

TENTAMEN: Dataanalys och statistik för I2, TMS136

Onsdagen den 25 oktober kl. 8.30-12.30 på M.

Jour: Jenny Andersson, ankn 5317

Hjälpmedel: Utdelad formelsamling med tabeller, BETA, på kursen använd ordlista och typgodkänd räknedosa.

Poängberäkning: Uppgifterna är av flervalstyp, där endast ett alternativ är rätt. Korrekt besvarad uppgift ger 2 poäng, obesvarad uppgift (vet inte) ger 0 poäng och felaktigt besvarad uppgift ger -0.5 poäng (flera ifyllda alternativ ger automatiskt -0.5 poäng). Inlämnade lösningar kommer ej tas hänsyn till vid rättningen. Fyll i och lämna in denna sida.

Svar: Lägg ut i studieportalen efter tentamens slut.

Uppgift	a	b	c	d	e	Vet inte	Poäng
1	☐	☐	☐	☒	☐	☐	
2	☐	☐	☐	☒	☐	☐	
3	☒	☐	☐	☐	☐	☐	
4	☒	☐	☐	☐	☐	☐	
5	☐	☐	☒	☐	☐	☐	
6	☐	☐	☒	☐	☐	☐	
7	☐	☐	☒	☐	☐	☐	
8	☒	☐	☐	☐	☐	☐	
9	☐	☐	☒	☐	☐	☐	
10	☐	☒	☐	☐	☐	☐	
11	☐	☐	☒	☐	☐	☐	
12	☐	☐	☒	☐	☐	☐	
13	☐	☐	☐	☒	☐	☐	
14	☐	☐	☐	☐	☒	☐	
15	☐	☐	☒	☐	☐	☐	

1d Normalfördelningen är symmetrisk så medelvärde=median.

2d Låt Y vara poängtal och y gräns för 5:a. Beräkna y så att

$$0.07 = P(Y \geq y).$$

Standardisering ger,

$$P(Y \geq y) = P\left(\frac{Y - 75}{5} \geq \frac{y - 75}{5}\right) = P\left(Z \geq \frac{y - 75}{5}\right),$$

där $Z \sim N(0, 1)$ eftersom $Y \sim N(75, 25)$. Ur tabell fås $P(Z \geq 1.48) = 0.0694$ vilket ger

$$\frac{y - 75}{5} = 1.48 \Rightarrow y = 75 + 5 \cdot 1.48 = 82.40.$$

Om man har valt 1.47 från tabellen istället (strax under 7%) så blir svaret i princip detsamma, nämligen 82.35.

3a Ett 90% konfidensintervall för den sanna andelen p ges av

$$\hat{p} \pm z_{0.95} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}.$$

Vi är 90% säkra på att p är i detta intervall, d.v.s. vi är 90% säkra på att p avviker med högst $z_{0.95} \sqrt{(\hat{p}(1 - \hat{p}))/n}$ från $\hat{p}$. Alltså

$$z_{0.95} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \leq 0.01.$$

Lös ut n ,

$$n \geq \left(\frac{z_{0.95}}{0.01}\right)^2 \hat{p}(1 - \hat{p}).$$

Problemet är nu att $\hat{p}(1 - \hat{p})$ är okänt innan undersökningen är gjord, men det är lätt att visa att $\hat{p}(1 - \hat{p}) \leq 1/4$ (maximum för $\hat{p} = 1/2$). Tillsammans med $z_{0.95} = 1.645$ så ger detta att

$$n \geq 6765.$$

4a Ett 90% konfidensintervall för den sanna andelen p ges av

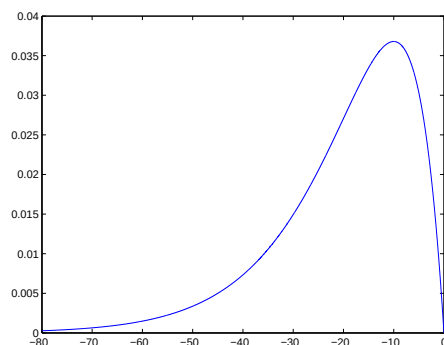
$$\hat{p} \pm z_{0.95} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}.$$

Med $\hat{p} = 0.80$, $n = 1000$ och $z_{0.95} = 1.645$ så blir det sökta konfidensintervallet

$$0.80 \pm 1.645 \sqrt{\frac{0.8 \cdot 0.2}{1000}} = 0.80 \pm 0.0208,$$

eller (0.77, 0.83).

- 5c** En vänsterskev fördelning är långsvansad till vänster, d.v.s. har en täthetsfunktion av följande typ:



Medianen delar tätheten med en vertikal linje så att arean till vänster är lika stor som den till höger. Medelvärde är balanspunkt för täthetsfunktionen. Medelvärdet kommer alltså påverkas mer än medianen av observationerna till vänster och blir mindre än medianen.

- 6c** Låt A vara händelsen att det inte snöar i Boston en viss dag i december och låt B vara händelsen att det inte snöar i Moskva en viss dag i december. Sannolikheten att det inte snöar i någon av städerna en viss dag är

$$P(A \cap B) = P(A)P(B) = (1 - 0.4)(1 - 0.6) = 0.18,$$

eftersom vädret i städerna är oberoende.

- 7c** Låt X_i , $i = 1, 2, \dots, 900$ vara vikten hos en låda. Givet är att vikten hos en låda har väntevärde $\mu = 40$ pounds och standardavvikelse $\sigma = 6$ pounds. Centrala gränsvärdessatsen ger att totala vikten på lådorna $Y = \sum_{i=1}^{900} X_i$ är approximativt $N(900 \cdot 40, 900 \cdot 6^2)$. Alltså, om $Z \sim N(0, 1)$ som vanligt, är sannolikheten att lådorna väger mer än 36450 pounds

$$P(Y > 36450) = P\left(Z > \frac{36450 - 400 \cdot 900}{\sqrt{900 \cdot 6^2}}\right) = P(Z > 2.5) = 1 - 0.9938 = 0.0062.$$

- 8a** Testa $H_0 : \delta = 0$ mot $H_a : \delta > 0$, där δ är väntevärdet på vikten före minus vikten efter. Givet medelviktförändringen $\bar{d} = 2.5$ kg, standardavvikelsen $s_d = 0.5$ kg och antalet män $n = 20$ är teststatisikan

$$t^* = \frac{\bar{d}}{s_d / \sqrt{n}} = 22.36.$$

Detta ligger långt över alla värden i t_{19} -tabellen så p-värdet är < 0.005 .

9c Skattningen av segertiden 1980 i minuter är:

$$\hat{Y}(1980) = 1029 - 0.453 \cdot 1980 = 132.06.$$

Omvandling till timmar:min:sek ger 2:12:04.

- 10b** Det går inte att urskilja något mönster i plot 1 och plot 3, medan plot 2 verkar ha linjärt växande data och plot 4 verkar ha växande varians. Alltså tyder det på att 1 och 3 är stationära.
- 11c** Alternativ a och b skulle varit sanna om intervallet var ett prediktionsintervall, d är inte sann eftersom vi är bara 95% säkra på att μ är i intervallet och ett 90% konfidensintervall är kortare än ett 95% så e är inte heller sann. Alternativ c däremot ger en korrekt tolkning av konfidensintervallet.
- 12c** Fall c är det enda då populationerna är beroende eftersom det är samma cyklister som studeras för båda typerna av måltid. Hade man valt två olika grupper som skulle testa var sin måltid så hade de varit oberoende.
- 13d** Test och konfidensintervall är ekvivalenta i den meningen att om testet i uppgiften förkastar nollhypotesen på signifikansnivå 0.05 så kommer ett 95% konfidensintervall för $\mu_1 - \mu_2$ inte att innehålla 0. Om testet däremot inte kunde förkasta nollhypotesen så innehåller intervallet 0. Alltså är d korrekt och c är inte korrekt. Alternativ a är inte korrekt eftersom ett 90% konfidensintervall är kortare än ett 95% konfidensintervall och om det senare inte innehåller 0 så kommer det förra inte heller göra det. Ett 99% konfidensintervall är längre än ett 95% så om det senare inte innehåller 0 så vet vi inte alls om 99% intervallet innehåller 0 eller inte, d.v.s. b är inte alltid sant.
- 14e** Ingen av alternativen a till d talar om ifall modellen passar eftersom de alla kan ge "bra" resultat även om data inte följer en rät linje.
- 15c** Alternativ a och c är ekvivalenta sätt att formulera hypoteserna i detta fall och nollhypotesen kan förkastas eftersom p-värdet är litet. Naturligtvis kan då inte c vara sann eftersom den säger motsatsen till b. Uttalandena d och e om Pearsonresidualerna är båda sanna.