STOCKHOLMS UNIVERSITET
Statistiska institutionen
Jessica Franzén

# TENTAMEN I STATISTIKENS GRUNDER 2

**2017-10-26**

---

**Skrivtid:** 15.00-20.00
**Godkända hjälpmedel:** Miniräknare.
Tentamen består av fem uppgifter. För full poäng på en uppgift krävs tydliga, utförliga och väl motiverade lösningar.

---

## Uppgift 1

a) Förklara vad som menas med en samplingfördelning. Ge exempel.

b) Man utgår från ett givet konfidensintervall. Om sedan stickprovsstorleken $n$ ökar samt konfidensgraden ökar, allt annat lika, kan man då veta vad som händer med längden på konfidensintervallet? Motivera.

c) Förklara begreppet "styrka" i ett hypotestest.

d) Hitta på två variabler som du tror är korrelerade och där det också finns ett kausalt samband. Hitta sedan på två nya variabler som du tror är korrelerade men där ett kasualt samband saknas. Förklara hur du tänker.

e) Om kovariansen mellan $X$ och $Y$ är negativ, vilken av $V(X+Y)$ och $V(X-Y)$ blir då störst? Motivera.

## Uppgift 2

Ett större företag vill undersöka bland sina anställda om det finns något intresse för att gå över till flextid. Ett slumpmässigt urval av 200 anställda tillfrågades och resultaten visas i tabellen.

| | Män | Kvinnor |
|---|---|---|
| Positiv till flextid | 50 | 90 |
| Negativ till flextid | 30 | 30 |

a) Är inställningen till flextid beroende av kön? Utför ett lämpligt test och använd signifikansnivån 2,5 procent.

b) Är andelen som är positiva till flextid bland företagets alla anställda större än 0.65? Testa med lämpligt test. Ange testets p-värde och diskutera resultatet.

**Uppgift 3**

Man vill jämföra studenternas kunskaper i statistik efter två likvärdiga grundkurser i statistik vid två olika universitet. Antal kursdeltagare är över 500 vid båda universiteten. En månad efter avslutad grundkurs låter man 25 slumpmässigt utvalda studenter från respektive universitet göra samma statistiktest. Testet består av 7 uppgifter där varje uppgift ger maximalt 1 poäng. Maxpoängen på testet är med andra ord 7 poäng. Poängmedelvärde och standardavvikelse på testet uträknat från de 25 studenterna vid respektive universitet framgår i tabellen nedan.

| Universitet A | Universitet B |
|---|---|
| $\bar{x} = 3.51$ | $\bar{x} = 3.92$ |
| $s = 0.94$ | $s = 0.85$ |

a) Baserat på testresultaten, är studenternas kunskapsnivå i statistik högre bland studenterna på universitet B än bland studenterna på universitet A? Utför ett lämpligt test på signifikansnivån 0.05. Redogör för nödvändiga antaganden för att kunna utföra testet.

b) Beräkna två 95%-iga konfidensintervall för medelpoängen på testet, ett för respektive universitet.

**Uppgift 4**

Antal liter efterfrågad mjölk per vecka producerade av mejeriet Skogskossan är normalfördelad med väntevärde 56 000 liter och standardavvikelse 5500 liter. Mejeriet Skogskossan klassificerar den veckovisa efterfrågan som "Låg" (0-50 000 liter), "Medel" (50 000-62 000 liter) eller "Hög" (62 000 liter och mer).

a) Räkna ut sannolikheterna för Låg, Medel och Hög efterfrågan av mejeriet Skogskossans mjölk.

b) Mejeriet Skogskossan överväger vilken produktionsvolym som är den bästa. De kan lägga sin produktionsvolym på 40, 50 eller 60 tusen liter mjölk per vecka. Veckovinsten (tusentals kronor) vid olika produktionsvolymer och efterfrågenivåer framgår av tabellen nedan. Hjälp Skogskossan att avgöra lämplig produktionsvolym genom att använda tabellen och resultatet i a).

| | | Efterfrågan | | |
|---|---|---|---|---|
| | | Låg (0-50000 liter) | Medel (50000-62000 liter) | Hög (62000 liter och mer) |
| | 40 | 20 | 20 | 20 |
| Produktionsvolym | 50 | 15 | 35 | 35 |
| | 60 | 0 | 30 | 100 |

c) För ett annat mejeri, Skogskalven, har man under det gångna årets 52 veckor registrerat antal liter efterfrågad mjölk per vecka uppdelad i de tre kategorerina "Låg" (0-50 000 liter), "Medel" (50 000-62 000 liter) eller "Hög" (62 000 liter och mer). Antal veckor med Låg, Medel respektive Hög efterfrågan framgår av tabellen nedan. Testa med lämpligt test om fördelningen över Skogskalvens efterfrågan det gångna året är densamma som Skogskossans fördelning som räknades ut i a). Använd signifikansnivån 2.5%.

| | Låg | Medel | Hög | Totalt |
|---|---|---|---|---|
| Antal veckor | 5 | 39 | 8 | 52 |

**Uppgift 5**

Vi har en $\chi^2$-fördelad variabel, närmare bestämt $X \sim \chi^2(10)$. Väntevärde och varians för en $\chi^2$-fördelad variabel är $\mu = v$ och $\sigma^2 = 2v$ där $v$ är antal frihetsgrader. Vidare har vi en normalfördelad variabel, närmare bestämt $Y \sim N(5, 2^2)$. $X$ och $Y$ är oberoende.

a) Ange ungefärligt $P(X > 4)$.

b) Man tar ett stickprov från $\chi^2$-fördelningen ovan där $n = 40$. Beräkna $P(\overline{X} > 4)$. Jämför resultaten i a) och b). Är det någon skillnad i resultat?, förklara varför eller varför inte och rita en skiss över aktuella fördelningar.

c) Bestäm hur stort $n$ minst måste vara för att $P(\overline{X} > 11) < 0.01$.

d) Bestäm väntevärde och varians för $K = 2X + 5Y$.

Statistiska institutionen

Stockholms universitet

# Rättningsblad

**Datum:** 26/10/17

**Sal:** Värtasalen

**Tenta:** Statistikens grunder 2

**Kurs:** Statistikens grunder 2

**ANONYMKOD:** SG-0008

☒ Jag godkänner att min tenta får läggas ut anonymt på hemsidan som studentsvar.

**OBS! SKRIV ÄVEN PÅ BAKSIDAN AV SKRIVBLADEN**

**Markera besvarade uppgifter med kryss**

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | Antal inl. blad |
|---|---|---|---|---|---|---|---|---|---|
| X | X | X | X | X | | | | | 4 2 |
| Lär.ant. 20 | 20 | 17 | 19 | 15 | | | | | |

| POÄNG | BETYG | Lärarens sign. |
|---|---|---|
| 91 | A | JF |

**SU, STATISTIK**

**Skrivsal:** Värtasalen **Anonymkod:** S6-0008 **Blad nr:** 1

Uppg. 1

a) En samplingfördelning är en sannolikhetsfördelning för urvalskarakteristika då dessa kan ses som stokastiska variabler. Exempel på urvalskarakteristika som kan vara samplingfördelade är urvalsmedelvärde samt urvalsvarians. Hur dessa fördelningar ser ut beror på de värden som erhålls i urvalet.

4

b) Nej, det kan man inte. Ökar man stickprovsstorleken $n$, minskar man bredden på konfidensintervallet. Ökar man däremot konfidensgraden, ökar man istället bredden på konfidensintervallet. Med andra ord ger dessa motsatt effekt. För att veta vad som händer med längden/bredden på konfidensintervallet måste man få veta värdena för dessa handlingar.

4

c) Ett tests styrka betecknas $1-\beta$ där $\beta$ står för sannolikheten att acceptera nollhypotesen $H_0$ trots att denna i verkligheten är falsk. Man eftersträvar höga värden för testets styrka, dvs låga värden på $\beta$ (som antar värden mellan 0 och 1).

Teststyrka: $1-\beta = 1-P(H_0 \text{ accepteras} \mid H_0 \text{ falsk})$

4

d) Två variabler som är korrelerade är inkomst och konsumtion som dessutom ger ett kausalt samband (orsak-verkan samband) då variabeln inkomst har en direkt påverkan på variabeln konsumtion. Dvs. sker en förändring i inkomst, kommer även en förändring i konsumtion uppstå/ras.

Två variabler som är korrelerade är trycket i en hunds ögon där varje öga givits en trycksänkande medicin. Det saknas dock ett kausalt samband, då korrelationen snarare beror på att mätvärdena är tagna från samma hund.

4

e) $V(X-Y)$ blir störst då beroende råder och

$V(X-Y) = V(X) + V(Y) - 2Cov(X,Y)$ kommer att resultera i att den sista termen blir positiv medan det för $V(X+Y) = V(X) + V(Y) + 2Cov(X,Y)$ kommer resultera i att den sista termen blir negativ. Detta gäller pga:

$-(-) = +$ och $+(-) = -$ $V(X-Y)$ blir då störst!

4

Uppg. 2. a) Oberoende test

Observerade frekvenser

| | Män | Kvinnor | |
|---|---|---|---|
| Positiv | 50 | 90 | 140 |
| Negativ | 30 | 30 | 60 |
| | 80 | 120 | 200 |

Förväntade frekvenser

| | Män | Kvinnor | |
|---|---|---|---|
| Positiv | 56 | 84 | 140 |
| Negativ | 24 | 36 | 60 |
| | 80 | 120 | 200 |

$H_0$: oberoende råder mellan inställning till flextid och kön

$H_A$: Beroende råder mellan inställning till flextid och kön

Testvariabel: Då alla förväntade frekvenser $E(n) \geq 5$ gäller

$$\chi^2 = \sum \frac{(n-E(n))^2}{E(n)} \underset{approx}{\sim} \chi^2(\nu)$$ om $H_0$ är sann, där $\nu = $ (antal kategorier $-1$)(antal rader $-1$), dvs i detta fall $\nu = 1$

Kritiskt värde: $\chi^2_{0,025}(1) = 5,024$

Beslutsregel: Förkasta $H_0$ om $\chi^2_{obs} > \chi^2_{0,025}(1) = 5,024$

$$\chi^2_{obs} = \frac{(50-56)^2}{56} + \frac{(90-84)^2}{84} + \frac{(30-24)^2}{24} + \frac{(30-36)^2}{36} \approx 3,571$$

Svar: $H_0$ kan inte förkastas på signifikansnivån 2,5 procent då

$\chi^2_{obs} = 3,571 < 5,024 = \chi^2_{0,025}(1)$. Dvs det går inte att bevisa beroende.

16

b) $p$ = "andelen positiva till flextid i urvalet"

$\pi$ = "andelen positiva till flextid bland företagets anställda"

$$p = \frac{50+90}{200} = \frac{140}{200} = 0,7$$

$n(1-p) \geq 5$

$np \geq 5$

$H_0: \pi = 0,65$

$H_A: \pi > 0,65$

Testvariabel:
Då urvalet är stort ($n > 30$) approximeras en normalfördelning enligt CGS dvs.

$$Z = \frac{p-\pi}{\sqrt{\frac{\pi(1-\pi)}{n}}} \underset{approx}{\sim} N(0,1)$$ om $H_0$ är sann

Högersidigt test med signifikansnivån 5% ger beslutsregeln:

$H_0$ förkastas om $Z_{obs} > Z_{0,05} = 1,6449$.

→ Forts.

**SU, STATISTIK**

**Skrivsal:** Värtasalen **Anonymkod:** S6-0008 **Blad nr:** 2

Uppg. 2 b) Forts.

$$Z_{obs} = \frac{0,7-0,65}{\sqrt{\frac{0,65 \cdot 0,35}{200}}} \approx 1,4825$$

$\boxed{\text{$H_0$ förkastas inte på signifikansnivån 5\% då; } Z_{obs} = 1,4825 < 1,6449 = Z_{0,05}}$

p-värde: $P(Z > Z_{obs}) = P(Z > 1,48) = 1 - P(Z < 1,48) = 0,06944$

$Z_{obs} = 1,4825 \approx 1,48$

Svar: Testets p-värde är 0,06944, dvs den lägsta signifikansnivån som $H_0$ kan förkastas på. Detta går i linje med resultatet som jag erhöll när jag testade hypotesen på signifikansnivån 5% och inte kunde förkasta $H_0$. Detta pga att man förkastar $H_0$ på alla signifikansnivåer större än det p-värde vi räknat ut. Man kan därför den förkasta hypotesen på 10% signifikansnivå. Däremot tvingas man acceptera $H_0$ på alla signifikansnivåer lägre än p-värdet. Man måste därför acceptera, dvs. inte förkasta, $H_0$ på 5% signifikansnivå, vilket också resultatet från testet jag utförde visade!

10

Uppg. 3.

a) A = "kunskapsnivå Universitet A"
B = "kunskapsnivå Universitet B"

$n_A = n_B = 25$

$\frac{N-n}{N-1} = \frac{500-25}{499} \approx 0,952$

dvs tas inte med! K

$$\boxed{\begin{array}{l} H_0: \mu_A = \mu_B \\ H_1: \mu_A < \mu_B \end{array}}$$

Antagande:
Vi måste anta oberoende observationer.
Normalförd. pop.

(Urvalsstorleken är inte stor ($n_A = 25$ och $n_B = 25$) men då tabellen saknar värden för $t_{0,05}(48)$ ~~måste normal approximation användas~~. Då testet för $t_{0,05}(50)$ (som finns i tabell) visar på samma resultat (att $H_0$ inte förkastas på sign.nivån 5%) anses resultaten rimliga.)

Testvariabel:

t-förd. $\cancel{Z} = \frac{\bar{x}_A - \bar{x}_B - D_0}{\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}} \underset{\text{appr}}{\sim} N(0,1)$ om $H_0$ är sann.

Beslutsregel:
Förkasta $H_0$ om $Z_{obs} < Z_{0,05} = -1,6449$

$\cancel{Z}_{obs} = \frac{3,51 - 3,92 - 0}{\sqrt{\frac{0,94^2}{25} + \frac{0,85^2}{25}}} \approx -1,6176$

Svar: $H_0$ förkastas inte på signifikansnivån 5% då $Z_{obs} = -1,6176 > -1,6449 = Z_{0,05}$. Dvs kan vi inte bevisa att kunskapsnivån är högre bland studenter vid universitet B.

För $t_{0,05}(50)$ gäller

$t_{obs} = \frac{\bar{x}_A - \bar{x}_B - D_0}{\sqrt{s_p^2 \left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$ K

Antaganden:
Oberoende obs
N. förd pop.
$\sigma^2 = \sigma_A^2 = \sigma_B^2$
R

Förkasta $H_0$ om $|t_{obs}| > 1,676$

$|t_{obs}| = \left| \frac{3,51 - 3,92 - 0}{\sqrt{0,80305\left(\frac{1}{25} + \frac{1}{25}\right)}} \right| \approx 1,618$

K

7

$H_0$ förkastas inte!

**SU, STATISTIK**

**Skrivsal:** Värtasalen **Anonymkod:** SG-0008 **Blad nr:** 3

Uppg. 3. b) Universitet A

Oberoende obs samt N.förd populationer antas för de båda universiteten

$\bar{X} \pm t_{0,025}(24) \cdot \frac{0,S}{\sqrt{n}}$

$3,51 \pm 2,064 \cdot \frac{0,94}{\sqrt{25}}$

$3,51 \pm 0,388032$

$[3,121968; 3,898032]$

Svar: Ett 95% k.i. för medelpoängen på testet för Universitet A ges av $[3,121968; 3,898032]$ N

Universitet B

$3,92 \pm t_{0,025}(24) \cdot \frac{0,85}{\sqrt{25}}$

$3,92 \pm 0,35088$

$[3,56912; 4,27088]$

Svar: Ett 95% k.i. för medelpoängen på testet för universitet B ges av $[3,56912; 4,27088]$ 4

10

Uppg. 4. a) $P(\text{Låg efterfrågan}) = P(X<50000) = P\left(Z < \frac{50000-56000}{5500}\right) =$

X="efterfrågan" $= P(Z<-1,09) = \Phi(-1,09) = 1-\Phi(1,09) = 0,13786$

$E(X)=56000$

$\sigma_X = 5500$

$X \sim N(56000, 5500^2)$

$P(\text{medel efterfrågan}) = P(50000<X<62000) =$
$= P\left(\frac{50000-56000}{5500} < Z < \frac{62000-56000}{5500}\right) = P(-1,09<X<1,09) =$
$= 2\Phi(1,09)-1 = 0,72428$

$P(\text{Hög efterfrågan}) = P(X>62000) = P\left(Z > \frac{62000-56000}{5500}\right) =$
$= 1-P(Z<1,09) = 0,13786$

$P(\text{Låg efterfrågan}) + P(\text{medel efterfrågan}) + P(\text{Hög efterfrågan}) = 1$

3

Uppg. 4 b) Beslut under risk

| | | Efterfrågan | | | |
|---|---|---|---|---|---|
| | | Låg $p=0,13786$ | Medel $p=0,72428$ | Hög $p=0,13786$ | Förväntad nytta |
| Produktionsvolym | 40 | 20 | 20 | 20 | $20\cdot0,13786+20\cdot0,72428+20\cdot0,13786=20$ |
| | 50 | 15 | 35 | 35 | $15\cdot0,13786+35\cdot0,72428+35\cdot0,13786=32,2428$ |
| | 60 | 0 | 30 | 100 | $0\cdot0,13786+30\cdot0,72428+100\cdot0,13786=35,5144$ |

Svar: Största förväntade nytta får vi om produktionsvolymen är 60 tusen liter mjölk per vecka. Därför är det lämpligt att välja detta alternativ.

8

c) Goodness-of-fit test.

| | $n_i$ | $E(n_i)$ |
|---|---|---|
| L | 5 | 7,16872 |
| M | 39 | 37,66256 |
| H | 8 | 7,16872 |
| | 52 | |

Alla $E(n_i) \geq 5$ så approx ok!

$H_0$: Fördelningen för skogsbolagens efterfrågan följer den för skogsklassen. dvs 0,13 0,72 0,13

$H_A$: Fördelningen för skogsbolagens efterfrågan följer inte den för skogsklassen.

Testvariabel:

$$\chi^2 = \sum \frac{(n_i - E(n_i))^2}{E(n_i)} \underset{\text{approx}}{\sim} \chi^2(k-1) \text{ om } H_0 \text{ är sann.}$$

$k-1 = 3-1 = 2$

Beslutsregel:

Förkasta $H_0$ om $\chi^2_{obs} > \chi^2_{0,025}(2) = 7,378$

$$\chi^2_{obs} = \frac{(5-7,16872)^2}{7,16872} + \frac{(39-37,66256)^2}{37,66256} + \frac{(8-7,16872)^2}{7,16872} \approx 0,800$$

Svar: $H_0$ kan inte förkastas på signifikansnivån 2,5 % då $\chi^2_{obs} = 0,801 < 7,378 = \chi^2_{0,025}(2)$

8

**SU, STATISTIK**

**Skrivsal:** Värtasalen **Anonymkod:** S6-0008 **Blad nr:** 4

Uppg. 5.

$\mu = v = 10 \quad \sigma^2 = 2v = 20$

$X \sim \chi^2(v)$

$v = 10$

a) $P(X > 4) = P\left(Z > \frac{4-10}{\sqrt{20}}\right) = P(Z > -1{,}34) =$

$= 1 - P(Z < -1{,}34) = 1 - \Phi(-1{,}34) = 1 - (1 - \Phi(1{,}34)) =$

$= \Phi(1{,}34) = 0{,}90988$ Svar: $P(X>4) = 0{,}90988$

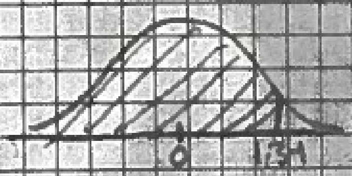
0

b) $P(\bar{X} > 4) = P\left(Z > \frac{4-10}{\sqrt{\frac{20}{40}}}\right) = P(Z > -8{,}49) =$

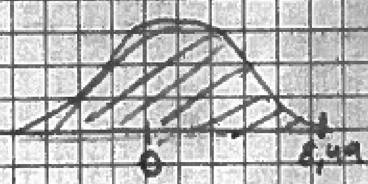
$= 1 - P(Z < -8{,}49) = 1 - (1 - \Phi(8{,}49)) = \Phi(8{,}49) \approx 1$

Svar: Det är skillnad i resultat då vi ser att slh att få ett urvalsmedelvärde större än 4 i ett urval om 40 är nästan 1 medan slh att få ett värde på X större än 4 är 0,9. Detta är inte konstigt då man vid enskilda mätningar får en större varians, i detta fall $\sigma^2 = 20$ än vid mätningar med större urval, i detta fall $\sigma^2_{\bar{x}} = \frac{20}{40}$. Med andra ord gör detta att variationerna på värdena vi får ut när vi drar ett urval är mindre än de som vi får när vi drar enskilda mätningar, detta visas i skisserna nedan.

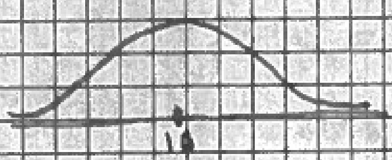
$P(X>4)$




$P(\bar{X}>4)$

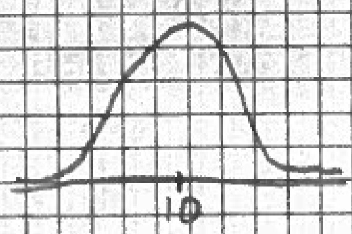



$\sigma^2 = 20$




6

$\sigma^2_{\bar{x}} = \frac{20}{40}$



Uppg. 5.

c) $P(\bar{X} > 11) = P\left(Z > \frac{11-10}{\sqrt{\frac{20}{n}}}\right) = P\left(Z > \frac{1}{\sqrt{\frac{20}{n}}}\right)$

$P(Z = 2{,}3263) = 0{,}01$

$P(Z > 2{,}3263) < 0{,}01$

$\downarrow$

$Z > 2{,}3263$

$Z > \frac{1}{\sqrt{\frac{20}{n}}}$

$2{,}3263 = \frac{1}{\sqrt{\frac{20}{n}}}$

$\frac{1}{2{,}3263} = \sqrt{\frac{20}{n}}$

$0{,}429867171 = \frac{4{,}472135955}{\sqrt{n}}$

$\sqrt{n} = 10{,}40352987$

$n \approx 108{,}23$ avrunda uppåt dvs $\boxed{n \geq 109}$

6

Svar: $n$ måste vara minst 109 $(n \geq 109)$ för att $P(\bar{X} > 11) < 0{,}01$

d) $E(K) = E(2X + 5Y) = 2E(X) + 5E(Y) =$
$= 2 \cdot \mu_X + 5 \cdot \mu_Y = 2 \cdot 10 + 5 \cdot 5 = 20 + 25 = 45$

Oberoende X och Y

$V(K) = V(2X + 5Y) = 2^2 V(X) + 5^2 V(Y) =$
$= 4V(X) + 25V(Y) = 4 \cdot (\cancel{2} \cdot 20) + 25 \cdot (2^2) =$
$= 4 \cdot \cancel{40} + 25 \cdot 4 = 160 + 100 = 260$

20

ingen kovarians

Svar: $E(K) = 45$ och $V(K) = 260$

3