signifikansnivån α om $p < \alpha$ dvs. om testvariabeln får sitt värde i det kritiska området $(0, F^{-1}_{\chi^2(k)}(\alpha))$.*
- *Om alternativet till nollhypotesen är tvåsidigt så gäller*
  - *p-värdet är $2\min(F_{\chi^2(k)}(c_*), 1 - F_{\chi^2(k)}(c_*))$.*
  - *Nollhypotesen förkastas på signifikansnivån α om $p < \alpha$ dvs. om testvariabeln får sitt värde i det kritiska området $(0, F^{-1}_{\chi^2(k)}(\frac{\alpha}{2})) \cup (F^{-1}_{\chi^2(k)}(1-\frac{\alpha}{2}), \infty)$.*

Exempel: Stickprovsvariansens fördelning

Om X_j, $j = 1, n$ *är ett stickprov av en* $\mathrm{N}(\mu, \sigma^2)$ *fördelad slumpvariabel så har* $\frac{(n-1)S^2}{\sigma^2}$ *fördelningen* $\chi^2(n-1)$. *Men vad händer om vi tar ett stickprov av en slumpvariabel* X *som är jämnt fördelad i intervallet* $[0, 1]$ *så att* $\mathrm{Var}(X) = \frac{1}{12}$?

Som nollhypotes tar vi att $\frac{(n-1)S^2}{\sigma^2}$ *fortfarande är* $\chi^2(n-1)$*-fördelad, vi väljer* $n = 5$ *och räknar variansen för* 100 *stickprov. Klasserna väljer vi som intervallen* $[0, 2)$, $[2, 4)$, $[4, 6)$, $[6, 8)$ *och* $[8, \infty)$ *och resultaten blir följande då vi ser efter i vilket intervall* $\frac{(5-1)s^2}{\frac{1}{12}}$ *hamnar:*

A_k	$[0, 2)$	$[2, 4)$	$[4, 6)$	$[6, 8)$	$[8, \infty)$
O_k	*16*	*41*	*25*	*16*	*2*

Sannolikheten att en $\chi^2(5-1)$*-fördelad slumpvariabel ligger i intervallet* $[a_{k-1}, a_k)$ *är* $F_{\chi^2(4)}(a_k) - F_{\chi^2(4)}(a_{k-1})$ *och de här sannolikheterna blir*

A_k	$[0, 2)$	$[2, 4)$	$[4, 6)$	$[6, 8)$	$[8, \infty)$
p_k	*0.264241*	*0.329753*	*0.206858*	*0.107570*	*0.091578*

Exempel: Stickprovsvariansens fördelning, forts.

Värdet av testvariabeln $C = \sum_{k=1}^{5} \frac{(O_k - 100 \cdot p_k)^2}{100 \cdot p_k}$ *blir nu*

$$c = \frac{(16 - 26.4241)^2}{26.4241} + \frac{(41 - 32.9753)^2}{32.9753} + \frac{(25 - 20.6858)^2}{20.6858} + \frac{(16 - 10.757)^2}{10.757} + \frac{(2 - 9.1578)^2}{9.1578} = 15.115.$$

Eftersom C *är ungefär* $\chi^2(5-1)$*-fördelad och endast stora värden på* C *motsäger nollhypotesen så blir testets p-värde*

$$p = \Pr(C \geq 15.115) = 1 - F_{\chi^2(4)}(15.115) = 0.0045.$$

Det här betyder att det finns skäl att förkasta nollhypotesen och om vi skulle ha räknat variansen för ännu flera stickprov skulle det här ha blivit ännu tydligare.

Exempel

Vi vill testa om sannolikheten att få en krona då man singlar en viss slant faktiskt är 0.5. *Hur många gånger måste vi singla slanten för att sannolikheten att nollhypotesen* $\mathrm{H}_0 : p = 0.5$ *förkastas på signifikansnivån* 0.05 *är åtminstone* 0.9 *om* $p \geq 0.52$*?*
Eftersom vi vill räkna ut en övre gräns för antalet kast räcker det att anta att $p = 0.52$. *Vi singlar alltså slant* n *gångar och andelen kronor blir då* $\hat{p}$. *Testvariabeln är (för normalapproximation)*

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}},$$

där $p_0 = 0.5$. *Eftersom signifikansnivån är vald till* 0.05 *och alternativet till nollhypotesen är tvåsidigt så är de kritiska värdena* $\pm z_{0.025} = \mp F^{-1}_{\mathrm{N}(0,1)}(0.025) = \pm 1.96$, *dvs. nollhypotesen förkastas om* $z > 1.96$ *eller* $z < -1.96$.

Exempel, forts.

Om nu i verkligheten $p = p_1 = 0.52$ *så är* $\dfrac{\hat{p} - p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}} \sim_a \mathrm{N}(0,1)$, *och vi får*

$$\Pr\left(\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} > 1.96\right) = \Pr\left(\hat{p} > p_0 + 1.96\sqrt{\frac{p_0(1-p_0)}{n}}\right)$$

$$= \Pr\left(\frac{\hat{p} - p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}} > \frac{p_0 + 1.96\sqrt{\frac{p_0(1-p_0)}{n}} - p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}}\right)$$

$$= \Pr\left(\frac{\hat{p} - p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}} > 1.96\sqrt{\frac{p_0(1-p_0)}{p_1(1-p_1)}} + \frac{p_0 - p_1}{\sqrt{p_1(1-p_1)}}\sqrt{n}\right)$$

$$\approx \Pr\left(\frac{\hat{p} - p_1}{\sqrt{\frac{p_1(1-p_1)}{n}}} > 1.962 - 0.04\sqrt{n}\right).$$

Exempel, forts.

Vi får också ett motsvarande uttryck för $\Pr\left(\frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} < -1.96\right)$ *men eftersom det räcker att få en nedre gräns för* n *och eftersom det är rimligt att anta att den senare sannolikheten är mycket liten så blir kravet att*

$$\Pr(Z > 1.96 - 0.04\sqrt{n}) \geq 0.9$$

vilket betyder att

$$1.962 - 0.04\sqrt{n} \lessapprox -1.28$$

eftersom $F^{-1}_{\mathrm{N}(0,1)}(1-0.9) \approx -1.28$ *och vi får villkoret*

$$n \gtrapprox \left(\frac{1.962+1.28}{0.04}\right)^2 = 6569.1,$$

vilket betyder att det är skäl att välja $n \geq 6600$.

Exempel, forts.

Om nu $n \geq 6600$ *så visar en räkning att*

$$\Pr\left(\frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} < -1.96\right)$$

$$= \Pr(Z < -1.96 - 0.04\sqrt{n}) < \Pr(Z < -1.96 - 1.962 - 1.28) \approx 10^{-7},$$

så det var helt korrekt att strunta i denna term.

Korrelation

Korrelationen eller korrelationskoefficienten mellan slumpvariablerna X och Y är

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

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

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

Obs!

Om X och Y är slumpvariabler med ändlig men positiv varians och a, b, c och d är tal (med $a \neq 0$ och $c \neq 0$) så är

$$\mathrm{Cor}(aX + b, cY + d) = \mathrm{sign}(ac)\mathrm{Cor}(X, Y).$$

Varför? Eftersom $\mathrm{Cor}(U, V) = \mathrm{Cor}(V, U)$ *så räcker det att visa att* $\mathrm{Cor}(aX + b, Y) = \mathrm{sign}(a)\mathrm{Cor}(X, Y)$ *för då är*

$$\begin{aligned}\mathrm{Cor}(aX + b, cY + d) &= \mathrm{sign}(a)\mathrm{Cor}(X, cY + d) = \mathrm{sign}(a)\mathrm{Cor}(cY + d, X) \\ &= \mathrm{sign}(a)\mathrm{sign}(c)\mathrm{Cor}(Y, X) = \mathrm{sign}(ac)\mathrm{Cor}(X, Y)\end{aligned}$$

Eftersom $\mathrm{E}(aX + b) = a\mathrm{E}(X) + b$ *så är* $\mathrm{Var}(aX + b) = \mathrm{E}((aX + b - a\mathrm{E}(X) - b)^2) = a^2\mathrm{Var}(X)$ *och*

$$\begin{aligned}\mathrm{Cov}(aX + b, Y) &= \mathrm{E}((aX + b - a\mathrm{E}(X) - b)(Y - \mathrm{E}(Y))) \\ &= a\mathrm{E}((X - \mathrm{E}(X))(Y - \mathrm{E}(Y))) = a\mathrm{Cov}(X, Y),\end{aligned}$$

så att

$$\mathrm{Cor}(aX+b, Y) = \frac{a\mathrm{Cov}(X, Y)}{\sqrt{a^2\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{a}{|a|}\mathrm{Cor}(X, Y) = \mathrm{sign}(a)\mathrm{Cor}(X, Y).$$

Stickprovskorrelationskoefficientens fördelning

- *Ifall* (X_j, Y_j), $i = 1, \ldots, n$ *är ett stickprov av en slumpvariabel* (X, Y) *där* X *och* Y *är oberoende, så att* $\rho_{xy} = 0$, *och den ena av slumpvariablerna är normalfördelad och den andra är kontinuerlig 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 bestämmer* b_0 *och* b_1 *så att*

$$\sum_{j=1}^{n} (y_j - b_0 - b_1 x_j)^2$$

är så liten som möjligt så blir svaret

$$b_1 = \frac{\sum_{j=1}^{n}(x_j - \overline{x})(y_j - \overline{y})}{\sum_{j=1}^{n}(x_j - \overline{x})^2} = \frac{s_{xy}}{s_x^2} \quad \textit{och} \quad b_0 = \overline{y} - b_1\,\overline{x},$$

där $\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$ *och* $\overline{y} = \frac{1}{n}\sum_{j=1}^{n} y_j$.
Varför? Vi kan skriva kvadratsumman $\sum_{j=1}^{n} (y_j - b_0 - b_1 x_j)^2$ *i formen* $f(\tilde{b}_0, b_1) = \sum_{j=1}^{n} \Big((y_j - \overline{y}) - \tilde{b}_0 - b_1(x_j - \overline{x})\Big)^2$, *och villkoret att den partiella derivatan med avseende på* $\tilde{b}_0$ *är* 0 *ger* $2n\tilde{b}_0 = 0$, *dvs.* $\tilde{b}_0 = 0$ *och villkoret att den partiella derivatan med avseende på* b_1 *är* 0 *ger* $2\sum_{j=1}^{n}\big(b_1(x_j - \overline{x}) - (y_j - \overline{y})\big)(x_j - \overline{x}) = 0$ *och denna ekvation ger uttrycket för* b_1.

Obs

I räkningarna ovan förekommer inga slumpvariabler men vi kan bra tänka oss att sambandet mellan variabler x och y är $y = \beta_0 + \beta_1 x$ *men då man mäter värdena av y-variabeln så förekommer det slumpmässiga fel som leder till att de uppmätta värdena blir*

$$y_j = \beta_0 + \beta_1 x_j + \varepsilon_j, \quad j = 1, \ldots, n$$

där ε_j *är slumpvariabler. Det faktum att man minimerar* $\sum_{j=1}^{n} (y_j - b_0 - b_1 x_j)^2$ *(och inte något annat uttryck) är förnuftigt om man antar att det inte förekommer några fel i* x_j*-värdena och att alla avvikelser från en rät linje beror på felaktiga* y_j*-värden. Att man sedan minimerar en kvadratsumma och inte tex. absolutbelopp är förnuftigt om man antar att slumpvariablerna* ε_j *är normalfördelade.*

Exempel: Regressionslinje

Vi har följande observationer

x	1.0	1.9	2.7	3.2	3.8	4.7	5.1	5.5
y	-0.8	-0.4	-0.0	0.9	1.2	1.3	1.7	2.1

Först räknar vi medelvärdena och de är

$$\overline{x} = 3.4875,$$
$$\overline{y} = 0.75.$$

Sedan skall vi räkna stickprovsvariansen av x och stickprovskovariansen av variablerna x och y och vi får

$$s_x^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \overline{x})^2 = 2.5184,$$
$$s_{xy} = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \overline{x})(y_j - \overline{y}) = 1.6121.$$

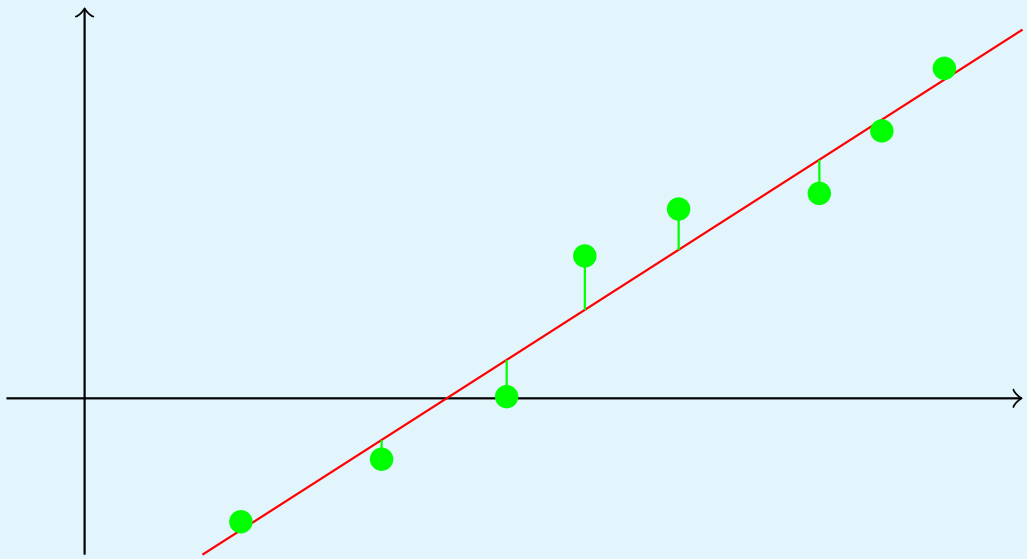
Exempel: Regressionslinje, forts.

Det här betyder att

$$b_1 = \frac{s_{xy}}{s_x^2} = 0.64015,$$

$$b_0 = \overline{y} - b_1\overline{x} = -1.4825.$$

Punkterna och linjen ser ut på följande sätt:



Regression

- *Vi antar att slumpvariabeln Y förutom på slumpen beror på variabeln x så att*

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

där ε är en slumpvariabel som vi antar att är 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 vi vanligen antar 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 vi använda*

$$W_0 = \frac{B_0 - \beta_0}{\sqrt{S^2\left(\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}\right)}} \sim \mathrm{t}(n-2),$$

$$W_1 = \frac{B_1 - \beta_1}{\sqrt{\frac{S^2}{(n-1)s_x^2}}} \sim \mathrm{t}(n-2).$$

Ett 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

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

Det senare resultatet visar att test av nollhypoteserna $\beta_1 = 0$ och $\rho_{xy} = 0$ ger samma resulat (då man antar normalfördelning).

Ett samband mellan estimatorerna, varför?

Eftersom $B_1 = \frac{S_{xy}}{s_x^2}$, $B_0 = \overline{Y} - B_1\overline{x}$, $S^2 = \frac{1}{n-2}\sum_{j=1}^{n}(Y_j - B_0 - B_1 x_j)^2$ *och* $S_{xy} = R_{xy}\sqrt{s_x^2 S_y^2}$ *så är*

$$(n-2)S^2 = \sum_{j=1}^{n}\big(B_0 + B_1 x_j - y_j\big)^2 = \sum_{j=1}^{n}\big(B_1(x_j - \overline{x}) - (y_j - \overline{y})\big)^2$$

$$= B_1^2\sum_{j=1}^{n}(x_j - \overline{x})^2 - 2B_1\sum_{j=1}^{n}(x_j - \overline{x})(y_j - \overline{y}) + \sum_{j=1}^{n}(y_j - \overline{y})^2$$

$$= (n-1)\big(B_1^2 s_x^2 - 2B_1 S_{xy} + S_y^2\big) = (n-1)\left(\frac{S_{xy}^2 s_x^2}{S_x^4} - 2\frac{S_{xy}^2}{s_x^2} + S_y^2\right)$$

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

så att

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

Ett samband mellan estimatorerna, varför?, forts.

En följd av det här är att

$$\frac{B_1}{\sqrt{\frac{S^2}{(n-1)s_x^2}}} = \frac{S_{xy}}{s_x^2\sqrt{\frac{(n-1)S_y^2(1-R_{xy}^2)}{(n-2)(n-1)s_x^2}}} = \frac{S_{xy}}{\sqrt{\frac{s_x^2 S_y^2(1-R_{xy}^2)}{n-2}}}$$

$$= \frac{R_{xy}\sqrt{n-2}}{\sqrt{1-R_{xy}^2}}.$$

Exempel: Trafikolyckor

Enligt statistikcentralen var antalet förolyckade personer i trafikolyckor under åren 2004–2013 följande

2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
375	379	336	380	344	279	272	292	255	248

I dethär fallet är det ändamålsenligt att som x-variabel ta årtalet från vilket vi subtraherar 2015 så att tabellen ser ut på följande sätt:

x	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2
y	375	379	336	380	344	279	272	292	255	248

Från det här stickprovet kan vi räkna följande estimat:

$\overline{x}$	$\overline{y}$	s_x^2	s_y^2	s_{xy}
-6.5	316	9.1667	2772.8889	-145.5556

Exempel: Trafikolyckor, regressionslinjen

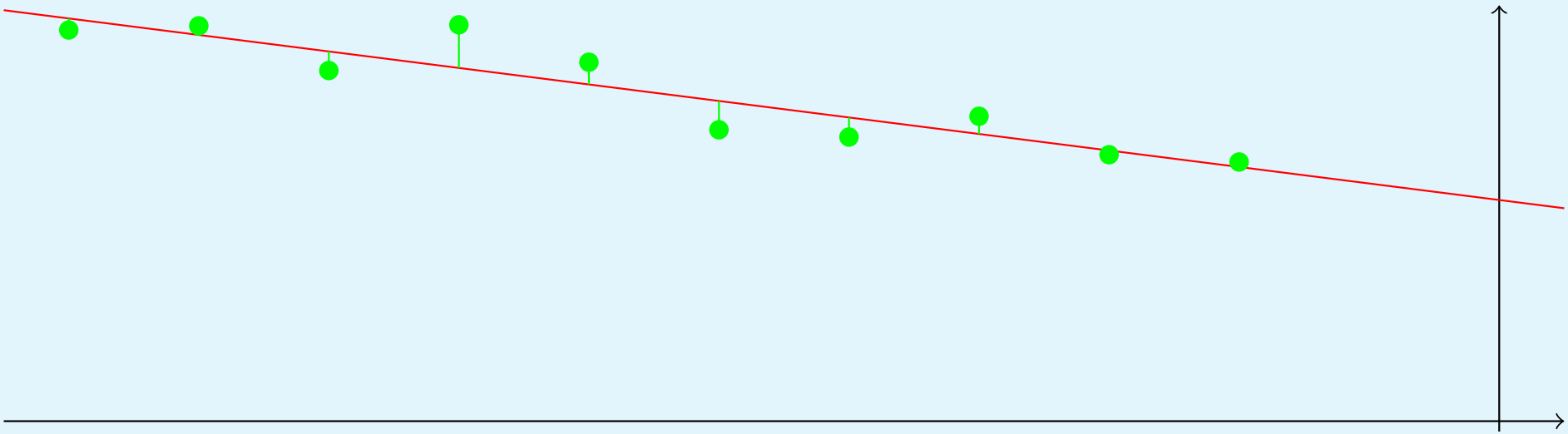
Nu får vi följande estimat för parametrarna i regressionsmodellen $Y_j = \beta_0 + \beta_1 x_j + \varepsilon_j$*:*

$$b_1 = \frac{s_{xy}}{s_x^2} = -15.879,$$

$$b_0 = \overline{y} - b_1 \overline{x} = 212.79,$$

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = -0.91297.$$

Linjen och datapunkterna ser ut på följande sätt:



Exempel: Trafikolyckor, β_1

Vi kan räkna ett estimat för restvariansen antingen direkt med formeln

$$s^2 = \frac{1}{10-2}\sum_{j=1}^{10}(y_j - b_0 - b_1 x_j)^2,$$

men i allmänhet är det enklare att använda formeln

$$s^2 = \frac{n-1}{n-2}s_y^2(1-r_{xy}^2) = \frac{9}{8}\cdot 2772.8889\cdot\left(1-(-0.91297)^2\right) = 519.35.$$

Nu kan vi testa nollhypotesen $\beta_1 = 0$ och då är testvariabeln

$$W_1 = \frac{B_1 - 0}{\sqrt{\frac{S^2}{(n-1)s_x^2}}} \sim \mathrm{t}(10-2),$$

och den här testvariabeln får värdet

$$w_1 = \frac{-15.879}{\sqrt{\frac{519.35}{9\cdot 9.1667}}} = -6.3287.$$

Exempel: Trafikolyckor, β_1, forts.

Eftersom nollhypotesen är $\beta_1 = 0$ (och inte tex. $\beta_1 \geq 0$ vilket man väl kunde motivera) så blir p-värdet

$$p = 2F_{\mathrm{t}(8)}(-6.3287) = 0.000226,$$

Exempel: Trafikolyckor, β_0

Eftersom vi subtraherade 2015 från årtalen är β_0 väntevärdet av antalet förolyckade i trafikolyckor år 2015.
Om vi vill testa hypotesen $\beta_0 \geq 240$ så använder vi som testvariabel

$$W_0 = \frac{B_0 - \beta_0}{\sqrt{S^2\left(\frac{1}{n} + \frac{\bar{x}^2}{(n-1)s_x^2}\right)}} \sim \mathrm{t}(n-2).$$

När vi sätter in de tal vi tidigare räknat ut i den här formeln så får vi

Exempel: Trafikolyckor, β_0, forts.

$$w_0 = \frac{212.79 - 240}{\sqrt{519.35\left(\frac{1}{10} + \frac{(-6.5)^2}{(10-1)9.1667}\right)}} = -1.5261$$

Eftersom nollhypotesen var $\beta_0 \geq 240$ *så är det endast stora negativa värden på testvariabeln som motsäger nollhypotesen, dvs. alternativet är ensidigt så p-värdet blir*

$$p = F_{\mathrm{t}(8)}(-1.5261) = 0.082749,$$

och vi förkastar inte nollhypotesen ens på signifikansnivån 0.05.

Exempel: Trafikolyckor, konfidensintervall för parametrarna

Konfidensintervall för parametrarna β_0 och β_1 definieras och beräknas på samma sätt som konfidensintervall för väntevärdet av en normalfördelad slumpvariabel. Om vi tex. skall bestämma ett 99% konfidensintervall för parametern β_1 så konstatterar vi först att eftersom

$$W_1 = \frac{B_1 - \beta_1}{\sqrt{\frac{S^2}{(n-1)s_x^2}}} \sim \mathrm{t}(n-2)$$

och $F^{-1}_{\mathrm{t}(8)}(0.995) = -F^{-1}_{\mathrm{t}(8)}(0.005) = 3.3554$ *så är*

$$\Pr\left(-3.3554 \leq \frac{B_1 - \beta_1}{\sqrt{\frac{S^2}{(n-1)s_x^2}}} \leq 3.3554\right) = 1 - 0.005 - 0.005 = 0.99.$$

Eftersom $-3.3554 \leq \frac{B_1 - \beta_1}{\sqrt{\frac{S^2}{(n-1)s_x^2}}} \leq 3.3554$ *om och endast om*

Exempel: Trafikolyckor, konfidensintervall för parametrarna, forts.

$$B_1 - 3.3554\sqrt{\frac{S^2}{(n-1)s_x^2}} \leq \beta_1 \leq B_1 + 3.3554\sqrt{\frac{S^2}{(n-1)s_x^2}} \text{ så är}$$

$$\Pr\left(\beta_1 \in \left[B_1 - 3.3554\sqrt{\frac{S^2}{(n-1)s_x^2}}, B_1 + 3.3554\sqrt{\frac{S^2}{(n-1)s_x^2}}\right]\right) = 0.99.$$

När vi sätter in de tal vi räknat ut tidigare så får vi som konfidensintervall med konfidensgraden 99%

$$\left[-15.879 - 3.3554\sqrt{\frac{519.35}{9 \cdot 9.1667}}, -15.8791 + 3.3554\sqrt{\frac{519.35}{9 \cdot 9.1667}}\right]$$
$$= \left[-24.295, -7.4628\right].$$

Betingade fördelningar av normalfördelningar, förklaringsgrad

Om (X, Y) är normalfördelad så är

$$(Y|X = x) \sim \mathrm{N}\left(\mu_Y + \rho_{XY}\frac{\sigma_Y}{\sigma_X}(x - \mu_X), (1 - \rho_{XY}^2)\sigma_Y^2\right),$$

dvs.

$$\mathrm{E}(Y|X = x) = \mu_Y + \rho_{XY}\frac{\sigma_Y}{\sigma_X}(x - \mu_X) = \beta_0 + \beta_1 x,$$

där

$$\beta_1 = \rho_{XY}\frac{\sigma_Y}{\sigma_X} = \frac{\sigma_{XY}}{\sigma_X^2},$$

$$\beta_0 = \mu_Y - \beta_1\mu_X.$$

Med minsta kvadratmetoden får vi estimat för parametrarna β_0 och β_1 som är

$$b_1 = r_{xy}\frac{s_y}{s_x} = \frac{s_{xy}}{s_x^2},$$

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

Betingade fördelningar av normalfördelningar, förklaringsgrad, forts.

Om (X, Y) är normalfördelad och $X = x$ så kan vi alltså skriva

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

där

$$\varepsilon \sim \mathrm{N}(0, (1 - \rho_{XY}^2)\sigma_Y^2).$$

Här är alltså restvariansen $(1 - \rho_{XY}^2)\sigma_Y^2$ den del av variansen av Y som inte kan förklaras med beroendet på X och den del av variansen av Y som kan förklaras med beroendet på X är

$$\frac{\rho_{XY}^2 \sigma_Y^2}{\sigma_Y^2} = \rho_{XY}^2.$$

Analogt med detta säger vi att talet r_{xy}^2, som är ett estimat av ρ_{XY}^2 är regressionsmodellens $Y_j = b_0 + b_1 x_j$ **förklaringsgrad**.

Interpolering och 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* $\tilde{x}_j = x_j - x_0$ *och sedan i normal ordning räkna ut estimat och göra hypotesprövningar för* β_0 *i regressionsmodellen* $Y = \beta_0 + \beta_1 \tilde{x} + \varepsilon$.

Logistisk regression

Antag att vi av friska och insjukanade personer mätt följande koncentrationer av fibrinogen i blodet:

Friska	*2.52*	*2.56*	*2.19*	*2.18*	*3.41*	*2.46*	*3.22*	*2.21*
Friska	*3.15*	*2.60*	*2.29*	*2.35*				
Insjuknade	*5.06*	*3.34*	*2.38*	*3.53*	*2.09*	*3.93*		

Om nu fibrinogenkoncentrationen i blodet på en viss person är 3.1 *så vad är sannolikheten att hen är frisk?*
Här antar vi alltså att sannolikheten att en person är frisk på något sätt beror på fibrinogenkoncentrationen, som vi betecknar med x*, dvs.* $\Pr(\text{"Personen är frisk"}) = p(x)$*. Nu är det inte förnuftigt att anta att detta samband är linjärt för d å går det lätt så att* $p(x)$ *får värden som inte ligger i intervallet* $[0, 1]$*. En bättre idé är att använda odds och anta att*

$$\log\left(\frac{p(x)}{1-p(x)}\right) = c_0 + c_1 x \quad \textit{dvs.} \quad p(x) = \frac{e^{c_0+c_1 x}}{1+e^{c_0+c_1 x}}.$$

För att estimera c_0 *och* c_1 *använder vi Maximum likelihood metoden.*

Logistisk regression, forts.

Låt nu f_i, $i = 1, \ldots, n_1$ vara koncentrationerna hos de friska personerna och s_i, $i = 1, \ldots, n_2$ koncentrationerna hos de insjuknade personerna. Låt nu $L(c_0, c_1)$ vara sannolikheten, med de antaganden vi gjort, att de friska är friska och den sjuka är sjuka, eller (eftersom $1 - p(x) = \frac{1}{1+e^{c_0+c_1 x}}$)

$$\mathcal{L}(c_0, c_1) = \frac{e^{c_0+c_1 t_1} \cdot \ldots \cdot e^{c_0+c_1 t_{n_1}}}{(1 + e^{c_0+c_1 t_1}) \cdot \ldots \cdot (1 + e^{c_0+c_1 t_{n_1}})} \cdot \frac{1}{(1 + e^{c_0+c_1 s_1}) \cdot \ldots \cdot (1 + e^{c_0+c_1 s_{n_2}})}.$$

Det är inte helt enkelt att bestämma den punkt i vilken denna funktion uppnår sitt största värde men med numeriska metoder får vi $c_0 \approx 5.4$ och $c_1 \approx -1.6$ så att $p(3.1) \approx 0.6$.