

Hypotesprövning, regression

På förra föreläsningen undersötte vi tre metoder för hypotesprövning. Två typer av fel kan uppträda vid hypotesprövning.

- Typ I-fel: Att förkasta H_0 trots att H_0 är sann. Sannolikheten för denna feltyp är lika med felrisken α .
- Typ II-fel: Att inte förkasta H_0 trots att H_0 är falsk.

Typ II-fel är svårare att kvantifiera. Hur stort detta fel är beror på det sanna värdet på parametern θ , vilket är okänt om H_0 är falsk. Som en försiktighetsåtgärd bör vi därför designa ett hypotestest så att typ I-felet, det som vi kan mäta, är det allvarigare felet. Vi exemplifierar detta i nästa exempel. För att även kunna kvantifiera typ II-felet så inför vi styrkefunktionen.

Def: Styrkefunktionen för ett test definieras som $h(\theta) = IP(H_0 \text{ förkastas} | \theta \text{ är det sanna parametervärdet})$

För $H_0: \theta = \theta_0$ så gäller $h(\theta_0) \leq \alpha$. Ett bra test har dessutom egenskapen att $h(\theta)$ är stort för alla θ -värden som uppfyller mothypotesen. Vi visar med ett exempel hur styrkefunktionen kan användas i praktiken.

Ex: Ett utandningsprov visar alkoholhalten $X_i = \mu + \varepsilon_i$ i blodet, där μ är den sanna alkoholhalten och $\varepsilon_i \in N(0, \sigma)$ är mätfelet. För att avgöra om en person är rattfull sätter vi $H_0: \mu \leq 0,2$ (oskyldig) och $H_1: \mu > 0,2$ (rattfull). Vi väljer μ i H_0 till 0,2 för att detta är den högsta tillåtna alkoholhalten (worst case scenario). Typ I-fel motsvarar här att döma en oskyldig person för rattfylleri medan typ II-fel motsvarar att frälsa ett rattfull person (enligt gängse juridiska principer är det första felet värre). Om vi utför n utandningsprov så är $\bar{X} \in N(\mu, \sigma/\sqrt{n})$. Vi antar att H_0 är sann och vill förkasta H_0 om $\bar{X}$ är stort. Det kritiska området $C = (a, \infty)$ uppfyller

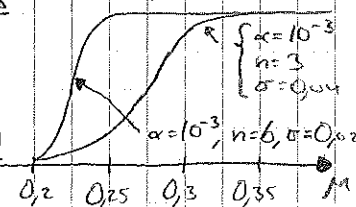
$$\alpha = IP(H_0 \text{ förkastas} | H_0 \text{ sann}) = IP(\bar{X} \in C) = IP(\bar{X} > a) = IP\left(\frac{\bar{X} - 0,2}{\sigma/\sqrt{n}} > \frac{a - 0,2}{\sigma/\sqrt{n}}\right)$$

så $a = 0,2 + \lambda_\alpha \sigma/\sqrt{n}$. Nu när det kritiska området är specificerat, så kan vi bestämma styrkefunktionen för $\mu > 0,2$

$$h(\mu) = IP(\bar{X} > a | \mu \text{ är det sanna parametervärdet}) \\ = IP\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} > \frac{a - \mu}{\sigma/\sqrt{n}}\right) = 1 - \Phi\left(\frac{a - \mu}{\sigma/\sqrt{n}}\right) \underset{a = 0,2 + \lambda_\alpha \sigma/\sqrt{n}}{=} 1 - \Phi\left(\frac{0,2 - \mu}{\sigma/\sqrt{n}} + \lambda_\alpha\right)$$

Med $n=3$, $\sigma=0,04$ och $\alpha=0,001$, så blir $h(0,3)=0,89$ och $h(0,33)=0,99$. Så för $\mu \geq 0,33$ är det mycket osannolikt att H_0 inte förkastas, dvs en person med mer än 0,33 promille i blodet kommer med minst 99% sannolikhet att befinnas

skyldig till rattfylleri. Om vi anser att testet är för dåligt, så kan vi välja att göra fler utandningsprov eller hitta en mer tillförlitlig mätmetod. Med $n=6$ och $\sigma=0,02$, så blir $h(0,245)=0,99$ så personer med mer än 0,245 promille i blodet kommer att befinnas skyldiga med minst 99% sannolikhet.

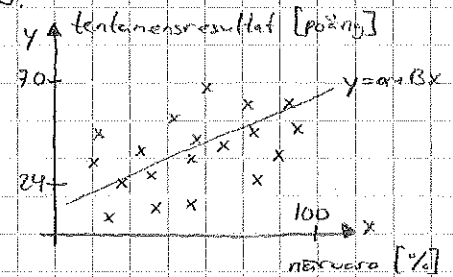


Vi avslutar med att notera att vi kan konstruera hypotestest för icke normalfördelade stickprov genom att använda normalapproximation av stickprovsvariablerna (jfr exemplet med Miljöpartiets valparandel).

Nästa föreläsning fortsätter vi med en viktig klass av hypotestest som alla involverar χ^2 -fördelningen. Vi avslutar denna föreläsning med en introduktion till regressionsanalys som är den kanske mest använda statistiska metoden. Ett vanligt problem är att avgöra om det finns något samband mellan två olika storheter. Finns det exempelvis något samband mellan studenters närvaro på föreläsningarna och deras resultat på den följande tentan? Linjär regression är ett metod för att finna sådana (linjära) samband. Vi har n par av värden

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$$

där $x_1, \dots, x_n$ är givna storheten (föreläsning-närvaro) och $y_1, \dots, y_n$ är observationer av oberoende s.v. $Y_1, \dots, Y_n$ där $Y_i \in N(\mu_i, \sigma)$ (tentamenresultat). Observera att σ antas vara oberoende av x . Vidare antas väntevärdena μ_i vara linjärt beroende av x_i , dvs $\mu_i = \alpha + \beta x_i$ och den rätta linjen $y = \alpha + \beta x$ kallas för den teoretiska regressionslinjen. Parametrarna α och β är okända, men vi kan skatta dem med hjälp av MK-metoden. Vi vill minimera kvadratsumman



$$Q(\alpha, \beta) = \sum_{i=1}^n (y_i - \mu_i)^2 = \sum_{i=1}^n (y_i - \alpha - \beta x_i)^2,$$

dvs summan av kvadraterna av avstånden ε_i i lodrät led mellan observationerna och regressionslinjen. Derivering av Q m.a.p. α och β ger:

$$\frac{\partial Q}{\partial \alpha} = -2 \sum_{i=1}^n (y_i - \alpha - \beta x_i) = 0 \quad \text{och} \quad \frac{\partial Q}{\partial \beta} = -2 \sum_{i=1}^n x_i (y_i - \alpha - \beta x_i) = 0$$

$$\Rightarrow \text{algebra} \quad B_{obs}^* = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad \text{och} \quad \alpha_{obs}^* = \bar{y} - B_{obs}^* \bar{x}$$

Notera att B_{obs}^* är kvoten av streckprovs kovariansen mellan $x_1, \dots, x_n$ och $y_1, \dots, y_n$ och streckprovs variansen för $x_1, \dots, x_n$. Med hjälp av punktskattningarna α_{obs}^* och B_{obs}^* , så får vi en skattad regressionslinje $y = \alpha_{obs}^* + B_{obs}^* x$. Vi undersöker om punktskattningen av β är väntevärdesriktig och effektiv. Vi kan skriva B_{obs}^* som en linjärkombination av observationerna, dvs

$$B_{obs}^* = \sum_{i=1}^n c_i y_i, \quad \text{där} \quad c_i = (x_i - \bar{x}) / \sum_{i=1}^n (x_i - \bar{x})^2$$

$$\Rightarrow E(B^*) = E\left(\sum_{i=1}^n c_i Y_i\right) = \sum_{i=1}^n c_i \mu_i = \sum_{i=1}^n c_i (\alpha + \beta x_i) = \left[\sum_{i=1}^n c_i \alpha = 0, \sum_{i=1}^n c_i x_i = 1\right] = \beta$$

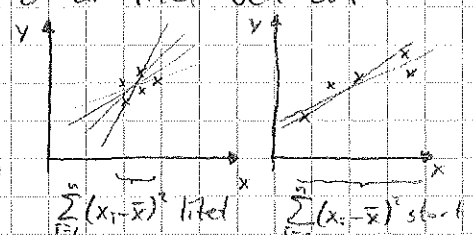
så skattningen är väntevärdesriktig. Vidare gäller

$$V(B^*) = V\left(\sum_{i=1}^n c_i Y_i\right) = \sum_{i=1}^n c_i^2 \sigma^2 = \left[\sum_{i=1}^n c_i^2 = 1 / \sum_{i=1}^n (x_i - \bar{x})^2\right] = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Punktskattningen B_{obs}^* blir alltså mer effektiv om σ är litet och om spridningen hos x_i -värdena är stor.

Eftersom B_{obs}^* är en linjärkombination av ett normalfördelat streckprov, så kan vi enkelt bestämma konfidsensintervall för β . I fallet då σ är känt för vi ett konfidsensintervall för β med konfidsensgrad p ges av

$$I_\beta = (B_{obs}^* - \lambda_{p/2} D, B_{obs}^* + \lambda_{p/2} D),$$



där $D = \sigma / \sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}$. Se boken för fallet med okänt σ . Vi kan använda detta konfidsensintervall för att göra hypotestest med $H_0: \beta = 0$ och $H_1: \beta \neq 0$. Om I_β innehåller origo så kan finns något signifikant samband mellan x och y .