free-learning

Statistik

Ladda ner kompendiet gratis på Studentia.se

STUDENTIA FREELEARNING STUDENTIA FREELEARNING

Björn Lantz

Lär lätt!

Statistik - Kompendium

Studentia

www.studentia.se

Lär lätt! Statistik - Kompendium
© 2006 Björn Lantz och Studentia
Ladda ner kompendiet gratis på www.studentia.se
ISBN 87-7681-080-1

Studentia
www.studentia.se

Innehållsförteckning

1. Introduktion till statistik

1.1 Inledning

Statistik används huvudsakligen till två olika saker:

- Att beskriva en datamängd i olika avseenden (beskrivande statistik).
- Att analysera verkligheten med hjälp av stickprov (analytisk statistik).

Grunden för statistik är numerisk information – data – av olika slag. Datan kan vara empirisk (insamlad från ”verkligheten”) eller teoretisk (baserad på antaganden). Med hjälp av sådan data kan verkligheten beskrivas och/eller analyseras på olika sätt med hjälp av statistiska metoder.

Om variabelvärdena i en viss mängd data representerar kvantiteter – d.v.s. svarar på frågan ”hur många?” – så sägs datan vara av *kvantitativ* karaktär. Motsatsen är om variabelvärdena representerar kategorier, t.ex. kön, färger eller placeringar. I detta fall sägs datan vara *kvalitativ*. Kvalitativ data kan delas in i *nominaldata*, där de olika kategorierna inte kännetecknas av någon inbördes ordning, eller *ordinaldata*, där de olika kategorierna kan rangordnas.

Observera att kvalitativ data mycket väl kan vara av sifferkaraktär. Till exempel är betyg på en skala 1-10 ordinaldata, medan postnummer är nominaldata. Data blir inte kvantitativ för att de olika kategorierna har namn som råkar består av siffror. Det krävs att datan faktiskt har en reell matematisk betydelse för att den ska vara kvantitativ. Det är viktigt att man har klart för sig vilken typ av data man arbetar med, eftersom valet av statistiska metoder och modeller styrs av detta.

1.2 Stolpdiagram och fördelning

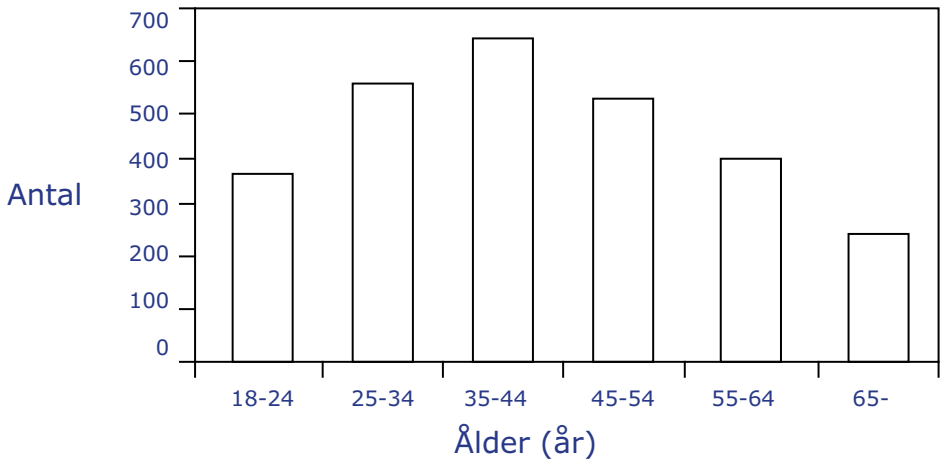
”En bild säger mer än tusen ord” lyder ett gammalt talesätt. Ett vanligt sätt att beskriva en datamängd, där de enskilda observationerna naturligt kan grupperas eller där det endast finns ett fåtal olika värden (*utfall*) för de enskilda observationerna, är att använda ett stolpdiagram. I ett stolpdiagram illustrerar höjden av varje enskild stolpe en frekvens, d.v.s. ett antal observationer. Hur antalet observationer i en datamängd är fördelat på olika utfall kallas för datamängdens *frekvensfördelning*. Stolpdiagram är ofta ett mycket användbart sätt att skaffa sig en bild av en viss datamängds frekvensfördelning.

Anta till exempel att 2689 personer har blivit av med sina körkort under en viss period, och att dessa personer åldersmässigt är fördelade på det sätt som framgår av tabellen nedan.

Ålderskategori	Antal personer
18-24	356
25-34	555
35-44	643
45-54	529
55-64	387
65-	219

Ett exempel på hur ett stolpdiagram kan beskriva denna datamängd finns i figur 1.1.

Figur 1.1
Stolpdiagram med frekvensfördelning

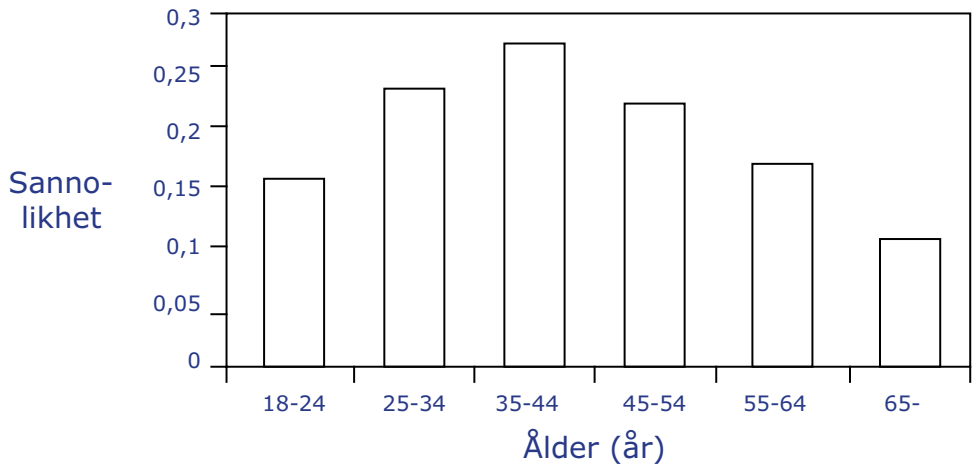


Samma typ av diagram kan också användas för att visa också hur sannolikt det är att en slumpmässigt vald person som har blivit av med körkortet hör till en viss ålderskategori. I tabellen nedan har frekvenserna räknats om till sannolikheter för de olika kategorierna genom att dividera var och en av dem med det totala antalet personer i datamängden.

Ålderskategori	Sannolikhet
18-24	356/2689=0,1324
25-34	555/2689=0,2064
35-44	643/2689=0,2391
45-54	529/2689=0,1967
55-64	387/2689=0,1439
65-	219/2689=0,0814

Generellt kallas en variabel som får sitt värde via en slumpmässig process för *slumpvariabel*. En fullständig beskrivning av hur sannolika de olika möjliga utfallen för en viss slumpvariabel är kallas för slumpvariabelns *sannolikhetsfördelning*, eller bara *fördelning*. I figur 1.2 nedan illustreras sannolikhetsfördelningen för exemplet ovan med ett stolpdiagram. Notera att stolparnas inbördes förhållande är exakt detsamma som i figur 1.1.

Figur 1.2
Stolpdiagram med sannolikhetsfördelning



1.3 Centraltendens

Kvantitativa mått används ofta för att beskriva en datamängd. För att beskriva datamängdens centraltendens (”tyngdpunkt”) används ofta måttet *medelvärde*. (Det aritmetiska) medelvärdet för ett stickprov bestående av n observationer $x_1, x_2, \ldots, x_n$ från en större datamängd betecknas med $\bar{x}$ och definieras som

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

där Σ betyder ”summa”. Det sanna medelvärdet för en större datamängd – en *population* – bestående av N observationer $x_1, x_2, \ldots, x_N$, betecknas med den grekiska bokstaven μ (uttalas ”my”) och definieras som

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Man inser enkelt att medelvärde bara kan beräknas för data av kvantitativ karaktär. (Fundera själv på vad nytta du skulle ha av medelvärdet av spelarnas nummer i ditt favoritlag i fotboll!) När man arbetar med ordinaldata används därför ofta *median* som mått på centraltendensen. Medianen definieras som den observation under vilken 50% av den totala rangordnade datamängden ligger. I praktiken innebär medianen ”den mittersta” observationen när antalet observationer är udda. Om antalet observationer är jämnt definieras medianen som det aritmetiska medelvärdet av de två mittersta observationerna.

Medianen kallas också för den 50:e *percentilen*, eftersom just 50% av observationerna har lägre värde. På samma sätt är 99:e percentilen den observation under vilken 99% av de observerade värdena ligger. Den 25:e respektive 75:e percentilen kallas dessutom för 1:a (eller ”nedre”) respektive 3:e (eller ”övre”) *kvartilen*, eftersom 1 respektive 3 kvartar (fjärdedelar) av den totala datamängden har lägre värde. Medianen är således samma sak som 2:a kvartilen.

När nominaldata ska beskrivas kan den inte rangordnas, vilket innebär att man inte kan tala om någon median, percentil eller kvartil för datamängden. Som mått på centraltendens används då istället datamängdens *typvärde*, vilket är samma sak som det vanligaste förekommande värdet. Om flera värden är lika vanligt förekommande i en datamängd, och inget annat värde är mer vanligt förekommande, så är samtliga dessa värden typvärden för datamängden.

Observera att även om man inte kan tala om medelvärde för en nominaldatamängd så går det utmärkt att beskriva kvantitativ data med typvärde. På samma sätt har en mängd kvantitativ data en median, även om en ordinaldatamängd inte har något medelvärde. En ordinaldatamängd har däremot alltid ett typvärde (eller flera). Man kan alltså säga att det finns tre ”nivåer” av data, nämligen

1. kvantitativ data
2. ordinaldata, och
3. nominaldata

där man för varje nivå även kan använda de metoder som finns tillgängliga för data på lägre nivå, men aldrig metoder som hör till data på högre nivå.

1.4 Spridning

Två datamängder som har samma centraltendens kan ändå skilja sig åt avsevärt vad gäller de enskilda observationernas spridning. Jämför t.ex. följande datamängder:

Datamängd I: 5, 6, 6, 6, 6, 7.
Datamängd II: 1, 4, 6, 6, 8, 11.

Både I och II består av 6 observationer och har samma medelvärde, median och typvärde, nämligen 6. Men II kännetecknas av betydligt större spridning kring sin centraltendens. För att beskriva en datamängd är därför mått på spridningen kring centraltendensen ofta användbart.

Det finns ett antal olika mått på spridning. Om datamängden är lägst på ordinalnivån så kan man tala om *intervallet* – differensen mellan lägsta och högsta värdet i datamängden, eller *interkvartila intervallet* – differensen mellan 1:a och 3:e kvartilen. Ett betydligt mer användbart mått på den genomsnittliga spridning kring ett medelvärde när man arbetar med en kvantitativ datamängd är emellertid *standardavvikelsen* för datamängden. Standardavvikelsen för en population bestående av N observationer $x_1, x_2, \ldots, x_N$, betecknas med den grekiska bokstaven σ (uttalas ”sigma”) och definieras som

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}}$$

Ju större spridningen är kring medelvärdet, desto större blir alltså värdet på standardavvikelsen.

Om datamängden utgörs av ett stickprov bestående av n observationer $x_1, x_2, \ldots, x_N$, från en population med okänt μ betecknas stickprovets standardavvikelse med s och definieras som

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}}$$

Anledningen till att man dividerar med n - 1 istället för n när man arbetar med stickprovsdata är att man måste justera för osäkerhet som beror på det faktum att vi *skattar* populationens sanna medelvärde μ med stickprovets medelvärde $\overline{x}$ i formeln. Man säger då att s beräknas med n - 1 frihetsgrader.

Ett mått som är relaterat till standardavvikelsen är *variansen*, som helt enkelt är standardavvikelsen i kvadrat. Variansen för en population respektive ett stickprov blir alltså

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

respektive

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

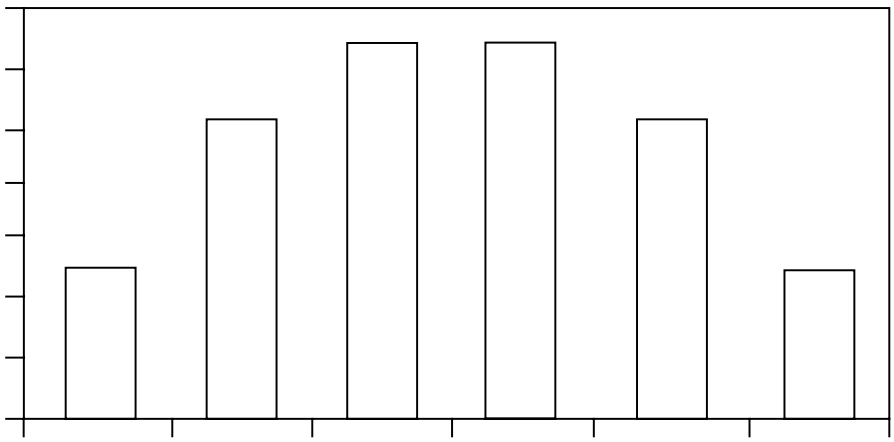
Det går ofta att dra ganska långtgående slutsatser om hur en datamängd ”ser ut” bara utifrån information om medelvärdet och standardavvikelsen. Den matematiska tesen *Chebyshevs teorem* säger t.ex. att det för alla datamängder, *oavsett hur de är fördelade*, gäller att minst 3 / 4 av de enskilda observationerna i datamängden kommer att ligga inom två standardavvikelser från datamängdens medelvärde. Teoremet säger också att minst 8 / 9 av de enskilda observationerna kommer att ligga inom tre standardavvikelser från medelvärdet.

Om man vet om att datamängden är rimligt ”klockformad”, d.v.s. hyfsat symmetrisk kring en topp, vilket fallet ofta är i många verkliga situationer, så säger den s.k. *empiriska regeln* att ungefär 2 / 3 av de enskilda observationerna i datamängden kommer att ligga inom en standardavvikelse från datamängdens medelvärde. Dessutom kommer ungefär 19 / 20 av de enskilda observationerna kommer att ligga inom två standardavvikelser från medelvärdet, och praktiskt taget alla att ligga inom tre standardavvikelser.

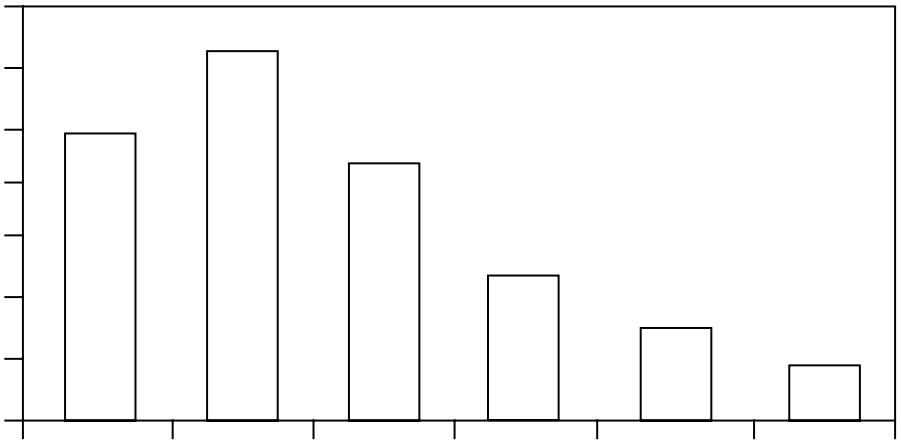
1.5 Skevhet

En fördelning kan vara *symmetrisk* (se figur 1.3). Ofta är fördelningar emellertid *skeva*, vilket innebär att spridningen inte är likadan på båda sidor om medelvärdet. En fördelning som är skev åt höger kommer i grafisk illustration att vara utsträckt åt höger (se figur 1.4), och tvärtom. I en fördelning som är skev åt höger kommer medelvärdet dessutom att vara högre än medianen, och tvärtom.

Figur 1.3
En symmetrisk fördelning



Figur 1.4
En fördelning som är skev åt höger



När en datamängd ska beskrivas är det ofta meningsfullt att inkludera ett mått på skevhet. Skevheten för en population bestående av N observationer $x_1, x_2, \ldots, x_N$, med medelvärdet μ och standardavvikelsen s beräknas normalt som

$$\text{Skevhet} = \sum_{i=1}^{N}\left[\frac{x_i - \mu}{\sigma}\right]^3 / N$$

När värdet på skevhet är positivt så innebär det att populationen ifråga är skev åt höger, och tvärtom. En helt symmetrisk fördelning har värdet 0 på skevheten.

1.6 Några exempel

Exempel 1.1

Under en följd av 12 dagar observerades följande antal dagliga köpare av bilar hos en viss bilhandlare: 1, 4, 1, 2, 6, 3, 2, 3, 2, 1, 4 och 2.

a) Vilken typ av data (nivå) är detta?
b) Vad är medelvärdet?
c) Vad är medianen?
d) Vad är typvärdet?
e) Beräkna σ.
f) Beräkna σ^2.
g) Beräkna s.
h) Beräkna s^2.
i) Beräkna skevheten.
j) Illustrera datamängden med ett stolpdiagram.
k) Kontrollera om Chebyshevs teorem stämmer.

Electrolux www.electrolux.se

At Electrolux we share a common goal – to make everyday life easier and more enjoyable. Our people work to reach this goal and they all contribute to the Group´s success.
With a presence in more than 100 countries Electrolux is a truly international company, which gives opportunities to succeed and professionally grow.

Electrolux focuses on developing its people, helping them to grow personally and professionally. To attract and retain highly talented personnel is considered to be at the core of our company strategy.

We consider talent as one of our most valuable assets. We believe that managing and developing this asset is a prerequisite for success. In Electrolux talent management is a strategic priority. Our vision for talent management is to have a strong pool of outstanding employees and a performance culture that strives for excellence.

Lösning

a) Kvantitativ data.

b) Medelvärde: $\frac{1+4+1+2+6+3+2+3+2+1+4+2}{12} = 2,583$

c) Median: 2

d) Typvärde: 2

e) $\sigma = \sqrt{\frac{(1-2,583)^2+(4-2,583)^2+(1-2,583)^2+(2-2,583)^2+...+(2-2,583)^2}{12}} = 1,441$

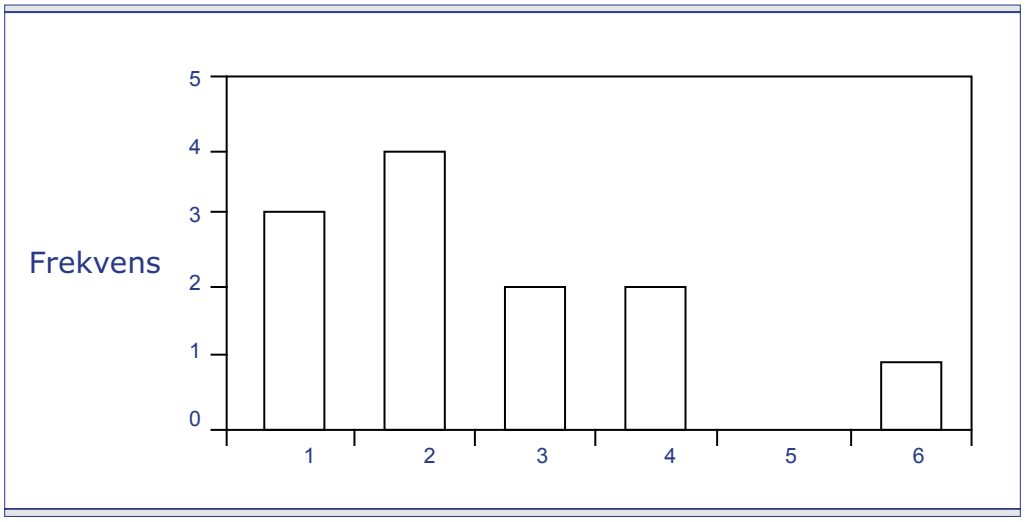
f) $\sigma^2 = 1,441^2 = 2,076$

g) $s = \sqrt{\frac{(1-2,583)^2+(4-2,583)^2+(1-2,583)^2+(2-2,583)^2+...+(2-2,583)^2}{11}} = 1,505$

h) $s = 1,505^2 = 2,265$

i) Skevhet $= \left[\frac{1-2,583}{1,441}\right]^3 / 3 + \left[\frac{4-2,583}{1,441}\right]^3 / 3 + ... + \left[\frac{2-2,583}{1,441}\right]^3 / 3 = 0,920$

j)

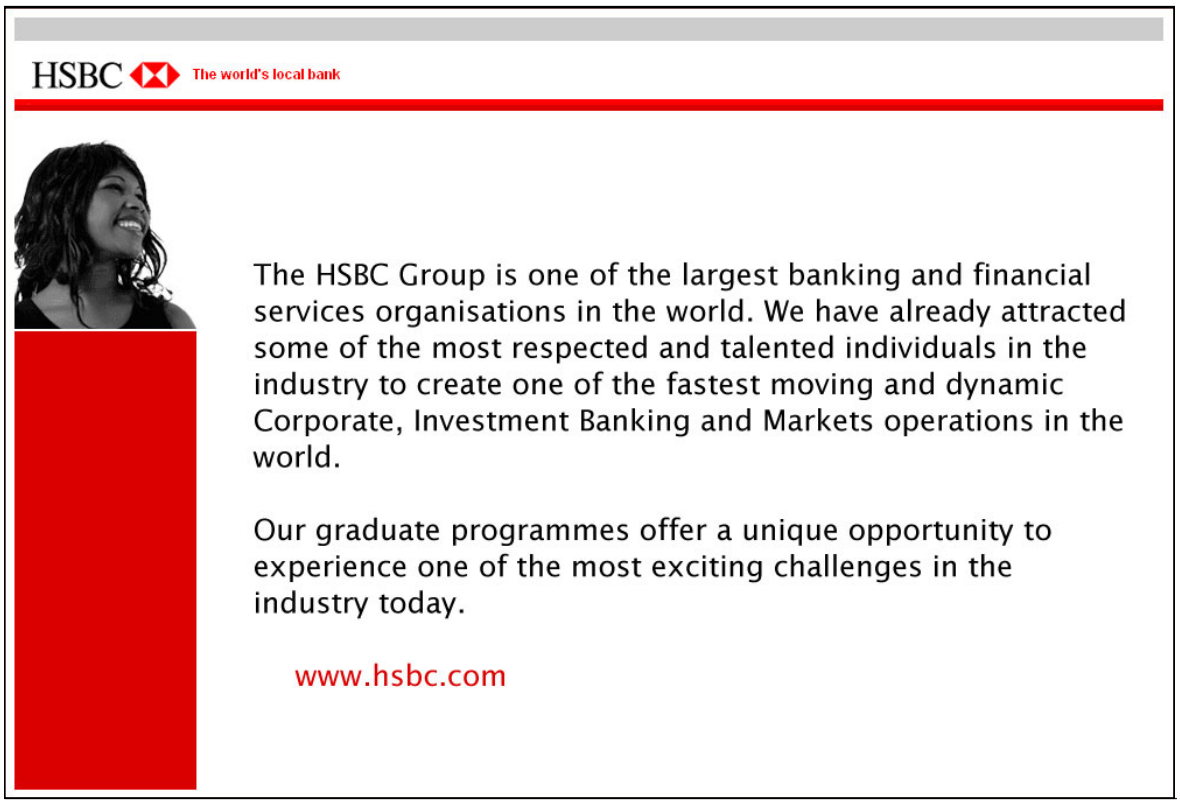


k) I intervallet som börjar vid 2,583 - 2 · 1,441 = -0,299 och slutar vid 2,583 + 2 · 1,441 = 5,465 täcker in 11 av de 12 observationerna, d.v.s. teoremet stämmer.

Exempel 1.2

Antalet barn per hushåll undersöktes i en viss stad, och resultatet framgår av tabellen nedan.

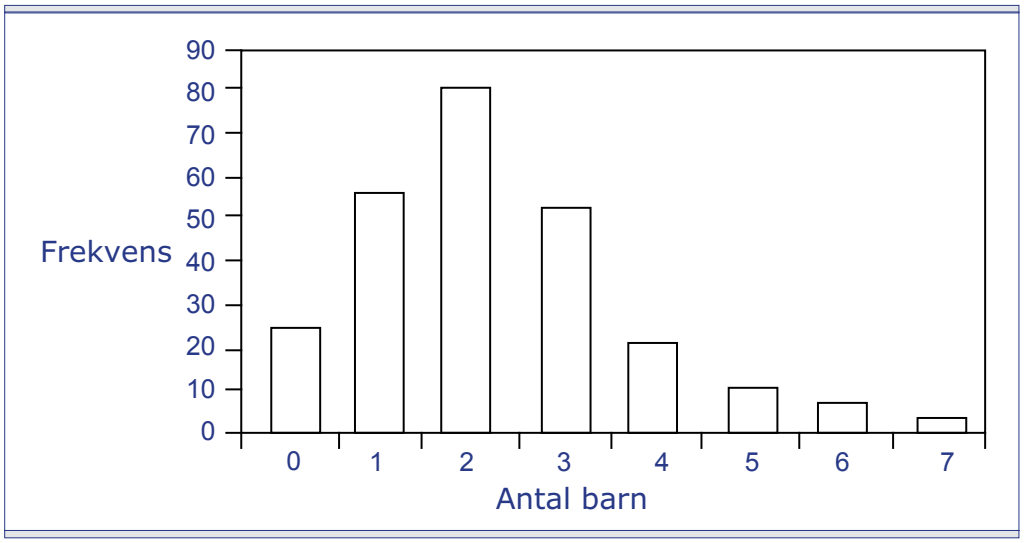
Antal barn	Frekvens
0	27
1	55
2	84
3	51
4	21
5	8
6	3
7	1



a) Illustrera datamängden med ett stolpdiagram.
b) Beräkna medelvärdet,
c) medianen och
d) standardavvikelsen för antalet barn per hushåll i staden.
e) Kontrollera om empiriska regeln stämmer.

Lösning

a)



b) $\mu = \frac{27 \cdot 0 + 55 \cdot 1 + 84 \cdot 2 + ... + 1 \cdot 7}{27 + 55 + 84 + ... + 1} = 2,1$

c) Median: Sorterar vi de enskilda observationerna i ordning kommer nr 1 – 27 att vara 0:or, nr 28 – 82 att vara 1:or, nr 83 – 166 att vara 2:or o.s.v. Medianen måste alltså vara en 2:a.

d) $\sigma = \sqrt{\frac{27(0-2,1)^2 + 55(1-2,1)^2 + 84(2-2,1)^2 + (7-2,1)^2}{27 + 55 + 84 + ... + 1}} = 1,327$

e) Enligt empiriska regeln ska ungefär 2 / 3 av observationerna ligga i intervallet som börjar vid 2,1 - 1,327 = 0,773 och slutar vid 2,1 + 1,327 = 3,427. Vi ser att (55 + 84 + 51) / 250 = 0,76 , så det stämmer hyfsat. Dessutom ska ungefär 19 / 20 av observationerna ligga i intervallet som börjar vid 2,1 - 2 · 1,327 = -0,554 och slutar vid 2,1 + 2 · 1,327 = 4,754. Vi kan konstatera att det stämmer bra, eftersom (27 + 55 + 84 + 51 + 21) / 250 = 0,952. Bortom tre standardavvikelser från medelvärdet hamnar endast en familj – den med 7 barn – så vi drar slutsatsen att empiriska regeln stämmer.

2. Sannolikhetslära

2.1 Inledning

En *sannolikhet* är ett mått på hur troligt det är att ett skeende som kan resultera i flera olika möjliga utfall leder fram till ett visst utfall. Sannolikheten för att en händelse *A* inträffar skrivs *P(A)*. En sannolikhet uttrycks matematiskt som ett numeriskt värde mellan 0 och 1, eller mellan 0% och 100%. Om sannolikheten för ett visst utfall är 75% så betyder det att detta utfall inträffar 3 gånger av 4 i det långa loppet. Att sannolikheten för händelsen *A* är 0,4 skrivs alltså *P(A)* = 0,4.

En beskrivning av de olika utfall som är möjliga för en viss slumpvariabel kallas *utfallsrum*. När man till exempel kastar två tärningar består utfallsrummet av 36 olika utfall, eftersom varje tärning har 6 olika sidor. En *händelse* utgörs av en *delmängd* av ett utfallsrum. Om man kastar två tärningar och bara är intresserad av det totala antalet prickar så motsvarar de 36 utfallen med tärningarna endast 11 olika händelser eftersom många av utfallen innebär samma sammanlagda antal prickar. Om tärning 1 visar en 5:a och tärning 2 visar en 2:a så är det ett annat utfall än – men ändå samma händelse som – om tärning 1 hade visat en 3:a och tärning 2 en 4:a. I båda fallen är ju sammanlagda antalet prickar 7.

Komplementet till en händelse innebär att händelsen inte inträffar. Händelsen *A* har komplementet $\overline{A}$. Enligt *lagen om total sannolikhet* är sannolikheten alltid 1 att en händelse *eller* dess komplement inträffar. För varje händelse *A* gäller alltså att $P(A) + P(\overline{A}) = 1$. Detta till synes självklara konstaterande är mycket användbart, som vi kommer att se senare.

2.2 Union och snitt

Unionen mellan ett antal händelser innebär ett utfall där minst en av händelserna inträffar. Vi betecknar unionen mellan händelserna *A* och *B* som $A \cup B$.

Snittet mellan ett antal händelser innebär ett utfall där samtliga händelser inträffar. Vi betecknar snittet mellan händelserna *A* och *B* som $A \cap B$. När händelser inte har något snitt säger man att de är *ömsesidigt uteslutande*, eller *disjunkta*. För unionen och snittet mellan två händelser gäller sambandet

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

eller, om man så vill,

$$P(A \cap B) = P(A) + P(B) - P(A \cup B)$$

Dessutom gäller enligt lagen om total sannolikhet att

$$P(A \cup B) = 1 - P(\overline{A} \cap \overline{B})$$

2.3 Oberoende händelser

Om två sannolikheter är *oberoende* så har den ena händelsens eventuella inträffande ingen påverkan på sannolikheten för huruvida den andra händelsen faktiskt inträffar. Två händelser A och B är oberoende om

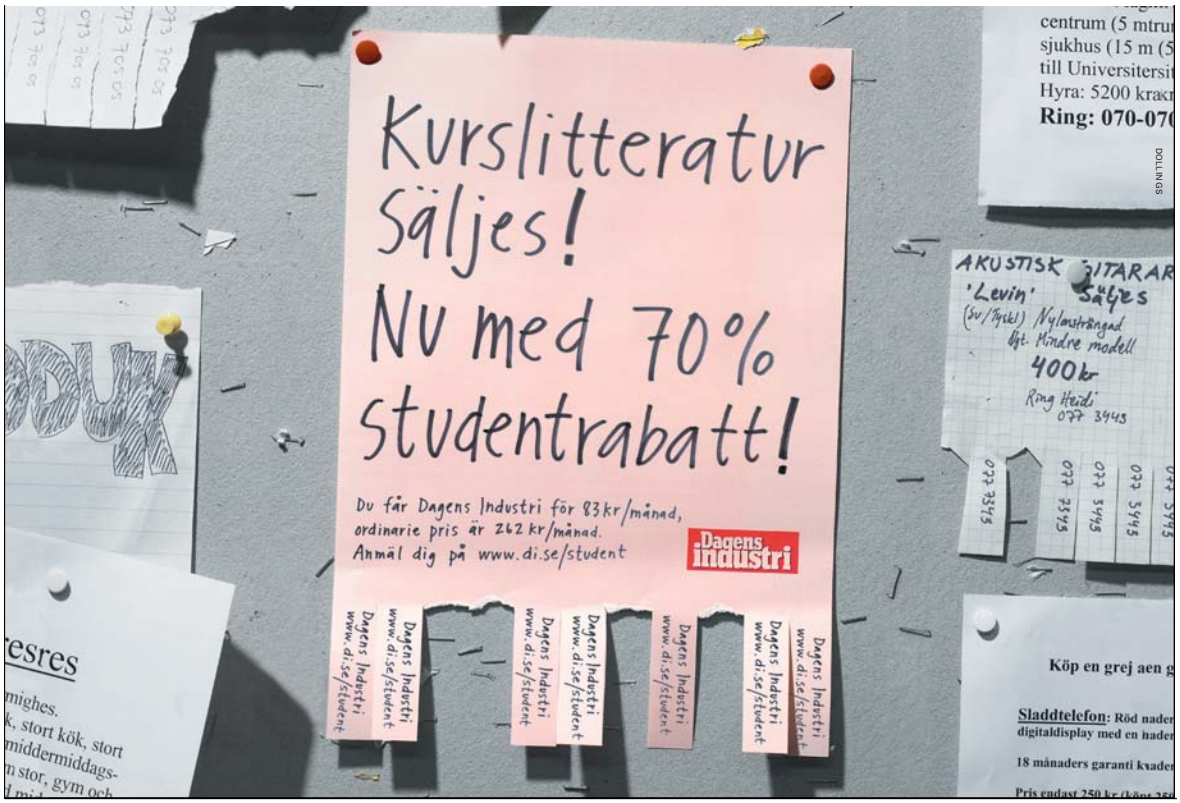
$$P(A \cap B) = P(A) \cdot P(B)$$

Exempel 2.1

Sannolikheten för att Yngve ska klara en viss tenta i statistik (händelse A) antas vara 80%. Sannolikheten för att han äter gröt till frukost när han ska ha tenta (händelse B) är 30%. Sannolikheten för att han antingen äter gröt eller klarar tentan (eller båda) är 86%. Råder det beroende mellan händelserna A och B?

Lösning

Nej, de är oberoende eftersom $0{,}8 + 0{,}3 - 0{,}86 = 0{,}8 \cdot 0{,}3 = 0{,}24$.



2.4 Betingade sannolikheter

En *betingad sannolikhet* är sannolikheten för att en viss händelse inträffar under förutsättning att en viss annan händelse inträffar. Den betingade sannolikheten för att händelse A inträffar givet att händelse B inträffar skrivs $P(A|B)$, och det gäller bland annat att

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

Två händelser A och B är oberoende om

$$P(A \mid B) = P(A)$$

Exempel 2.2

Visa att sannolikheten för att Yngve ska klara sin tenta i exempel 2.1 inte är beroende av huruvida han äter gröt till frukost med hjälp av oberoendesambandet för betingade sannolikheter.

Lösning

Det råder oberoende, eftersom $P(A \mid B) = \frac{0,24}{0,3} = 0,8 = P(A)$.

2.5 Bayes teorem

Med hjälp av de samband vi har definierat hittills kan ett antal andra användbara samband härledas, t.ex.

- $P(A) = P(A \cap B) + P(A \cap \overline{B})$ (enligt lagen om total sannolikhet)
- $P(A) = P(B) \cdot P(A \mid B) + P(\overline{B}) \cdot P(A \mid \overline{B})$ (enligt lagen om total sannolikhet)
- $P(B \mid A) = \frac{P(A \mid B) \cdot P(B)}{P(A \mid B) \cdot P(B) + P(A \mid \overline{B}) \cdot P(\overline{B})}$ (”Bayes teorem”).

Exempel 2.3

Anta att sannolikheten att Ericssons B-aktie har gått upp på börsen under en viss dag är 0,55. Anta vidare att sannolikheten att Yngves aktierådgivning har gett ett köptips på aktien samma dag som aktien gått upp är 0,4. Vad är sannolikheten att aktien gått upp och att Yngves aktierådgiving under samma dag inte gett aktien ett köptips?

Lösning

Eftersom $P(A) = P(A \cap B) + P(A \cap \overline{B})$ så gäller att $P(A \cap \overline{B}) = P(A) - P(A \cap B)$. Låt A vara händelsen ”aktien går upp” och B vara händelsen ”köptips gavs”. Vi har då sannolikheten $P(A \cap \overline{B}) = P(A) - P(A \cap B) = 0,55 - 0,4 = 0,15$.

Exempel 2.4

Yngve räknar med att Ericssons B-aktie har sannolikheten 0,8 att gå upp under nästa månad om man presenterar en mobiltelefon baserad på ny teknik. Om man inte presenterar någon sådan telefon tror Yngve att sannolikheten att aktien går upp är 0,4. Enligt Yngves säkra källor är sannolikheten att en tekniskt sett ny telefon presenteras under kommande år 0,25. Vad är sannolikheten att aktien går upp?

Lösning

Låt A vara händelsen ”aktien går upp” och B vara händelsen ”telefon presenteras”. Vi har då $P(A) = P(B) \cdot P(A \mid B) + P(\overline{B}) \cdot P(A \mid \overline{B}) = 0,25 \cdot 0,8 + (1 - 0,25) \cdot 0,4 = 0,5$.

Exempel 2.5

Att en aktie går upp kraftigt på börsen är sällsynt – sannolikheten för det är bara 0,3%. Yngve har därför utvecklat en prognosmodell för aktievärdering för att kunna hitta de kraftiga uppgångarna på börsen. När en aktie kommer att gå upp kraftigt förutspår Yngves prognosmodell det helt korrekt med 90% sannolikhet. Modellen är dock inte perfekt - när den tillämpas på en aktie som inte kommer att gå upp kraftigt så visar den ändå med 5% sannolikhet att aktien kommer att gå upp kraftigt. Hur sannolikt är det att en aktie går upp kraftigt om Yngves prognosmodell indikerar det?

Lösning

Låt B vara händelsen ”aktien går upp” och A vara händelsen ”modellen indikerar uppgång”. Vi har då sannolikheten

$$P(B \mid A) = \frac{P(A \mid B) \cdot P(B)}{P(A \mid B) \cdot P(B) + P(A \mid \overline{B}) \cdot P(\overline{B})} =$$

$$= \frac{0,9 \cdot 0,003}{0,9 \cdot 0,003 + 0,05 \cdot (1 - 0,003)} = 0,051$$

vilket innebär att Yngves modell faktiskt har fel ungefär 19 gånger av 20 när den indikerar kraftig uppgång.

2.6 Permutationer

Ett sätt på vilket ett visst antal element kan väljas ut i en viss ordning från en viss mängd element kallas för en *permutation*. Antalet olika permutationer som kan åstadkommas när r element ska väljas från en mängd bestående av n element betecknas $n\mathrm{P}r$ och beräknas med formeln

$$n\mathrm{P}r = \frac{n!}{(n-r)!}$$

där $n! = 1 \cdot 2 \cdot 3 \cdot \ldots \cdot n$.

Exempel 2.6

Per ska lackera sina tre bilar, en Opel, en Volvo och en Saab, i tre olika färger. Han kan välja mellan svart, blå, röd, vit, gul, grön och silvermetallic. Hur sannolikt är det att Opeln blir röd, Volvon svart och Saaben grön om Per väljer färger helt slumpmässigt?

Lösning

Eftersom det spelar roll vilken bil som får vilken färg rör det sig om permutationer. Per kan lackera sina bilar på

$$\frac{7!}{(7-3)!} = 210$$

olika sätt. Eftersom alla utfallen är lika sannolika är den sökta sannolikheten 1 / 210 = 0,00476.

2.7 Kombinationer

Om ordningen de utvalda elementen emellan saknar betydelse så kallas den utvalda mängden för en *kombination*. Antalet kombinationer som kan åstadkommas när *r* element ska väljas från en mängd bestående av *n* element betecknas *n*C*r* eller $\binom{n}{r}$ (vilket utläses ”n över r”) och beräknas med formeln

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

Exempel 2.7

I spelet Lotto dras 7 nummer av 35. Hur stor är sannolikheten att få 7 rätt på Lotto om man spelar en enkelrad bestående av 7 nummer?

Lösning

Eftersom det inte spelar någon roll i vilken ordning numren dras rör det sig om kombinationer. Det finns

$$\binom{35}{7} = \frac{35!}{7!(35-7)!} = 6724520$$

sådana kombinationer. Sannolikheten att en spelad enkelrad på Lotto ger 7 rätt är alltså 1 / 6724520 = 0,000000148.

3. Diskreta fördelningar

3.1 Inledning

Som framgick av kapitel 1 så kallar vi en variabel som får sitt värde via en slumpmässig process för *slumpvariabel*. En slumpvariabel kan vara *diskret* eller *kontinuerlig*. En diskret slumpvariabel kan bara anta vissa värden – oftast heltalsvärden – medan en kontinuerlig slumpvariabel kan anta vilket värde som helst inom ett visst intervall. En kontinuerlig slumpvariabel har alltså, till skillnad från en diskret, ett oändligt antal möjliga utfall. I detta kapitel ska vi titta på diskreta slumpvariabler.

Fördelningen, eller *fördelningsfunktionen*, för en diskret slumpvariabel X visar sannolikheten att den får det specifika värdet x, d.v.s. $P(X = x)$, för varje möjligt värde på x. Den *kumulativa fördelningsfunktionen* för en diskret slumpvariabel X betecknas med $F(x)$ och visar sannolikheten att den får *högst* det specifika värdet x, d.v.s. $F(x) = P(X \leq x)$, för varje möjligt värde på x. Observera att det gäller att

$$F(x) = P(X \leq x) = \sum_{i=0}^{x} P(X = x)$$

när X bara kan anta icke-negativa heltalsvärden. Observera också att det enligt lagen om total sannolikhet gäller att

$$\sum_{Alla\,x} P(X = x) = 1$$

3.2 Väntevärde och varians för en diskret slumpvariabel

Det förväntade värdet, eller *väntevärdet*, för en diskret slumpvariabel X betecknas med $E(X)$ och utgörs allmänt av dess sanna vägda medelvärde där de enskilda utfallens sannolikheter används som vikter. Det gäller alltså att

$$\mu = E(X) = \sum_{Alla\,x} x \cdot P(X = x)$$

Variansen för en diskret slumpvariabel X, betecknad med $V(X)$, är dess genomsnittliga kvadrerade avvikelse från medelvärdet. Vi har alltså

$$\sigma^2 = V(X) = \sum_{Alla\,x} (x - \mu)^2 \cdot P(X = x)$$

En ekvivalent, men ibland beräkningsmässigt enklare formel för denna varians, är

$$V(X) = E(X^2) - (E(X))^2$$

Standardavvikelsen för en diskret slumpvariabel X, betecknad med $S(X)$, är roten ur dess varians, d.v.s.

$$\sigma = S(X) = \sqrt{V(X)}$$

Exempel 3.1

Specificera fördelningsfunktionen för antalet sexor som kommer upp när man kastar två tärningar? Vad är medelvärdet och variansen?

Lösning

Låt X vara antalet sexor som kommer upp. Vi har då $P(X=2) = \frac{1}{6}\cdot\frac{1}{6} = \frac{1}{36} = 0,0278$ och $P(X=0) = \frac{5}{6}\cdot\frac{5}{6} = \frac{25}{36} = 0,6944$. Enligt lagen om total sannolikhet är således $P(X=1) = 1-(0,6944+0,0278) = 0,2778$. Vi har alltså fördelningsfunktionen:

Antal sexor	Sannolikhet
0	0,6944
1	0,2778
2	0,0278

StudentSMS erbjuder dig gratis ett omfattande verktyg för att underlätta din studievardag!

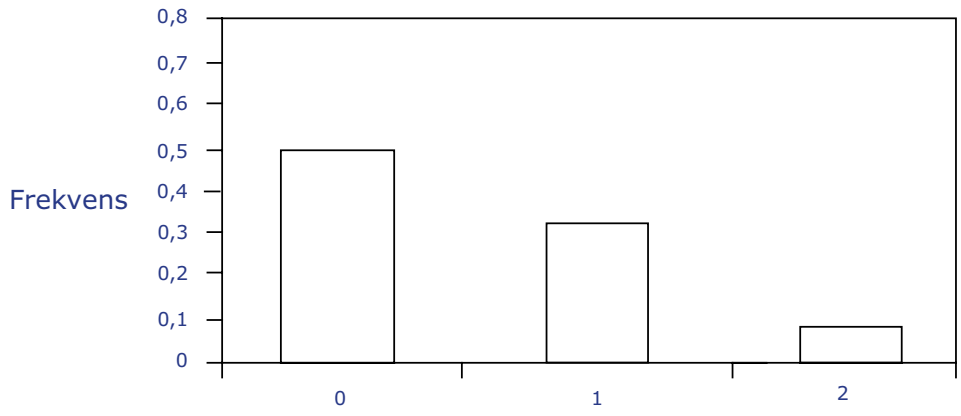
- Du får meddelande till mobilen från din lärare om plötsligt inställda lektioner, salbyten etc
- Effektivisera dina arbeten med dina kursare via vårt projektverktyg
- Få ordning och struktur på ditt studentliv med kalendern som också ger info om roliga events till mobilen
- StudentSMS hjälper dig att hitta bostad, extra jobb och exjobb

Allt detta och mycket, mycket mer får du gratis!
Registrera dig på: www.studentsms.se

Student SMS

Fördelningsfunktionen kan beskrivas med ett stolpdiagram som i figur 3.1. Notera särskilt att summan av stolparnas höjder – enligt lagen om total sannolikhet – blir 1.

Figur 3.1
Stolpdiagram med sannolikhetsfördelning



Medelvärdet för antalet sexor som kommer upp blir

$$E(X) = 0 \cdot 0,6944 + 1 \cdot 0,2778 + 2 \cdot 0,0278 = 0,3333$$

med variansen

$$V(X) = (0 - 0,3333)^2 \cdot 0,6944 + (1 - 0,3333)^2 \cdot 0,2778 + (2 - 0,3333)^2 \cdot 0,0278 = 0,2778$$

3.3 Binomialfördelningen

I exempel 3.1 kunde vi faktiskt ha beräknat väntevärdet och variansen för slumpvariabeln X på ett enklare sätt, eftersom X där uppfyllde vissa villkor. När ett skeende utfaller antingen ”positivt” eller ”negativt” med känd och konstant sannolikhet, och X är en slumpvariabel som räknar antalet positiva utfall i ett antal på varandra oberoende upprepningar av skeendet, då kommer X att följa *binomialfördelningen*. Sannolikheten är alltid 1 / 6 för att en tärning ska visa en sexa, och vi lät i exempel 3.1 X symbolisera antalet sexor efter två av varandra oberoende tärningskast. X var alltså en binomialfördelad slumpvariabel.

Vi betecknar med $X \sim B(n, p)$ att X är en binomialfördelad slumpvariabel som räknar antalet positiva utfall efter n oberoende försök där sannolikheten för positivt utfall i varje enskilt försök är p. I exempel 3.1 hade vi alltså $X \sim B(2, 0,1667)$. För $X \sim B(n, p)$ gäller medelvärdet

$$E(X) = n \cdot p$$

och variansen

$$V(X) = n \cdot p(1 - p)$$

Vidare kan $P(X = x)$ när $X \sim B(n, p)$ beräknas som

$$P(X = x) = \binom{n}{x} p^x (1-p)^{(n-x)}$$

Exempel 3.2

Verifiera $P(X = 1)$ samt $E(X)$ och $V(X)$ i exempel 3.1 med hjälp av binomialfördelningsformlerna.

Lösning

$$P(X = 1) = \frac{2!}{1!(2-1)!} \cdot 0{,}1667^1 (1-0{,}1667)^{(2-1)} = 0{,}2778$$

$$E(X) = 2 \cdot 0{,}1667 = 0{,}3333$$

$$V(X) = 2 \cdot 0{,}1667(1-0{,}1667) = 0{,}2778$$

3.4 Poissonfördelningen

Om $X \sim B(n, p)$ när p är mycket litet och n är mycket stort så kan talen blir så stora att miniräknare, eller t.o.m. datorer, inte klarar av dem. Anta t.ex. att vi har n = 100 000 och p = 0,00004. Fakulteten 100 000! blir ett alltför stort tal för att kunna beräknas med normala hjälpmedel, så vi kan inte använda den vanliga binomialfördelningsformeln för att beräkna $P(X = x)$. Däremot kan vi lätt beräkna det förväntade värdet av den binomialfördelade variabeln som $E(X) = n \cdot p = 100000 \cdot 0{,}00004 = 4$.

Det finns emellertid en lösning på problemet. När $X \sim B(n, p)$, n är stor och p är liten, men $E(X) = n \cdot p$ är ”lagom stor”, kommer slumpvariabeln X att approximativt följa den s.k. *poissonfördelningen*. Tumregelmässigt, om $n > 10$, $p < 0{,}1$ och $0{,}01 < E(X) < 50$, så kan binomialfördelningsformeln approximeras med

$$P(X = x) = \binom{n}{x} p^x (1-p)^{(n-x)} \approx \frac{e^{-\mu} \cdot \mu^x}{x!}$$

där $\mu = E(X) = n \cdot p$ och e = 2,718281828… (se miniräknaren). Generellt gäller att en slumpvariabel X kommer att följa poissonfördelningen om X symboliserar det antal gånger ett visst fenomen uppträder under ett avgränsat intervall i tiden eller rummet, och om sannolikheten för att fenomenet ska uppträda är densamma för alla intervaller av samma storlek och oberoende mellan fenomenets förekomster i olika intervaller råder. Variansen för en poissonfördelad variabel kommer att vara lika med väntevärdet, vilket enklast förklaras av att när p är mycket litet så är $(1-p)$ mycket nära 1, varvid $V(X) = n \cdot p(1-p)$ reduceras till $V(X) = n \cdot p = \mu$. En poissonfördelad slumpvariabel har alltså bara en parameter, μ. Att X är en poissonfördelad slumpvariabel med parametern μ betecknas $X \sim Po(\mu)$.

Exempel 3.3

En viss process i en dator som körs en gång per sekund dygnet runt åstadkommer att datorn havererar med sannolikheten 1 på en miljon. Hur sannolikt är det att datorn havererar exakt en gång per månad (30 dagar)?

Lösning

På en månad körs processen $n = 60 \cdot 60 \cdot 24 \cdot 30 = 2592000$ gånger. Varje gång är sannolikheten för haveri $p = 0{,}000001$. Det genomsnittliga antalet haverier per månad är alltså $\mu = 2592000 \cdot 0{,}000001 = 2{,}592$. Om vi låter X beteckna antalet haverier per månad så vet vi att $X \sim Po(2{,}592)$. Vi kan nu beräkna den sökta sannolikheten

$$P(X = 1) = \frac{e^{-2,592} \cdot 2,592^1}{1!} = 0,194$$

3.5 Hypergeometriska fördelningen

Binomialfördelningen förutsatte att samma sannolikhet p gällde för positivt utfall i vart och ett av de n skeendena. När n element tas från en mängd som totalt innehåller N element, och där man på förhand vet att R element är ”positiva” och således att $N-R$ är ”negativa”, så kommer sannolikheten för positivt utfall för det enskilda elementet att variera beroende på vilka element som redan tagits. En slumpvariabel X som räknar antalet positiva utfall kommer då inte att följa binomialfördelningen. Istället kommer X att följa den *hypergeometriska fördelningen*, betecknat med $X \sim HG(n, R, N)$. För denna variabel gäller väntevärdet

$$E(X) = n \cdot p$$

och variansen

$$V(X) = n \cdot p(1-p)(N-n)/(N-1)$$

där $p = R / N$. För $X \sim HG(n, R, N)$ beräknas $P(X = x)$ med hjälp av formeln

$$P(X = x) = \frac{\binom{R}{x} \cdot \binom{N-R}{n-x}}{\binom{N}{n}}$$

Exempel 3.4

Vad är medelvärdet och variansen för antalet ess på given i vanlig mörkpoker? Hur stor är sannolikheten att du får triss i ess på given?

3.6 Geometriska fördelningen

Om vi har ett skeende med den konstanta sannolikheten p för positivt utfall, och slumpvariabeln X istället räknar antalet gånger skeendet måste upprepas för att man ska få ett positivt utfall, då kommer X att följa den *geometriska fördelningen*. Detta betecknas med $X \sim G(p)$. För $X \sim G(p)$ gäller väntevärdet

$$E(X) = 1 / p$$

och variansen

$$V(X) = (1 - p) / p^2$$

För $X \sim G(p)$ beräknas $P(X = x)$ med hjälp av formeln

$$P(X = x) = p(1 - p)^{x-1}$$

Exempel 3.5

Yngve går en kurs i statistik. Han vet att han har 75% chans att klara tentan på kursen. Vad är väntevärdet och variansen för det antal gånger Yngve kommer att behöva skriva tentan. Hur sannolikt är det att han kommer att behöva skriva tentan exakt tre gånger?

Lösning

Låt X vara det antal gånger Yngve behöver skriva tentan tills han klarar den. Vi har då $X \sim G(0{,}75)$ och kan beräkna

$$E(X) = 1 / 0{,}75 = 1{,}3333$$

$$V(X) = (1 - 0{,}75) / 0{,}75^2 = 0{,}4444$$

$$P(X = 3) = 0{,}75(1 - 0{,}75)^{3-1} = 0{,}0469$$

3.7 Negativa binomialfördelningen

I geometriska fördelningen räknas antalet gånger ett skeende med den konstanta sannolikheten p för positivt utfall måste upprepas till dess att ett positivt utfall erhålls. Ibland är man emellertid intresserad av att fortsätta tills man har erhållit två eller fler positiva utfall. Om slumpvariabeln X istället räknar antalet gånger skeendet måste upprepas för att man ska få s positiva utfall, då kommer X att följa den *negativa binomialfördelningen*. Detta betecknas med $X \sim NB(s, p)$. För $X \sim NB(s, p)$ gäller väntevärdet

$$E(X) = s / p$$

och variansen

$$V(X) = s(1-p) / p^2$$

För $X \sim NB(s, p)$ beräknas $P(X = x)$ med hjälp av formeln

$$P(X = x) = p^s \cdot (1-p)^{x-s} \cdot \binom{x-1}{s-1}$$

Exempel 3.6

Sannolikheten för att Yngve ska göra mål när han skjuter en straffspark är 0,6. Hur många straffsparkar måste han i genomsnitt skjuta för att han ska komma upp i 3 mål? Vad är variansen? Hur sannolikt är det att han måste skjuta exakt 7 straffar för att göra 3 mål?

Lösning

Låt X vara det antal straffar Yngve behöver skjuta tills han gjort 3 mål. Vi har då $X \sim NB(3, 0{,}6)$ och kan beräkna

$$E(X) = 3 / 0,6 = 5$$

$$V(X) = 3(1-0,6) / 0,6^2 = 3,3333$$

$$P(X = 7) = 0,6^3 \cdot (1-0,6)^{7-3} \cdot \left(\frac{6!}{2!(6-2)!} \right) = 0,0829$$

3.8 Additions- och multiplikationsformler

Om c är en konstant och X och Y är slumpvariabler så gäller sambanden

$$E(X + Y) = E(X) + E(Y)$$

$$E(c \cdot X) = c \cdot E(X)$$

$$V(c \cdot X) = c^2 \cdot V(X)$$

$$V(c + X) = V(X)$$

Om X och Y dessutom är oberoende så gäller sambanden

$$E(X \cdot Y) = E(X) \cdot E(Y)$$

$$V(X + Y) = V(X) + V(Y)$$

Det sista sambandet är mycket användbart. I ord säger det alltså att *variansen för summan av två slumpvariabler är lika med summan av slumpvariablernas varianser.* Man säger därvid att varianser är *additiva*.

Exempel 3.7

Yngve och hans bror Steve arbetar som bilförsäljare på olika bilfirmor. Medelvärdet för Yngves försäljning är 1,2 bilar per dag med standardavvikelsen 0,3 bilar. Steve säljer 1,6 bilar med standardavvikelsen 0,8 bilar. Vad är medelvärdet och standardavvikelsen för brödernas sammanlagda dagliga bilförsäljning?

Lösning

Låt X vara Yngves och Y Steves dagliga bilförsäljning. För deras sammanlagda dagliga prestation har vi då medelvärdet

$$E(X+Y) = E(X) + E(Y) = 1,2 + 1,6 = 2,8$$

och variansen

$$V(X+Y) = V(X) + V(Y) = 0,3^2 + 0,8^2 = 0,7300$$

och således standardavvikelsen $\sqrt{0,7300} = 0,8544$. Observera att standardavvikelser *inte* är additiva, d.v.s. $0,3 + 0,8 \neq 0,8544$. Man måste alltid gå vägen över varianserna för att kunna utnyttja additiviteten.

SCANIA

Jobba hos Scania

Scania kan erbjuda unika möjligheter för dig som är ung och studerar – både under studietiden och efter examen. Missa inte att söka Sommarjobb, Verkstadspraktik för teknologer, Trainee eller Utvecklingsingenjörsprogrammet.

Vi spänner över ett brett register, från sommarskolan för grundskolestuderande till industriforskarprogrammet, som bygger på samarbete mellan Scania och Sveriges tekniska högskolor. Däremellan finns Industrigymnasiet, sommarjobb, examensjobb och Scanias traineeprogram, för att nämna några av alla de aktiviteter Scania erbjuder studenter.

www.scania.se

4. Kontinuerliga fördelningar

4.1 Inledning

I det här kapitlet ska vi studera kontinuerliga slumpvariabler. En kontinuerlig variabel får sitt värde som resultat av en mätning som i princip kan göras med hur stor noggrannhet som helst. Därför går det inte att specificera ett utfallsrum i termer av ett antal specifika utfall på samma sätt som för en diskret variabel. Utfallsrummet för en kontinuerlig variabel måste istället uttryckas som det intervall inom vilket variabelns värde kan hamna.

Fördelningen för en kontinuerlig slumpvariabel X åskådliggörs ofta med dess *sannolikhetskurva* $f(x)$, som ibland även kallas *täthetsfunktion*. För $f(x)$ gäller följande:

1. $\int_{-\infty}^{\infty} f(x)dx = 1$
2. $P(X < a) = \int_{-\infty}^{a} f(x)dx$
3. $P(a < X < b) = \int_{a}^{b} f(x)dx$.

Uttryckt i ord så innebär det följande:

1. Den totala ytan under en sannolikhetskurva är alltid lika med 1 eftersom kurvan är den kontinuerliga motsvarigheten till det diskreta stolpdiagrammet med relativa frekvenser där stolparnas höjder kan summeras till 1. Tankemässigt kan man således tänka sig en sannolikhetskurva som ett stolpdiagram med ett oändligt antal stolpar där varje stolpe är oändligt smal.

2. Sannolikheten att en kontinuerlig slumpvariabel X får ett värde lägre än ett specifikt värde a motsvaras av ytan under sannolikhetskurvan från lägsta möjliga värde på x upp till och med a.

3. Sannolikheten att en kontinuerlig slumpvariabel X får ett värde högre än a men lägre än ett annat specifikt värde b motsvaras av ytan under sannolikhetskurvan mellan a och b.

4.2 Exponentialfördelningen

I föregående kapitel såg vi hur den geometriska fördelningen kunde användas för att räkna på vilket genomsnittligt antal gånger ett visst skeende måste upprepas för att skeendet ska resultera i ett positivt utfall när sannolikheten för positivt utfall i varje enskilt skeende p är känd. Den kontinuerliga motsvarigheten till detta problem är att räkna på den förväntade återstående tiden till en viss händelse som inträffar slumpmässigt i tiden där händelsens genomsnittliga intensitet är känd.

Om vi låter slumpvariabeln X symbolisera den återstående tiden mätt i någon tidsenhet till en händelse som i medeltal inträffar med medelvärdet μ gånger per tidsenhet så kommer X att följa *exponentialfördelningen*, vilket vi betecknar med $X \sim EXP(\mu)$.

För $X \sim EXP(\mu)$ så gäller medelvärdet

$$E(X) = 1/\mu$$

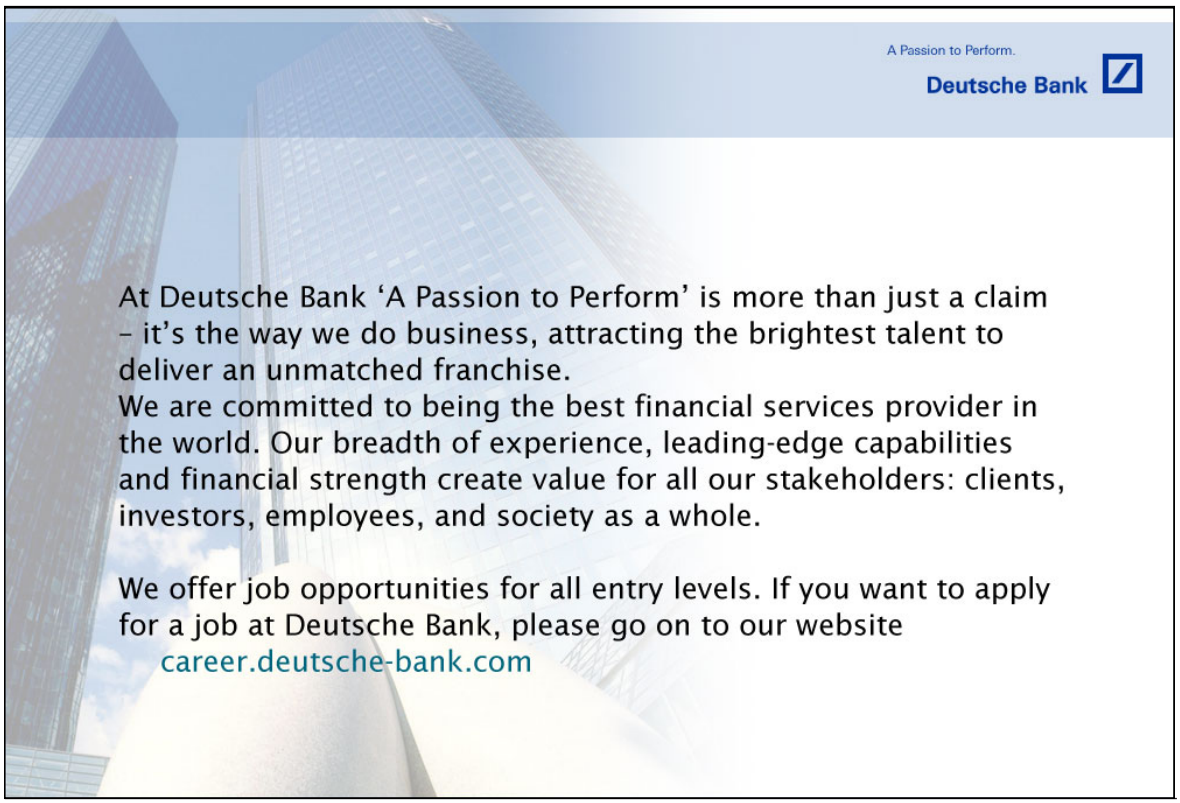
och variansen

$$V(X) = 1/\mu^2$$

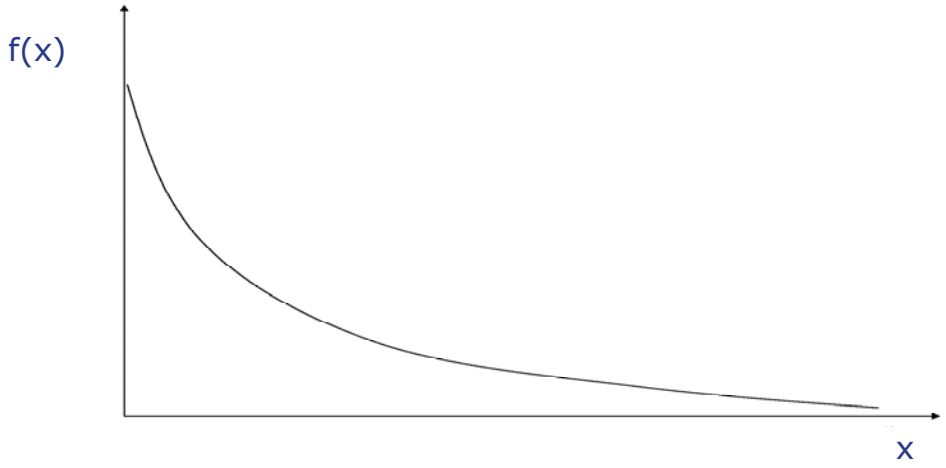
Vidare kan $P(X > x)$ när $X \sim \text{EXP}(\mu)$ beräknas som

$$P(X > x) = e^{(-x \cdot \mu)}$$

Exponentialfördelningens sannolikhetskurva framgår av figur 4.1.



Figur 4.1
Exponentialfördelningens sannolikhetskurva



Exempel 4.1

Anta att en viss maskin har en så hög driftssäkerhet att den endast måste repareras i genomsnitt 0,5 gånger per år. Hur sannolikt är det då att den återstående tiden till reparation vid ett visst tillfälle kommer att vara längre än 1,5 år?

Lösning

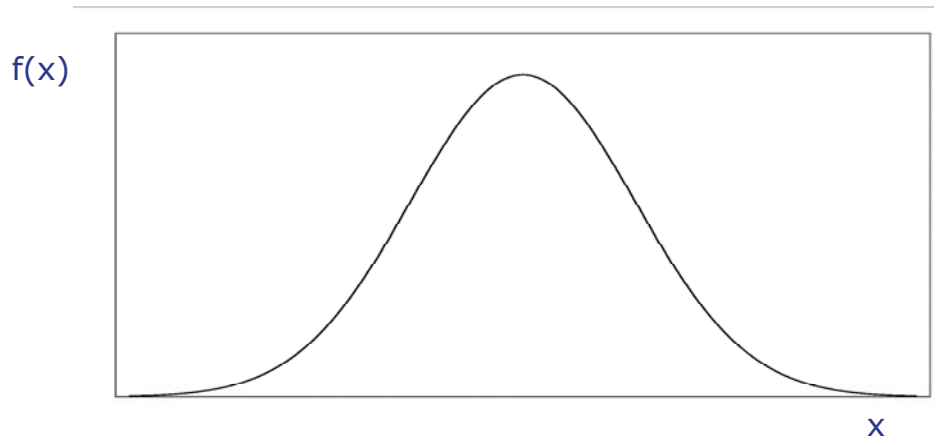
Vi låter slumpvariabeln X symbolisera tiden mellan varje reparation. Vi har då $E(X) = 1/0,5 = 2$, och således $X \sim EXP(0,5)$. Vi kan nu beräkna den sökta sannolikheten som

$$P(X > 1,5) = e^{-1,5 \cdot 0,5} = 0,4723$$

4.3 Normalfördelningen

Om en slumpvariabel X definieras som summan av n slumpvariabler med en viss godtycklig fördelning samt medelvärdet μ och standardavvikelsen σ så kommer X att följa *normalfördelningen* med medelvärdet $\mu \cdot n$ och standardavvikelsen $\sigma \cdot \sqrt{n}$ om det är så att n är stort. Detta faktum kallas *centrala gränsvärdessatsen* (CGS) och är med stor säkerhet den viktigaste statistiska lag du någonsin kommer att stöta på. Oavsett vilken slags fördelning de enskilda slumpvariablerna i summan kommer från så kommer deras summa alltså alltid att vara en slumpvariabel som är fördelad på samma sätt om antalet enskilda slumpvariabler är stort. Normalfördelningen är kontinuerlig och sannolikhetskurvan för en normalfördelad slumpvariabel är ”symmetriskt klockformad”, vilket framgår av figur 4.2.

Figur 4.2
Normalfördelningens sannolikhetskurva



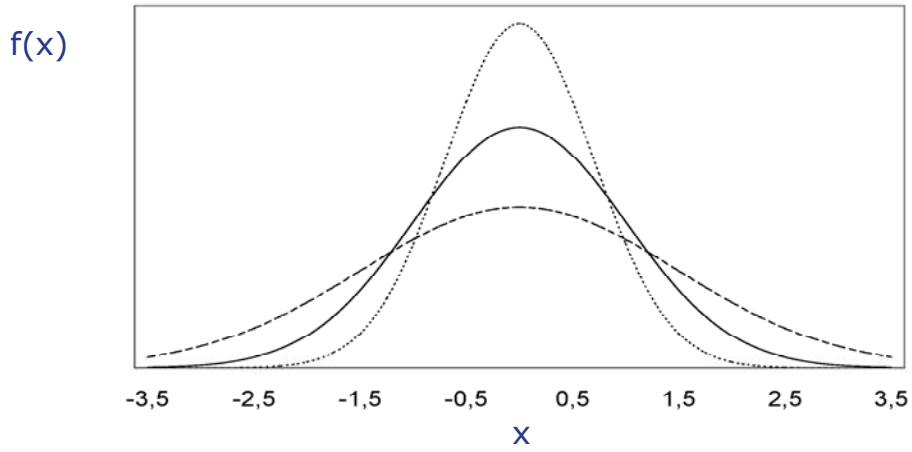
Vad menar vi då med att n ska vara stort? Jo, ju mer ”olik” de enskilda slumpvariablernas fördelning är en normalfördelning, desto större måste n vara för att deras summa ska vara en normalfördelad slumpvariabel. Om den ursprungliga fördelningen är relativt lik en normalfördelning räcker det normalt att $n \geq 5$. Om man inte känner till vilken den ursprungliga fördelningen är, eller har anledning att anta att den är mycket sned eller på annat sätt är olik en normalfördelning, så krävs att n är betydligt större. En vanlig tumregel är att summan av 30 eller fler slumpvariabler från en och samma fördelning alltid kan antas vara en normalfördelad slumpvariabel.

Att CGS är viktig kommer att framgå tydligt längre fram när vi tittar på hur man kan analysera stickprovsmedelvärden. Eftersom täljaren i ett stickprovsmedelvärde är just en variabelsumma där de enskilda variabelvärdena kommer från samma fördelning så kan ett stickprovsmedelvärde ofta antas vara just en normalfördelad slumpvariabel oavsett vilken typ av fördelning den population som stickprovet är hämtat från kännetecknas av. När man kan anta att en viss slumpvariabel är normalfördelad underlättas nämligen analysen av variabeln avsevärt, vilket vi kommer att se i detta avsnitt.

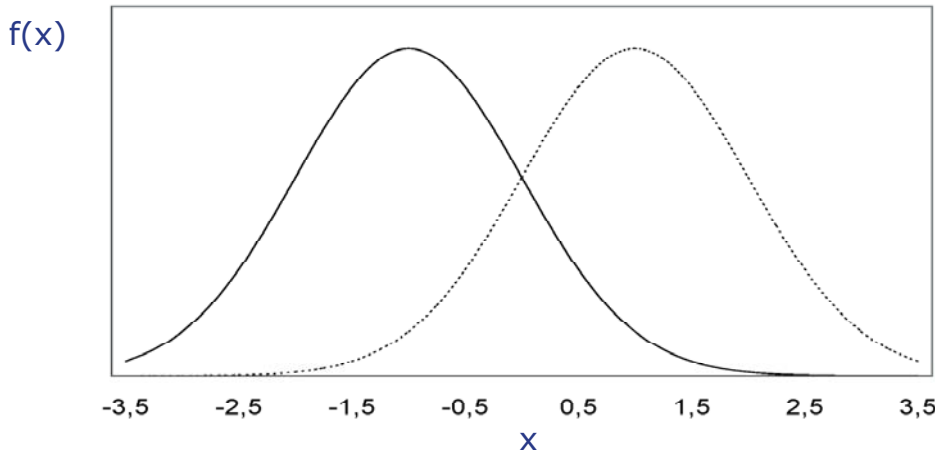
Det är dock inte alls bara teoretiska variabelsummor som är normalfördelade. Många olika slag av slumpvariabler ”ute i verkligheten” följer approximativt (ungefärligen) normalfördelning. Generellt gäller att om värdet på en slumpvariabel påverkas av många oberoende faktorer, där ingen enskild faktor har stor inverkan i förhållande till någon annan på slumpvariabelns värde, då kommer slumpvariabeln att vara approximativt normalfördelad.

Att X är en normalfördelad slumpvariabel med medelvärdet μ och standardavvikelsen σ betecknar vi med $X \sim N(\mu, \sigma)$. Teoretiskt sett finns alltså ett oändligt antal normalfördelningar – en för varje kombination av värden på μ och σ. Detta illustreras i figurerna 4.3 och 4.4. I figur 4.3 finns sannolikhetskurvorna för tre normalfördelningar med $\mu = 0$ men olika σ. I figur 4.4 finns sannolikhetskurvorna för två normalfördelningar med samma σ men med $\mu = -1$ respektive $\mu = 1$.

Figur 4.3
Tre normalfördelningar med samma medelvärde men med olika standardavvikelser.



Figur 4.4
Tre normalfördelningar med olika medelvärden men med samma standardavvikelser.



4.4 Standardnormalfördelningen

Den enskilt viktigaste normalfördelningen är *standardnormalfördelningen* som har medelvärdet 0 och standardavvikelsen 1. Den är standard såtillvida att den enkelt kan användas för att analysera varje annan normalfördelad slumpvariabel, vilket vi kommer att titta på i de två nästföljande avsnitten. Den standardnormalfördelade slumpvariabeln är så viktig att den har ett eget ”namn”, nämligen Z. Vi definierar alltså $Z \sim N(0, 1)$. Med hjälp av tabell A1 i appendix kan man enkelt utläsa ytan under standardnormalfördelningens sannolikhetskurva för $Z > z$ för $z > 0$. Denna yta är då samma sak som $P(Z > z)$.

Observera att det gäller att

$$P(Z < -a) = P(Z > a) = 1 - P(Z < a)$$

och att

$$P(a < Z < b) = 1 - P(Z < a) - P(Z > b)$$

för godtyckliga värden på a och b.

Exempel 4.2

Ta med hjälp av tabell A1 fram sannolikheterna $P(Z > 0,4)$, $P(Z > -1,63)$ och $P(0,38 < Z < 1,12)$.

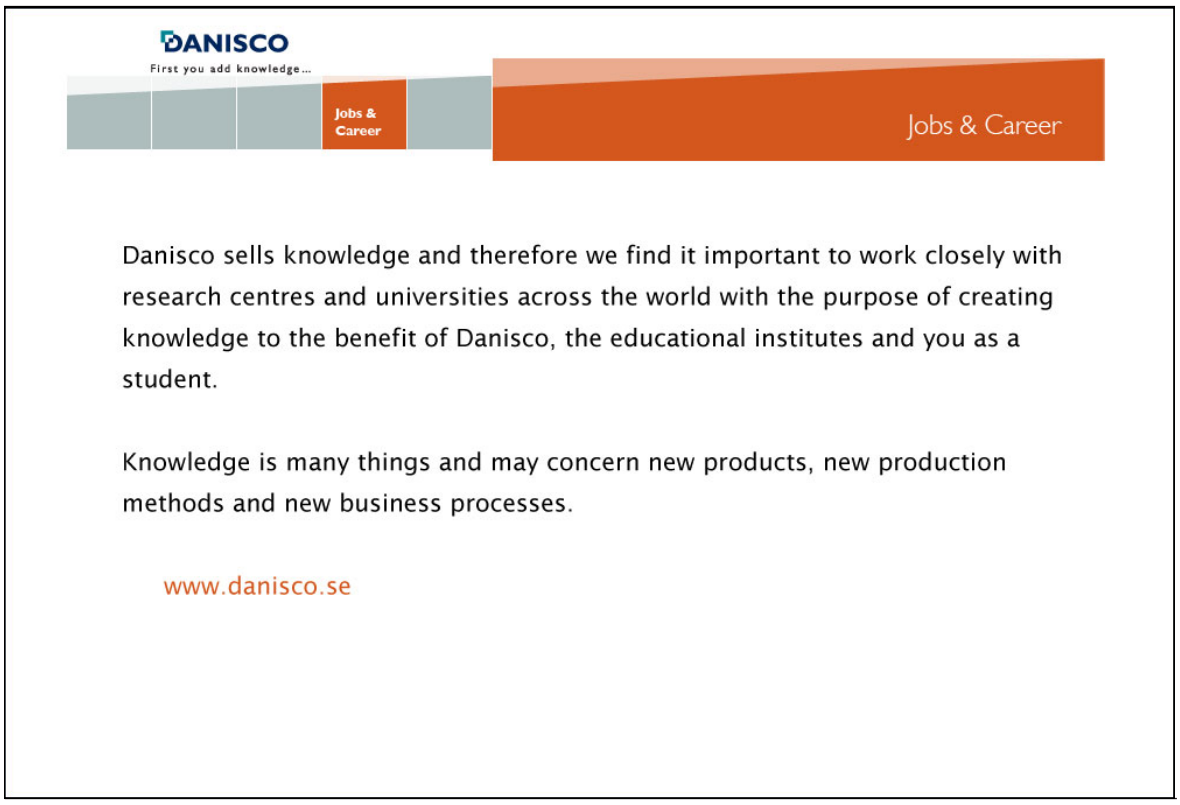
Lösning

$$P(Z > 0,4) = 0,3446$$

$$P(Z > -1,63) = 1 - P(Z > 1,63) = 1 - 0,0516 = 0,9484$$

$$P(0,38 < Z < 1,12) = 1 - P(Z < 0,38) - P(Z > 1,12) = 1 - (1 - P(Z > 0,38)) - P(Z > 1,12) =$$

$$= 1 - (1 - 0,3520) - 0,1314 = 0,2206$$



4.5 Transformering till standardnormalfördelning

Det finns ett oändligt antal normalfördelade slumpvariabler, men varje slumpvariabel $X \sim N(\mu, \sigma)$ kan *transformeras* till $Z \sim N(0, 1)$ vilket ger oss möjlighet att analysera X med hjälp av tabell A1. Denna transformering görs på följande sätt:

$$Z = \frac{X - \mu}{\sigma}$$

Exempel 4.3

Anta att $X \sim N(42, 5)$. Använd transformering till Z för att finna $P(X > 50)$.

Lösning

$$P(X > 50) = P\left(\frac{X - \mu}{\sigma} > \frac{50 - \mu}{\sigma}\right) = P\left(Z > \frac{50 - 42}{5}\right) = P(Z > 1,6) = 0,0548$$

4.6 Transformering från standardnormalfördelning

I föregående avsnitt transformerade vi en godtycklig normalfördelad slumpvariabel X till Z så att tabell A1 skulle kunna användas för att få fram en sannolikhet. Men transformering kan också göras åt andra hållet. Genom att arrangera om uttrycket för transformation från X till Z ovan så får vi nämligen

$$X = \mu + Z \cdot \sigma$$

Således kan vi transformera ett specifikt värde på z från tabell A1 till ett specifikt värde på x för varje $X \sim N(\mu, \sigma)$ med hjälp av sambandet

$$X = \mu + z \cdot \sigma$$

Exempel 4.4

Anta att $X \sim N(42, 5)$. Vilket värde på x innebär att $P(X > x) = 0,33$?

Lösning

Från tabell A1 utläser vi att $P(X > x) = 0,33$ gäller för $z = 0,44$. Vi får då

$$x = \mu + z \cdot \sigma = 42 + 0,44 \cdot 5 = 44,2$$

Alltså gäller det att $P(X > 44,2) = 0,33$.

4.7 Normalfördelningsapproximation av binomialfördelningen

Som vi såg i föregående kapitel om diskreta fördelningar så är binomialfördelningen svårhanterlig när n är stor. Om det samtidigt gäller att n är stor och p är liten så kunde vi approximera binomialfördelningen med en poissonfördelning. Om n är stor och p inte är mycket liten (eller mycket stor) så kan vi istället utnyttja det faktum att binomialfördelningen närmar sig normalfördelningen när n blir större. När n är stort kan vi därför använda approximationen

$$X \sim B(n, p) \approx X \sim N\left(n \cdot p, \sqrt{n \cdot p(1-p)}\right)$$

En vanlig tumregel är att denna approximation får användas om det samtidigt gäller att $n \cdot p \geq 5$ och $n(1-p) \geq 5$. Eftersom normalfördelningen är kontinuerlig och binomialfördelningen är diskret måste man dock göra en s.k. *kontinuitetskorrigering*, vilket innebär att man låter varje diskret värde motsvaras av ett kontinuerligt intervall ±0,5 kring det diskreta värdet.

Exempel 4.5

Yngve kastar en tärning 1200 gånger. Vad är sannolikheten att han får minst 220 sexor?

Lösning

Sannolikheten för en sexa i ett enskilt kast är $1 / 6 = 0{,}1667$. Om vi låter X beteckna antalet sexor så vet vi att $X \sim B(1200, 0{,}1667)$. Tumregelmässigt får vi approximera X med normalfördelning eftersom

$$n \cdot p = 1200 \cdot 0{,}1667 = 200 \geq 5$$

och

$$n(1-p) = 1200(1-0{,}1667) = 1000 \geq 5$$

Vi kan alltså använda approximationen

$$X \sim B(1200,\ 0{,}1667) \approx X \sim N(200,\ 12{,}91) \text{ eftersom } \sqrt{1200 \cdot 0{,}1667 \cdot 0{,}8333} = 12{,}91$$

$P(X \geq 220)$ måste dock kontinuitetskorrigeras till $P(X > 219{,}5)$ när normalfördelningen ska användas. Vi får då sannolikheten

$$P(X > 219{,}5) = P\left(Z > \frac{219{,}5 - 200}{12{,}91}\right) = P(Z > 1{,}51) = 0{,}0655$$

4.8 Normalfördelningsapproximation av poissonfördelningen

Även poissonfördelningen kan ofta approximeras med normalfördelningen, vilket gör att man kan undvika besvärliga beräkningar med poissonformeln. En tumregel för att få använda denna approximation är att poissonfördelningens väntevärde ska vara minst 10, men approximationen bli bättre ju större väntevärdet är. Approximationen blir då

$$X \sim Po(\mu) \approx X \sim N(\mu, \sqrt{\mu})$$

Kontinuitetskorrigering måste dock göras eftersom poissonfördelningen är diskret.

Exempel 4.6

Yngve möter i genomsnitt 18 människor på sin dagliga morgonpromenad. Vad är sannolikheten att han möter minst 25 människor en viss morgon?

Lösning

Om vi låter slumpvariabeln X beteckna antalet möten så kommer $X \sim Po(18)$. Eftersom 18 är större än 10 kan vi approximera X med normalfördelningen så att $X \sim N(18, \sqrt{18})$. Vi kontinuitetskorrigerar och kan sedan beräkna

$$P(X \geq 24,5) = P\left(Z > \frac{24,5-18}{\sqrt{18}}\right) = P(Z > 1,53) = 0,0630$$

4.9 Fördelningen för ett stickprovsmedelvärde

Ett stickprovsmedelvärde $\bar{X}$ kan ses som en slumpvariabel, eftersom upprepad stickprovstagning kommer att resultera i olika faktiska värden på $\bar{x}$ vid olika tillfällen p.g.a. att slumpen styr vilka element som hamnar i stickprovet. Om ett slumpmässigt stickprov om totalt *n* observationer tas från en population med medelvärdet μ och standardavvikelsen σ så kommer stickprovsmedelvärdet $\bar{X}$ enligt CGS att vara en normalfördelad slumpvariabel med medelvärdet μ och standardavvikelsen $\sigma \cdot \sqrt{n}$ om *n* är stort. Om *n* är litet men ursprungspopulationen är någorlunda lik en normalfördelning så kommer också stickprovsmedelvärdet att vara en normalfördelad slumpvariabel med medelvärdet μ och standardavvikelsen $\sigma \cdot \sqrt{n}$. Storheten $\sigma \cdot \sqrt{n}$ kallas i stickprovsanalys ofta för ”medelfelet” för att inte förväxlas med ursprungspopulationens standardavvikelse σ.

När stickprovsmedelvärdet $\bar{X}$ är en normalfördelad slumpvariabel så kan den naturligtvis transformeras till *Z*, d.v.s.

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Observera att man dividerar med medelfelet $\sigma \cdot \sqrt{n}$, eftersom det är det som är standardavvikelsen för slumpvariabeln $\bar{X}$.

I praktiken när man arbetar med stickprov så känner man nästan aldrig till ursprungspopulationens sanna standardavvikelse σ, vilket innebär att man inte kan göra *Z*-transformationen ovan, men om man vet att populationen är approximativt normalfördelad kan man istället beräkna stickprovets standardavvikelse och använda den istället för σ i transformationen. Resultatet blir en slumpvariabel som kommer att följa den s.k. *t-fördelningen*. Slumpvariabeln *t*, definierad som

$$t = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$

kommer alltså att vara en *t*-fördelad slumpvariabel med $n - 1$ frihetsgrader (*df*). För slumpvariabeln *t* gäller väntevärdet

$$E(t) = 0$$

för $df > 1$ och variansen

$$V(t) = \frac{fg}{fg - 2}$$

för $df > 2$. Sannolikhetskurvan för *t*-fördelningen blir mer och mer lik sannolikhetskurvan för *Z*-fördelningen ju större *df* är, vilket framgår av tabell A2 där värden på *t* för olika antal frihetsgrader och olika återstående andelar av ytan under kurvan i högra svansen åskådliggörs.

Exempel 4.7

Anta att ett stickprov med $n = 12$ som tas från en normalfördelad population med okänd σ har $\mu = 35$. Om $s = 9$, vad är sannolikheten att stickprovet visar $\bar{x} \geq 45$.

Lösning

Vi kan beräkna

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} = \frac{45 - 35}{9 / \sqrt{12}} = 3,85$$

med 12 – 1 = 11 *df*. I tabell A2 ser vi att det för $df = 11$ gäller att $P(t > 3,1058) = 0,005$. Eftersom 3,85 > 3,1085 så är sannolikheten alltså mindre än 0,005 att ett stickprov under de aktuella omständigheterna kommer att visa $\bar{x} \geq 45$.

5. Konfidensintervall

5.1 Inledning

Statistikämnets grundpelare kan sägas vara skattning av det sanna värdet för någon populationsparameter, t.ex. ett medelvärde, med hjälp av slumpmässiga stickprov från populationen. (Stickprov antas i denna bok alltid vara slumpmässiga om inget annat sägs.) En *punktskattning* innebär just att skatta värdet på en sann populationsparameter med motsvarande parametervärde för ett stickprov från populationen. Om värdet på den aktuella parametern i stickprovet *i genomsnitt* kan förväntas vara lika med den sanna populationsparameterns värde så sägs stickprovsparameterns värde vara *väntevärdesriktigt*. Om det finns något som åstadkommer en systematisk avvikelse för stickprovsparameterns värde i förhållande till populationsparameterns sanna värde så är den alltså inte väntevärdesriktig.

Problemet är att även ett väntevärdesriktigt stickprovsvärde normalt är en slumpvariabel, vilket innebär att det värde man får med hjälp av stickprovet normalt kommer att avvika i lägre eller högre grad från populationens sanna värde. För att få en bild av hur osäker en punktskattning faktiskt är i en viss situation kompletteras den ofta med ett intervall kring det punktskattade värdet inom vilket den sanna populationsparametern ligger med en viss sannolikhet (*konfidens*). Ett sådant intervall kallas *konfidensintervall*, och skattning av en sann populationsparameter med hjälp av ett konfidensintervall brukar kallas *intervallskattning*. Begreppet konfidens innebär alltså ”sannolikheten att man har rätt” när man t.ex. skattar en populationsparameter med hjälp av ett intervall. Ett begrepp som är nära relaterat till konfidens är *signifikans*, vilket helt enkelt är komplementet till konfidens, således alltså ”sannolikheten att man har fel”. Signifikans är ett mycket viktigt begrepp i den statistiska inferensen, och vi kommer att använda det flitigt längre fram i denna bok. Signifikansnivå brukar generellt betecknas med α (uttalas ”alfa”). Konfidensnivån i en viss situation är således $1 - \alpha$.

I detta kapitel ska vi nu titta närmare på teorin bakom konfidensintervall och på hur några olika slag av konfidensintervall konstrueras.

5.2 Konfidensintervall för populationsmedelvärde när σ är känd

Om ett stickprov med storleken n tas från en population med standardavvikelsen σ så kommer stickprovsmedelvärdet att vara en normalfördelad slumpvariabel med standardavvikelsen (medelfelet) $\sigma/\sqrt{n}$ om populationen kan antas vara approximativt normalfördelad *eller* om stickprovsstorleken n är stor. Hur utnyttjar vi den kunskapen för att konstruera ett konfidensintervall kring stickprovets medelvärde med t.ex. 95% konfidens?

Av tabell A1 kan vi utläsa att $P(Z > z) = 0{,}025$ för $z = 1{,}96$. Det innebär att sannolikheten att en normalfördelad slumpvariabel får ett värde som ligger mer än 1,96 standardavvikelser ovanför medelvärdet är 2,5%. Eftersom normalfördelningen är symmetrisk ryms därför tydligen 95% av utfallsrummet för en normalfördelad slumpvariabel inom ett intervall motsvarande 1,96 standardavvikelser åt respektive håll från medelvärdet. Om $\bar{X}$ kan antas vara en normalfördelad slumpvariabel så kommer 95% av alla värden på $\bar{X}$ som är möjliga att få alltså att ligga i intervallet $\mu \pm 1{,}96 \cdot \sigma/\sqrt{n}$. Med andra ord – sannolikheten att det stickprovsmedelvärde $\bar{x}$ vi får för ett visst stickprov kommer med 95% sannolikhet att ligga i intervallet $\mu \pm 1{,}96 \cdot \sigma/\sqrt{n}$. Om $\bar{x}$ med 95% sannolikhet finns i intervallet $\mu \pm 1{,}96 \cdot \sigma/\sqrt{n}$ så ligger $\bar{x}$ alltså mindre än $\mu \pm 1{,}96 \cdot \sigma/\sqrt{n}$ åt endera

hållet från μ, och i så fall kommer μ att täckas in av intervallet $\overline{x} \pm 1{,}96 \cdot \sigma / \sqrt{n}$ med just 95% sannolikhet. Intervallet $\overline{x} \pm 1{,}96 \cdot \sigma / \sqrt{n}$ är alltså ett 95% konfidensintervall för μ, eftersom μ med 95% sannolikhet kommer att finnas i detta intervall.

Vill man konstruera intervallet med en annan konfidensnivå väljer man helt enkelt ett annat *z*-värde från tabell A1.

Generellt när $\overline{X}$ är normalfördelad och σ är känd så beräknas ett konfidensintervall för μ med konfidensnivån $1 - \alpha$ som

$$\overline{x} \pm z_{\alpha/2} \cdot \sigma / \sqrt{n}$$

där $z_{\alpha/2}$ är det *z*-värde från tabell A1 för vilket $P(Z > z_{\alpha/2}) = \alpha / 2$.

Exempel 5.1

Ett stickprov med $n = 18$ från en approximativt normalfördelad population med $\sigma = 75$ visar att $\overline{x} = 225$. Intervallskatta μ med 99% konfidens.

Lösning

Vi ska använda konfidensnivån $1 - \alpha = 0{,}99$, vilket innebär att $\alpha = 1 - 0{,}99 = 0{,}01$. Från tabell A1 får vi då $z_{\alpha/2} = z_{0{,}005} = 2{,}58$, vilket innebär att konfidensintervallet blir

$$\overline{x} \pm z_{\alpha/2} \cdot \sigma / \sqrt{n} = 225 \pm 2{,}58 \cdot 75 / \sqrt{18} = 225 \pm 45{,}61 \text{ ,}$$

d.v.s. [179,39 , 270,61].

5.3 Konfidensintervall för populationsmedelvärde när σ inte är känd

I de flesta praktiska situationer är σ okänd när μ ska skattas för en viss population. Som framgick av föregående kapitel kan σ då skattas med *s* om populationen är approximativt normalfördelad, varvid *t*-fördelningen kan användas istället för *Z*-fördelningen för att analysera situationen. Logiken för konstruktion av konfidensintervall är exakt densamma som när σ är känd – den enda skillnaden är att konstanten $z_{\alpha/2}$ ersätts med motsvarande konstant från *t*-fördelningen, d.v.s. $t_{\alpha/2}$ med aktuellt antal *df*, eftersom medelfelet nu måste skrivas som $s / \sqrt{n}$. I och med att osäkerheten i skattningen därmed är större så blir konfidensintervallet vidare.

Generellt när $\overline{X}$ är normalfördelad och σ är okänd så beräknas ett konfidensintervall för μ med konfidensnivån $1 - \alpha$ som

$$\overline{x} \pm t_{\alpha/2} \cdot s / \sqrt{n}$$

där $t_{\alpha/2}$ är det *t*-värde från tabell A2 för vilket $P(t > t_{\alpha/2}) = \alpha / 2$ med $df = n - 1$.

Exempel 5.2

Ett stickprov med n = 18 från en approximativt normalfördelad population visar att $\overline{x}$ = 225 och s = 75. Intervallskatta μ med 99% konfidens.

Lösning

Vi har df = 18 – 1 = 17, och vi ska använda konfidensnivån $1 - \alpha = 0{,}99$, vilket innebär att $\alpha = 1 - 0{,}99 = 0{,}01$. Från tabell A2 får vi då $t_{\alpha/2} = t_{0{,}005} = 2{,}8982$, vilket innebär att konfidensintervallet blir

$$\overline{x} \pm t_{\alpha/2} \cdot s / \sqrt{n} = 225 \pm 2{,}8982 \cdot 75 / \sqrt{18} = 225 \pm 51{,}23$$

d.v.s. [173,77 , 276,23].

5.4 Konfidensintervall för populationsproportion

Ofta är man intresserad av att kartlägga vilken *andel* av en population som uppfyller ett visst villkor. Ett vanligt exempel är undersökningar som syftar till att skatta andelen röstberättigade medborgare som skulle rösta på ett visst parti om det vore val idag. Den sanna andelen av en population som uppfyller ett visst villkor betecknar vi med p, eftersom det är konceptuellt sett samma sak som sannolikheten att ett slumpmässigt utvalt element från populationen uppfyller det aktuella villkoret. Det är alltså precis samma p som vi har sett tidigare i binomialfördelningen, och det är rent teoretiskt just binomialfördelningen som ska ligga till grund för ett konfidensintervall kring den *stickprovsproportion*, d.v.s. andelen element i stickprovet som uppfyller det aktuella villkoret. En stickprovsproportion betecknar vi med $\hat{p}$.

Problemet är att binomialfördelningen inte är symmetrisk, och att det är komplicerat att beräkna konfidensintervall baserat på denna fördelning. Som vi har sett tidigare kan binomialfördelningen emellertid approximeras med normalfördelningen när stickprovsstorleken är stor, vilket tumregelmässigt är fallet när $n \cdot p \geq 5$ och $n(1 - p) \geq 5$. Eftersom p är okänd så använder vi istället tumregeln att det ska gälla att $n \cdot \hat{p} \geq 5$ och $n(1 - \hat{p}) \geq 5$.

Medelfelet för p blir teoretiskt sett

$$\sqrt{n \cdot p(1-p)} / n = \sqrt{p(1-p)/n}$$

men vi förlorar en frihetsgrad när p skattas med $\hat{p}$, så för att skattningen av medelfelet ska bli väntevärdesriktig måste vi dividera med $n - 1$ istället för med n, och resultatet blir att vi skattar medelfelet för $\hat{p}$ med uttrycket $\sqrt{\hat{p}(1-\hat{p})/(n-1)}$.

Ett generellt konfidensintervall för p med konfidensnivån $1 - \alpha$ kan därför beräknas som

$$\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\hat{p}(1-\hat{p})/(n-1)}$$

där $z_{\alpha/2}$ är det z-värde från tabell A1 för vilket $P(Z > z_{\alpha/2}) = \alpha / 2$.

Exempel 5.3

I en undersökning där 240 slumpvis utvalda svenskar deltog visade sig 54 sympatisera med moderaterna. Ange ett 95% konfidensintervall för den sanna andelen moderata sympatisörer i Sverige vid det aktuella tillfället.

Lösning

Vi har punktskattningen $\hat{p} = 54 / 240 = 0{,}225$. Vi ska använda konfidensnivån $1 - \alpha = 0{,}95$, vilket innebär att $\alpha = 1 - 0{,}95 = 0{,}05$. Från tabell A1 får vi då $z_{\alpha/2} = z_{0{,}025} = 1{,}96$, vilket innebär att konfidensintervallet blir

$$\hat{p} \pm z_{\alpha/2} \cdot \sqrt{\hat{p}(1-\hat{p})/(n-1)} = 0{,}225 \pm 1{,}96 \cdot \sqrt{0{,}225(1-0{,}225)/(240-1)} = 0{,}225 \pm 0{,}053$$

d.v.s. [0,172 , 0,278].

5.5 Konfidensintervall för ändliga populationer

Anta att det stickprov om n element vi har tagit utgör en relativt stor andel av en total ändlig population som vi vet uppgår till N element. Jämfört med de tidigare fallen, där vi implicit har antagit att populationen är oändlig (eller i alla fall mycket stor i förhållande till stickprovsstorleken) så kommer osäkerheten nu att vara mindre.

Man kan visa att ett korrekt sätt att hantera denna situation är att justera den aktuella stickprovsparameterns medelfel genom att multiplicera det med faktorn $\sqrt{(N-n)/N}$. Det innebär konkret att ett konfidensintervall blir snävare ju större andel av den totala populationen som ingår i

stickprovet. Om hela populationen ingår så har vi en *totalundersökning*, och då blir medelfelet 0 efter den nämnda justeringen, vilket är logiskt eftersom någon osäkerhet då inte finns.

Tumregelmässigt bör man alltid göra denna justering om stickprovsstorleken utgör mer än 5% av den totala populationen.

Exempel 5.4

Ett stickprov med $n = 12$ från en approximativt normalfördelad population bestående av 20 element visar att $\bar{x}$ och $s = 1,5$. Intervallskatta μ med 95% konfidens.

Lösning

Vi har $df = 12 - 1 = 11$, och $\alpha = 1 - 0,95 = 0,05$. Medelfelet blir efter justering

$$\sqrt{(N-n)/N} \cdot s/\sqrt{n} = \sqrt{(20-12)/20} \cdot 1,5/\sqrt{12} = 0,1732$$

Från tabell A2 får vi då $t_{\alpha/2} = t_{0,025} = 2,2010$, vilket innebär att konfidensintervallet blir

$$7,3 \pm 2,2010 \cdot 0,1732 = 7,3 \pm 0,3812$$

d.v.s. [6,9188, 7,6812].

5.6 Att bestämma stickprovsstorlek

När $\bar{X}$ är normalfördelad och σ är känd så beräknas ett konfidensintervall för μ som $\bar{x} \pm z_{\alpha/2} \cdot \sigma/\sqrt{n}$. Storheten $z_{\alpha/2} \cdot \sigma/\sqrt{n}$ kallas vanligen för skattningens *felmarginal*, eftersom den visar hur stor punktskattningen avvikelse från sanna medelvärdet som högst kan vara med aktuell konfidensnivå. Vi betecknar felmarginalen med *B*, och från

$$B = z_{\alpha/2} \cdot \sigma/\sqrt{n}$$

kan vi bryta ut *n*, vilket ger

$$n = (z_{\alpha/2} \cdot \sigma/B)^2$$

Om vi känner till, eller kan åstadkomma en rimlig skattning av, σ så kan detta uttryck utnyttjas för att bestämma minsta möjliga stickprovsstorlek vid en på förhand bestämd signifikansnivå och önskad felmarginal.

På samma sätt kan vi utnyttja vetskapen om att felmarginalen för en populationsproportion är

$$B = z_{\alpha/2} \cdot \sqrt{p(1-p)/n}$$

från vilket vi på samma sätt som ovan kan bryta ut *n*, vilket ger

$$n = (z_{\alpha/2} \cdot \sqrt{p(1-p)}/B)^2$$

som en formel för att kunna beräkna det n som uppfyller kraven på precision och säkerhet förutsatt att den vanliga normalfördelningsapproximationen får användas. Observera dock att en rimlig gissning av p måste göras för att man ska kunna använda denna formel. Om en sådan inte enkelt låter sig göras bör man använda försiktighetsprincipen och sätta $p = 0{,}5$ när formeln tillämpas.

Exempel 5.5

Yngve vill skatta den genomsnittliga längden på sina tio tusen röda rosor i trädgården med 95% konfidens och en felmarginal på högst 2,5 cm. Han har anledning att tro att $\sigma = 15$ cm. Hur många rosor bör Yngve ha med i stickprovet?

Lösning

$$n = (z_{\alpha/2} \cdot \sigma / B)^2 = (1{,}96 \cdot 15 / 2{,}5)^2 = 138{,}3$$

d.v.s. minst 139 rosor bör ingå i stickprovet.

Exempel 5.6

Yngve har ingen aning om hur stor andelen taggiga rosor i hans trädgård är, varför han vill skatta denna andel med 90% konfidens och en felmarginal på högst 0,03. Hur stort stickprov bör han ta?

Lösning

$$n = (z_{\alpha/2} \cdot \sqrt{p(1-p)} / B)^2 = (1{,}65 \cdot \sqrt{0{,}5(1-0{,}5)} / 0{,}03)^2 = 756{,}25$$

d.v.s. minst 757 rosor bör ingå i stickprovet.

QUINTILES TRANSNATIONAL

Students and New Graduates

Quintiles Transnational Corp. realizes that as a student or recent graduate, you have many available career options. As such, we would like to invite you to explore the vast opportunities we have to offer. Career opportunities with Quintiles Transnational Corp. provide a work environment shared by those who value hard work and will create a learning environment drawing on the wisdom, expertise, experience, and talents of our unequaled staff.

For more information on identifying career opportunities, please go to www.quintiles.com

6. Hypotestest

6.1 Inledning

I föregående kapitel tittade vi på hur stickprov från en population kan användas för att skatta en populationsparameter. Detta kan ses som en naturlig övergång från den beskrivande statistiken till den analytiska statistiken. I detta kapitel ska vi ta steget fullt ut och introducera principen om *statistisk inferens*, d.v.s. hur man med hjälp av stickprov kan dra konkreta slutsatser om värdet på olika populationsparametrar.

Grunden för den statistiska inferensen är *hypotestesten*. En hypotestest är precis vad ordet säger – en test av en hypotes. Med hypotes avses ett initialt antagande om värdet på en populationsparameter. Så länge tillräckligt starka bevis inte har presenterats mot detta initiala antagande så måste man anta att antagandet *kan* vara sant. Det initiala antagandet kallas därför vanligen för *nollhypotes*, och betecknas med H0. Jämför gärna med den klassiska rättsprincipen ”hellre fria än fälla” som innebär att en åtalad ska frikännas om åklagaren inte har kunnat presentera tillräckligt starka bevis (”bortom rimligt tvivel”) mot den åtalade. Ett frikännande betyder inte att domstolen konstaterar att den åtalade *är* oskyldig, bara att sannolikheten inte är tillräckligt stor för att man ska kunna dra slutsatsen att denne är skyldig.

Först om domstolen finner bevisningen såpass stark att det är bevisat bortom rimligt tvivel att den åtalade är skyldig, då fälls denne. Enligt samma logik genomförs en hypotestest. När vi testar en nollhypotes så tar vi ett stickprov och beräknar sannolikheten att få just de stickprovsdata vi faktiskt fick om det är så att nollhypotesen är sann. Om denna sannolikhet är tillräckligt låg så *förkastar* vi nollhypotesen och drar därmed slutsatsen att komplementet till nollhypotesen – den s.k. *mothypotesen*, betecknad med H1 – är sann. En hypotestest baseras alltså på att man har formulerat en nollhypotes och dess mothypotes, och att nollhypotesen kan testas med statistiska metoder.

Exempel 6.1

Yngve vill använda stickprov för att se om andelen röda rosor i hans trädgård kan antas utgöra mer än 50% av det totala antalet rosor. Hur ska han formulera sina hypoteser?

Lösning

$$\text{H0: } p \leq 0{,}5$$
$$\text{H1: } p > 0{,}5$$

En sak bör dock särskilt observeras här: Om H0 i denna situation inte kan förkastas så har Yngve *inte* bevisat att andelen röda rosor *är* högst 50%, bara att risken att det *kan vara* så är för stor för han ska kunna påstå motsatsen. Detta är ett klassiskt fel som man tyvärr ofta ser folk göra när de använder hypotestest. Den enda gången man kan dra en konkret slutsats på basis av en hypotestest är när man *kan* förkasta H0, varvid man drar slutsatsen att H1 är sann. Kan H0 däremot inte förkastas så kan man inte dra någon annan slutsats än att bevisen inte var tillräckligt starka för att H0 skulle kunna förkastas.

När hypoteser formuleras måste H0 och H1 uppenbarligen vara varandras komplement. Det är dock viktigt att tänka på att det är H0 som måste innehålla alternativet med likhet, och resonemanget i föregående stycke förklarar på sätt och vis varför. Man kan nämligen aldrig bevisa med hjälp av stickprov att en viss populationsparameter är *lika med* ett visst värde – för att kunna dra en sådan slutsats måste man alltid ta till en totalundersökning. Men om stickprovsdatan talar mot H0 tillräckligt mycket ska man dra slutsatsen att det är H1 som är sann. H1 måste därför alltid formuleras som en olikhet, och alternativet med likhet ska därför alltid återfinnas i H0.

6.2 Fel av typ I och typ II

När man testar en hypotes finns i de flesta fall en risk att man gör fel. Jämför återigen med rättsprincipen ”hellre fria än fälla” där beviskravet är ”skyldig bortom rimligt tvivel”. Även om domstolen finner det troligt att en åtalad är skyldig så ska denne frikännas om det finns rimligt tvivel på dennes skuld. Och tvärtom – även om domstolen bara får fälla när den finner att det bortom rimligt tvivel är så att en åtalad är skyldig så händer det trots allt ibland att oskyldiga döms till straff. Uppenbarligen finns det två typer av fel som kan göras: Man kan frikänna en person som egentligen var skyldig, och man kan fälla en person som egentligen var oskyldig.

När det gäller hypotestest talar man om fel av typ I och fel av typ II, vilka har sina direkta motsvarigheter i exemplet i stycket ovan. Att förkasta H0 och därmed konkludera att H1 är sann innebär ett typ I-fel om det är så att H0 är sann. Att istället låta bli att förkasta en nollhypotes som egentligen inte var sann innebär ett typ II-fel. I de flesta fall kan ett typ I-fel sägas vara allvarligare än ett typ II-fel, eftersom man i det första fallet faktiskt drar en konkret slutsats. Detta illustreras av principen ”hellre fria än fälla”, eftersom ett typ I-fel motsvarar att döma en oskyldig, medan ett typ II-fel motsvaras av att frikänna en skyldig.

När man ska genomföra en hypotestest bestämmer man en konkret gräns för hur stor sannolikheten för ett typ I-fel får vara – den s.k. *signifikansnivån* α. Idén är sedan att analysera den stickprovsdata man har med lämpliga metoder och därmed beräkna sannolikheten för att få minst så pass extrema data som man faktiskt har fått om det är så att H0 *faktiskt* är sann. Denna sannolikhet kallas *p-värde*. Beslutsregeln är sedan enkel: Om testets p-värde är lägre än α, då förkastas H0.

Sannolikheten för att en viss test leder till ett typ II-fel brukar betecknas med β. Sannolikheten $1 - \beta$, d.v.s. sannolikheten att en viss test undviker typ II-fel, brukar kallas testens *styrka*. Vi kommer att återkomma till detta begrepp längre fram.

6.3 Test av ett populationsmedelvärde när σ är känd

Anta vi vet att det för en viss population gäller att $\sigma = 5$, att vi har tagit ett stickprov med $n = 50$ och $\bar{x} = 201{,}3$, och att vi nu vill genomföra hypotestesten

$$\text{H0: } \mu \leq 200$$
$$\text{H1: } \mu > 200$$

Vad är sannolikheten att få minst ett så pass extremt stickprovsmedelvärde som 201,3 när $n = 50$ och $\sigma = 5$ om det är så att sanna medelvärdet är 200? Det tar vi reda på med en vanlig Z-transformation, eftersom vi får anta att $\bar{x}$ är en normalfördelad slumpvariabel med medelfelet $5/\sqrt{50}$ eftersom stickprovet är stort:

$$P(Z > z) = P\left(Z > \frac{201,3 - 200}{5/\sqrt{50}}\right) = P(Z > 1,84) = 0,0329$$

Sannolikheten för att få minst ett så pass extremt stickprovsmedelvärde om sanna medelvärdet faktiskt är 200 är alltså 0,0329. Det är det som är det här testets p-värde. Om vi på förhand hade bestämt oss för att arbeta med signifikansnivån 0,05, d.v.s. vi vill vara minst 95% säkra på att ha rätt om vi förkastar H0, så ska vi i detta fall förkasta H0 och dra slutsatsen att H1 är sann, eftersom $0,0329 < 0,05$.

I detta fall använde vi Z som *testvariabel*. Som vi kommer att se längre fram finns olika slag av testvariabler, och valet av testvariabel beror på vilken typ av situation man står inför.

Faktum är att vi egentligen inte behövde räkna ut p-värdet för att kunna genomföra den här hypotestesten. Beslutsregeln är ju att förkasta H0 om p-värdet är lägre än α, och om vi vet vilket värde på testvariabeln som exakt motsvarar ett p-värde lika med α (det s.k. *kritiska värdet* för testvariabeln) så vet vi att H0 ska förkastas om testvariabelns framräknade värde är mer extremt (d.v.s. avlägset från 0) än dess kritiska värde. Även om vi då inte känner till det exakta p-värdet så vet vi att p-värdet i alla fall är lägre än α. I exemplet ovan säger man att testet görs i normalfördelningens högra svans, eftersom högre värden på Z ger lägre p-värden när H1 är av karaktären ”större än”.

Generellt kan sambandet mellan testens H1 och sättet på vilket den genomförs sammanfattas enligt följande:

Om H1 är {1. ”större än”, 2. ”mindre än”, 3. ”skild från”} så genomförs hypotestesten i {1. högra, 2. vänstra, 3. både högra och vänstra} svansen av testvariabelns fördelning, vilket innebär att testens p-värde blir lägre ju {1. större, 2. mindre, 3. mer extremt åt endera hållet} testvariabelns värde är.

Vilket är då det kritiska *Z*-värdet för detta test, d.v.s. vilket är det värde som det med stickprovsdata framräknade *Z*-värdet måste överstiga för att p-värdet ska vara lägre än α? Enligt tabell A1 går gränsen för de mest extrema 5% i *Z*-fördelningens högra svans vid värdet 1,65, d.v.s. $z_{0,05} = 1,65$. Eftersom vi med hjälp av det stickprov vi har tagit kan beräkna testvärdet

$$z = \frac{201,3 - 200}{5/\sqrt{50}} = 1,84$$

så *vet* vi att testets p-värde är lägre än $\alpha = 0,05$ eftersom $z > z_{\alpha}$, d.v.s. 1,84 > 1,65. Detta innebär att H0 ska förkastas, och det exakta p-värdet saknar egentligen betydelse. Båda sätten att genomföra hypotestester ger naturligtvis alltid samma slutsats, men beroende på situationen så är det ena sättet ofta enklare att använda än det andra. Detta kommer att framgå av den fortsatta framställningen.

6.4 Test av ett populationsmedelvärde när σ inte är känd

I praktiken är σ nästan alltid okänd när μ okänd. Vi kan då inte använda *Z* som testvariabel som i avsnitt 6.3. För att kunna göra en hypotestest av μ måste vi nu istället använda oss av stickprovets standardavvikelse *s* och basera testet på *t*-fördelningen, förutsatt att populationen som stickprovet kommer från kan antas vara approximativt normalfördelad. Vi räknar då fram ett *t*-värde för den stickprovsdata vi har med uttrycket

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}}$$

där μ är den konstant som vi testar mot i våra hypoteser. Sedan jämförs testets *t*-värde med t_a, det kritiska t-värdet för den aktuella signifikansnivån och antalet *df*. Om testets *t*-värde är mer extremt än det kritiska *t*-värdet så förkastas H0.

Exempel 6.2

Ett stickprov med $n = 17$ från en approximativt normalfördelad population visar att $\overline{x}$ och $s = 5,2$.

- Kan man med $\alpha = 0,05$ dra slutsatsen att $\mu = 25$ med $\alpha = 0,05$?
- Kan man med $\alpha = 0,05$ dra slutsatsen att $\mu \neq 24$ med $\alpha = 0,05$

Lösning

Vi har hypoteserna

$$\begin{aligned} &\text{H0: } \mu \leq 25 \\ &\text{H1: } \mu > 25 \end{aligned}$$

vilket innebär att testet genomförs i *t*-fördelningens högra svans.

Kritiskt *t* för $df = 17 - 1 = 16$ och signifikansnivån 0,05 kan från tabell A2 avläsas till $t_{0,05} = 1,7459$. Testets *t*-värde beräknas till

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} = \frac{26,92 - 25}{5,2 / \sqrt{17}} = 1,52$$

vilket innebär att H0: $\mu \leq 25$ inte kan förkastas då $1,52 < 1,7459$. Observera återigen att detta *inte* betyder att vi på något sätt skulle ha ”bevisat” att $\mu \leq 25$, bara att bevisen mot denna hypotes inte är tillräckligt starka för att vi ska kunna förkasta den.

Om vi istället vill kontrollera om det går att dra slutsatsen att $\mu \neq 3$ så får vi hypoteserna

$$\begin{aligned} &\text{H0: } \mu = 24 \\ &\text{H1: } \mu \neq 24 \end{aligned}$$

vilket innebär att testet genomförs i *t*-fördelningens *båda* svansar (en s.k. *tvåsidig test*). Detta beror på att tillräckligt extrema avvikelser i stickprovmedelvärde såväl uppåt som nedåt från de 25 som H0 antar leder till att H0 ska förkastas.

Vi får därmed två kritiska *t*-värden med $\alpha = 0,05$, nämligen $t_{\alpha/2} = t_{0,025} = 2,1199$ och $-t_{\alpha/2} = -t_{0,025} = -2,1199$. Testets *t*-värde blir

$$t = \frac{26,92 - 24}{5,2 / \sqrt{17}} = 2,32$$

och eftersom testets *t*-värde är mer extremt än ett av de kritiska *t*-värdena, d.v.s. $2,1199 < 2,32$, så ska H0: $\mu = 24$ förkastas. Vi har tillräckligt starka bevis för att kunna dra slutsatsen att $\mu \neq 24$.

6.5 Test av en populationsproportion – stort stickprov

Vid hypotestest av hur den sanna populationsproportionen *p* förhåller sig till ett visst värde p_0 så kan normalfördelningsapproximation användas när stickprovsstorleken är stor. Testvariabeln blir då *Z*, och testets *z*-värde beräknas med hjälp av stickprovsdata med formeln

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0) / n}}$$

där $\hat{p}$ är stickprovsproportionen och *n* är stickprovsstorleken.

Exempel 6.3

I en undersökning där 240 slumpvis utvalda svenskar deltog visade sig 54 sympatisera med moderaterna. Kan man med 99% säkerhet dra slutsatsen att den sanna andelen moderata sympatisörer i Sverige vid det aktuella tillfället är lägre än 30%?

Lösning

Vi har hypoteserna

$$\text{H0: } p \geq 0{,}30$$
$$\text{H1: } p < 0{,}30$$

och $\hat{p} = 54 / 240 = 0{,}225$. Testet utförs i Z-fördelningens vänstra svans, vilket innebär att det kritiska Z-värdet enligt tabell A1 blir $-z_{\alpha} = -z_{0{,}01} = -2{,}33$. Testets Z-värde blir

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} = \frac{0{,}225 - 0{,}30}{\sqrt{0{,}30(1-0{,}30)/240}} = \frac{-0{,}075}{0{,}0296} = -2{,}53$$

varvid vi förkastar H0 och drar slutsatsen att $p < 0{,}30$ eftersom $-2{,}33 > -2{,}53$.

6.6 Test av en populationsproportion – litet stickprov

Vid hypotestest av hur den sanna populationsproportionen p förhåller sig till ett visst värde p_0 så måste binomialfördelningen användas när stickprovsstorleken är liten. Vi beräknar då p-värdet för det aktuella testet direkt såsom sannolikheten att faktiskt erhålla just det antal positiva utfall som vi faktiskt fick eller ett ännu mer extremt antal, i den totala mängden utförda försök om det är så att H0 faktiskt är sann. Är detta p-värde lägre än signifikansnivån så förkastas H0.

Exempel 6.4

En skojare på torget har tre koppar på ett bord. Under en av kopparna finns en kula, och folk slår vad om pengar med skojaren att de kan gissa under vilken kopp kulan finns när skojaren har flyttat runt kopparna ett tag. Gissar man på rätt kopp så får man dubbla insatsen tillbaks, annars får man ingenting. När folk gissat på en kopp vänds denna upp, och om kulan inte fanns där så visar skojaren under vilken av de båda andra kopparna kulan fanns för att demonstrera att det inte var något fusk.

Du har observerat att folk väldigt ofta gissar på fel kopp, och bestämmer dig för att göra ett experiment. Du spelar spelet med skojaren 7 gånger och blundar varje gång medan han flyttar runt kopparna för att sedan välja en kopp slumpmässigt. Två av de 7 gångerna gissade du rätt. Kan man på basis av denna data med $\alpha = 0{,}1$ dra slutsatsen att spelet är orättvist för spelarna?

Lösning

Vi låter p vara sannolikheten att spelaren gissar på rätt kopp vid slumpmässigt val av kopp. Vi får då hypoteserna

$$\text{H0: } p \geq 1/3$$
$$\text{H1: } p < 1/3$$

Sannolikheten att få högst 2 positiva utfall vid 7 successiva oberoende försök när $p = 1/3$ beräknas med binomialfördelningsformeln till

$$P(X \leq 2) = \sum_{x=0}^{2} \binom{7}{x} (1/3)^x (1-(1/3))^{(7-x)} = 0{,}5707$$

Eftersom $0{,}1 < 0{,}5707$ så har vi inte tillräckligt starka bevis för att kunna förkasta H0 och därmed kunna dra slutsatsen att spelet är orättvist för spelarna.

6.7 Test av en populationsvarians

Ibland är det spridningen i en population man vill kunna uttala sig om. Den fördelning som normalt används för att testa en populationsvarians när populationen i sig kan antas vara approximativt normalfördelad är chitvå-fördelningen, eller som den oftare skrivs med hjälp av det grekiska alfabetet, χ^2-fördelningen. χ^2 är i sig en kontinuerlig slumpvariabel vars fördelning formellt sett är sannolikhetsfördelningen för summan av flera oberoende Z-fördelade slumpvariabler som kvadrerats, därav tvåan i χ^2. I tabell A3 illustreras, på samma sätt som för t-fördelningen i tabell A2, de kritiska värdena för χ^2-fördelningen vid olika antal frihetsgrader och olika värden på α.

När en hypotestest behandlar en populationsvarians, t.ex. om vi har H0: $\sigma^2 = \sigma_0^2$, så är det alltså χ^2 som är testvariabel. Testets χ^2-värde beräknas enligt formeln

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2}$$

vilket relateras till kritiskt χ^2-värde med $n - 1$ *df* vid aktuell signifikansnivå α. Om testets χ^2-värde är mer extremt än det kritiska χ^2-värdet så förkastas H0.

Vidare kan ett konfidensintervall för σ^2 för en normalfördelad population på nivån $1 - \alpha$ definieras som

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

med $n - 1$ *df*.

χ^2-fördelningen används även vid andra typer av tester, vilket vi kommer att se längre fram i denna bok.

Exempel 6.5

I en produktionsprocess bör outputen bland annat kännetecknas av en varians under 75 för att allt ska anses vara normalt. Ett stickprov med $n = 10$ ger en stickprovsvarians på 26,4. Kan man med $\alpha = 0{,}05$ dra slutsatsen att sanna variansen understiger 75? Beräkna också ett 95% konfidensintervall för sanna variansen.

Lösning

Vi har hypoteserna

H0: $\sigma^2 \geq 75$
H1: $\sigma^2 < 75$

Testet genomförs i χ^2-fördelningens vänstra svans, eftersom H1 är av karaktären ”mindre än”. Testets χ^2-värde blir

$$\chi^2 = \frac{(10-1)\cdot 26{,}4}{75} = 3{,}168$$

och det kritiska χ^2-värdet blir det värde som avgränsar de lägsta 5%, d.v.s. de högsta 95%, i χ^2-fördelningen med $10 - 1 = 9$ *df*, vilket kan avläsas från tabell A3 till $\chi^2_{0{,}95} = 3{,}325$. Detta innebär att H0 förkastas, eftersom $3{,}325 > 3{,}168$.

Ett 95% konfidensintervall beräknas enligt formeln som

$$\left[\frac{(10-1)\cdot 26{,}4}{19{,}023}, \frac{(10-1)\cdot 26{,}4}{2{,}700}\right] = [12{,}49,\ 88{,}00]$$

6.8 Test av två populationsvarianser

Ibland är man intresserad av att testa hur två varianserna från två oberoende populationer förhåller sig till varandra. Om vi kallar de båda populationsvarianserna σ_1^2 och σ_2^2 så finns det konceptuellt sett tre tänkbara nollhypoteser att testa, nämligen H0: $\sigma_1^2 = \sigma_2^2$, H0: $\sigma_1^2 \leq \sigma_2^2$ och H0: $\sigma_1^2 \geq \sigma_2^2$. Speciellt den förstnämnda har stor betydelse, vilket kommer att framgå längre fram i detta kapitel.

Denna typ av test kallas för *F*-test och baseras på det faktum att kvoten mellan stickprovsvarianserna s_1^2 och s_2^2, baserade på stickprovsstorlekarna n_1^2 och n_2^2, följer den s.k. *F*-fördelningen (eller Fishers *F*-fördelning) om populationerna är approximativt normalfördelade. Slumpvariabeln *F* definieras alltså som

$$F = \frac{s_1^2}{s_2^2}$$

med $n_1^2 - 1$ *df* i täljaren och $n_2^2 - 1$ *df* i nämnaren. Om *F*–värdet för viss testdata är större än 1 så indikerar det att det kan vara så att $\sigma_1^2 > \sigma_2^2$, och tvärtom. Test av hypotesen H0: $\sigma_1^2 \leq \sigma_2^2$ utförs alltså i *F*-fördelningens högra svans, och tvärtom.

Kritiska värden för *F*-fördelningens högra svans finns i tabell A4 – A9 för olika värden på α där kolumnrubrikerna representerar täljarens *df* och radrubrikerna motsvarar nämnarens *df*. Kritiska värden för vänstra svansen fås enkelt via sambandet

$$F_{1-\alpha} = \frac{1}{F_\alpha}$$

Exempel 6.6

Antag att vi har tagit stickprov med $n_1^2 = 11$ och $n_2^2 = 15$ från två oberoende populationer som kan antas vara approximativt normalfördelade. Antag vidare att vi har $s_1^2 = 18$ och $s_2^2 = 41$. Kan vi då med $\alpha = 0{,}05$ dra slutsatsen att $\sigma_1^2 < \sigma_2^2$?

Lösning

Vi har hypoteserna

$$\text{H0: } \sigma_1^2 \geq \sigma_2^2$$
$$\text{H1: } \sigma_1^2 < \sigma_2^2$$

och testet görs således i *F*-fördelningens vänstra svans med 11 – 1 = 10 *df* i täljaren och 15 – 1 = 14 *df* i nämnaren. Kritiskt *F*-värde blir då

$$F_{1-\alpha} = \frac{1}{F_\alpha} = \frac{1}{2,60} = 0,3846$$

Testets *F*-värde blir

$$F = \frac{18}{41} = 0,439$$

vilket innebär att H0 inte kan förkastas eftersom 0,3849 < 0,439.

6.9 Test av två populationsmedelvärden – lika standardavvikelser

En *t*-test kan också användas för att med hjälp av stickprovsdata analysera skillnad i medelvärde mellan två av varandra oberoende populationer där båda populationerna är approximativt normalfördelade. Även här finns konceptuellt sett tre olika slags nollhypoteser som kan testas, nämligen H0: $\mu_1 - \mu_2 = D$, H0: $\mu_1 - \mu_2 \leq D$, och H0: $\mu_1 - \mu_2 \geq D$, där *D* är just den differens mellan de sanna populationsmedelvärdena som man vill testa för. Man är ofta intresserad av att använda $D = 0$, vilket innebär att man testar om det finns någon skillnad mellan de båda populationernas medelvärden över huvud taget.

När vi ska testa någon av dessa nollhypoteser så måste vi först bestämma oss för om vi ska anta att de båda populationernas standardavvikelser ska antas vara lika eller ej. Detta kan man med fördel *F*-testa med hjälp av den stickprovsdata man har enligt föregående avsnitt (kom ihåg att variansen bara är kvadraten på standardavvikelsen). Om man kan anta att standardavvikelserna är lika så börjar man med att ta fram den sammanvägda stickprovsstandardavvikelsen s_p såsom

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

varefter testets *t*-värde beräknas som

$$t = \frac{(\overline{x}_1 - \overline{x}_2) - D}{s_p \sqrt{(1/n_1 + 1/n_2)}}$$

med $n_{1+}n_2 - 2$ *df*. Det kritiska *t*-värdet hämtas som vanligt från tabell A2. Ett $1 - \alpha$ konfidensintervall för den sanna skillnaden $\mu_1 - \mu_2$ kan sedan beräknas som

$$(\overline{x}_1 - \overline{x}_2) \pm t_{\alpha/2} s_p \sqrt{(1/n_1 + 1/n_2)}$$

med $n_1 + n_2 - 2$ *df*.

Exempel 6.7

Stickprov från två olika populationer, som kan antas vara approximativt normalfördelade, med stickprovsstorlekarna $n_1 = 8$ och $n_2 = 7$ ger att $\overline{x}_1 = 18{,}6$, $\overline{x}_2 = 19{,}9$, $s_1 = 1{,}6$ och $s_2 = 0{,}9$. Kan man med $\alpha = 0{,}05$ dra slutsatsen att $\mu_1 < \mu_2$ om man antar att populationernas sanna standardavvikelser är lika? Beräkna ett konfidensintervall för $\mu_1 = \mu_2$.

Lösning

Vi har hypoteserna

H0: $\mu_1 - \mu_2 \geq 0$
H1: $\mu_1 - \mu_2 < 0$ (samma sak som $\mu_1 < \mu_2$)

Den sammanvägda stickprovsstandardavvikelsen s_p blir

$$s_p = \sqrt{\frac{(8-1)1,6^2 + (7-1)0,9^2}{8+7-2}} = 1,3237$$

varefter testets *t*-värde kan beräknas

$$t = \frac{(18,6_1 - 19,9) - 0}{1,3237\sqrt{(1/8 + 1/7)}} = \frac{-1,3}{0,6851} = -1,8976$$

H1 är av karaktären ”mindre än” vilket innebär att testet görs i *t*-fördelningens vänstra svans. Kritiskt *t*-värde för $\alpha = 0{,}05$ och $8 + 7 - 2 = 13$ *df* är enligt tabell A2 -1,7709. Eftersom -1,7709 > -1,8976 förkastas H0. Ett 95% konfidensintervall blir

$$(18,6 - 19,9) \pm 2,1604 \cdot 1,3237\sqrt{(1/8 + 1/7)} = -1,3 \pm 1,48 = \text{ [-2,78 , 0,18]}$$

6.10 Test av två populationsmedelvärden – olika standardavvikelser

När de sanna populationsstandardavvikelserna kan antas vara olika används respektive stickprovs-standardavvikelse direkt i beräkningen av testets *t*-värde:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - D}{\sqrt{s_1^2 / n_1 + s_2^2 / n_2}}$$

Antalet frihetsgrader beräknas sedan till

$$df = \frac{(s_1^2 / n_1 + s_2^2 / n_2)^2}{(s_1^2 / n_1)^2 / (n_1 - 1) + (s_2^2 / n_2)^2 / (n_2 - 1)}$$

vilket alltid ska avrundas nedåt om resultatet inte blir ett heltal. Ett $1 - \alpha$ konfidensintervall för den sanna skillnaden $\mu_1 - \mu_2$ kan sedan beräknas som

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} s_p \sqrt{(1/n_1 + 1/n_2)}$$

JPMorgan
The 360° career.
www.jpmorgan.com
Get the European Perspective

We believe that JPMorgan is the most challenging and rewarding career choice a talented graduate can make. We call this the 360° career because it is a total package of earning power, job satisfaction and personal development.

We take graduates into a range of different businesses from Investment Banking to Technology. Our training programmes combine on-the-job learning with top-quality classroom instruction and practical experience gained in different parts of the business.

Exempel 6.8

Räkna om exempel 6.7 under förutsättning att de sanna populationsstandardavvikelserna inte antas vara lika.

Lösning

Som innan har vi hypoteserna

$$\text{H0: } \mu_1 - \mu_2 \geq 0$$
$$\text{H1: } \mu_1 - \mu_2 < 0$$

Testets *t*-värde kan nu beräknas till

$$t = \frac{(18,6_1 - 19,9) - 0}{\sqrt{(1,6^2/8 + 0,9^2/7)}} = \frac{-1,3}{0,6601} = -1,9695$$

och antalet frihetsgrader blir

$$df = \frac{(1,6^2/8 + 0,9^2/7)^2}{(1,6^2/8)^2/(8-1) + (0,9^2/7)^2/(7-1)} = \frac{0,1898}{0,0168} = 11,2976$$

som ska avrundas nedåt till 11. Kritiskt *t*-värde avläses från tabell A2 till -1,7959 och H0 ska därmed förkastas eftersom -1,7959 > -1,9695. Ett 95% konfidensintervall blir nu

$$(18,6 - 19,9) \pm 2,2010\sqrt{(1,6^2/8 + 0,9^2/7)} = -1,3 \pm 1,45 = \text{ [-2,75 , 0,15]}$$

6.11 Test av parvisa observationer

I många situationer är det naturligt att göra parvisa observationer. Ett enkelt exempel är ”före-efter”-analyser. Det mest exakta sättet att analysera hur en viss parameter påverkas av en viss behandling är att mäta parameterns värde hos *n* slumpmässigt utvalda objekt före behandlingen och sedan mäta igen efter behandlingen för alla *n* objekten. På så vis kan man dra slutsatser om sanna parametervärdet före behandlingen i förhållande till sanna parametervärdet efter behandlingen.

Generellt innebär parvisa observationer att man för ett stickprov om *n* element observerar värdet för samma parameter vid två olika tillfällen. Vid analys av parvisa skillnader utgår man från medelvärdet av de *n* enskilda parvisa observationernas differenser. Vid analys av skillnad mellan två oberoende grupper utgår man, som vi såg i de båda föregående avsnitten, istället från differensen mellan stickprovsmedelvärdena för de båda grupperna som helhet. Som tidigare finns konceptuellt sett tre olika slags nollhypoteser som kan testas, nämligen H0: $\mu_1 - \mu_2 = D$, H0: $\mu_1 - \mu_2 \leq D$, och H0: $\mu_1 - \mu_2 \geq D$, där *D* är just den differens mellan de sanna populationsmedelvärdena som man vill testa för.

I princip görs detta genom att göra en enkel *t*-test på differenserna. Man beräknar alltså den genomsnittliga differensen $\bar{d}$ för de *n* observationerna samt dess standardavvikelse s_d. Testets *t*-värde beräknas sedan som

$$t = \frac{(\bar{d} - D)}{(s_d / \sqrt{n})}$$

med $n - 1$ *df*. Ett $1 - \alpha$ konfidensintervall för den sanna skillnaden $\mu_1 - \mu_2$ kan sedan beräknas som

$$\bar{d} \pm t_{\alpha/2}(s_d / \sqrt{n})$$

Exempel 6.9

En viss affärskedja med sex butiker genomför en reklamkampanj. Omsättningen veckan före respektive veckan efter kampanjen i de olika butikerna framgår av nedanstående tabell:

Butik nr	Före	Efter
1	2,6	3,1
2	2,2	2,3
3	3,7	4,4
4	2,8	2,7
5	2,7	2,9
6	3,1	3,5

Kan man med $\alpha = 0{,}05$ dra slutsatsen att reklamkampanjen har ökat den genomsnittliga omsättningen i butikskedjan? Beräkna även ett 95% konfidensintervall för skillnaden i omsättning före och efter reklamkampanjen.

Lösning

Låt μ vara den sanna omsättningen veckan före, och μ_2 veckan efter. Vi får då hypoteserna

H0: $\mu_2 - \mu_2 \geq 0$
H1: $\mu_2 - \mu_2 < 0$

Differenserna framgår av tabellen nedan:

Butik nr	Före	Efter	Differens
1	2,6	3,1	-0,5
2	2,2	2,3	-0,1
3	3,7	4,4	-0,7
4	2,8	2,7	-0,1
5	2,7	2,9	-0,2
6	3,1	3,5	-0,4

Medelvärdet av differenserna blir

$$\overline{d} = \frac{-0,5 + (-0,1) + ... + (-0,4)}{6} = -0,3$$

med standardavvikelsen

$$s_d = \sqrt{\frac{(-0,5 - (-0,3))^2 + ... + (-0,5 - (-0,4))^2}{6 - 1}} = 0,2898$$

Testets *t*-värde blir då

$$t = \frac{(-0,3 - 0)}{(0,2898 / \sqrt{6})} = -2,5357$$

med 6 – 1 = 5 *df*. Kritiskt *t*-värde avläses från tabell A2 till -2,0150 och H0 ska därmed förkastas eftersom -1,0150 > -2,5357. Ett 95% konfidensintervall blir

$$-0,3 \pm 2,5706 \cdot 0,2898 / \sqrt{6} = -0,3 \pm 0,304 = \text{ [-0,604 , 0,004]}$$

6.12 Test av två populationsproportioner – stora stickprov

Skillnaden mellan två populationsproportioner testas – precis som skillnaden mellan en populationsproportion och en konstant – med *Z* som testvariabel när stickprovsstorleken är stor. Ett vanligt förekommande tillämpningsområde för denna typ av test som de flesta svenskar känner till är opinionsundersökningarna baserade på frågan ”Vad skulle du rösta på om det vore val idag?”. Vid resultatredovisningen av en sådan undersökning sägs ofta att t.ex. en ökning i ett visst parti väljarsympatier är statistiskt säkerställd. Med det avses att man kunde förkasta nollhypotesen att den sanna proportionen som skulle ha röstat på det aktuella partiet var oförändrad eller hade minskat. De tre olika slags nollhypoteser som kan testas är H0: $p_1 - p_2 = D$, H0: $p_1 - p_2 \leq D$, och H0: $p_1 - p_2 \geq D$, där *D* är just den differens mellan de sanna populationsproportionerna som man vill testa för.

Om vi har stickprov från två populationer (eller två olika stickprov från samma population tagna vid olika tillfällen) med stickprovsstorlekarna n_1 och n_2, och där observerat stickprovsproportionerna $\hat{p}_1$ och $\hat{p}_2$, så kan vi beräkna testets *Z*-värde som

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - D}{\sqrt{\hat{p}_1(1-\hat{p}_1)/(n_1 - 1) + \hat{p}_2(1-\hat{p}_2)/(n_2 - 1)}}$$

varefter kritiskt *Z*-värde hämtas från tabell A1 för den valda signifikansnivån. Ett 1 – α konfidens–intervall för den sanna skillnaden $p_1 - p_2$ kan sedan beräknas som

$$(\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2}\sqrt{\hat{p}_1(1-\hat{p}_1)/(n_1 - 1) + \hat{p}_2(1-\hat{p}_2)/(n_2 - 1)}$$

BARCLAYS CAPITAL

Branding

Barclays Capital is a world leading investment bank. It is also a young organisation that has grown rapidly over the last eight years. With the support of a parent bank with a balance sheet of over £520($983) billion, we have an unusual combination of history and dynamic youth. With offices in 22 countries and over 8,000 employees, we are continuing to expand every year.

Internships

An internship is an excellent way for you to gain an understanding of Barclays Capital. There is a range of different internship opportunities for students and they most commonly take place over the summer. Lasting about ten weeks, they are available at both Analyst and Associate level. An internship or placement may lead to an offer of permanent employment on successful completion of your degree.

www.barclayscapital.com/campusrecruitment

Exempel 6.10

17 av 135 slumpmässigt utvalda och undersökta bilar som reparerat avgassystemet i verkstad A fick anmärkning vid kontrollbesiktningen. 24 av 103 undersökta bilar som reparerats i verkstad B fick också anmärkning. kan man dra slutsatsen (på nivån α = 0,05) att den sanna andelen bilar med anmärkning är mer än 2 procentenheter högre i verkstad B? Beräkna ett 95% konfidensintervall för den sanna skillnaden.

Lösning

Vi har hypoteserna

$$\text{H0: } p_1 - p_2 \geq 0$$
$$\text{H1: } p_1 - p_2 < 0$$

och stickprovsproportionerna $\hat{p}_1$ = 17/135 = 0,1259 och $\hat{p}_2$ = 24/103 = 0,2330.

Testets *Z*-värde blir

$$Z = \frac{(0,1259 - 0,233) - (-0,02)}{\sqrt{0,1259(1-0,1259)/(135-1) + 0,233(1-0,233)/(103-1)}}$$

$$= \frac{-0,0871}{0,0507} = -1,72$$

Kritiskt *Z*-värde finns i *Z*-fördelningens vänstra svans eftersom H1 är av karaktären ”mindre än”, och det kan enligt tabell A1 avläsas till -1,64. H0 förkastas därmed eftersom -1,64 > -1,72. Ett 95% konfidensintervall för den sanna skillnaden $p_1 - p_2$ är

$$(0,1259 - 0,233) \pm 1,96 \cdot 0,0507 = -0,1071 \pm 0,0994 \quad = [-0,2065 \text{ , } -0,0077]$$

7. Variansanalys

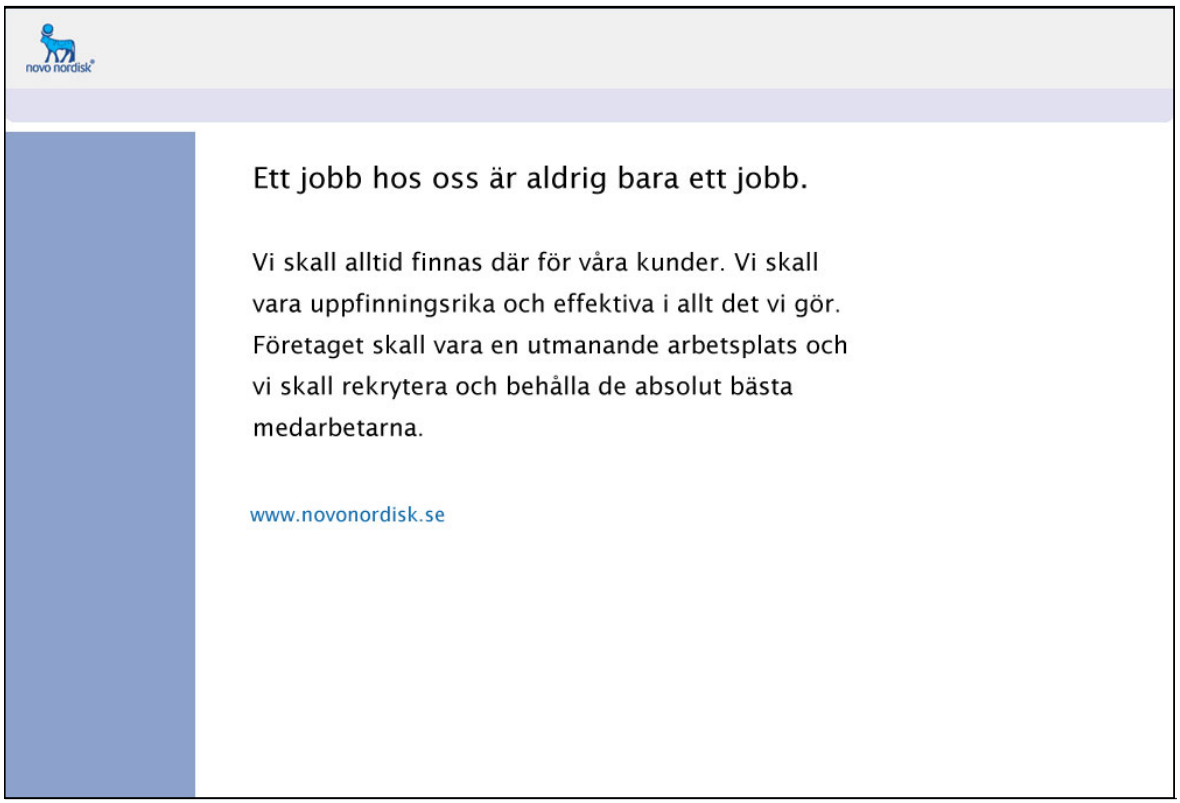
7.1 Inledning

I föregående kapitel visade vi hur nollhypotesen att två populationsmedelvärden är lika kan testas med hjälp av *t*-test. Ofta vill man emellertid jämföra fler än två populationsmedelvärden med varandra för att se om något av dem avviker. Om man har *r* oberoende populationer, och vill testa nollhypotesen att samtliga dessa populationer har samma sanna medelvärde, så blir de formella hypoteserna:

$$H0: \mu_1 = \mu_2 = ... = \mu_r$$
$$H1: \text{Minst två av } \mu_1 = \mu_2 = ... = \mu_r \text{ är olika}$$

Observera att mothypotesen som alltid definieras som komplementet till nollhypotesen. Mothypotesen täcker i detta fall alltså in alla tänkbara situationer där samtliga populationsmedelvärden inte är lika. Det enda slutsats som kan dras om det efter ett test visar sig att H0 ska förkastas är därför att minst två av de sanna populationsmedelvärdena är olika – men ingenting annat. Vi kan alltså inte utan att genomföra fler tester säga något om hur många sådana olikheter som finns, eller vilka av de testade populationsmedelvärdena som ska antas vara olika.

Den typ av test som normalt används för att testa om tre (eller fler) populationsmedelvärden kan antas vara lika är variansanalys (på engelska: ANalysis Of VAriance, ANOVA). Denna metod baseras på antagandet att de *r* populationer som stickproven kommer från är normalfördelade med lika varianser. I en variansanalys delas den totala variansen för hela den mängd enheter som ingår i de tre (eller fler) stickproven upp i varians *inom* respektive population och varians *mellan* populationerna.



Man kan nämligen visa att ju större skillnad i medelvärde som råder mellan olika populationer med samma varians, desto större kommer variationen *mellan* populationerna att vara jämfört med variationen *inom* populationerna. Om det då för den stickprovsdata man har är så att variationen mellan stickproven är stor jämfört med variationen inom stickproven så indikerar det alltså att medelvärdena för populationerna är olika. Detta är den teoretiska grunden för en variansanalys.

7.2 Enkel variansanalys

I en enkel (”ensidig”) variansanalys testas just nollhypotesen H0: $\mu_1 = \mu_2 = ... = \mu_r$. Testen baseras på antagandet att alla *r* populationer som ingår i testen är approximativt normalfördelade med lika varianser.

Generellt beräknas en stickprovsvarians såsom summan av ett antal kvadrerade differenser – en s.k. kvadratsumma – som dividerats med sitt antal frihetsgrader:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

Kvadratsumman för varians inom populationerna, betecknad med SS_W (”Sums of Squares Within”) beräknas genom att beräkna populationsvisa kvadratsummor baserade på de enskilda populationernas medelvärden, varefter summering för samtliga populationer sker. Om vi har *r* populationer med stickprovsstorlekarna n_1, n_2, ..., n_r, så definieras alltså SS_W som

$$SS_W = \sum_{i=1}^{r}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)^2$$

där x_{ij} är observation *j* inom stickprovet från population *i*, och där $\overline{x}_i$ är stickprovsmedelvärdet för population *i*. Stickprovsvariansen inom populationerna får vi sedan genom att dela denna kvadratsumma med antalet frihetsgrader. Vi har en total stickprovsstorlek $n = n_1 + n_2 + ... + n_r$, och vi skattade *r* parametrar (medelvärdet i respektive population) med stickprovsdata för att kunna beräkna kvadratsummorna. Antalet frihetsgrader är därför $n - r$. Variansen inom populationerna, betecknad med MS_W (Mean Square Within) är alltså

$$MS_W = SS_W / (n - r)$$

Kvadratsumman för varians mellan populationerna, betecknad med SS_B (”Sums of Squares Between) fås med liknande resonemang via uttrycket

$$SS_B = \sum_{i=1}^{r} n_i(\overline{x}_i - \overline{x})^2$$

Genom att dividera SS_B med sitt antal frihetsgrader får vi stickprovsvariansen mellan populationerna.

Men hur många frihetsgrader har denna varians? Ja, vi vet att frihetsgrader är additiva, att den totala stickprovsvariansen hade $(n - 1)$ frihetsgrader, och att stickprovsvariansen inom populationena hade $(n - r)$ frihetsgrader. Således måste stickprovsvariansen mellan populationerna ha $(n - 1) - (n - r) = (r - 1)$ frihetsgrader. Variansen mellan populationerna, betecknad med MS_B (Mean Square Between) är alltså

$$MS_B = SS_B /(r-1)$$

Hur används då dessa varianser för att testa nollhypotesen att alla r populationsmedelvärden är lika? Jo, som vi har sett kommer kvoten mellan två oberoende stickprovsvarianser, där de sanna varianserna antas vara lika och populationerna normalfördelade, att följa F-fördelningen. Denna kvot blir därför vår testvariabel. Vårt testvärde beräknas därmed som

$$F = \frac{MS_B}{MS_W}$$

med $r - 1$ och $n - r$ frihetsgrader.

När alla populationsmedelvärden är lika så kommer variansen inom populationerna att vara relativt sett stor jämfört med variansen mellan populationerna. Vi är därför intresserade av det fall där stickprovsvariansen mellan populationerna är relativt sett stor jämfört med stickprovsvariansen inom populationerna, eftersom det tyder på att de sanna populationsmedelvärdena är olika. Det kritiska området av F-fördelningen i en variansanalys måste således vara dess högra svans.

Exempel 7.1

Från produktionslinjerna 1, 2 och 3 på en bilfabrik har stickprov om 5, 4 respektive 4 enheter tagits och antalet defekter per enhet har undersökts. Resultatet framgår av tabellen nedan.

Linje 1	Linje 2	Linje 3
29	41	36
31	38	36
37	39	32
34	40	33
36		

Kan man dra slutsatsen med $\alpha = 0{,}05$ att de olika produktionslinjerna har olika medelvärde för parametern ”antal defekter per enhet”?

Lösning

Vi har tre populationer med total stickprovsstorlek $n = 5 + 4 + 4 = 13$ och hypoteserna

H0: $\mu_1 = \mu_2 = \mu_3$
H1: Minst två av μ_1, μ_2, μ_3 är olika

Vi beräknar först de enkla stickprovsmedelvärdena:

$$\overline{x}_1 = 33{,}4$$
$$\overline{x}_2 = 39{,}5$$
$$\overline{x}_3 = 34{,}25$$
$$\overline{x} = 35{,}54$$

Kvadratsummorna SS_W *och* SS_B kan sedan beräknas:

$$SS_W = (29-33{,}4)^2 + (31-33{,}4)^2 + (37-33{,}4)^2 + (34-33{,}4)^2 + (36-33{,}4)^2 +$$
$$+(41-39{,}5)^2 + (38-39{,}5)^2 + (39-39{,}5)^2 + (40-39{,}5)^2 +$$
$$+(36-34{,}25)^2 + (36-34{,}25)^2 + (32-34{,}25)^2 + (33-34{,}25)^2 = 62{,}95$$

$$SS_B = 5(33{,}4-35{,}54)^2 + 4(39{,}5-35{,}54)^2 + 4(34{,}25-35{,}54)^2 = 92{,}28$$

VOLVO

Volvo – Med dig vid ratten

Volvo strävar efter att vara en världsledande leverantör av kommersiella transportlösningar. För att vi ska nå dit måste vi värna våra anställda och den kompetens de besitter.

Alla medarbetare inom Volvo tillhör ett arbetslag. De tar aktivt del av företagets utveckling, förändringar och framtid. Vi uppmuntrar detta genom att arbeta med energi, passion och respekt för individen.

På den här webbplatsen får du veta mer om vad Volvo har att erbjuda sina medarbetare.

www.volvo.com

Vi får då varianserna

$$MS_W = 62,95/(13-3) = 6,295$$

och

$$MS_B = 92,28/(3-1) = 46,14$$

Testets *F*-värde kan nu beräknas

$$F = \frac{46,14}{6,295} = 7,33$$

med 3 – 1 = 2 och 13 – 3 = 10 frihetsgrader.

Kritiskt *F*-värde avläses från tabell A5 till 4,10. Eftersom 4,10 < 7,33 förkastas H0, och vi drar slutsatsen att minst ett populationsmedelvärde skiljer sig från övriga.

7.3 Uppföljning av enkel variansanalys

Om man har förkastat H0 i en enkel variansanalys så inställer sig ofta frågan om på vad sätt de aktuella populationernas medelvärden inte är lika. Spontant tycker man kanske då att man kan göra *t*-tester för att se vilka skillnader som är signifikanta, men eftersom variansanalysen inkluderar all data i en och samma ”körning” så måste den uppföljande testen också göra det för att vara giltig. Det ligger utanför denna boks syfte att gå igenom testmetoder som uppfyller detta kriterium i detalj, men några exempel på användbara tester i detta sammanhang är Tukeys test och Bonferronimetoden. Läsaren hänvisas till mer avancerad litteratur för beskrivning av dessa.

7.4 Andra typer av variansanalys

Variansanalys kan användas till mycket mer än att bara testa nollhypotesen att ett antal populationer har samma medelvärde. Till exempel kan man testa om andra faktorer än den bara faktorn som definierade populationerna har betydelse för utfallet på stickprovsvariabeln, och/eller om interaktionen mellan olika faktorer har betydelse för utfallet. När man använder variansanalys för att testa två (eller flera) olika faktorers inverkan så talar man om två- (eller fler-) vägs variansanalys. Läsaren hänvisas även här till mer avancerad litteratur.

8. Regressionsanalys

8.1 Inledning

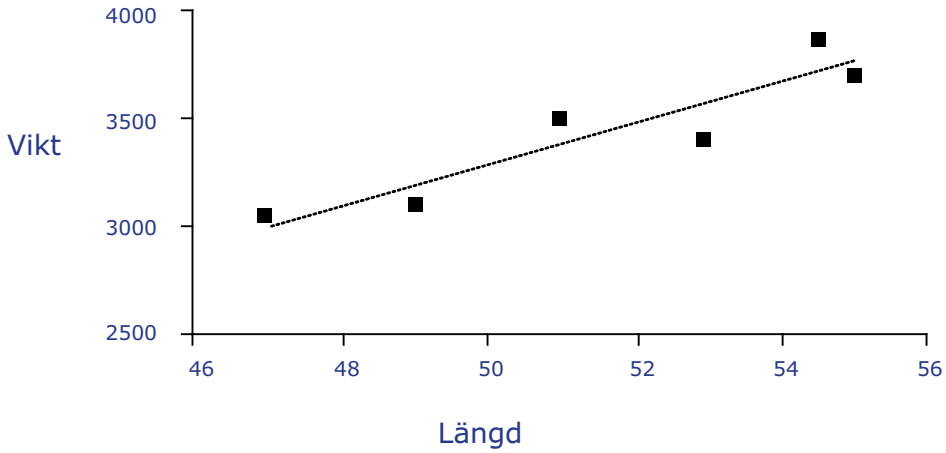
En viktig del av den analytiska statistiken är *sambandsanalys*, d.v.s. att analysera om, och i så fall hur, slumpvariabler *samvarierar* (korrelerar). Ett sådant samband kan analyseras i linjära och/eller icke-linjära termer, och det kan inkludera två eller flera olika slumpvariabler. I det allra enklaste fallet, som vi ska inleda med, studerar vi linjär samvariation mellan två slumpvariabler X och Y. Vi antar alltså att det går att beskriva sambandet mellan X och Y med räta linjens ekvation $Y = \beta_0 + \beta_1 X$, där β_0 är skärningspunkten och β_1 riktningskoefficienten när det linjära sambandet illustreras i ett koordinatsystem.

Ett vanligt exempel på två slumpvariabler som korrelerar linjärt är längden och vikten hos nyfödda barn. I tabellen nedan visas information om längd och vikt för de sex barn som föddes på ett sjukhus en viss dag. Informationen i denna tabell kommer att användas som grund för ett genomgående exempel i detta kapitel.

Barn nr	Längd (cm)	Vikt (g)
1	47	3040
2	49	3100
3	51	3500
4	53	3420
5	54,5	3870
6	55	3710

I fig. 8.1 åskådliggörs sambandet mellan längd och vikt i ett koordinatsystem. Det framgår att ett linjärt samband, illustrerat av den prickade linjen, verkar finnas.

Figur 8.1
Ett linjärt samband



Även om X och Y är slumpvariabler som samvarierar linjärt innebär det dock nästan aldrig att sambandet är *perfekt* linjärt, utan att det ofta finns mindre, eller i enskilda fall även större avvikelser från det linjära sambandet. I figur 8.1 illustreras detta genom att prickarna i diagrammet inte bokstavligt talat ligger på linjen, men väl ganska nära. Vi skriver därför ofta det linjära sambandet mellan X och Y som $Y = \beta_0 + \beta_1 X + \varepsilon$ där ε (den grekiska bokstaven ”eta”) symboliserar slumpavvikelserna i det linjära sambandet.

T-Mobile culture and values

T-Mobile Jobs
Find your calling.™

Thinking Big...
T-Mobile's success and growth has been phenomenal and we're adding to our ranks, with 25,000 employees nationwide. Our potential is only limited by our imagination...which, with the kind of thinkers we have around here, is pretty much limitless. Click here to find out more about our economic history, quick facts, and press releases.

...**And Acting Small**
As we grow, we've made a commitment to not lose sight of the reason for our success in the first place...our customers and the people who work here.

www.t-mobile.com/jobs

8.2 Covarians och korrelationskoefficienten

Det enklaste statistiska måttet på linjärt samband mellan två slumpvariabler *X* och *Y* är variablernas *covarians*. Den sanna covariansen *Cov(X,Y)* för ett par av slumpvariabler *X* och *Y* definieras som det förväntade värdet av produkten av respektive slumpvariabels avvikelse från dess medelvärde, d.v.s.

$$Cov(X,Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

där μ_X är det sanna populationsmedelvärdet för *X*, och μ_Y det sanna populationsmedelvärdet för *Y*.

När *Cov*(*X,Y*) har ett positivt värde så innebär det att när *X* ökar så ökar (i genomsnitt) även *Y*, medan ett negativt värde på *Cov*(*X,Y*) innebär att när *X* ökar så minskar (i genomsnitt) *Y*. När *Cov*(*X,Y*) är noll så finns inget linjärt samband mellan variablerna.

En mer användbar tolkning erhålls om *Cov*(*X,Y*) divideras med produkten av de båda slumpvariablernas standardavvikelser. Då erhålls den sanna *korrelationskoefficienten*, betecknad med ρ (uttalas ”rå”) för *X* och *Y*. För korrelationskoefficienten

$$\rho = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

där σ_X är den sanna populationsstandardavvikelsen för *X*, och σ_Y den sanna populationsstandardavvikelsen för *Y*, gäller nämligen att dess värde alltid hamnar mellan 1 och -1. Om ρ har ett värde nära 0 så finns inget linjärt samband mellan *X* och *Y*. Om ρ ligger nära 1 så är det linjära sambandet mellan *X* och *Y* starkt positivt. Om ρ å andra sidan ligger nära -1 så är det linjära sambandet mellan *X* och *Y* starkt negativt.

I praktisk statistisk analys beräknas naturligtvis dessa och andra relaterade mått rörande variabelsambandet med stickprovsdata. Precis som för variansanalys baseras denna analys på kvadratsummor. Vi definierar därför *X*-värdenas kvadratsumma SS_X, *Y*-värdenas kvadratsumma SS_Y och produktsumman SS_{XY} på följande sätt baserat på *n* par av observerade värden för *X* och *Y*:

$$SS_X = \sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

$$SS_Y = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{\left(\sum_{i=1}^{n} y_i\right)^2}{n}$$

$$SS_{XY} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

Covariansen för *X* och *Y* skattas med hjälp av stickprovsdata som SS_{XY} /(n – 1). Från den vanliga definitionen av stickprovsstandardavvikelsen *s* inser man att σ_X skattas med hjälp av stickprovsdata som $\sqrt{SS_X/(n-1)}$, och σ_Y analogt som $\sqrt{SS_Y/(n-1)}$. Det innebär att den med stickprovsdata

skattade korrelationskoefficienten för X och Y, betecknad med r, blir

$$r = \frac{SS_{XY}/(n-1)}{\sqrt{SS_X/(n-1)}\sqrt{SS_Y/(n-1)}} = \frac{SS_{XY}}{\sqrt{SS_X SS_Y}}$$

Kvadraten på korrelationskoefficienten, r^2, har också ett speciellt namn och en speciell betydelse. Den kallas determinationskoefficient, och kan ses som ett deskriptivt mått som visar ”andelen variation i de observerade värdena på Y som förklaras av det bästa linjära sambandet mellan X och Y”. Vi ska dock inte gå in i detalj på detta mått här. Däremot kan r och r^2 användas för att testa någon av nollhypoteserna H0: $\rho = 0$, H0: $\rho \leq 0$ eller H0: $\rho \geq 0$. Om någon av dessa hypoteser kan förkastas så dras ju slutsatsen att det finns ett sant linjärt samband mellan variablerna. Testet baseras på t-fördelningen, och testvariabelns värde beräknas som

$$t = \frac{r}{\sqrt{(1-r^2)/(n-2)}}$$

med $n - 2$ *df*.

Exempel 8.1

Utgå från tabelldatan som fig. 8.1 baserades på, och anta att dessa kan ses som ett stickprov från populationen ”nyfödda barn i allmänhet”. Beräkna SS_X, SS_Y och SS_{XY}, och använd dessa för att beräkna r. Testa sedan om man med $\alpha = 0{,}01$ kan dra slutsatsen att det verkligen finns ett positivt linjärt samband mellan längd och vikt för nyfödda barn.

Lösning

$$SS_X = 47^2 + 49^2 + \ldots + 55^2 - \frac{(47+49+\ldots+55)^2}{6} = 50{,}21$$

$$SS_Y = 3040^2 + 3100^2 + \ldots + 3710^2 - \frac{(3040+3100+\ldots+3710)^2}{6} = 537400$$

$$SS_{XY} = 47\cdot 3040 + 49\cdot 3100 + \ldots + 55\cdot 3710 - \frac{(47+49+\ldots+55)(3040+3100+\ldots+3710)}{6} = 4825$$

varefter vi kan beräkna

$$r = \frac{4825}{\sqrt{50{,}21\cdot 537400}} = 0{,}9289$$

Vi får hypoteserna

$$\begin{aligned} &\text{H0: } \rho \leq 0 \\ &\text{H1: } \rho > 0 \end{aligned}$$

och testets *t*-värde

$$t = \frac{0,9289}{\sqrt{(1-0,9289^2)/(6-2)}} = 5,02$$

med 6 – 2 = 4 *df*. Testet görs i *t*-fördelningens högra svans, varvid kritiskt *t*-värde avläses från tabell A2 till 3,7469. Således förkastas H0 eftersom 3,7469 < 5,02. Vi drar slutsatsen att det finns ett positivt linjärt samband mellan längd och vikt för nyfödda barn.

8.3 Minstakvadrat-metoden för en regressionslinje

När man med hjälp av stickprov har konstaterat att det finns ett signifikant linjärt samband mellan två variabler *X* och *Y* så är man ofta intresserad av att kartlägga närmare hur detta samband ser ut. Den linje som ”bäst” beskriver det sanna linjära sambandet kallas *regressionslinje*, och skrivs som tidigare sagts $Y = \beta_0 + \beta_1 X + \varepsilon$ när *Y* kan antas vara *beroende* av *X*. Enligt den s.k. *minstakvadrat-metoden* åstadkommer vi skattningar av β_0 och β_1, vilka betecknas med b_0 och b_1, med hjälp av kvadrat- och produktsummorna som vi använde tidigare:

$$b_1 = \frac{SS_{XY}}{SS_X}$$

$$b_0 = \overline{y} - b_1\overline{x}$$

På det sättet får vi den skattade regressionslinjen $Y = b_0 + b_1 X + e$, där e symboliserar de ”fel” som kan observeras när linjen $b_0 + b_1 X$ anpassas till de n datapunkterna.

De antaganden som måste göras för att man ska kunna använda minstakvadrat-metoden är ganska komplicerade, så vi ska inte förklara dem i detalj. De är

- ε är en normalfördelad slumpvariabel med medelvärdet 0.
- ε har en varians som är konstant för olika X-värden, d.v.s. datapunkternas genomsnittliga avvikelser från regressionslinjen påverkas inte av att X ändras.
- Det råder oberoende mellan alla ε -värden

När vi säger att resultatet av minstakvadrat-metoden är den ”bästa” linjen så menar vi mer konkret att det är den linje för vilken summan av alla kvadrerade avvikelser i Y-led mellan linjen och de enskilda punkterna, d.v.s. summan av alla kvadrerade ”fel”, minimeras. Om vi definierar kvadratsumman SS_E som

$$SS_E = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2 = SS_Y - \frac{(SS_{XY})^2}{SS_X}$$

så innebär minstakvadrat-metoden formellt att SS_E minimeras.

Exempel 8.2

Utgå från exempel 8.1, och skatta regressionslinjen med minstakvadrat-metoden. Beräkna också SS_E.

Lösning

$$b_1 = \frac{4825}{50,21} = 96,1$$

$$b_0 = \frac{3040 + 3100 + ... + 3710}{6} - 96,1 \cdot \frac{47 + 49 + ... + 55}{6} = -1517,1$$

Den skattade regressionslinjen blir alltså $Y = -1517,1 + 96,1X + e$.

$$SS_E = SS_Y - \frac{(SS_{XY})^2}{SS_X} = 537400 - \frac{4825^2}{50,21} = 73734,9$$

8.4 Konfidensintervall för regressionsparametrarna

Skattningarna b_0 och b_1 är naturligtvis punktskattningar av de sanna parametrarna β_0 och β_1. Det naturliga nästa steget i analysen är därmed att göra motsvarande intervallskattningar. Med andra ord, hur konstrueras konfidensintervall för β_0 och β_1?

För att kunna svara på det måste vi veta hur medelfelen för b_0 och b_1 beräknas. Vi kallar dessa medelfel för $s(b_0)$ och $s(b_1)$, och man kan visa att formlerna för dessa storheter blir

$$s(b_0) = \sqrt{\frac{SS_E \sum_{i=1}^{n} x_i^2}{n(n-2)SS_X}}$$

och

$$s(b_1) = \sqrt{\frac{SS_E}{(n-2)SS_X}}$$

Med dessa medelfel kan konfidensintervall för β_0 och β_1 enkelt konstrueras. Ett $(1 - \alpha)$ konfiden–sintervall för β_0 blir

$$b_0 \pm t_{\alpha/2} s(b_0)$$

med $n - 2$ *df*.

Ett $(1 - \alpha)$ konfidensintervall för β_1 blir

$$b_1 \pm t_{\alpha/2} s(b_1)$$

med $n - 2$ *df*.

Exempel 8.3

Utgå från exempel 8.2, och beräkna 95% konfidensintervall för de sanna regressionslinje-parametrarna.

Lösning

Vi börjar med att beräkna medelfelen för skattningarna. De blir

$$s(b_0) = \sqrt{\frac{73734,9(47^2 + 49^2 + ... + 55^2)}{6(6-2)50,21}} = 989,9$$

och

$$s(b_1) = \sqrt{\frac{73734,9}{(6-2)50,21}} = 19,16$$

Konfidensintervallen blir då

$$-1517{,}1 \pm 2{,}7765 \cdot 989{,}9 = \text{ [- 4265,56 , 1231,36]}$$

och

$$96{,}1 \pm 2{,}7765 \cdot 19{,}16 = \text{ [42,90 , 149,30]}$$

8.5 Hypotestest för regressionsparametrarna

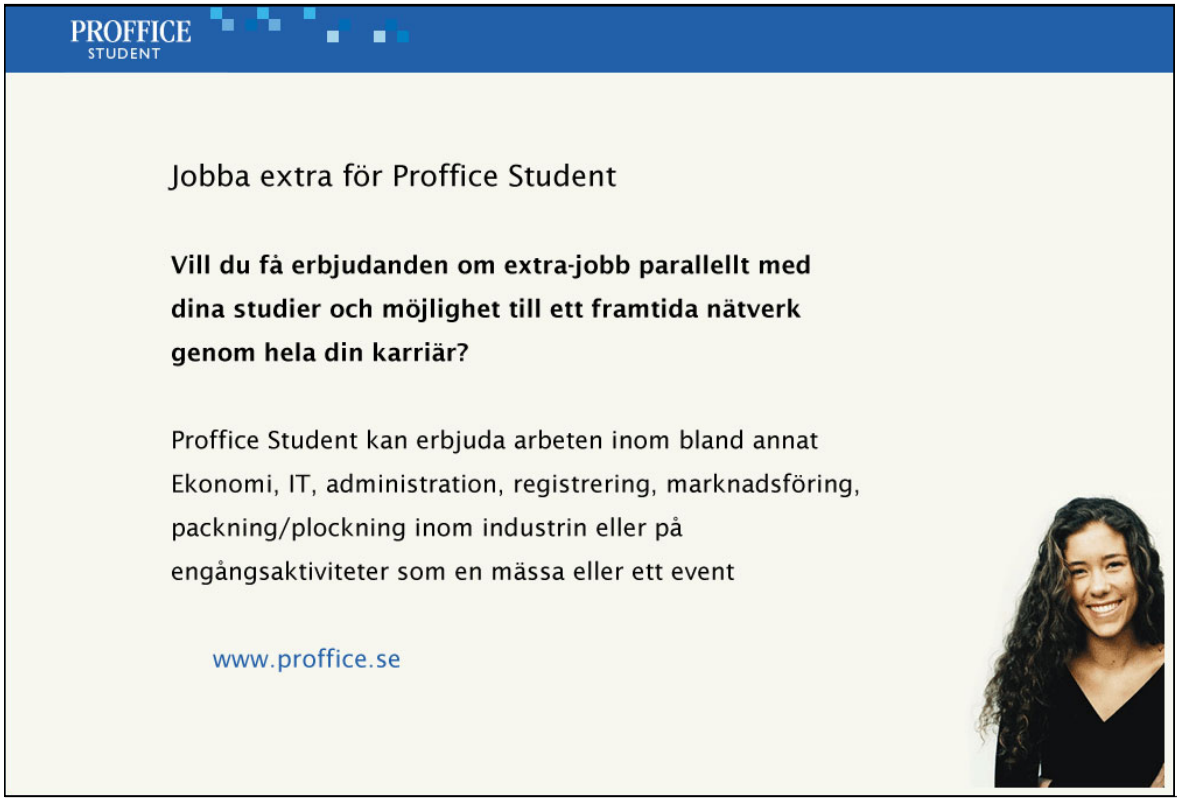
Den enda *slutsats* vi har dragit om regressionssambandet i det genomgående exemplet så här långt är att det finns ett positivt samband, i och med att vi i exempel 8.1 kunde dra slutsatsen att $r > 0$. Det är ofta en väsentlig del av en regressionsanalys att testa om de sanna parametrarna, särskilt β_1, är signifikant skilda från 0. Detta görs med enkla *t*-tester. Testernas *t*-värden beräknas då enligt

$$t = \frac{b_0}{s(b_0)}$$

för parametern β_0, och med

$$t = \frac{b_1}{s(b_1)}$$

för parametern β_1, i båda fallen med $n - 2$ *df*.



Exempel 8.4

Utgå från exempel 8.3. Testa om båda parametrarna är signifikant skilda från 0, i båda fallen med $\alpha = 0{,}01$.

Lösning

Vi har först nollhypotesen H0: $\beta_0 = 0$. Testets *t*-värde blir

$$t = \frac{b_1}{s(b_1)}$$

Testet är tvåsidigt, och enligt tabell A2 är de kritiska *t*-värdena ± 2,7765. H0 kan alltså inte förkastas, eftersom -2,7765 < -1,53 < 2,7765.

Vi har sedan nollhypotesen H0: $\beta_1 = 0$. Testets *t*-värde blir

$$t = \frac{-1517{,}1}{989{,}9} = -1{,}53$$

Testet är även här tvåsidigt, och de kritiska *t*-värdena är fortfarande ± 2,7765. H0 förkastas då, eftersom 2,7765 < 5,02. Vi drar slutsatsen att parametern β_1 är signifikant skild från 0.

8.6 Prediktionsintervall vid extrapolering av regressionslinjen

Ett syfte med regressionsanalys är ofta att ta fram en prognosmodell, d.v.s. en modell med vilken man på basis av hur en viss variabel har utvecklat sig historiskt försöker förutspå hur den kommer att utveckla sig i framtiden. Denna del av den statistiska analysen kallas *tidsserieanalys*. Den grundläggande idén är att helt enkelt att anta att det linjära samband man har modellerat kommer att fortsätta gälla. När man extrapolerar en regressionslinje (d.v.s. ”drar ut linjen” så att den hamnar ”utanför” den datamängd man använde för att ta fram regressionslinjen) så måste man dock alltid vara medveten om att det linjära sambandet kanske inte gäller längre.

Ett punktskattat *y*-värde, betecknat med $\hat{y}$, baserat på ett specifikt *x*-värde fås enkelt genom att sätta in det specifika *x*-värdet i regressionslinjens ekvation, d.v.s.

$$\hat{y} = b_0 + b_1 x$$

Osäkerheten i en skattning av denna typ beror på hur långt linjen extrapoleras, och vi kan beräkna ett *prediktionsintervall* inom vilket *y*-värdet faktiskt kommer att ligga med sannolikheten $1 - \alpha$ såsom

$$\hat{y} \pm t_{\alpha/2}\sqrt{\left(\frac{SS_E}{(n-2)}\right)\left(1+\frac{1}{n}+\frac{(x-\overline{x})^2}{SS_X}\right)}$$

med *n*-2 *df*.

Exempel 8.5

Utgå från exempel 8.1 – 8.4. Beräkna ett 95% prediktionsintervall för vikten på ett barn som är 56 cm långt.

Lösning

Vi har stickprovsmedelvärdet för längden

$$\bar{x} = (47 + 49 + ... + 55) / 6 = 51{,}18$$

och punktskattningen för vikten vid längden 56 cm blir

$$\hat{y} = -1517{,}1 + 96{,}1 \cdot 56 = 3864{,}5 \text{ gram}$$

Vi får sedan prediktionsintervallet

$$3864{,}5 \pm 2{,}7765\sqrt{\left(\frac{73734{,}9}{(6-2)}\right)\left(1+\frac{1}{6}+\frac{(56-51{,}18)^2}{51{,}58}\right)} = [3385,\ 4344] \text{ gram,}$$

inom vilket ett barn med längden 56 cm ligger med 95% säkerhet om det linjära sambandet kunde antas fortsätta gälla även för högre x-värden än dem vi använde för att ta fram regressionslinjen.

8.7 Multipel regression

Vi har hittills antagit att värdet på slumpvariabeln Y beror linjärt på värdet av *en* slumpvariabel X. Logiken bakom analysen går att generalisera så att Y kan antas vara linjärt beroende av värdet på k olika slumpvariabler $X_1, X_2, ..., X_k$. Regressionsmodellen blir då

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$$

och vi talar då om *multipel* linjär regressionsanalys till skillnad från *enkel* linjär regressionsanalys som baseras på endast en oberoende variabel X. Det primära problemet i en multipel regressionsanalys är naturligtvis att skatta $\beta_1, \beta_2, ..., \beta_k$. Vi ska inte gå in i detalj på hur man räknar för att ta fram dessa skattningar, men den grundläggande logiken är densamma: Det gäller att hitta det linjära uttryck som minimerar summan av de kvadrerade avvikelserna. Man använder normalt sett alltid datorstöd när man arbetar med multipel regression.

8.8 Polynom regression

En regressionsmodell behöver inte nödvändigtvis baseras på antaganden om linjärt samband. En enkel plott av den data man har kan indikera att sambandet mellan X och Y är icke-linjärt, och då måste man naturligtvis ta hänsyn till detta. Om det t.ex. är rimligt att anta att sambandet kan beskrivas med en andragradsfunktion så kommer regressionsmodellen att få formen

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

Regressionsanalys baserade på icke-linjära modeller brukar betecknas *polynom* regression. Även här används normalt alltid datorstöd för beräkningar.

8.9 Dummyvariabler

En användbar teknik i multipel regression är att använda en variabel som symboliserar huruvida ett visst *villkor* är uppfyllt eller ej. Ett enkelt exempel kan vara om vi använder regressionsanalys för att bestämma hur de mätbara variablerna vikt och cylindervolym för bilar påverkar bilarnas acceleration. Om vissa av bilarna då har turbo så kan vi använda en tredje variabel för att ta hänsyn till detta. Vi låter först variablerna X_1 och X_2 symbolisera vikt och cylindervolym på vanligt sätt. Sedan inför vi variabeln X_3 som får symbolisera närvaron av turbo. För bilar som har turbo så får variabeln X_3 då värdet 1, för övriga bilar får X_3 värdet 0. Med regressionsmodellen

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

så kommer parametern β_3 då att visa hur bilarnas acceleration i medeltal påverkas av det faktum att turbo är installerat, allt annat lika. Variabeln X_3 kallas för *dummyvariabel*, eller *binär* variabel, i denna modell, eftersom den symboliserar huruvida ett visst villkor är uppfyllt och därmed endast kan anta värdet 0 eller 1. Räknetekniskt fungerar allting på precis samma sätt oavsett om en variabel är en dummyvariabel eller en vanlig (kvantitativ) variabel.

SCRIBONA

Att arbete på Scribona

Scribona erbjuder en modern arbetsplats för medarbetare som vill överträffa kundernas förväntningar.

Vi som arbetar på Scribona är resultatinriktade människor som vill utvecklas i vår yrkesroll och som gillar utmaningar. Vi trivs att arbeta i grupp och sätter gruppens gemensamma resultat framför våra egna.

Scribona präglas av hög kompetens med spetskunskap inom många olika teknik- och produktområden. Vi har program för kompetensutveckling med bland annat kurser, regelbundna utvecklingssamtal och projekt för intern kompetensöverföring.

www.scribona.se

8.10 Multicolinjäritet

När man arbetar med multipel regression så antar man generellt att de olika *X*-variablerna inte korrelerar med varandra. Två variabler X_1 och X_2 korrelerar perfekt linjärt om sambandet dem emellan går att uttrycka

$$X_1 = a_1 + a_2 X_2$$

där a_1 och a_2 är konstanter. X_1 och X_2 sägs då vara *colinjära*. När inbördes colinjäritet finns mellan flera *X*-variabler i en multipel regressionsmodell kännetecknas modellen av *multicolinjäritet*. Multicolinjäritet är ett problem i multipel regression, eftersom *X*-variablerna modellmässigt var och en för sig antas bidra med information om hur *Y* varierar. Om X_1 då fullt ut förklaras av värdet på X_2 så bidrar ju inte X_1 med någon förklarande effekt för *Y*. Tvärtom kommer närvaron av X_1 i regressionsmodellen att skapa snedvridande effekter i regressionsanalysen. Exempel på effekter som kan uppstå är följande:

- Konfidensintervallen för regressionsmodellens parametrar blir mycket stora.
- En eller flera parametrar i regressionsmodellen får värden som uppenbart är orimliga med avseende på storlek och/eller tecken.
- Uteslutande av en variabel från modellen ger stor inverkan på värdet för andra variablers parametrar.

När man misstänker multicolinjäritet är det viktigt att man försöker utreda vilka variabler som är colinjära, och utesluter variabler från modellen så att inga colinjära variabelrelationer återstår.

8.11 *F*-test av regressionssamband

Det generella sättet att testa huruvida en regressionsmodell är signifikant, d.v.s. om regressionssambandet mellan *Y* och de aktuella *X*-variablerna är tillräckligt ”bra” för att man ska kunna hävda att modellen faktiskt förklarar sambandet mellan variablerna, är att göra en *F*-test. Generellt innebär detta att man genomför hypotestesten

H0: $\beta_1 = \beta_2 = ... = \beta_k = 0$
H1: Minst en av $\beta_1, \beta_2, ..., \beta_k$ är skild från 0

för en regressionsmodell med *k X*-variabler. Testet görs i *F*-fördelningens högra svans. I fallet med en enda *X*-variabel så är denna test ekvivalent med den *t*-test vi såg tidigare i kapitlet, och vi kommer att illustrera tillvägagångssättet baserat på enkel linjär regressionsanalys.

Testets *F*-värde beräknas då som

$$F = \frac{SS_Y - SS_E}{SS_E/(n-2)}$$

med 1 och $n - 2$ frihetsgrader.

Exempel 8.6

Utgå från exempel 8.1 – 8.5. Kontrollera med en *F*-test om regressionsmodellen är signifikant på nivån 1%.

Lösning

Vi beräknar *F*-värdet för testen:

$$F = \frac{537400 - 73734,9}{73734,9/(6-2)} = 25,15$$

Kritiskt *F* för $\alpha = 0,01$ samt 1 och 6 – 2 = 4 frihetsgrader är enligt tabell A7 lika med 21,20. Således kan H0 förkastas eftersom 21,20 < 25,15.

9. Chitvå-tester

9.1 Inledning

I kapitlet om hypotestest introducerades chitvå-fördelningen, vilken användes för att testa hur en populationsvarians förhöll sig till ett visst värde. Chitvå-fördelningen kan även användas till andra typer av tester som är mycket användbara. I detta kapitel ska vi titta på några sådana tester.

Den grundläggande idén är densamma i samtliga tester. Vi har stickprovsdata i termer av *observerade frekvenser* – d.v.s. antal faktiskt observerade utfall – för var och en av ett antal möjliga kategorier. Vår data är alltså alltid kvalitativ eller diskret kvantitativ. En chitvå-test baseras sedan på att man beräknar *förväntade frekvenser* för var och en av de möjliga kategorierna under förutsättning att den nollhypotes man vill testa faktiskt är sann. Om de observerade frekvenserna skiljer sig tillräckligt mycket från de förväntade så drar man slutsatsen att nollhypotesen är falsk. Den testvariabel som utnyttjas är genomgående χ^2. Varje sådan test görs i chitvå-fördelningens högra svans.

Som kommer att framgå längre fram är det ibland så att det är oklart hur många kategorier man bör arbeta med, och/eller att användaren själv kan definiera kategorierna. Det är då viktigt att tänka på att en chitvå-test förutsätter att den förväntade frekvensen i var och en av de kategorier som används är tillräcklig stor. En tumregel är att den förväntade frekvensen inte ska understiga 5 i någon kategori.

9.2 Test av anpassningsgrad – diskret fördelning

Som har framgått av tidigare kapitel görs ofta ett antagande om vilken fördelning stickprovsdata hämtats från. Det är ofta möjligt att testa om stickprovsdata är tillräckligt väl anpassad till en viss teoretisk fördelning för att ett sådant antagande ska kunna anses vara rimligt eller ej. Man talar då om ett *test av anpassningsgrad*. (på engelska: test of goodness-of-fit).

Vid test av anpassningsgrad gäller generellt att antalet frihetsgrader i respektive fall är lika med antalet kategorier minus 1. Om man måste skatta parametervärden av något slag med hjälp av stickprovsdata för att kunna genomföra testen reduceras antalet frihetsgrader därutöver med 1 per sådan parameter.

Testets χ^2-värde beräknas som

$$\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{e_i}$$

där k är antalet kategorier, o_i den observerade frekvensen för kategori i, och e_i den förväntade frekvensen för kategori i under förutsättning att nollhypotesen är sann.

Exempel 9.1

I en marknadsundersökning fick 140 slumpmässigt utvalda personer blindtesta fyra olika sorters läsk, betecknade A, B, C och D, varefter de bestämde vilken läsk de föredrog. Resultatet framgår av nedanstående tabell.

Läsk	A	B	C	D
Observerad frekvens	34	45	42	19

Testa om man med $\alpha = 0{,}01$ kan dra slutsatsen att smaken är olika.

Lösning

Vi har hypoteserna

H0: Smaken är lika
H1: Smaken är olika

Om smaken faktiskt är lika, som H0 säger, så måste den förväntade frekvensen vara densamma i alla fyra kategorierna. Med 140 ”försökspersoner” blir det således 140 / 4 = 35 per kategori. Vi kan således beräkna testets χ^2-värde som

$$\chi^2 = \frac{(34-35)^2}{35} + \frac{(45-35)^2}{35} + \frac{(42-35)^2}{35} + \frac{(19-35)^2}{35} = 11{,}6$$

Kritiskt χ^2-värde med 4 – 1 = 3 frihetsgrader är 11,3449 enligt tabell A3. Eftersom 11,3449 < 11,6 förkastas H0.

Exempel 9.2

Antalet betjänade kunder i en viss skönhetssalong per timme under en vecka har observerats under de 55 öppettimmar salongen haft under en vecka. Resultatet framgår av tabellen nedan.

Antal kunder	0	1	2	3	4	5
Observerad frekvens	3	7	11	14	15	5

Testa med α = 0,05 om poissonfördelning kan antas föreligga för slumpvariabeln ”antal betjänade kunder”.

Lösning

Vi har hypoteserna

H0: Poissonfördelning föreligger
H1: Poissonfördelning föreligger inte

Om poissonfördelning föreligger så måste vi ha fördelningens väntevärde (dess enda parameter) för att kunna beräkna förväntade frekvenser. Hypoteserna säger inget om vad detta väntevärde ska antas vara, så vi måste beräkna det genomsnittliga antalet kunder per timme med hjälp av stickprovsdatan till

$$\overline{x} = (0 \cdot 3 + 1 \cdot 7 + 2 \cdot 11 + 3 \cdot 14 + 4 \cdot 15 + 5 \cdot 5) / 55 = 2,8364$$

varvid vi har förlorat en frihetsgrad. Med hjälp av den vanliga poissonformeln (se kapitel 3) beräknar vi sedan de förväntade frekvenserna. Resultatet framgår av nedanstående tabell.

Antal kunder	0 eller 1	2	3	4	5 eller fler
Observerad frekvens	10	11	14	15	5
Förväntad frekvens	12,37	12,97	12,26	8,70	8,69

Observera att vi måste slå ihop kategorierna 0 och 1 kunder till en kategori, liksom alla kategorier över 4 kunder , eftersom vi annars hade haft kategorier med förväntad frekvens under 5.

Vi kan nu beräkna testets χ^2-värde som

$$\chi^2 = \frac{(10-12,37)^2}{12,37} + \frac{(11-12,97)^2}{12,97} + \frac{(14-12,26)^2}{12,26} + \frac{(15-18,7)^2}{8,7} + \frac{(5-8,69)^2}{8,69} = 7,13$$

Kritiskt χ^2-värde med 5 – 1 – 1 = 3 frihetsgrader är 7,81472 enligt tabell A3. Eftersom 7,81472 > 7,13 kan H0 inte förkastas med $\alpha = 0,05$.

9.3 Test av anpassningsgrad – kontinuerlig fördelning

I exempel 9.2 i ovanstående avsnitt var det relativt enkelt att definiera de kategorier som skulle användas, eftersom det rörde sig om en *diskret* fördelning. Det var då bara en fråga om att se till så att inget av de möjliga diskreta utfallen hade en förväntad frekvens som understeg 5.

Om man vill testa om stickprovsdata kan antas komma från en *kontinuerlig* fördelning finns per definition ett oändligt antal utfall, och man måste som användare då själv dela in utfallsrummet i intervaller för att åstadkomma de kategorier som chitvå-testen kräver. Huvudregeln är att man ska försöka definiera kategorierna så att den förväntade frekvensen i respektive kategori blir ungefär lika. När det gäller antalet kategorier så får man i varje enskilt fall göra en avvägning mellan det positiva faktum att många kategorier innebär fler frihetsgrader, och det negativa faktum att fler kategorier leder till att förväntad frekvens per kategori sjunker.

Exempel 9.3

Ett slumpmässigt urval om 40 kunder i en butik kartlades i termer av bland annat den totalsumma de handlade för. Resultatet (i euro per kund) framgår nedan:

31,70	6,00	32,80	22,80
22,70	31,60	28,00	33,10
34,20	38,60	36,20	24,90
29,10	33,10	49,60	61,00
34,00	23,60	26,80	31,70
15,70	9,90	17,90	56,60
46,50	43,10	35,70	29,50
38,90	52,90	27,80	35,80
12,40	42,10	37,60	35,10
17,00	32,10	43,70	24,20

Kan köpesumman per kund antas vara en normalfördelad slumpvariabel? Använd $\alpha = 0,1$.

Lösning

Vi har hypoteserna

H0: Normalfördelning föreligger
H1: Normalfördelning föreligger inte.

För att kunna konstruera intervaller som är ungefär lika sannolika krävs att vi har information om medelvärde och standardavvikelse. Från stickprovsdatan får vi $\overline{x}$ = 32,15 och s = 12,00, vilka vi använder som skattningar av μ och σ. Vi skattar alltså två parametrar med stickprovsdata, vilket kostar två frihetsgrader.

Hur många kategorier ska vi använda? Eftersom normalfördelningen är symmetrisk kan det vara vettigt med ett jämnt antal. Fler än 8 är dock inte ens teoretiskt möjligt, eftersom 40 / 8 = 5, vilket är absolut minimum för förväntad frekvens per kategori. För säkerhets skull kan vi använda 6 kategorier, så har vi lite marginal vad gäller förväntade frekvenser. I så fall ska kategorierna definieras så att varje kategori kan förväntas innehålla ungefär 1 / 6 = 0,1667 av den totala stickprovsstorleken. Från tabell A1 kan vi avläsa att $P(Z > 1) = 0{,}1587$ och således vet vi också att $P(Z < -1) = 0{,}1587$. Vi ser också att $P(0 < Z < 0{,}44) = 0{,}3300 - 0{,}1585 = 0{,}1713$ och således även att $P(0 > Z > -0{,}44) = 0{,}1713$. Avslutningsvis har vi att $P(0{,}44 < Z < 1) = 0{,}5 - 0{,}1713 - 0{,}1585 = 0{,}17$ och därmed även att $P(-0{,}44 > Z > -1) = 0{,}17$. Därmed har vi delat in normalfördelningen i sex intervaller av ungefär samma storlek. Med omvänd Z-transformation hittar vi enkelt de fem gränserna för indelning i dessa sex kategorier enligt H0:

1. $32{,}15 - 1 \cdot 12 = 20{,}15$

2. $32{,}15 - 0{,}44 \cdot 12 = 26{,}87$

3. $32{,}15$

4. $32{,}15 + 0{,}44 \cdot 12 = 37{,}43$

5. $32{,}15 + 1 \cdot 12 = 44{,}15$

Vi kan nu räkna observerade och förväntade frekvenser för de olika kategorierna sammanfatta det hela i tabellform:

Intervall	Observerade frekvenser	Förväntade frekvenser
< 20,15	6	0,1587·40 = 6,348
20,15 – 26,87	6	0,1713·40 = 6,852
26,87 – 32,15	8	0,17·40 = 6,8
32,15 – 37,43	9	0,17·40 = 6,8
37,43 – 44,15	6	0,1713·40 = 6,852
> 44,15	5	0,1587·40 = 6,348

Vi kan nu beräkna testets χ^2-värde som

$$\chi^2 = \frac{(6-6,348)^2}{6,348} + \frac{(6-6,852)^2}{6,852} + \frac{(8-6,8)^2}{6,8} + \frac{(9-6,8)^2}{6,8} + \frac{(6-6,852)^2}{6,852} + \frac{(5-6,348)^2}{6,348} = 1,44$$

Kritiskt χ^2-värde med $\alpha = 0,1$ och med 6 – 1 –2 = 3 frihetsgrader är 6,25139 enligt tabell A3. Eftersom 6,25139 > 1,44 kan H0 inte förkastas.

9.4 Korstabellanalys

Ett vanligt sätt att presentera resultatet av en undersökning där respondenterna får svara på flera olika frågor är att använda korstabeller för att ge en översiktlig bild av hur de som svarar på en viss fråga tenderar att svara på en annan fråga. Om 100 slumpmässigt utvalda människor tillfrågades om dels kön och dels om sin uppfattning när det gäller EU så skulle resultatet kunna beskrivas som i korstabellen nedan.

	Män	Kvinnor
Positiv till EU	34	15
Negativ till EU	29	22

Kan man på basis av denna tabell utan vidare dra slutsatsen att män generellt är mer positiva till EU? Nej, naturligtvis inte utan att genomföra någon form av analys som pekar på att det är så.

Standardmetoden för att analysera huruvida en klassificeringskategori (i detta fall: kön) är *oberoende* av en annan (i detta fall: uppfattning om EU) är att använda en chitvå-test. Testen baseras på den sannolikhetsteoretiska definitionen av oberoende. I kapitel 2 definierades två händelser A och B som oberoende om $P(A \cap B) = P(A) \cdot P(B)$.

Notera att varje ”ruta” i korstabellen hör till en viss rad och en viss kolumn. Om klassificering skategorierna är oberoende enligt $P(A \cap B) = P(A) \cdot P(B)$ så ska den förväntade frekvensen i en viss ruta vara lika med summan av radens observerade frekvenser multiplicerat med summan av kolumnens observerade frekvenser dividerat med den totala stickprovsstorleken. När man på det sättet har tagit fram observerade och förväntade frekvenser så kan man använda en chitvå-test för att testa nollhypotesen att klassificeringskategorierna är oberoende på det vanliga sättet. Den enda skillnaden jämfört med tidigare är att antalet frihetsgrader beräknas som

$$df = (r - 1)(c - 1)$$

där r är lika med antalet rader och c är lika med antalet kolumner i korstabellen.

Det finns ingen övre gräns för hur många rader och/eller kolumner som korstabellen kan bestå av för att en chitvå-test ska vara adekvat som analysverktyg.

Exempel 9.4

Testa med $\alpha = 0{,}05$ om det finns beroende mellan kön och uppfattning om EU när 100 slumpvis tillfrågade personer tillfrågats och observerade frekvenser är:

	Män	Kvinnor
Positiv till EU	34	15
Negativ till EU	29	22

Lösning

Vi har $n = 100$ och hypoteserna

H0: Det råder oberoende mellan kön och uppfattning om EU
H1: Det råder inte oberoende mellan kön och uppfattning om EU.

Vi beräknar rad-, kolumn- och totalsumma:

	Män	Kvinnor	Total
Positiv till EU	34	15	49
Negativ till EU	29	22	51
Total	63	37	100

Förväntade frekvenser under antagandet att H0 är sann blir då:

	Män	Kvinnor
Positiv till EU	49 · 63/100=30,87	49 · 37/100=18,13
Negativ till EU	51 · 63/100=32,13	51 · 37/100=18,87

Vi kan nu beräkna testets χ^2-värde som

$$\chi^2 = \frac{(34-30,87)^2}{30,87} + \frac{(15-18,13)^2}{18,13} + \frac{(29-32,13)^2}{32,13} + \frac{(22-18,87)^2}{18,87} = 1,68$$

Kritiskt χ^2-värde med $\alpha = 0,05$ och med $(2 - 1)(2 - 1) = 1$ frihetsgrad är enligt tabell A3 lika med 3,84146. Eftersom 3,84146 > 1,68 kan H0 inte förkastas.

10. Icke-parametriska metoder

10.1 Inledning

Icke-parametriska (eller parameterfria/fördelningsfria) metoder skiljer sig från parametriska metoder såtillvida att inget särskilt antagande om någon populationsparameter behöver göras. I t.ex. en *t*-test eller en variansanalys antas bland annat att populationsparametern som undersöks är approximativt normalfördelad. En fråga som ofta uppstår är vad man gör om normalfördelning inte kan antas råda, och svaret på den frågan är att man då helt enkelt inte får använda *t*-test eller variansanalys. Istället får man använda metoder som inte kräver någon specifik fördelning.

I detta kapitel ska vi titta på några alternativ till *t*-test, variansanalys och regressionsanalys som kan användas när normalfördelningsantagandet inte är uppfyllt. Dessa metoder baseras inte på något starkare antagande än att observationerna är möjliga att *rangordna*. Det innebär att de ersätter de metoder vi har sett tidigare i denna bok vid två tillfällen:

1. När datan man arbetar med är på intervall- eller kvotskalan, men där ett antagande om normalfördelning kan antas vara felaktigt.

2. När datan man arbetar med är på ordinalskalan.

Särskilt viktig är punkt 2 ovan. Många olika slag av undersökningar baseras på att respondenter får avge svar ”på en skala” på frågor. Den data man då erhåller är på ordinalskalan, vilket innebär att parametriska metoder inte får användas för dataanalysen. Just detta fel ser man ändå ofta att folk gör. Fundera därför alltid på vilka antaganden en viss analysmodell baseras på innan du använder den, och välj därefter en modell som inte kräver mer av den data du har än vad datan faktiskt uppfyller.

10.2 Teckentest

Som ett alternativ till en enkel *t*-test (se avsnitt 6.4) kan en *teckentest* användas. Här testas nollhypotesen att en populationsmedian M_d är högst, lägst, eller lika med ett visst värde M_0. Testen baseras formellt på binomialfördelningen, men vi kommer i denna bok endast att studera hur testen görs under normalfördelningsapproximation, vilket i detta fall förutsätter att stickprovsstorleken $n > 10$.

Tillvägagångssättet för en teckentest är följande: Bestäm först värdet för variabeln *S*:

- Om H0: $M_d \geq M_0$ så är *S* lika med antalet observationer i stickprovet som överstiger M_0.

- Om H0: $M_d < M_0$ så är *S* lika med antalet observationer i stickprovet som understiger M_0.

- Om H0: $M_d = M_0$ så är *S* lika med det största av a) antalet observationer i stickprovet som överstiger M_0 respektive b) antalet observationer i stickprovet som understiger M_0

Beräkna sedan testets *Z*-värde som

$$z = \frac{(S - 0,5) - 0,5n}{0,5\sqrt{n}}$$

Exempel 10.1

En tillverkare av glödlampor hävdar att medianvärdet för deras glödlampors livslängd är minst 3 hela år. Ett stickprov om 30 glödlampor av den aktuella typen tas och deras livslängd observeras. Resultatet framgår av tabellen nedan.

Livslängd (år)	0-1	1-2	2-3	3-4	4-5	5-6	6-7
Observerad frekvens	11	8	6	3	1	0	1

Kan man på basis av detta stickprov hävda att tillverkaren har fel med $\alpha = 0,01$?

Lösning

Vi har hypoteserna

H0: $M_d \geq 3$
H1: $M_d < 3$

Antalet av de 30 observerade livslängderna som överstiger 3 år är 5 st. Vi beräknar alltså testets *Z*-värde till

$$z = \frac{(5 - 0,5) - 0,5 \cdot 30}{0,5\sqrt{30}} = -3,47$$

Eftersom H1 är av karaktären ”mindre än” genomförs testet i *Z*-fördelningens vänstra svans. Kritiskt *Z*-värde för en ensidig test för $\alpha = 0,01$ kan då avläsas från tabell A1 till -2,33. Eftersom -2,33 > -3,47 förkastas H0.

10.3 Wilcoxons teckenrangtest

Som alternativ till den parvisa *t*-testen (se avsnitt 6.11) kan *Wilcoxons teckenrangtest* användas. Här testas nollhypotesen att skillnaden mellan två populationsmedianer, M_1 och M_2, är högst, lägst, eller lika med noll. När antalet observerade par är litet baseras testen på en fördelning som inte tas upp i denna bok, men när $n > 20$ kan normalfördelningsapproximation användas.

Tillvägagångssättet är att beräkna differensen mellan vart och ett av de *n* dataparen, utesluta de datapar som har differensen 0 från testen, och därpå rangordna differensernas absolutvärden. Datapar med differensen 0 utesluts från testen. Därefter beräknas storheterna T^- och T^+ där

- T^- = summan av rangvärdena (”platssiffrorna”) för de datapar som har negativ differens
- T^+= summan av rangvärdena för de datapar som har positiv differens

När flera datapar har samma absoluta differens så får samtliga ett rangvärde motsvarande deras genomsnittliga rangvärde.

Värdet på variabeln T definieras sedan på följande sätt:

- Om H0: $M_1 \leq M_2$ så är $T = T^-$
- Om H0: $M_1 > M_2$ så är $T = T^+$ så är $T = T^+$
- Om H0: $M_1 = M_2$ så är T = det minsta av T och T^+

Testets Z-värde beräknas som

$$z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}}$$

Exempel 10.2

28 slumpmässigt valda kunder fick svara på hur positivt inställda de var till en viss nöjespark när de gick in i parken. Samma kunder svarade även på samma fråga när de lämnade parken. Svaren avgavs på en skala 1 – 7 där 7 betydde att man var maximalt positiv. Resultatet framgår av tabellen nedan.

Kund	1	2	3	4	5	6	7	8	9	10	11	12	13	14
In	3	4	5	3	4	5	7	5	3	4	5	7	4	4
Ut	5	5	6	7	5	4	6	6	5	4	5	7	6	3

Kund	15	16	17	18	19	20	21	22	23	24	25	26	27	28
In	4	3	4	5	1	3	3	4	5	3	5	6	7	3
Ut	2	2	4	5	5	5	4	6	6	4	7	6	5	7

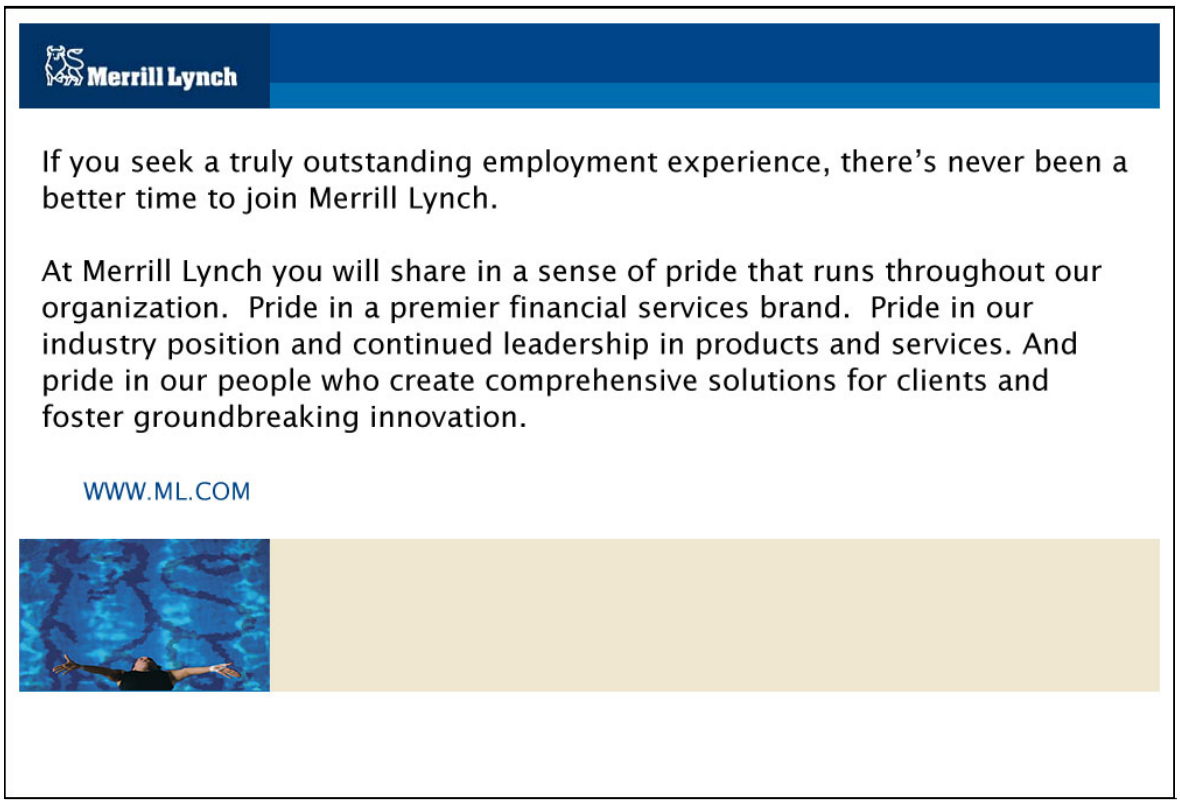
Kan man med α = 0,05 dra slutsatsen att kunderna är mer nöjda när de lämnar parken?

Lösning

Vi får hypoteserna

$$\text{H0: } M_1 \geq M_2$$
$$\text{H1: } M_1 < M_2$$

Vi beräknar de absoluta differenserna mellan varje enskild observation, exkluderar de sex observationerna med 0 i differens, och ger varje absolut differens sitt rangvärde. Vi har 11 observationer med den absoluta differensen 1, vilket innebär att dessa 11 observationer har det genomsnittliga rangvärdet 6 (medelvärdet av platssiffrorna 1 till 11). Vi har sedan 8 observationer med den absoluta differensen 2, vilket ger dessa observationer det genomsnittliga rangvärdet 15,5 (medelvärdet av platssiffrorna 12 till 19), och så vidare. De rangvärden som hör till en negativ differens summeras sedan till T^-, medan de rangvärden som hör till en positiv differens summeras till T^+. Alltsammans sammanfattas i nedanstående tabell.



Kund	Före	Efter	Differens	Absolut differens	Rangvärde	Rangvärde för positiv differens	Rangvärde för negativ differens
1	3	5	-2	2	15,5		15,5
2	4	5	-1	1	6		6
3	5	6	-1	1	6		6
4	3	7	-4	4	21		21
5	4	5	-1	1	6		6
6	5	4	1	1	6	6	
7	7	6	1	1	6	6	
8	5	6	-1	1	6		6
9	3	5	-2	2	15,5		15,5
10	4	4	0				
11	5	5	0				
12	7	7	0				
13	4	6	-2	2	15,5	0	15,5
14	4	3	1	1	6	6	
15	4	2	2	2	15,5	15,5	
16	3	2	1	1	6	6	
17	4	4	0				
18	5	5	0				
19	1	5	-4	4	21		21
20	3	5	-2	2	15,5		15,5
21	3	4	-1	1	6		6
22	4	6	-2	2	15,5		15,5
23	5	6	-1	1	6		6
24	3	4	-1	1	6		6
25	5	7	-2	2	15,5		15,5
26	6	6	0				
27	7	5	2	2	15,5	15,5	
28	3	7	-4	4	21		21
						$T^+ = 55$	$T^- = 198$

De sex observationer med 0 i differens utesluts från testen, som alltså baseras på $n = 28 - 6 = 22$, vilket innebär att vi kan använda normalfördelningsapproximation. Vi har en nollhypotes av karaktären $M_1 \geq M_2$ så $T = T^+ = 55$. Testets Z-värde blir då

$$z = \frac{T - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} = \frac{55 - \frac{22(22+1)}{4}}{\sqrt{\frac{22(22+1)(2\cdot 22+1)}{24}}} = \frac{-71,5}{30,8} = -2,32$$

Kritiskt *Z*-värde för $\alpha = 0{,}05$ i vänstra svansen är enligt tabell A1 lika med -1,64. Eftersom -1,64 > -2,32 så förkastas H0.

10.4 Mann-Whitneys test

Som alternativ till *t*-testen för två oberoende populationer (se avsnitt 6.4) kan *Mann-Whitneys test* (kallas ibland Wilcoxons rangsummetest – icke att förväxla med Wilcoxons teckenrangtest som illustrerades ovan) användas. Här testas nollhypotesen att skillnaden mellan två populationsmedianer, M_1 och M_2, är högst, lägst, eller lika med noll. När antalet observationer från minst en av populationerna är litet baseras testen på en fördelning som inte tas upp i denna bok, men när $n_1 > 10$ och $n_2 > 10$ kan normalfördelningsapproximation användas.

Tillvägagångssättet är följande: Uttryck först hypoteserna så att $n_1 \leq n_2$. Rangordna sedan de $n_1 + n_2$ observationerna från lägsta till högsta värde. När flera observationer har samma värde så får samtliga ett rangvärde motsvarande deras genomsnittliga rangvärde. Värdet på variabeln *T* definieras nu som summan av rangvärden för stickprovet från population 1.

Testets *Z*-värde beräknas sedan som

$$z = \frac{\frac{n_1(n_1+1)+n_1n_2}{2} - T}{\sqrt{\frac{n_1n_2(n_1+n_2+1)}{12}}}$$

Exempel 10.3

24 slumpmässigt utvalda personer (11 män och 13 kvinnor) fick bedöma på en skala 1 – 10 (där 10 är bäst) vad de tyckte om kvaliteten på en viss utbildning de hade genomgått. Resultatet framgår av tabellen nedan.

Män	9	7	6	7	5	5	7	8	7	4	5		
Kvinnor	4	4	6	5	8	7	4	5	6	5	3	4	5

Kan man med $\alpha = 0{,}01$ dra slutsatsen att männen i medeltal är mer nöjda än kvinnorna?

Lösning

Vi sätter männen som population 1 eftersom 11 < 13 och får hypoteserna

$$\text{H0: } M_1 \leq M_2$$
$$\text{H1: } M_1 > M_2$$

Vi ger därefter var och en av de 11 + 13 = 24 observationerna sitt rangvärde. Vi har en 3:a som lägsta värde, den får rangvärdet 1. Vi har därefter 5 st. 4:or på platserna 2 till 6. Deras genomsnittliga rangvärde är alltså 4, och så vidare. Resultatet framgår av tabellen nedan.

Population	Värde	Rangvärde
1	9	24
1	7	19
1	6	15
1	7	19
1	5	10
1	5	10
1	7	19
1	8	22,5
1	7	19
1	4	4
1	5	10
2	4	4
2	4	4
2	6	15
2	5	10
2	8	22,5
2	7	19
2	4	4
2	5	10
2	6	15
2	5	10
2	3	1
2	4	4
2	5	10

Vi summerar rangvärdena för observationerna från population 1:

$$T = 24 + 19 + ... + 10 = 171,5$$

Sedan beräknas testets *Z*-värde:

$$z = \frac{\frac{11(11+1)+11\cdot 13}{2} - 171{,}5}{\sqrt{\frac{11\cdot 13(11+13+1)}{12}}} = \frac{-34}{17{,}26} = -1{,}97$$

Kritiskt *Z*-värde för $\alpha = 0{,}01$ är enligt tabell A1 lika med -2,33. H0 kan således inte förkastas, eftersom -2,33 < -1,97.

10.5 Kruskal-Wallis test

Som alternativ till envägs variansanalys för 3 eller fler oberoende populationer (se kapitel 7) kan *Kruskal-Wallis test* användas. Här testas nollhypotesen att skillnaden mellan tre populationsmedianer är lika med noll. När antalet observationer från minst en av populationerna är litet baseras testen på en fördelning som inte tas upp i denna bok, men när stickprovsstorleken från var och en av de populationer som undersöks är åtminstone 5 kan chitvå-fördelningsapproximation användas. Denna test påminner mycket om Mann-Whitneys test (de är rent tekniskt identiska om antalet populationer som undersöks är två).

Antag att vi har stickprovsdata på minst ordinalskalan från *k* olika populationer. Precis som i Mann-Whitneys test inleder man då med att tilldela var och en av de $n = n_1 + n_2 + + ... + n_k$ observationerna ett rangvärde. Även här är det så att när flera observationer har samma värde så får samtliga ett rangvärde motsvarande deras genomsnittliga rangvärde. Vi definierar sedan T_1 som summan av rangvärden från stickprov 1, T_2 som summan av rangvärden från stickprov 2 och så vidare upp till T_k som summan av rangvärden från stickprov *k*. Därefter beräknas värdet för testvariabeln *H* som

$$H = \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{T_i^2}{n_i} - 3(n+1)$$

som approximativt följer en chitvå-fördelning med $k - 1$ frihetsgrader.

Exempel 10.4

En butik delar in sina kunder i tre kategorier: yngre, medelålders och äldre. Man vill nu testa om kunderna i medeltal kan antas vara inne i butiken lika länge. Man mäter därför hur länge 5 slumpmässigt valda kunder från respektive kategori är inne i butiken. Resultatet (i minuter) framgår av tabellen nedan.

Yngre	Medelålders	Äldre
4	6	13
7	8	3
3	14	5
4	8	6
11	18	5

Kan man med $\alpha = 0{,}05$ dra slutsatsen att någon ålderskategori i medeltal är inne längre i butiken än någon annan?

Lösning

Vi ger var och en av de 5 + 5 + 5 = 15 observationerna sitt rangvärde. Vi har 2 st. 3:or som lägsta värde – deras genomsnittliga rangvärde är alltså 1,5. Vi har därefter 2 st. 4:or på platserna 3 och 4. Deras genomsnittliga rangvärde är alltså 3,5, och så vidare. Resultatet framgår av tabellen nedan.

Population	Värde	Rangvärde
1	4	3,5
1	7	9
1	3	1,5
1	4	3,5
1	11	12
2	6	7,5
2	8	10,5
2	14	14
2	8	10,5
2	18	15
3	13	13
3	3	1,5
3	5	5,5
3	6	7,5
3	5	5,5

Vi summerar därefter rangvärdena för var och en av populationerna:

$$T_1 = 3,5 + 9 + 1,5 + 3,5 + 12 = 29,5$$

$$T_2 = 7,5 + 10,5 + 14 + 10,5 + 15 = 57,5$$

$$T_3 = 13 + 1,5 + 5,5 + 7,5 + 5,5 = 33$$

Vi kan nu beräkna värdet för testvariabeln H:

$$H = \frac{12}{15(15+1)}\left(\frac{29,5^2}{5} + \frac{57,5^2}{5} + \frac{33^2}{5}\right) - 3(15+1) = 4,655$$

Kritiskt χ^2-värde för $\alpha = 0,05$ med 3 – 1 = 2 *df* är enligt tabell A3 lika med 5,99148. Således kan H0 inte förkastas, eftersom 5,99148 > 4,655.

10.6 Spearmans rank-korrelationstest

Som alternativ till den ”vanliga” korrelationskoefficienten vid sambandsanalys för två slumpvariabler *X* och *Y* (se avsnitt 8.2) kan *Spearmans rank-korrelationstest* användas. Denna förutsätter endast att *X*- och *Y*-värdena är rangordningsbara, d.v.s. de kan vara på ordinalskalan eller de kan vara kvantitativ data med godtycklig fördelning.

Idén är helt enkelt att beräkna korrelationskoefficienten *r* med den vanliga formeln, men på basis av observationernas rangvärden istället för observationerna i sig. På samma sätt som innan så blir det så att när flera observationer har samma värde så får samtliga ett rangvärde motsvarande deras genomsnittliga rangvärde.

Tolkningen av denna korrelationskoefficient blir densamma som för den vanliga korrelationskoef ficienten. För mindre värden på *n* kommer *r* dock att följa en fördelning som inte tas upp i denna bok. För större värden på *n* (tumregel: $n > 10$) kan däremot *t*-fördelningen användas för att analysera nollhypotesen att den sanna korrelationskoefficienten är högst, lägst, eller lika med noll. Testets *t*-värde beräknas då enligt formeln

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

med $n - 2$ frihetsgrader.

Exempel 10.5

10 studenter tillfrågas om hur många timmar de har arbetar med en viss inlämningsuppgift, för att kunna analysera om det finns något samband mellan arbetsinsats och betyg. Betyg gavs på en skala 1 – 5 där 5 är bäst. Resultatet framgår av tabellen nedan.

Student	Timmar	Betyg
1	3	2
2	5	4
3	4	3
4	8	3
5	7	4
6	9	4
7	6	3
8	6	2
9	10	5
10	7	5

Finns det någon signifikant korrelation mellan betyg och arbetsinsats i tid räknat? Använd $\alpha = 0{,}05$.

Lösning

Vi beräknar först rangvärdena för variablerna ”timmar” respektive ”betyg” var för sig för respektive student:

Student	Rangvärde Timmar	Rangvärde Betyg
1	1	1,5
2	3	7
3	2	4
4	8	4
5	6,5	7
6	9	7
7	4,5	4
8	4,5	1,5
9	10	9,5
10	6,5	9,5

Electrolux www.electrolux.se

At Electrolux we share a common goal – to make everyday life easier and more enjoyable. Our people work to reach this goal and they all contribute to the Group´s success.
With a presence in more than 100 countries Electrolux is a truly international company, which gives opportunities to succeed and professionally grow.

Electrolux focuses on developing its people, helping them to grow personally and professionally. To attract and retain highly talented personnel is considered to be at the core of our company strategy.

We consider talent as one of our most valuable assets. We believe that managing and developing this asset is a prerequisite for success. In Electrolux talent management is a strategic priority. Our vision for talent management is to have a strong pool of outstanding employees and a performance culture that strives for excellence.

Beräkning av korrelationskoefficienten med den vanliga formeln (se avsnitt 8.2) ger $r = 0{,}63$ (pröva själv!). Stickprovsdatan indikerar alltså en viss korrelation. Vi vill testa om den är signifikant positiv, och får då hypoteserna

$$\begin{aligned} &\text{H0: } \rho \leq 0 \\ &\text{H1: } \rho > 0 \end{aligned}$$

Testets *t*-värde beräknas till

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0{,}63\sqrt{10-2}}{\sqrt{1-0{,}63^2}} = 2{,}29$$

Kritiskt *t*-värde för 10 – 2 = 8 *df* är enligt tabell A2 lika med 1,8595. Således kan H0 förkastas, eftersom 1,85 < 2,29.

Tabell A1: Standardnormalfördelningen

Tabellen visar $P(Z > z) = \int_{z}^{\infty} f(z)dz$ när $Z \sim N(0, 1)$.

Exempel: När $Z \sim N(0, 1)$ så blir $P(Z > 1{,}42) = 0{,}0778$

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,0	0,5000	0,4960	0,4920	0,4880	0,4840	0,4801	0,4761	0,4721	0,4681	0,4641
0,1	0,4602	0,4562	0,4522	0,4483	0,4443	0,4404	0,4364	0,4325	0,4286	0,4247
0,2	0,4207	0,4168	0,4129	0,4090	0,4052	0,4013	0,3974	0,3936	0,3897	0,3859
0,3	0,3821	0,3783	0,3745	0,3707	0,3669	0,3632	0,3594	0,3557	0,3520	0,3483
0,4	0,3446	0,3409	0,3372	0,3336	0,3300	0,3264	0,3228	0,3192	0,3156	0,3121
0,5	0,3085	0,3050	0,3015	0,2981	0,2946	0,2912	0,2877	0,2843	0,2810	0,2776
0,6	0,2743	0,2709	0,2676	0,2643	0,2611	0,2578	0,2546	0,2514	0,2483	0,2451
0,7	0,2420	0,2389	0,2358	0,2327	0,2296	0,2266	0,2236	0,2206	0,2177	0,2148
0,8	0,2119	0,2090	0,2061	0,2033	0,2005	0,1977	0,1949	0,1922	0,1894	0,1867
0,9	0,1841	0,1814	0,1788	0,1762	0,1736	0,1711	0,1685	0,1660	0,1635	0,1611
1,0	0,1587	0,1562	0,1539	0,1515	0,1492	0,1469	0,1446	0,1423	0,1401	0,1379
1,1	0,1357	0,1335	0,1314	0,1292	0,1271	0,1251	0,1230	0,1210	0,1190	0,1170
1,2	0,1151	0,1131	0,1112	0,1093	0,1075	0,1056	0,1038	0,1020	0,1003	0,0985
1,3	0,0968	0,0951	0,0934	0,0918	0,0901	0,0885	0,0869	0,0853	0,0838	0,0823
1,4	0,0808	0,0793	0,0778	0,0764	0,0749	0,0735	0,0721	0,0708	0,0694	0,0681
1,5	0,0668	0,0655	0,0643	0,0630	0,0618	0,0606	0,0594	0,0582	0,0571	0,0559
1,6	0,0548	0,0537	0,0526	0,0516	0,0505	0,0495	0,0485	0,0475	0,0465	0,0455
1,7	0,0446	0,0436	0,0427	0,0418	0,0409	0,0401	0,0392	0,0384	0,0375	0,0367
1,8	0,0359	0,0351	0,0344	0,0336	0,0329	0,0322	0,0314	0,0307	0,0301	0,0294
1,9	0,0287	0,0281	0,0274	0,0268	0,0262	0,0256	0,0250	0,0244	0,0239	0,0233
2,0	0,0228	0,0222	0,0217	0,0212	0,0207	0,0202	0,0197	0,0192	0,0188	0,0183
2,1	0,0179	0,0174	0,0170	0,0166	0,0162	0,0158	0,0154	0,0150	0,0146	0,0143
2,2	0,0139	0,0136	0,0132	0,0129	0,0125	0,0122	0,0119	0,0116	0,0113	0,0110
2,3	0,0107	0,0104	0,0102	0,0099	0,0096	0,0094	0,0091	0,0089	0,0087	0,0084
2,4	0,0082	0,0080	0,0078	0,0075	0,0073	0,0071	0,0069	0,0068	0,0066	0,0064
2,5	0,0062	0,0060	0,0059	0,0057	0,0055	0,0054	0,0052	0,0051	0,0049	0,0048
2,6	0,0047	0,0045	0,0044	0,0043	0,0041	0,0040	0,0039	0,0038	0,0037	0,0036
2,7	0,0035	0,0034	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026
2,8	0,0026	0,0025	0,0024	0,0023	0,0023	0,0022	0,0021	0,0021	0,0020	0,0019
2,9	0,0019	0,0018	0,0018	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014
3,0	0,0013	0,0013	0,0013	0,0012	0,0012	0,0011	0,0011	0,0011	0,0010	0,0010

Tabell A2: *t*-fördelningen

Tabellen visar kritiska värden för *t*-fördelningen.

Exempel: För 8 frihetsgrader så är $0{,}01 > P(t > 3) > 0{,}005$ eftersom $2{,}8965 < 3 < 3{,}3554$.

df	t0,1	t0,05	t0,025	t0,01	t0,005	t0,001
1	3,0777	6,3137	12,706	31,821	63,656	318,29
2	1,8856	2,9200	4,3027	6,9645	9,9250	22,328
3	1,6377	2,3534	3,1824	4,5407	5,8408	10,214
4	1,5332	2,1318	2,7765	3,7469	4,6041	7,1729
5	1,4759	2,0150	2,5706	3,3649	4,0321	5,8935
6	1,4398	1,9432	2,4469	3,1427	3,7074	5,2075
7	1,4149	1,8946	2,3646	2,9979	3,4995	4,7853
8	1,3968	1,8595	2,3060	2,8965	3,3554	4,5008
9	1,3830	1,8331	2,2622	2,8214	3,2498	4,2969
10	1,3722	1,8125	2,2281	2,7638	3,1693	4,1437
11	1,3634	1,7959	2,2010	2,7181	3,1058	4,0248
12	1,3562	1,7823	2,1788	2,6810	3,0545	3,9296
13	1,3502	1,7709	2,1604	2,6503	3,0123	3,8520
14	1,3450	1,7613	2,1448	2,6245	2,9768	3,7874
15	1,3406	1,7531	2,1315	2,6025	2,9467	3,7329
16	1,3368	1,7459	2,1199	2,5835	2,9208	3,6861
17	1,3334	1,7396	2,1098	2,5669	2,8982	3,6458
18	1,3304	1,7341	2,1009	2,5524	2,8784	3,6105
19	1,3277	1,7291	2,0930	2,5395	2,8609	3,5793
20	1,3253	1,7247	2,0860	2,5280	2,8453	3,5518
21	1,3232	1,7207	2,0796	2,5176	2,8314	3,5271
22	1,3212	1,7171	2,0739	2,5083	2,8188	3,5050
23	1,3195	1,7139	2,0687	2,4999	2,8073	3,4850
24	1,3178	1,7109	2,0639	2,4922	2,7970	3,4668
25	1,3163	1,7081	2,0595	2,4851	2,7874	3,4502
26	1,3150	1,7056	2,0555	2,4786	2,7787	3,4350
27	1,3137	1,7033	2,0518	2,4727	2,7707	3,4210
28	1,3125	1,7011	2,0484	2,4671	2,7633	3,4082
29	1,3114	1,6991	2,0452	2,4620	2,7564	3,3963
30	1,3104	1,6973	2,0423	2,4573	2,7500	3,3852
35	1,3062	1,6896	2,0301	2,4377	2,7238	3,3400
40	1,3031	1,6839	2,0211	2,4233	2,7045	3,3069
45	1,3007	1,6794	2,0141	2,4121	2,6896	3,2815
50	1,2987	1,6759	2,0086	2,4033	2,6778	3,2614
60	1,2958	1,6706	2,0003	2,3901	2,6603	3,2317
80	1,2922	1,6641	1,9901	2,3739	2,6387	3,1952
100	1,2901	1,6602	1,9840	2,3642	2,6259	3,1738
∞	1,2815	1,6448	1,9600	2,3264	2,5758	3,0902

Tabell A3: χ^2 -fördelningen

Tabellen visar kritiska värden för χ^2-fördelningen.

Exempel: För 8 frihetsgrader så är $0,05 > P(\chi^2 > 16) > 0,025$ eftersom 15,5073 < 16 < 17,5345.

df	$\chi^2_{0,995}$	$\chi^2_{0,99}$	$\chi^2_{0,975}$	$\chi^2_{0,95}$	$\chi^2_{0,9}$	$\chi^2_{0,1}$	$\chi^2_{0,05}$	$\chi^2_{0,025}$	$\chi^2_{0,01}$	$\chi^2_{0,005}$
1	0,00004	0,00016	0,00098	0,00393	0,01579	2,70554	3,84146	5,02390	6,63489	7,87940
2	0,01002	0,02010	0,05064	0,10259	0,21072	4,60518	5,99148	7,37778	9,21035	10,5965
3	0,07172	0,11483	0,21579	0,35185	0,58438	6,25139	7,81472	9,34840	11,3449	12,8381
4	0,20698	0,29711	0,48442	0,71072	1,06362	7,77943	9,48773	11,1433	13,2767	14,8602
5	0,41175	0,55430	0,83121	1,14548	1,61031	9,23635	11,0705	12,8325	15,0863	16,7496
6	0,67573	0,87208	1,23734	1,63538	2,20413	10,6446	12,5916	14,4494	16,8119	18,5475
7	0,98925	1,23903	1,68986	2,16735	2,83311	12,0170	14,0671	16,0128	18,4753	20,2777
8	1,34440	1,64651	2,17972	2,73263	3,48954	13,3616	15,5073	17,5345	20,0902	21,9549
9	1,73491	2,08789	2,70039	3,32512	4,16816	14,6837	16,9190	19,0228	21,6660	23,5893
10	2,15585	2,55820	3,24696	3,94030	4,86518	15,9872	18,3070	20,4832	23,2093	25,1881
11	2,60320	3,05350	3,81574	4,57481	5,57779	17,2750	19,6752	21,9200	24,7250	26,7569
12	3,07379	3,57055	4,40378	5,22603	6,30380	18,5493	21,0261	23,3367	26,2170	28,2997
13	3,56504	4,10690	5,00874	5,89186	7,04150	19,8119	22,3620	24,7356	27,6882	29,8193
14	4,07466	4,66042	5,62872	6,57063	7,78954	21,0641	23,6848	26,1189	29,1412	31,3194
15	4,60087	5,22936	6,26212	7,26093	8,54675	22,3071	24,9958	27,4884	30,5780	32,8015
16	5,14216	5,81220	6,90766	7,96164	9,31224	23,5418	26,2962	28,8453	31,9999	34,2671
17	5,69727	6,40774	7,56418	8,67175	10,0852	24,7690	27,5871	30,1910	33,4087	35,7184
18	6,26477	7,01490	8,23074	9,39045	10,8649	25,9894	28,8693	31,5264	34,8052	37,1564
19	6,84392	7,63270	8,90651	10,1170	11,6509	27,2036	30,1435	32,8523	36,1908	38,5821
20	7,43381	8,26037	9,59077	10,8508	12,4426	28,4120	31,4104	34,1696	37,5663	39,9969
22	8,64268	9,54249	10,9823	12,3380	14,0415	30,8133	33,9245	36,7807	40,2894	42,7957
24	9,88620	10,8563	12,4011	13,8484	15,6587	33,1962	36,4150	39,3641	42,9798	45,5584
26	11,1602	12,1982	13,8439	15,3792	17,2919	35,5632	38,8851	41,9231	45,6416	48,2898
28	12,4613	13,5647	15,3079	16,9279	18,9392	37,9159	41,3372	44,4608	48,2782	50,9936
30	13,7867	14,9535	16,7908	18,4927	20,5992	40,2560	43,7730	46,9792	50,8922	53,6719
35	17,1917	18,5089	20,5694	22,4650	24,7966	46,0588	49,8018	53,2033	57,3420	60,2746
40	20,7066	22,1642	24,4331	26,5093	29,0505	51,8050	55,7585	59,3417	63,6908	66,7660
50	27,9908	29,7067	32,3574	34,7642	37,6886	63,1671	67,5048	71,4202	76,1538	79,4898
60	35,5344	37,4848	40,4817	43,1880	46,4589	74,3970	79,0820	83,2977	88,3794	91,9518
70	43,2753	45,4417	48,7575	51,7393	55,3289	85,5270	90,5313	95,0231	100,425	104,215
80	51,1719	53,5400	57,1532	60,3915	64,2778	96,5782	101,879	106,629	112,329	116,321
90	59,1963	61,7540	65,6466	69,1260	73,2911	107,565	113,145	118,136	124,116	128,299
100	67,3275	70,0650	74,2219	77,9294	82,3581	118,498	124,342	129,561	135,807	140,170

Tabell A4: *F*-fördelningen för α = 0,1

Tabellen visar kritiska värden för *F*-fördelningen när $\alpha = 0{,}1$.

Exempel: För 8 frihetsgrader i täljaren och 5 frihetsgrader i nämnaren så är P($F > 4$) < 0,1 eftersom 4 > 3,34

	Täljarens *df*									
Nämnarens *df*	1	2	3	4	5	6	7	8	9	10
1	39,86	49,50	53,59	55,83	57,24	58,20	58,91	59,44	59,86	60,19
2	8,53	9,00	9,16	9,24	9,29	9,33	9,35	9,37	9,38	9,39
3	5,54	5,46	5,39	5,34	5,31	5,28	5,27	5,25	5,24	5,23
4	4,54	4,32	4,19	4,11	4,05	4,01	3,98	3,95	3,94	3,92
5	4,06	3,78	3,62	3,52	3,45	3,40	3,37	3,34	3,32	3,30
6	3,78	3,46	3,29	3,18	3,11	3,05	3,01	2,98	2,96	2,94
7	3,59	3,26	3,07	2,96	2,88	2,83	2,78	2,75	2,72	2,70
8	3,46	3,11	2,92	2,81	2,73	2,67	2,62	2,59	2,56	2,54
9	3,36	3,01	2,81	2,69	2,61	2,55	2,51	2,47	2,44	2,42
10	3,29	2,92	2,73	2,61	2,52	2,46	2,41	2,38	2,35	2,32
11	3,23	2,86	2,66	2,54	2,45	2,39	2,34	2,30	2,27	2,25
12	3,18	2,81	2,61	2,48	2,39	2,33	2,28	2,24	2,21	2,19
13	3,14	2,76	2,56	2,43	2,35	2,28	2,23	2,20	2,16	2,14
14	3,10	2,73	2,52	2,39	2,31	2,24	2,19	2,15	2,12	2,10
15	3,07	2,70	2,49	2,36	2,27	2,21	2,16	2,12	2,09	2,06
16	3,05	2,67	2,46	2,33	2,24	2,18	2,13	2,09	2,06	2,03
17	3,03	2,64	2,44	2,31	2,22	2,15	2,10	2,06	2,03	2,00
18	3,01	2,62	2,42	2,29	2,20	2,13	2,08	2,04	2,00	1,98
19	2,99	2,61	2,40	2,27	2,18	2,11	2,06	2,02	1,98	1,96
20	2,97	2,59	2,38	2,25	2,16	2,09	2,04	2,00	1,96	1,94
22	2,95	2,56	2,35	2,22	2,13	2,06	2,01	1,97	1,93	1,90
24	2,93	2,54	2,33	2,19	2,10	2,04	1,98	1,94	1,91	1,88
26	2,91	2,52	2,31	2,17	2,08	2,01	1,96	1,92	1,88	1,86
28	2,89	2,50	2,29	2,16	2,06	2,00	1,94	1,90	1,87	1,84
30	2,88	2,49	2,28	2,14	2,05	1,98	1,93	1,88	1,85	1,82
35	2,85	2,46	2,25	2,11	2,02	1,95	1,90	1,85	1,82	1,79
40	2,84	2,44	2,23	2,09	2,00	1,93	1,87	1,83	1,79	1,76
50	2,81	2,41	2,20	2,06	1,97	1,90	1,84	1,80	1,76	1,73
60	2,79	2,39	2,18	2,04	1,95	1,87	1,82	1,77	1,74	1,71
70	2,78	2,38	2,16	2,03	1,93	1,86	1,80	1,76	1,72	1,69
80	2,77	2,37	2,15	2,02	1,92	1,85	1,79	1,75	1,71	1,68
90	2,76	2,36	2,15	2,01	1,91	1,84	1,78	1,74	1,70	1,67
100	2,76	2,36	2,14	2,00	1,91	1,83	1,78	1,73	1,69	1,66
9999	2,71	2,30	2,08	1,94	1,85	1,77	1,72	1,67	1,63	1,60

Tabell A5: *F*-fördelningen för α = 0,05.

Tabellen visar kritiska värden för *F*-fördelningen när α = 0,05.

Exempel: För 8 frihetsgrader i täljaren och 5 frihetsgrader i nämnaren så är P(F > 5) < 0,05 eftersom 5 > 4,82

	Täljarens *df*									
Nämnarens *df*	1	2	3	4	5	6	7	8	9	10
1	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	240,5	241,9
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25
26	4,23	3,37	2,98	2,74	2,59	2,47	2,39	2,32	2,27	2,22
28	4,20	3,34	2,95	2,71	2,56	2,45	2,36	2,29	2,24	2,19
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16
35	4,12	3,27	2,87	2,64	2,49	2,37	2,29	2,22	2,16	2,11
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08
50	4,03	3,18	2,79	2,56	2,40	2,29	2,20	2,13	2,07	2,03
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99
70	3,98	3,13	2,74	2,50	2,35	2,23	2,14	2,07	2,02	1,97
80	3,96	3,11	2,72	2,49	2,33	2,21	2,13	2,06	2,00	1,95
90	3,95	3,10	2,71	2,47	2,32	2,20	2,11	2,04	1,99	1,94
100	3,94	3,09	2,70	2,46	2,31	2,19	2,10	2,03	1,97	1,93
9999	3,84	3,00	2,60	2,37	2,21	2,10	2,01	1,94	1,88	1,83

Tabell A6: *F*-fördelningen för α = 0,025.

Tabellen visar kritiska värden för *F*-fördelningen när α = 0,025.

Exempel: För 8 frihetsgrader i täljaren och 5 frihetsgrader i nämnaren så är P(F > 7) < 0,025 eftersom 7 > 6,76.

	Täljarens *df*									
Nämnarens *df*	1	2	3	4	5	6	7	8	9	10
1	647,8	799,5	864,2	899,6	921,8	937,1	948,2	956,6	963,3	968,6
2	38,51	39,00	39,17	39,25	39,30	39,33	39,36	39,37	39,39	39,40
3	17,44	16,04	15,44	15,10	14,88	14,73	14,62	14,54	14,47	14,42
4	12,22	10,65	9,98	9,60	9,36	9,20	9,07	8,98	8,90	8,84
5	10,01	8,43	7,76	7,39	7,15	6,98	6,85	6,76	6,68	6,62
6	8,81	7,26	6,60	6,23	5,99	5,82	5,70	5,60	5,52	5,46
7	8,07	6,54	5,89	5,52	5,29	5,12	4,99	4,90	4,82	4,76
8	7,57	6,06	5,42	5,05	4,82	4,65	4,53	4,43	4,36	4,30
9	7,21	5,71	5,08	4,72	4,48	4,32	4,20	4,10	4,03	3,96
10	6,94	5,46	4,83	4,47	4,24	4,07	3,95	3,85	3,78	3,72
11	6,72	5,26	4,63	4,28	4,04	3,88	3,76	3,66	3,59	3,53
12	6,55	5,10	4,47	4,12	3,89	3,73	3,61	3,51	3,44	3,37
13	6,41	4,97	4,35	4,00	3,77	3,60	3,48	3,39	3,31	3,25
14	6,30	4,86	4,24	3,89	3,66	3,50	3,38	3,29	3,21	3,15
15	6,20	4,77	4,15	3,80	3,58	3,41	3,29	3,20	3,12	3,06
16	6,12	4,69	4,08	3,73	3,50	3,34	3,22	3,12	3,05	2,99
17	6,04	4,62	4,01	3,66	3,44	3,28	3,16	3,06	2,98	2,92
18	5,98	4,56	3,95	3,61	3,38	3,22	3,10	3,01	2,93	2,87
19	5,92	4,51	3,90	3,56	3,33	3,17	3,05	2,96	2,88	2,82
20	5,87	4,46	3,86	3,51	3,29	3,13	3,01	2,91	2,84	2,77
22	5,79	4,38	3,78	3,44	3,22	3,05	2,93	2,84	2,76	2,70
24	5,72	4,32	3,72	3,38	3,15	2,99	2,87	2,78	2,70	2,64
26	5,66	4,27	3,67	3,33	3,10	2,94	2,82	2,73	2,65	2,59
28	5,61	4,22	3,63	3,29	3,06	2,90	2,78	2,69	2,61	2,55
30	5,57	4,18	3,59	3,25	3,03	2,87	2,75	2,65	2,57	2,51
35	5,48	4,11	3,52	3,18	2,96	2,80	2,68	2,58	2,50	2,44
40	5,42	4,05	3,46	3,13	2,90	2,74	2,62	2,53	2,45	2,39
50	5,34	3,97	3,39	3,05	2,83	2,67	2,55	2,46	2,38	2,32
60	5,29	3,93	3,34	3,01	2,79	2,63	2,51	2,41	2,33	2,27
70	5,25	3,89	3,31	2,97	2,75	2,59	2,47	2,38	2,30	2,24
80	5,22	3,86	3,28	2,95	2,73	2,57	2,45	2,35	2,28	2,21
90	5,20	3,84	3,26	2,93	2,71	2,55	2,43	2,34	2,26	2,19
100	5,18	3,83	3,25	2,92	2,70	2,54	2,42	2,32	2,24	2,18
9999	5,02	3,69	3,12	2,79	2,57	2,41	2,29	2,19	2,11	2,05

Tabell A7: *F*-fördelningen för α = 0,01.

Tabellen visar kritiska värden för *F*-fördelningen när $\alpha = 0{,}01$.

Exempel: För 8 frihetsgrader i täljaren och 5 frihetsgrader i nämnaren så är P(F > 11) < 0,01 eftersom 11 > 10,29

	Täljarens *df*									
Nämnarens *df*	1	2	3	4	5	6	7	8	9	10
1	4052	4999	5404	5624	5764	5859	5928	5981	6022	6056
2	98,50	99,00	99,16	99,25	99,30	99,33	99,36	99,38	99,39	99,40
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69
17	8,40	6,11	5,19	4,67	4,34	4,10	3,93	3,79	3,68	3,59
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17
26	7,72	5,53	4,64	4,14	3,82	3,59	3,42	3,29	3,18	3,09
28	7,64	5,45	4,57	4,07	3,75	3,53	3,36	3,23	3,12	3,03
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98
35	7,42	5,27	4,40	3,91	3,59	3,37	3,20	3,07	2,96	2,88
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80
50	7,17	5,06	4,20	3,72	3,41	3,19	3,02	2,89	2,78	2,70
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63
70	7,01	4,92	4,07	3,60	3,29	3,07	2,91	2,78	2,67	2,59
80	6,96	4,88	4,04	3,56	3,26	3,04	2,87	2,74	2,64	2,55
90	6,93	4,85	4,01	3,53	3,23	3,01	2,84	2,72	2,61	2,52
100	6,90	4,82	3,98	3,51	3,21	2,99	2,82	2,69	2,59	2,50
9999	6,64	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32

Tabell A8: *F*-fördelningen för α = 0,005.

Tabellen visar kritiska värden för *F*-fördelningen när α = 0,005.

Exempel: För 8 frihetsgrader i täljaren och 5 frihetsgrader i nämnaren så är P(F > 14) < 0,005 eftersom 14 > 13,96.

	Täljarens *df*									
Nämnarens *df*	1	2	3	4	5	6	7	8	9	10
1	16212	19997	21614	22501	23056	23440	23715	23924	24091	24222
2	198,5	199,0	199,2	199,2	199,3	199,3	199,4	199,4	199,4	199,4
3	55,55	49,80	47,47	46,20	45,39	44,84	44,43	44,13	43,88	43,68
4	31,33	26,28	24,26	23,15	22,46	21,98	21,62	21,35	21,14	20,97
5	22,78	18,31	16,53	15,56	14,94	14,51	14,20	13,96	13,77	13,62
6	18,63	14,54	12,92	12,03	11,46	11,07	10,79	10,57	10,39	10,25
7	16,24	12,40	10,88	10,05	9,52	9,16	8,89	8,68	8,51	8,38
8	14,69	11,04	9,60	8,81	8,30	7,95	7,69	7,50	7,34	7,21
9	13,61	10,11	8,72	7,96	7,47	7,13	6,88	6,69	6,54	6,42
10	12,83	9,43	8,08	7,34	6,87	6,54	6,30	6,12	5,97	5,85
11	12,23	8,91	7,60	6,88	6,42	6,10	5,86	5,68	5,54	5,42
12	11,75	8,51	7,23	6,52	6,07	5,76	5,52	5,35	5,20	5,09
13	11,37	8,19	6,93	6,23	5,79	5,48	5,25	5,08	4,94	4,82
14	11,06	7,92	6,68	6,00	5,56	5,26	5,03	4,86	4,72	4,60
15	10,80	7,70	6,48	5,80	5,37	5,07	4,85	4,67	4,54	4,42
16	10,58	7,51	6,30	5,64	5,21	4,91	4,69	4,52	4,38	4,27
17	10,38	7,35	6,16	5,50	5,07	4,78	4,56	4,39	4,25	4,14
18	10,22	7,21	6,03	5,37	4,96	4,66	4,44	4,28	4,14	4,03
19	10,07	7,09	5,92	5,27	4,85	4,56	4,34	4,18	4,04	3,93
20	9,94	6,99	5,82	5,17	4,76	4,47	4,26	4,09	3,96	3,85
22	9,73	6,81	5,65	5,02	4,61	4,32	4,11	3,94	3,81	3,70
24	9,55	6,66	5,52	4,89	4,49	4,20	3,99	3,83	3,69	3,59
26	9,41	6,54	5,41	4,79	4,38	4,10	3,89	3,73	3,60	3,49
28	9,28	6,44	5,32	4,70	4,30	4,02	3,81	3,65	3,52	3,41
30	9,18	6,35	5,24	4,62	4,23	3,95	3,74	3,58	3,45	3,34
35	8,98	6,19	5,09	4,48	4,09	3,81	3,61	3,45	3,32	3,21
40	8,83	6,07	4,98	4,37	3,99	3,71	3,51	3,35	3,22	3,12
50	8,63	5,90	4,83	4,23	3,85	3,58	3,38	3,22	3,09	2,99
60	8,49	5,79	4,73	4,14	3,76	3,49	3,29	3,13	3,01	2,90
70	8,40	5,72	4,66	4,08	3,70	3,43	3,23	3,08	2,95	2,85
80	8,33	5,67	4,61	4,03	3,65	3,39	3,19	3,03	2,91	2,80
90	8,28	5,62	4,57	3,99	3,62	3,35	3,15	3,00	2,87	2,77
100	8,24	5,59	4,54	3,96	3,59	3,33	3,13	2,97	2,85	2,74
9999	7,88	5,30	4,28	3,72	3,35	3,09	2,90	2,74	2,62	2,52

Tabell A9: *F*-fördelningen för α = 0,001.

Tabellen visar kritiska värden för *F*-fördelningen när α = 0,001.

Exempel: För 8 frihetsgrader i täljaren och 5 frihetsgrader i nämnaren så är P(F > 28) < 0,001 eftersom 28 > 27,65.

	Täljarens *df*									
Nämnarens *df*	1	2	3	4	5	6	7	8	9	10
1	405312	499725	540257	562668	576496	586033	593185	597954	602245	605583
2	998,4	998,8	999,3	999,3	999,3	999,3	999,3	999,3	999,3	999,3
3	167,1	148,5	141,1	137,1	134,6	132,8	131,6	130,6	129,9	129,2
4	74,13	61,25	56,17	53,43	51,72	50,52	49,65	49,00	48,47	48,05
5	47,18	37,12	33,20	31,08	29,75	28,83	28,17	27,65	27,24	26,91
6	35,51	27,00	23,71	21,92	20,80	20,03	19,46	19,03	18,69	18,41
7	29,25	21,69	18,77	17,20	16,21	15,52	15,02	14,63	14,33	14,08
8	25,41	18,49	15,83	14,39	13,48	12,86	12,40	12,05	11,77	11,54
9	22,86	16,39	13,90	12,56	11,71	11,13	10,70	10,37	10,11	9,89
10	21,04	14,90	12,55	11,28	10,48	9,93	9,52	9,20	8,96	8,75
11	19,69	13,81	11,56	10,35	9,58	9,05	8,65	8,35	8,12	7,92
12	18,64	12,97	10,80	9,63	8,89	8,38	8,00	7,71	7,48	7,29
13	17,82	12,31	10,21	9,07	8,35	7,86	7,49	7,21	6,98	6,80
14	17,14	11,78	9,73	8,62	7,92	7,44	7,08	6,80	6,58	6,40
15	16,59	11,34	9,34	8,25	7,57	7,09	6,74	6,47	6,26	6,08
16	16,12	10,97	9,01	7,94	7,27	6,80	6,46	6,20	5,98	5,81
17	15,72	10,66	8,73	7,68	7,02	6,56	6,22	5,96	5,75	5,58
18	15,38	10,39	8,49	7,46	6,81	6,35	6,02	5,76	5,56	5,39
19	15,08	10,16	8,28	7,27	6,62	6,18	5,85	5,59	5,39	5,22
20	14,82	9,95	8,10	7,10	6,46	6,02	5,69	5,44	5,24	5,08
22	14,38	9,61	7,80	6,81	6,19	5,76	5,44	5,19	4,99	4,83
24	14,03	9,34	7,55	6,59	5,98	5,55	5,24	4,99	4,80	4,64
26	13,74	9,12	7,36	6,41	5,80	5,38	5,07	4,83	4,64	4,48
28	13,50	8,93	7,19	6,25	5,66	5,24	4,93	4,69	4,50	4,35
30	13,29	8,77	7,05	6,12	5,53	5,12	4,82	4,58	4,39	4,24
35	12,90	8,47	6,79	5,88	5,30	4,89	4,59	4,36	4,18	4,03
40	12,61	8,25	6,59	5,70	5,13	4,73	4,44	4,21	4,02	3,87
50	12,22	7,96	6,34	5,46	4,90	4,51	4,22	4,00	3,82	3,67
60	11,97	7,77	6,17	5,31	4,76	4,37	4,09	3,86	3,69	3,54
70	11,80	7,64	6,06	5,20	4,66	4,28	3,99	3,77	3,60	3,45
80	11,67	7,54	5,97	5,12	4,58	4,20	3,92	3,70	3,53	3,39
90	11,57	7,47	5,91	5,06	4,53	4,15	3,87	3,65	3,48	3,34
100	11,50	7,41	5,86	5,02	4,48	4,11	3,83	3,61	3,44	3,30
9999	10,83	6,91	5,42	4,62	4,10	3,74	3,47	3,27	3,10	2,96

	Täljarens *df*								
Nämnarens *df*	12	15	20	25	30	40	50	100	9999
1	610352	616074	620842	623703	626087	628471	630379	633240	636578
2	999,3	999,3	999,3	999,3	999,3	999,3	999,3	999,3	999,3
3	128,3	127,4	126,4	125,8	125,4	125,0	124,7	124,1	123,5
4	47,41	46,76	46,10	45,69	45,43	45,08	44,88	44,47	44,05
5	26,42	25,91	25,39	25,08	24,87	24,60	24,44	24,11	23,79
6	17,99	17,56	17,12	16,85	16,67	16,44	16,31	16,03	15,75
7	13,71	13,32	12,93	12,69	12,53	12,33	12,20	11,95	11,70
8	11,19	10,84	10,48	10,26	10,11	9,92	9,80	9,57	9,33
9	9,57	9,24	8,90	8,69	8,55	8,37	8,26	8,04	7,81
10	8,45	8,13	7,80	7,60	7,47	7,30	7,19	6,98	6,76
11	7,63	7,32	7,01	6,81	6,68	6,52	6,42	6,21	6,00
12	7,00	6,71	6,40	6,22	6,09	5,93	5,83	5,63	5,42
13	6,52	6,23	5,93	5,75	5,63	5,47	5,37	5,17	4,97
14	6,13	5,85	5,56	5,38	5,25	5,10	5,00	4,81	4,60
15	5,81	5,54	5,25	5,07	4,95	4,80	4,70	4,51	4,31
16	5,55	5,27	4,99	4,82	4,70	4,54	4,45	4,26	4,06
17	5,32	5,05	4,78	4,60	4,48	4,33	4,24	4,05	3,85
18	5,13	4,87	4,59	4,42	4,30	4,15	4,06	3,87	3,67
19	4,97	4,70	4,43	4,26	4,14	3,99	3,90	3,71	3,51
20	4,82	4,56	4,29	4,12	4,00	3,86	3,77	3,58	3,38
22	4,58	4,33	4,06	3,89	3,78	3,63	3,54	3,35	3,15
24	4,39	4,14	3,87	3,71	3,59	3,45	3,36	3,17	2,97
26	4,24	3,99	3,72	3,56	3,44	3,30	3,21	3,02	2,82
28	4,11	3,86	3,60	3,43	3,32	3,18	3,09	2,90	2,69
30	4,00	3,75	3,49	3,33	3,22	3,07	2,98	2,79	2,59
35	3,79	3,55	3,29	3,13	3,02	2,87	2,78	2,59	2,38
40	3,64	3,40	3,15	2,98	2,87	2,73	2,64	2,44	2,23
50	3,44	3,20	2,95	2,79	2,68	2,53	2,44	2,25	2,03
60	3,32	3,08	2,83	2,67	2,55	2,41	2,32	2,12	1,89
70	3,23	2,99	2,74	2,58	2,47	2,32	2,23	2,03	1,79
80	3,16	2,93	2,68	2,52	2,41	2,26	2,16	1,96	1,72
90	3,11	2,88	2,63	2,47	2,36	2,21	2,11	1,91	1,66
100	3,07	2,84	2,59	2,43	2,32	2,17	2,08	1,87	1,62
9999	2,74	2,51	2,27	2,11	1,99	1,84	1,73	1,49	1,00

Tabell B: Engelsk–svensk ordlista

Alternative hypothesis ..... Mothypotes
Analysis of variance ..... Variansanalys
Approximation ..... Approximation, ungefärlig beräkning
Arithmetic mean ..... Aritmetiskt medelvärde
Average ..... Genomsnitt

Bias ..... Systematiskt fel
Binomial distribution ..... Binomialfördelning

Central limit theorem ..... Centrala gränsvärdessatsen
Combinations ..... Kombinationer
Chi-square distribution ..... Chitvå-fördelning, -fördelning
Coefficient of determination ..... Determinationskoefficient
Conditional probability ..... Betingad sannolikhet
Confidence interval ..... Konfidensintervall
Contingency table ..... Korstabell, kontingenstabell
Continous distribution ..... Kontinuerlig fördelning
Correlation ..... Korrelation
Correlation coefficient ..... Korrelationskoefficient
Covariance ..... Kovarians, covarians
Cumulative probability ..... Kumulativ sannolikhet

Degreees of freedom ..... Frihetsgrader
Density function ..... Täthetsfunktion
Disjoint events ..... Ömsesidigt uteslutande händelser
Discrete distribution ..... Diskret fördelning
Distribution ..... Fördelning
Distribution function ..... Fördelningsfunktion

Error of type I ..... Typ 1-fel
Estimate ..... Estimat, skattning (substantiv)
Estimation ..... Estimering, skattning (verb)
Event ..... Händelse
Expected value ..... Väntevärde
Exponential distribution ..... Exponentialfördelningen

F-distribution ..... F-fördelningen

Geometric distribution ..... Geometrisk fördelning
Goodness of fit ..... Anpassningsgrad

Hypergeometric distribution ..... Hypergeometriska fördelningen
Hypothesis ..... Hypotes

Independent events ..... Oberoende händelser
Interquartile range ..... Interkvartila intervallet
Intersection ..... Snitt

Kurtosis ..... Toppighet

Linear regression ..... Linjär regression(-sanalys)

Mean ..... Medelvärde
Mean square ..... Medelkvadrat
Median ..... Median
Mode ..... Typvärde
Multicollinearity ..... Multicolinearitet
Multiple regression ..... Multipel regression(-sanalys)

Non-parametric ..... Ickeparametrisk, parameterfri, fördelningsfri
Normal distribution ..... Normalfördelning
Null hypothesis ..... Nollhypotes

One-sided test ..... Ensidigt test
Outcome ..... Utfall
Outlier ..... Extremt avvikande observation

Permutations ..... Permutationer
Point estimation ..... Punktskattning
Poisson distribution ..... Poissonfördelning
Pooled variance ..... Sammanvägd varians
Probability ..... Sannolikhet
Probability distribution ..... Sannolikhetsfördelning, fördelning
Probability function ..... Sannolikhetsfunktion
Probability space ..... Utfallsrum

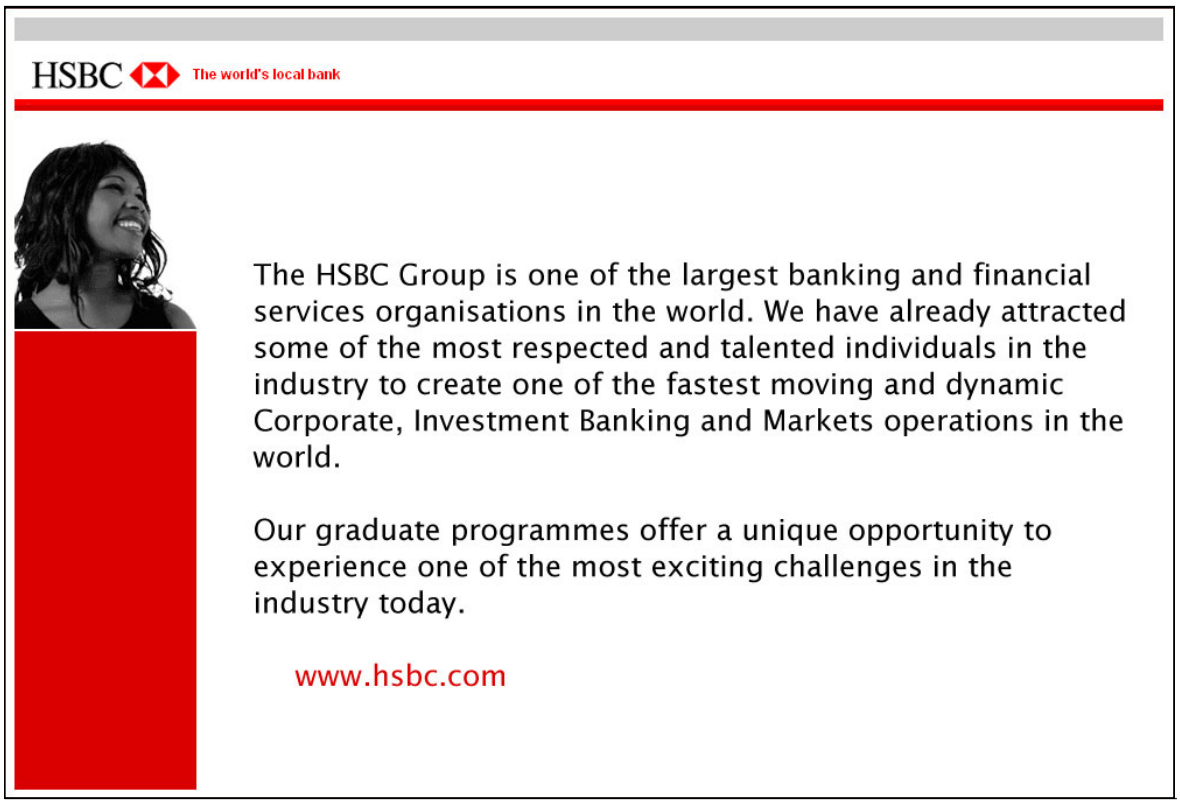
Quartile ..... Kvartil

Random variable ..... Slumpvariabel
Range ..... Intervall
Rank sum ..... Rangsumma
Reject ..... Förkasta

Sample ..... Stickprov, urval
Sample size ..... Stickprovsstorlek
Sampling distribution ..... Samplingfördelning, stickprovsfördelning
Skewness ..... Skevhet
Slope ..... Lutning, riktningskoefficient
Standard deviation ..... Standardavvikelse
Statistic ..... Statistiska, testvariabel

Unbiased ..... Väntevärdesriktig
Uniform distribution ..... Likformig fördelning
Union ..... Union

Variance ..... Varians



Index