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

G. Gripenberg

Aalto-universitetet

13 februari 2015

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* $\epsilon > 0$.

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

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

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

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

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* $\overline{x} = 0.8$.

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*

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

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 $\mathrm{N}(\mu, \sigma^2)$ *där* $\mu \geq 0$. *Testvariabeln skulle ha varit precis densamma men p-värdet skulle ha blivit*

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

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 $\mathrm{N}(\mu, \sigma^2)$ *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.*

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.

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*.*

Exempel: Singla slant

Antag att vi singlar slant 400 *gånger och får* 170 *klavor och* 230 *kronor. Som nollhypotes tar vi* $\mathrm{H}_0 : p = 0.5$ *där* $p = \Pr(T)$*. Om* Y *är antalet klavor så är* $Y \sim \mathrm{Binom}(n, p)$ *med* $n = 400$ *och* $p = 0.5$*. Det betyder att* $\frac{Y - np}{\sqrt{np(1-p)}} \sim_a \mathrm{N}(0,1)$ *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* $C \sim_a \chi^2(m-1)$.

Exempel: Stickprovsvariansens fördelning

Om X_j, $j = 1, n$ *är ett stickprov av en* $\text{N}(\mu, \sigma^2)$ *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* $\text{Var}(X) = \frac{1}{12}$*?*
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*

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

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* $\text{H}_0 : p = 0.5$ *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* $\pm z_{0.025} = \mp F^{-1}_{\text{N}(0,1)}(0.025) = \pm 1.96$, *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.

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*

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

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*

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

så att

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

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

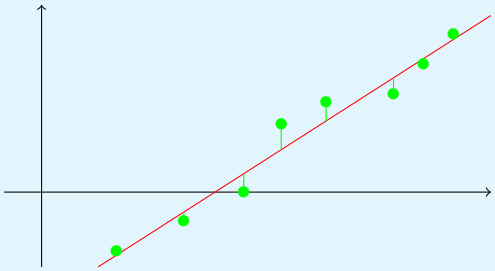
$$s_x^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \overline{x})^2 = 2.5184,$$
$$s_{xy} = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \overline{x})(y_j - \overline{y}) = 1.6121.$$

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:



Ett samband mellan estimatorerna, varför?

Eftersom $B_1 = \frac{S_{xy}}{s_x^2}$, $B_0 = \overline{Y} - B_1\overline{x}$, $S^2 = \frac{1}{n-2}\sum_{j=1}^{n}(Y_j - B_0 - B_1x_j)^2$ *och* $S_{xy} = R_{xy}\sqrt{s_x^2S_y^2}$ *så är*

$$\begin{aligned}(n-2)S^2 &= \sum_{j=1}^{n}\big(B_0 + B_1x_j - y_j\big)^2 = \sum_{j=1}^{n}\big(B_1(x_j - \overline{x}) - (y_j - \overline{y})\big)^2\\ &= B_1^2\sum_{j=1}^{n}(x_j - \overline{x})^2 - 2B_1\sum_{j=1}^{n}(x_j - \overline{x})(y_j - \overline{y}) + \sum_{j=1}^{n}(y_j - \overline{y})^2\\ &= (n-1)\big(B_1^2s_x^2 - 2B_1S_{xy} + S_y^2\big) = (n-1)\left(\frac{S_{xy}^2s_x^2}{S_x^4} - 2\frac{S_{xy}^2}{s_x^2} + S_y^2\right)\\ &= (n-1)\left(S_y^2 - R_{xy}^2S_y^2\right) = (n-1)S_y^2(1-R_{xy}^2),\end{aligned}$$

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

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

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:

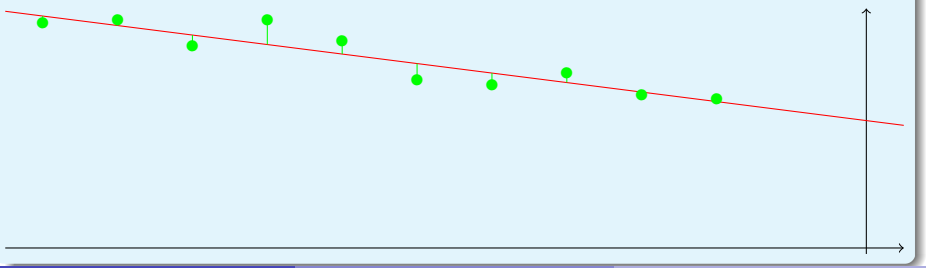
$\overline{x}$	$\overline{y}$	s_x^2	s_y^2	s_{xy}
-6.5	316	9.1667	2772.8889	-145.5556

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*

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

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 $-3.3554 \leq \dfrac{B_1 - \beta_1}{\sqrt{\frac{S^2}{(n-1)s_x^2}}} \leq 3.3554$ *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%

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

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 $p(3.1) \approx 0.6$.