MS-A0509 Grundkurs i sannolikhetskalkyl och statistik
Sammanfattning, 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 $X_1, \ldots, X_n$ 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!

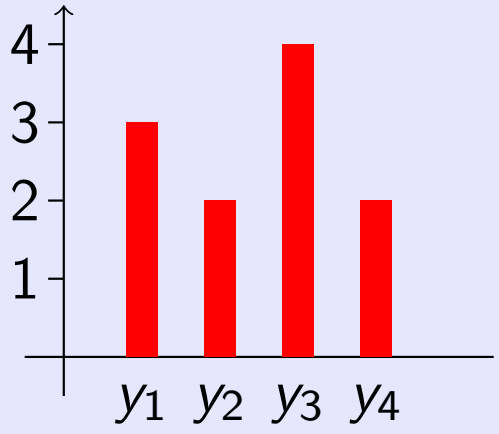
Antagandet att slumpvariablerna X_j i ett stickprov är oberoende förutsätter att vi använder "dragning med återläggning", men detta villkor uppfylls sällan! Det finns dessutom många andra större svårigheter när man i praktiken skall ta ett stickprov och detta är ett viktigt problem som inte behandlas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

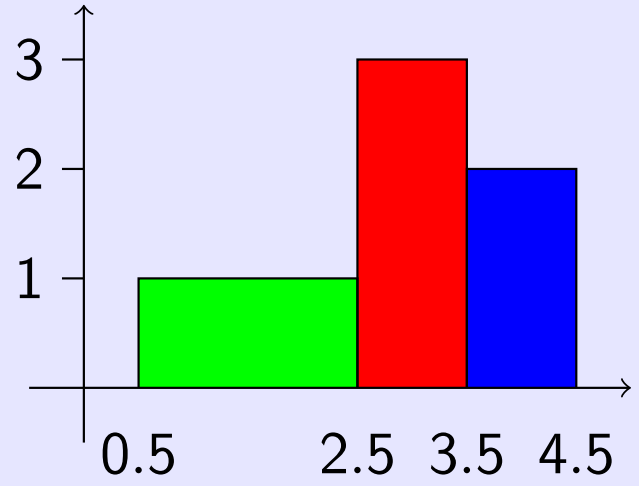
Stapeldiagram

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.*



Aritmetiskt medelvärde

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

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

och

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

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.

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.

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 $X \sim \mathrm{N}(\mu, \sigma^2)$

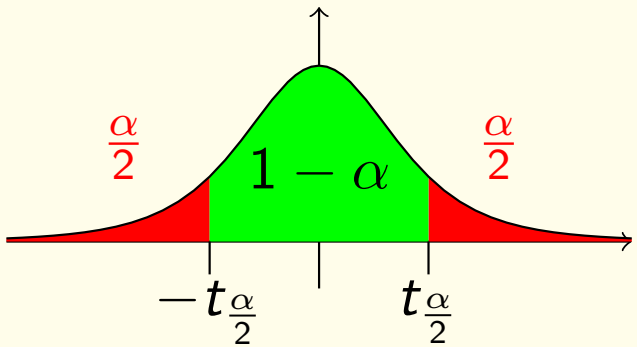
Om $X_1, X_2, \ldots, X_n$ är ett stickprov med medelvärde $\overline{X}$ och stickprovsvarians S^2 av en $\mathrm{N}(\mu, \sigma^2)$-fördelad slumpvariabel så är

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

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 $X \sim \mathrm{N}(\mu, \sigma^2)$

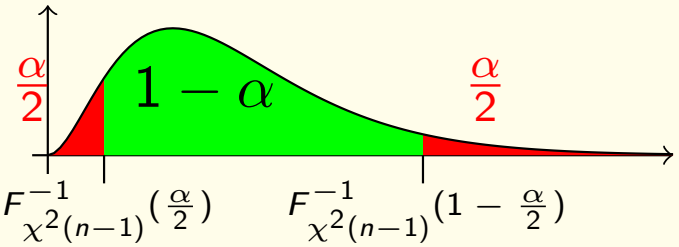
Om $X_1, X_2, \ldots, X_n$ är ett stickprov med stickprovsvarians S^2 av en $\mathrm{N}(\mu, \sigma^2)$ fördelad slumpvariabel så är

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

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.*

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*

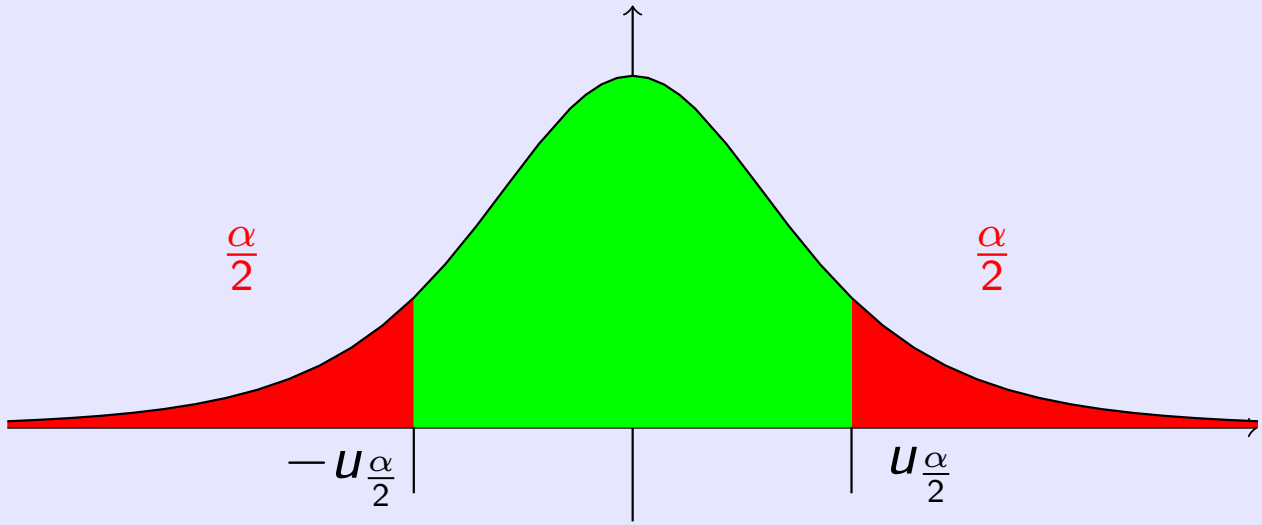
$$\frac{Y - np_0}{\sqrt{p_0(1-p_0)n}} \sim_a \mathrm{N}(0,1).$$

Obs!

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

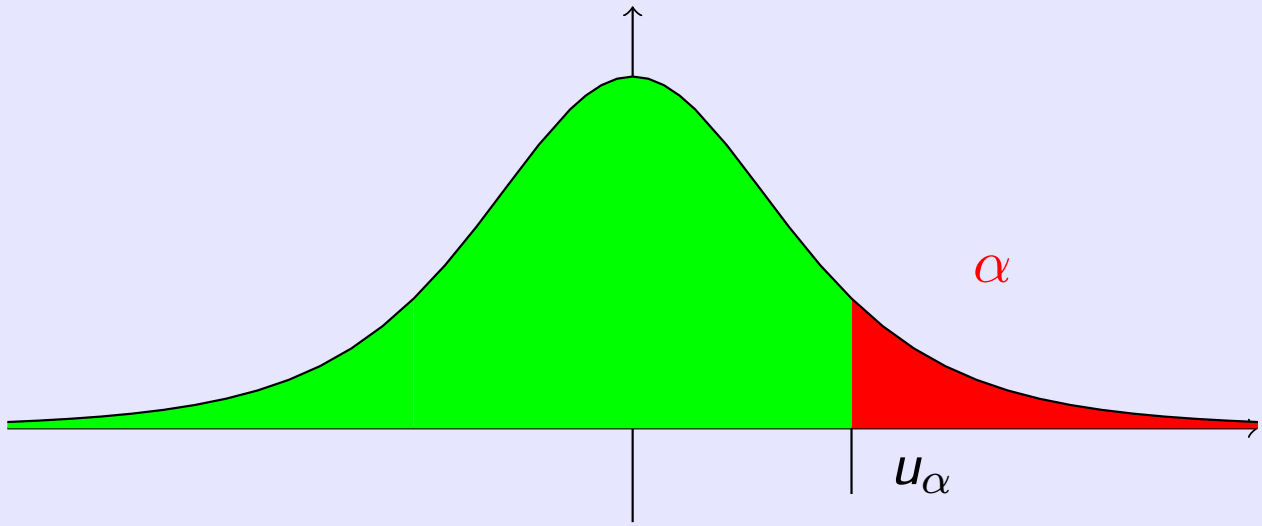
p-värde, kritiskt område, t(m)- eller N(0, 1)-testvariabel

- *Vi antar att testvariabeln U är $t(m)$- eller (approximativt)* $\mathrm{N}(0,1)$*-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* $u_{\frac{\alpha}{2}} = -F_U^{-1}(\frac{\alpha}{2}) = F_U^{-1}(1 - \frac{\alpha}{2})$.



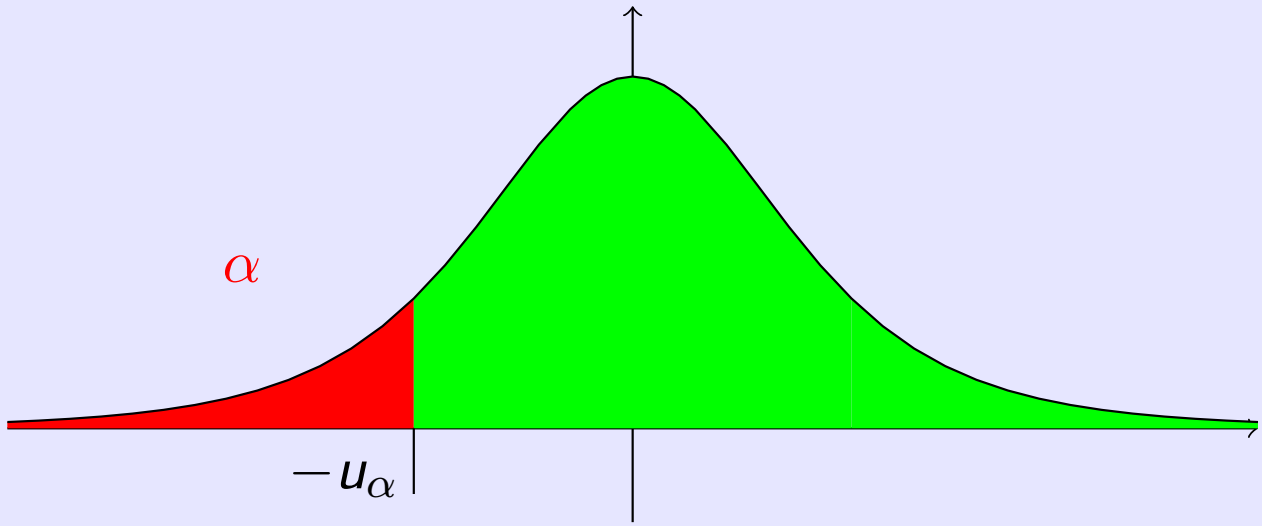
p-värde, kritiskt område, t(m)- eller N(0, 1)-testvariabel, forts.

- *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* $C = \frac{1}{\sigma^2}((n_x - 1)S_x^2 + (n_y - 1)S_y^2)$*.*

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

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

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 $\{ j : X_j \in A_k \}$.

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 $(0, F^{-1}_{\chi^2(k)}(\frac{\alpha}{2})) \cup (F^{-1}_{\chi^2(k)}(1-\frac{\alpha}{2}), \infty)$.*

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

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

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.

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).

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* $Y = \beta_0 + \beta_1 \tilde{x} + \varepsilon$.