

Svar och lösningar till tentamen i Dataanalys och statistik för I2 den 12:e mars 2004

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

- 1 Stationäritet innebär att fördelningen för data tagna vid olika tidpunkter är densamma. Det kan då inte vara någon trend i data för att de skall vara stationära. Att veta att data är stationära säger ingenting om vilken fördelning de följer bara att det är samma fördelning hela tiden. Oberoende och stationäritet är inte samma sak, oberoende är ett starkare krav. Det som är korrekt är att vi inte kan kolla oberoende utan kollar stationäritet istället eftersom det är det bästa vi klarar av.
- 2 Normalfördelningen är en kontinuerlig fördelning och sannolikheten att en kontinuerlig stokastisk variabel antar ett visst värde är alltid 0. Jämför med vad som händer om man integrerar en funktion över ett värde.
- 3 T-fördelningen är en symmetrisk fördelning precis som normalfördelningen. Skillnaden mellan dem är att t-fördelningen har tjockare svansar, d.v.s. större sannolikhet för stora eller små observationer. Ingen av påståendena är alltså rätt. Påstående a säger dessutom samma sak som påstående d.
- 4 Styrkan hos ett test är dess förmåga att förkasta nollhypotesen när alternativhypotesen är sann. Styrkan är en funktion av det sanna parametervärdet, här μ . Styrkan hos ett test kan ökas genom att öka antalet observationer i styckprovet eller genom att öka signifikansnivån. Alltså är bara alternativ b korrekt.
- 5 Stationäritet kan enbart kontrolleras genom att plotta data i någon form mot tid. Alltså är endast plot av rådata mot tid samt plot av glidande medelvärde mot tid metoder som kan användas.
- 6 Eftersom en binomialfördelad stokastisk variabel kan skrivas som summan av oberoende bernoullivariabler är centrala gränsvärdessatsen tillämplig. Övriga alternativ är inte ens korrekta påståenden och kan därmed inte vara svar på frågan. Normalfördelningen är alltid symmetrisk, men binomialfördelningen är bara symmetrisk då $p = 0.5$. Binomialfördelningen är definierad just med hjälp av ovan nämnda summa och inte med hjälp av normalfördelningen. Slutligen har inte alla fördelningar som har samma väntevärde samma fördelningsfunktion.
- 7 Konfidensintervallet som det handlar om är

$$\bar{x} \pm \frac{s}{\sqrt{n}} t_{n-1, (1+L)/2}.$$

Bredden på detta intervall är

$$2 \cdot \frac{s}{\sqrt{n}} t_{n-1, (1+L)/2}.$$

Ingenting händer med bredden när en konstant adderas till varje observation. Medelvärde kommer att öka med samma konstant, men vid beräkningen av s subtraheras medelvärdet från varje mätvärde så additionen av konstanten saknar betydelse. Övriga nämnda faktorer påverkar bredden. Storleken på stickprovet, n , ingår i uttrycket både självt och för att avgöra frihetsgraderna på t-fördelningen. Både skattningen av standardavvikelsen, s , och konfidensgraden, L , ingår i uttrycket för bredden.

- 8 Det går att inse med en gånag att det inte går att förkasta nollhypotesen. Skattningen av p är ju noll och mothypotesen är att p är större än 0.5. Man kan också välja att som vanligt beräkna p-värdet. Låt Y vara antal söndagsmorgonar med guld i mund bland 5. Då blir $Y \sim b(5, p)$. Vår skattning av Y är 0 så p-värdet beräknas som

$$p\text{-värde} = P(Y \geq 0) = 1.$$

- 9 Att få ≥ 10 poäng är samma sak som att ge ≥ 7 korrekta svar. I själva verket gäller att om X är antal korrekta svar och Y är antal poäng att

$$Y = 2X + (-0.5)(15 - X),$$

eftersom man får två poäng för varje rätt svar och minus en halv poäng för felaktigt svar. Från detta har vi alltså att $P(Y \geq 10) = P(X \geq 7)$ och eftersom $X \sim b(15, 1/5)$ är resten lätt:

$$P(X \geq 7) = 1 - P(X \leq 6) = 1 - 0.9819 = 0.0181$$

med hjälp av tabell A.1.

- 10 För att se om data är stationära måste man plotta mot tid så 1) är falsk. Det är inte medelvärdet utan medianen som syns i en boxplot så 2) är också falsk. Outliers däremot är boxplot ett bra verktyg för att upptäcka och här finns en med värde på ungefär 60 så 3) är sann. Normalfördelning kan vi inte alls se med boxplot, till det krävs helst en qq-plot eller möjligen ett histogram. Alltså är även 4) falsk.

- 11 Ett 99% prediktions intervall för en framtida längd är

$$\bar{y} \pm s \sqrt{1 + \frac{1}{n}} t_{n-1, 0.995}.$$

Antal observationer, n är enormt stort så istället för t-tabellen kan man använda normalfördelningstabell (eller kolla på t_∞ vilket är samma sak) så med insatta värden fås

$$180.2 \pm 6.5 \sqrt{1 + \frac{1}{17671}} \cdot 2.58 = 180.2 \pm 16.8.$$

- 12 Vi vill göra ett test för två oberoende populationer i binomialmodellen. Vi har mätt $y_1 = 33$, $n_1 = 50$, $y_2 = 27$ och $n_2 = 50$. Då y_1 och y_2 samt $n_1 - y_1$ och $n_2 - y_2$ alla är större än 10 går det att använda sig av den aktuella metoden i appendix 6.1. Med $\hat{p}_1 = y_1/n_1 = 0.66$, $\hat{p}_2 = y_2/n_2 = 0.54$ och $\hat{p} = (y_1 + y_2)/(n_1 + n_2)$ är det observerade värdet på teststatistikan

$$z^* = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p})(1/n_1 + 1/n_2)}} = 1.22$$

Om Z som vanligt är standardnormalfördelad så är

$$\text{p-värde} = P(Z \geq 1.22) = 1 - P(Z \leq 1.22) = 1 - 0.8888 = 0.11$$

Eftersom p-värdet är stort kan nollhypotesen inte förkastas.

- 13 När två populationer är beroende skall parat t-test användas. Fall 1) är beroende eftersom mätningarna av föroreningen sker i samma rör och halten i slutet av röret naturligtvis beror av halten i början av röret. När det handlar om bedövningsmedlet i 2) är det också beroende data eftersom det handlar om syskon som i hög grad har samma förutsättningar och bör reagera likartat. I 3) är populationerna oberoende eftersom det är två grupper om olika möss som undersöks. Slutligen är fall 4) också beroende eftersom det är samma text som skrivs på båda maskinerna och tiden beror på textens längd.

- 14 Här är det ett t-konfidensintervall som ska användas då σ är okänd, nämligen

$$\bar{y} \pm t_{n-1, 0.995} s / \sqrt{n} = \bar{y} \pm 2.771 s / \sqrt{28}.$$

- 15 Låt X vara företagets inkomster och Y dess utgifter under ett år. Variansen för skillnaden är $\sigma_X^2 + \sigma_Y^2 = 0.3^2 + 0.1^2 = 0.10$. Alltså är standardavvikelsen $\sqrt{0.10}$.