Statistik 1 för biologer, logopeder och psykologer

Paul Blomstedt

Innehåll

1 Inledning

Vad är statistik?

Statistik är inte läran om att föra statistik.

Statistiska undersökningar – målsättningar.

- Strävar i allmänhet efter att
  - beskriva
  - förklara
  - göra prognoser för
  - kontrollera

  olika fenomen.
- En förutsättning är att man ska kunna samla in information om fenomenet numerisk form.
- Man hoppas kunna skilja åt de *regelbundna* och de *slumpmässiga* karaktärsdragen i fenomenet.

Statistiska undersökningar – delmoment.

I en statistisk undersökning ingår i allmänhet följande tre moment:

- att insamla
- att sammanställa
- att dra slutsatser

av ett datamaterial.

2 Deskriptiv statistik

2.1 Variabler och datamaterial

Variabler

- En variabel är en storhet som varierar från individ till individ.
- Exempel på variabler är
  - längd
  - antal barn
  - utbildning
- En variabel kan anta olika värden.
- Ett datamaterial består av flera observationer på en eller flera variabler.

Kvantitativa och kvalitativa variabler

Kvantitativa variabler antar numeriska värden

- t.ex. ålder och vikt.

Kvalitativa variabler antar inte numeriska värden

- t.ex. utbildning och kön.

Kontinuerliga och diskreta variabler

Kontinuerliga variabler kan anta alla tänkbara värden i ett visst intervall

- t.ex. längd och vikt.

Diskreta kan endast anta vissa distinkta värden

- t.ex. antal barn och kön.

Skaltyper

Mätningar av variabler kan göras på olika skalnivåer:

- nominalskala
- ordinalskala
- intervallskala
- kvotskala

Skaltypen påverkar sättet att framställa och analysera datamaterialet.

Nominalskala

- Talar om för oss *vilken* klass en observation tillhör.
- Klasserna kan inte rangordnas sinsemellan.
- Exempel på observationer mätta på nominalskala:
  - kön
  - blodgrupp
  - hemstad.

Ordinalskala

- Talar om för oss vilken klass en observation tillhör samt om observationen har *mer* av en egenskap än en annan observation.
- Klasserna *kan* rangordnas sinsemellan
- Exempel på observationer mätta på ordinalskala:
  - militärgrad
  - klädstorlek: S, M, L, XL
  - vitsord i studentexamen.

Intervallskala

- Talar om för oss *hur mycket* en observation skiljer sig från en annan observation.
- Observationerna har numeriska värden men saknar en absolut nollpunkt.
- Exempel på observationer mätta på intervallskala:
  - temperatur mätt i Celsius eller Farenheit
  - vattenstånd i cm över en viss referenspunkt
  - datum.

Kvotskala

- Talar om för oss *hur många gånger* en observation har mer av en egenskap än en annan observation har.
- Numeriska värden som det är möjligt att bilda kvoter av.
- Exempel på observationer mätta på kvotskala:
  - temperatur mätt i Kelvin (där en absolut nollpunkt är definierad)
  - längd, bredd, vikt
  - ålder.

Fördelning av ett datamaterial

- För vidare analyser är det viktigt att bilda sig en uppfattning om det insamlade datamaterialet.
- Kan vara svårt att greppa utan någon form av sammanställning eller sammanfattning.
- Vi är ofta intresserade av hur observationerna är fördelade.

- Olika sätt att beskriva en fördelning:
  - tabulering
  - grafisk beskrivning
  - användning an central- och spridningsmått.

2.2 Tabulering och grafisk beskrivning

2.2.1 Diskreta observationer

Frekvenstabeller

- Fördelningen av ett diskret datamaterial kan presenteras i form av en frekvenstabell
- Anger i tabellform antalet (=frekvensen) observationer tillhörande respektive klasser.
- Ofta anges också de relativa frekvenserna.

Tabell 1: Exempel på en frekvenstabell.

åsikt	frekvens	rel. frekv.
A	12	$12/73 \approx 0.16$
B	29	$29/73 \approx 0.40$
C	14	$14/73 \approx 0.19$
D	7	$7/73 \approx 0.10$
E	11	$11/73 \approx 0.15$
sammanlagt	73	

- De relativa frekvenserna kan också anges som procentandelar.

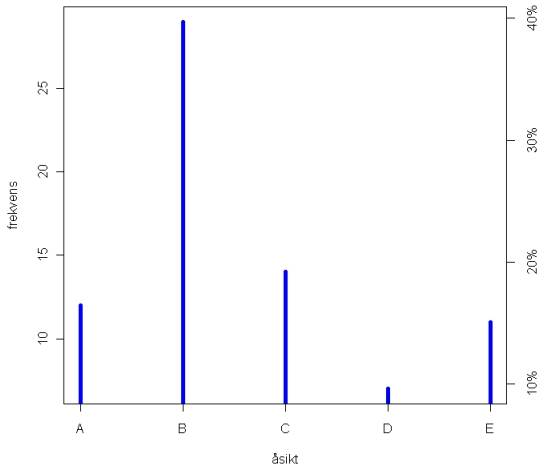
Stolpdiagram

- Diskreta fördelningar kan också presenteras grafiskt i form av stolpdiagram.
- Innehåller samma information som en frekvenstabell.
- En variant av stapeldiagram.

Korstabeller

Fördelningen av *bivariata* observarioner (två variabler observerade av samma individ) kan presenteras i form av en korstabell.

Figur 1: Exempel på ett solpdiagram.



Tabell 2: Korstabell.

		INTELLIGENS		
		låg	medel	hög
ANPASSNINGS-FÖRMÅGA	låg	26	43	20
	medel	54	96	45
	hög	22	45	24

2.2.2 Kontinuerliga observationer

Klassindelning

- För kontinuerliga variabler får vi sällan två observationer med exakt samma värde.
- Det blir därmed meningslöst att i tabellform räkna upp frekvenserna för alla observerade värden.
- För att kunna konstruera en frekvenstabell blir det nödvändigt med klassindelning av datamaterialet.

Histogram

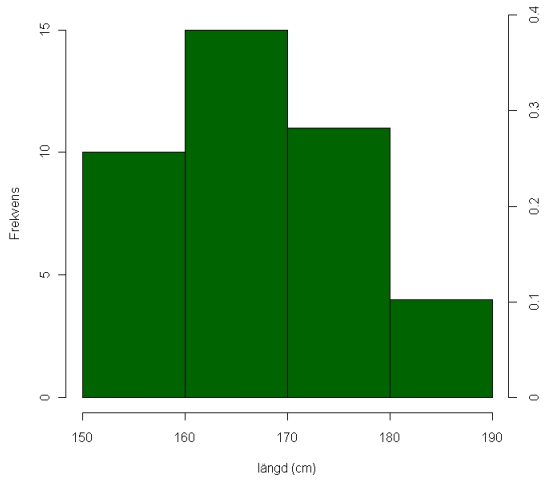
- Ett histogram är ett slags stapeldiagram där staplarna är fast i varandra.

Tabell 3: Frekvenstabell på klassindelat datamaterial över längder.

längd (cm)	frekvens	rel. frekv.
150–159	10	$10/39 \approx 0.26$
160–169	15	$15/39 \approx 0.38$
170–179	11	$11/39 \approx 0.28$
180–189	4	$4/39 \approx 0.10$
sammanlagt	39	

- Om klasserna är lika breda motsvarar staplarnas höjd klassfrekvenserna.
- Om klasserna *inte* är lika breda måste frekvenserna korrigeras i förhållande till klassbredden.

Figur 2: Histogram på datamaterialet över längder.

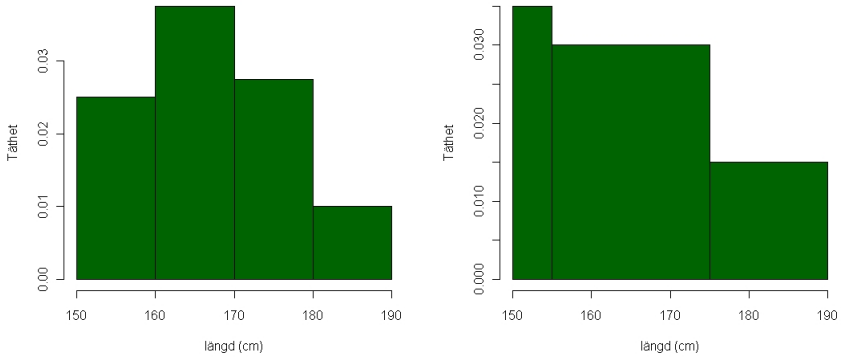


Korrigering av klassfrekvens

Kumulativ frekvens

- Ofta vill inte endast veta klassfrekvenserna utan även hur många observationer som är mindre än ett visst värde.
- Vi talar då om den kumulativa frekvensen

Figur 3: Histogram över samma datamaterial med jämn och ojämn klassindelning.



Tabell 4: Frekvenstabell med korrigerad klasfrekvens.

klass	klassbredd	frekv.	korrigerad frekv.
12.5–13.4	1	2	3
13.5–14.4	1	7	7
14.5–15.4	2	12	$12/2 = 6$
16.5–20.4	4	6	$6/4 = 1.5$
sammanlagt		28	

- Den kumulativa frekvensen av den första klassen är samma som klassfrekvensen.

Summapolygon

- Summaplygonet är en grafisk beskrivning av den kumulativa frekvensen.

- En annan benämning på summapolygonet är empirisk fördelningsfunktion.

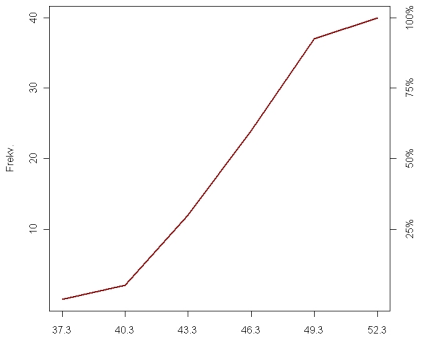
Andra typer av diagram

- En förteckning över vanliga diagramtyper:
  `http://sv.wikipedia.org/wiki/Diagram`

Tabell 5: Frekvenstabell med kumulativa frekvenser.

klass	frekv.	kumulativ frekv.	relativ kumulativ frekv.
37.3–40.2	2	2	$2/40 \cdot 100\% = 5.0\%$
40.3–43.2	10	$2 + 10 = 12$	$12/40 \cdot 100\% = 30.0\%$
43.3–46.2	12	$12 + 12 = 24$	$24/40 \cdot 100\% = 60.0\%$
46.3–49.2	13	$24 + 13 = 37$	$37/40 \cdot 100\% = 92.5\%$
49.3–52.2	3	$37 + 3 = 40$	100%

Figur 4: Summapolygon.



- Se även den engelskspråkiga sidan för lite utförligare förklaringar av de vanligaste diagramtyperna:
  `http://en.wikipedia.org/wiki/Chart`

3 Central- och spridningsmått

3.1 Centralmått

Typvärde

- Typvärdet är den klass / det värde som har den högsta frekvesnsen i en frekvenstabell.
- I ett stapeldiagram/histogram är den högsta stapeln vid typvärdet.
- Det kan finnas flera typvärden i ett datamaterial.
- Kan bestämmas för observationer mätta på alla skalnivåer.

Medelvärde

- Det aritmetiska melevärdet definieras som

$$\overline{x} = \frac{\text{summan av observationerna}}{\text{antalet observationer}} = \frac{\sum_{i=1}^{n} x_i}{n} .$$

- Medelvärdet kan endast räknas för intervall- och kvotskalevariabler.

Exempel. 1. Medelvärdet av talen 3, 6 , 8, 4, 7, 1 räknas som

$$\frac{3+6+8+4+7+1}{5} = 4.83 .$$

Viktat medelvärde

- För att räkna medelvärdet på ett datamaterial på basen av en frekvenstabell använder vi oss av ett viktat medelvärdet.
- Det viktade medelvärdet räknas genom att vikta (=multiplicera) variabelns varje värde med antalet gånger det har dykt upp i datamaterialet (=frekvensen).
- För ett kontinuerligt klassindelat datamaterial kan klassmedelpunkterna användas som värde för variabeln.

Exempel. 2. Vi betraktar följande frekvenstabell:

antal bilar per hushåll	frekvens
0	5
1	3
2	1
3	1
sammanlagt	10

Det viktade medlvärdet av datamaterialet är

$$\frac{5 \cdot 0 + 3 \cdot 1 + 1 \cdot 2 + 1 \cdot 3}{10} = 0.8 .$$

Viktat medelvärde med relativa frekvenser.

Har vi tillgång till de relativa frekvenserna får vi medeltalet direkt som en viktad *summa* av de observerade värdena.

Exempel. 3.

antal bilar per hushåll	frekvens	rel. frekv.
0	5	5/10=0.5
1	3	3/10=0.3
2	1	1/10=0.1
3	1	1/10=0.1
sammanlagt	10	

Det viktade medlvärdet räknas nu som

$$0.5 \cdot 0 + 0.3 \cdot 1 + 0.2 \cdot 1 + 0.2 \cdot 1 = 0.8 \, .$$

Jfr uträkningen i föregående exempel!

Median

- Medianen är det mittersta värdet i ett datamaterial som är ordnat från den minsta observationen till den största.
- Medianen lämpar sig för variabler av alla andra skaltyper förutom nominalskalan.
- Om datamaterialet innehåller ett jämnt antal observationer är medianen någondera av de två mittersta observationerna eller, om möjligt, medeltalet av dem.

Median, kvantiler och fraktiler

- För ett kontinuerligt datamaterial är medianen det värde där den relativa kumulativa frekvensen är 0.5 (dvs. 50%).
- På samma sätt kan även den undre kvartilen (rel. kum. frekv. 0.25) och övre kvartilen (rel. kum. frekv. 0.75) bestämmas.
- Mer allmänt kan fraktiler för vilken relativ kumulativ frekvens som helst betämmas.
- Om de relativa kumulativa frekvenserna är angivna i procent kallas fraktilerna ofta *percentiler*.

Grafisk bestämning av median och kvartiler

Medianen, kvantiler och fraktiler kan även bestämmas grafiskt från en empirisk fördelningsfunktion.

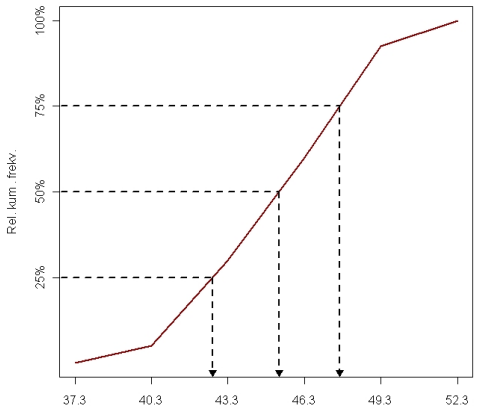
Boxplot

En boxplot (kallas även låddiagram, lådogram) sammanfattar grafiskt en fördelning i följande 5 punkter (räknat uppifrån ner):

- högsta värdet
- övre kvartilen
- medianen
- undre kvartilen
- mista värdet.

Boxplot lämpar sig speciellt bra för jämförelser av en och samma variabel mellan olika grupper eller experiment.

Figur 5: Grafisk bestämning av median och kvartiler.



3.2 Spridningsmått

Variationsvidd

- Variationsvidden anger avståndet mellan det mista och det största värdet i ett datamaterial.
- Variationsvidden definieras som

$$\text{variationsvidd} = \text{största värdet} - \text{minsta värdet}\,.$$

- Variationsvidden kan endast bestämmas för intervall- och kvotskalevariabler.

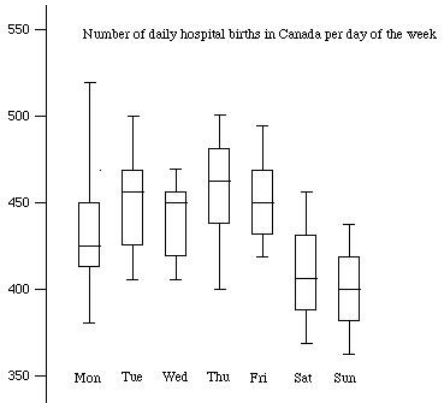
Kvartilavstånd

- Kvartilavståndet anger avståndet mellan den undre och den övre kvartilen.
- Kvartilavståndet definieras som

$$\text{kvartilavstånd} = \text{övre kvartilen} - \text{undre kvartilen}\,.$$

- Kvartilavståndet kan endast bestämmas för intervall- och kvotskalevariabler.

Figur 6: Boxplot.



Standardavvikelse och varians.

- Standardavvikelsen anger hur mycket observationerna i ett kvantitativt datamaterial i *genomsnitt* avviker från medelvärdet.
- Standardavvikelsen beräknas enligt

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

  där x_i är den i:te observationen, $\overline{x}$ är medelvärdet på alla observationer och n är antalet observationer.
- Lämnar vi bort kvadratroten i uttrycket ovan får vi variansen s^2.
- Standardavvikelsen anges i samma enhet som observationerna (t.ex. cm) medan variansen är en dimensionslös storhet.

Exempel. 4. Vi beräknar variansen och standardavvikelsen av talen

$$1, 3, 5, 8, 9, 10 \, .$$

Vi börjar med att räkna avvikelsen mellan varje observation och deras medelvärde 6. Då avvikelserna är

$$-5, -3, -1, 2, 3, 4$$

och det totala antalet observationer är 6, blir variansen

$$\frac{(-5)^2 + (-3)^2 + (-1)^2 + 2^2 + 3^2 + 4^2}{5} = 12.8$$

och standardavvikelsen $\sqrt{12.8} \approx 3.6$.

Variationskoefficient

- Variationskoefficienten är en normaliserad standardavvikelse som uttrycker hur många procent i genomsnitt observationerna avviker från medelvärdet.
- Variationskoefficienten definieras som

$$\text{variationskoefficient} = \frac{\text{standardavvikelse}}{\text{medelvärde}} \cdot 100\% = \frac{s}{\overline{x}} \cdot 100\% \, .$$

- Möjliggör jämförelser av standardavvikelser beräkande av datamaterial där observationerna är mätta i olika enheter.
- Används endast på icke-negativa data.

Om användninga av summatecknet

- Kort om använding av summatecknet $\sum$ som förekommer i en del av formlerna för central- och spridningsmått:
  `http://sv.wikipedia.org/wiki/Summatecken`

Fallgropar med vanliga central- och spridningsmått

`http://web.abo.fi/fak/mnf/mate/jc/statistik1/DeskriptivtExempel.pdf`

4 Korrelation och regression

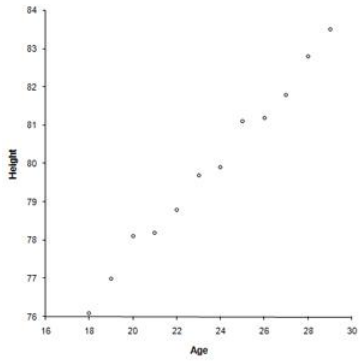
Spridningsdiagram

- Då ett datamaterial består av två (eller flera) variabler är man ofta intresserad av att veta om det finns ett samband mellan variablerna.
- Om variablerna är kvantitativa kan man bilda sig en uppfattning om ett eventuellt samband genom att grafiskt märka ut observationerna i ett koordinatsystem.
- En sådant diagram kallas kallas spridningsdiagram eller *scatter plot*.

Korrelation

- Korrelation anger styrkan och riktningen av sambandet mellan två variabler.
- Om vi anpassar en rät linje genom punkterna i ett spridningsdiagram anger korrelationen hur stor eller liten spridningen kring linjen är.
- Fast inte alla samband är linjära, ger korrelationen ändå *ofta* (men inte alltid!) en god uppfattning om huruvida ett samband finns eller inte.

Figur 7: Ett spridningsdiagram som visar sambandet mellan ålder och längd hos ett antal individer. Vi ser ett klart *linjärt* samband mellan variablerna.



Pearson's korrelationskoefficient

- Pearson's korrelationskoefficient definieras som

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{(n-1) \cdot s_x \cdot s_y} \ ,$$

  där s_x är standardavvikelsen av talen x_i, s_y är standardavvikelsen av talen y_i och n är antalet observationspar.

- Korrelationskoefficienten r får ett värde mellan -1 och 1:
  - om $r = -1$, befinner sig alla punkter på en linje med negativ lutning
  - om $r = 0$, finns inget linjärt samband mellan variablerna
  - om $r = 1$, befinner sig alla punkter på en linje med positiv lutning.

Regression

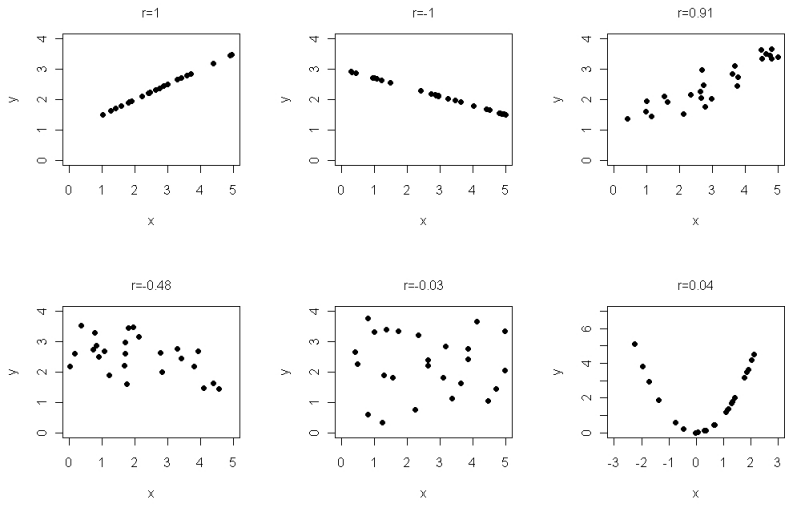
- Ett samband mellan två variabler kan vidare granskas genom regressionsanalys.
- Regression är en allmän benämning på modeller där värdet på en sk. *beroende* variabel förklaras med hjälp av en eller flera sk. *förklarande* variabler.
- Den vanligaste formen av regression är *linjär regression*, där sambandet mellan variablerna antas vara linjärt.

Enkel linjär regression

- Aningen förenklat kan ekvationen för enkel linjär regression (linjär regression med *en* förklarande variabel) skrivas som

$$y = a + b \cdot x \ ,$$

Figur 8: Datamaterial med olika grader av korrelation.

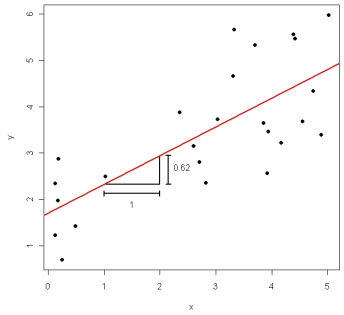


där y är den beroende variabeln och x är den förklarande variabeln.

- Konstanterna a (skärningspunkten med y-akseln) och b (riktningskoefficienten) skattas från datamaterialet men hjälp av *minsta kvadrat-metoden*.

Regressionskoefficienten

- Riktings- eller regressionscoefficienten b talar om för oss hur mycket den beroende variabeln y ändras då den förklarande variabeln x ökar med 1.
- T.ex. i den linjära modellen $y = a + b \cdot x = 1.70 + 0.62x$ nedan leder en ökning av enhetslängd i variabeln x till en ökning av 0.62 i variabeln y.



Förklaringsgrad

- Förklaringsgraden anger vilken andel (ofta i procent) av den beroende variablens värde i en regressionsmodell som kan förklaras med den förklarande variabeln.
- Förklaringsgraden betecknas R^2 och definieras som

$$R^2 = r^2 \cdot 100\% \,,$$

  där r är korrelationskoefficienten.
- Om korrelationen är -1 eller 1, blir $R^2 = 100\%$. Då kan datamaterialets y-värden fullständigt förklaras med x.
- Om korrelationen är 0 blir $R^2 = 0\%$. Då kan datamaterialets y-värden över huvudtaget inte förklaras med x.

En del vanliga misstolkningar

- Även om y delvis kan förklaras med x kan vi inte säga att y *förorsakas* av x.
- Ett synbarligen starkt samband mellan två orelaterade variabler kan uppstå om båda påverkas av en tredje variabel:
  - t.ex. kan man påvisa ett positivt samband mellan glassförsäljning och antalet drunkningsolyckor. Fast variablerna är totalt orelaterade påverkas de båda av årstiden.

Partiell korrelation och justering

Ibland kan ett samband mellan två variabler x och y på något sätt förvrängas av en tredje variabel z. Såväl korrelations- som regressionskoefficienten kan då få felaktiga värden.

- En korrelationskoefficient mellan x och y där inverkan av z har beaktats, kallas partiell korrelation och betecknas $r_{xy \cdot z}$ (se definition samt exempel `http://faculty.vassar.edu/lowry/ch3a.html`).
- Regressionsmodellen $y = a + bx$ justeras för inverkan av z genom att man lägger till z som förklarande variabel i modellen. I den nya modellen $y = a + bx + cz$ får koefficenten b ett värde där inverkan av z har beaktats.

Rangkorrelation

- Pearson's korrelationskoefficient (dvs. "vanlig" korrelation) lämpar sig endast för intervall- och kvotskalevariabler.

- För ordinalskalevariabler kan man i stället andvända sig av Spearman's rangkorrelationskoefficient

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n(n^2-1)} \ ,$$

där d_i är differensen mellan rangtalen av observationerna för individ i och n är antalet observationspar (individer).

- Ett motsvarande rangkorrelationsmått är Kendall's tau, se `http://en.wikipedia.org/wiki/Kendall_tau_rank_correlation_coefficient`.

Spearman's rangkorrelation

Exempel. 5. Två vinsmakare A och B ordnar sex vinprover i rangordning från den bästa (1) till den sämsta (6). Hur väl stämmer åsikterna överens?

vinprov	rang		d_i	d_i^2
	A	B		
a	5	3	2	4
b	2	1	1	1
c	1	2	-1	1
d	4	4	0	0
e	3	5	-2	4
f	6	6	0	0

Rangkorrelationen blir då

$$\begin{aligned} r_s &= 1 - \frac{6 \cdot 10}{6(36-1)} = 1 - \frac{2}{7} \\ &\approx 0.71 \ . \end{aligned}$$

5 Grunderna i sannolikhetslära

Satistik och sannolikhetslära

- Statistik handlar om att utvinna information från data. I praktiken inhehåller de data man analyserar oftast någon form av variabilitet eller osäkerhet. Denna variabilitet strävar man efter att kvantifiera med hjälp av sannolikhetslära.Sannolikhetslära är således nödvändigt för att förstå statistiska analysmetoder.

5.1 Grundbegrepp

Slumpmässiga försök

- Ett slumpmässigt försök är en företeelse som kan upprepas under likartade förhållanden.

- Resultatet kan inte anges i förväg även om man många gånger tidigare utfört samma försök.
- T.ex. att kasta tärning kan ses som ett slumpmässigt försök.

Utfall och händelser

- Ett utfall är resultatet av ett slumpmässigt försök.
- Alla tänkbara utfall tillsammans kallas utfallsrummet.
  - T.ex. i tärningskast är utfallsrummet $\{1, 2, 3, 4, 5, 6\}$.
- En händelse är en samling ufall.
  - T.ex. "vi får ett udda tal" $= \{1, 3, 5\}$
  - Händelser betecknas ofta med stora bokstäver $A, B, C, \ldots$
  - Ibland kan det vara nyttigt att visualisera kombinationer av händelser med hjälp av sk. Venn diagram.

Sannolikhet

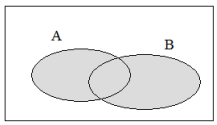
- Sannolikhet är ett tal mellan 0 och 1 som förknippas med en händelse eller kombination av händelser.
  - 0 betyder att händelsen är *omöjlig*.
  - 1 betyder att händelsen är *säker*.
- Sannolikheten för händelsen A betecknas $P(A)$.
- Sannolikheten för händelsen A eller B betecknas $P(A \cup B)$.
- Alternativt kan man explicit skriva ut $P(A$ inträffar), $P(A$ eller $B)\ldots$

Tolkning av sannolikhet

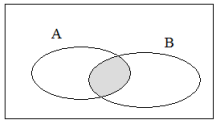
Sannolikhet kan tolkas på flera olika sätt:

- Enligt den klassiska tolkningen definieras sannolikhet som $P(A) = \dfrac{\text{"antalet för } A \text{ gynnsamma utfall"}}{\text{"totala antalet utfall"}}$.
- Den empiriska (eller *frekventistiska*) tolkningen av sannolikhet är $P(A) =$ den relativa frekvensen för A i ett stort antal försök.
- Det finns även en subjektiv tolkning där $P(A) =$ subjektiv uppfattning om hur trolig händelsen A är.

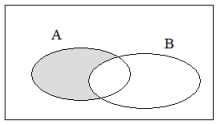
Figur 9: Venndiagram med union av två händelser: $A \cup B =$ "A eller B".



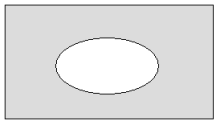
Figur 10: Snitt av två händelser: $A \cap B =$ "A och B".



Figur 11: Differens av två händelser: $A \backslash B =$ "A men inte B".



Figur 12: Komplementhändelse: $A^c =$ "inte A".



5.2 Egenskaper och räkneregler

Komplement.

- Om händelsen E består av hela utfallsrummet får vi $P(E) = 1$.
- Sannolikheten för komplementhändelsen till A, dvs. händelsen $A^c =$ "inte A" är $P(A^c) = 1 - P(A)$.
- Sannolikheten för komplementet till utfallsrummet är $P(E^c) = 1 - P(E) = 0$.
  - Utfallsrummets komplement E^c (den omöjliga händelsen) betecknas ofta $\emptyset$.

Additionssatsen

- Sannolikheten för att händelsen A eller B inträffar är

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

- Ordet "eller" förknippas i sannolikhetskalkyl med *addition*.
- För att inte räkna de gemensamma utfallen för A och B två gånger subtraherar vi $P(A \cap B)$ från summan av sannolikheterna $P(A)$ och $P(B)$.

Exempel. 6. Av ett tillverkat parti enheter har 2% fel vikt, 4% fel färg och 1% både fel vikt och färg. Vi räknar sannolikheten att en slumpmässigt vald enhet har antingen fel vikt eller fel färg (eller båda)?

Låt $A =$ "enheten har fel vikt" och $B =$ "enheten har fel färg". De angivna sannolikheterna är då $P(A) = 0.02$, $P(B) = 0.04$ och $P(A \cap B) = 0.01$. Från additionssatsen får vi

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) = 0.02 + 0.04 - 0.01 = 0.05\,.$$

Oförenliga händelser

- Om händelserna A och B är oförenliga utesluter de varandra (endast en av dem åt gången kan inträffa).
- Sannolikheten för "A och B" blir då $P(A \cap B) = 0$.
- Additionssatsen för oförenliga händelser förenklas till

$$P(A \text{ eller } B) = P(A \cup B) = P(A) + P(B).$$

- T.ex. sannolikheten att få resultatet 2 eller 3 vid *ett* tärningskast är

$$P(\text{vi får 2}) + P(\text{vi får 3}) = \frac{1}{6} + \frac{1}{6} = \frac{1}{3}$$

eftersom resultaten utesluter varandra.

Betingad sannolikhet

- Den betingade sannolikheten för A givet B betyder att vi begränsar oss till att endast betrakta sådana utfall som tillhör händelsen B.
- Vi kan också tänka oss den betingade sannolikheten för A givet B som sannolikheten för A då vi vet att B har inträffat.
- Sannolikheten för A påverkas av vår kunskap om händelsen B.
- Den betingade sannolikheten för A givet B definieras som

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

  där vi antar att $P(B) > 0$.

- Den betingade sannolikheten kan även tolkas som:

$$P(A|\,B) = \frac{\text{"antalet för } A \text{ gynnsamma utfall i } B\text{"}}{\text{"totala antalet utfall i } B\text{"}}\ .$$

Exempel. 7. Vi drar slumpmässigt ett kort ur en kortlek på 52 kort. Vi frågar oss först vad sannoikheten är att det dragna kortet är en kung. Vi betecknar $A =$ "kung". Eftersom det finns sammanlagt 4 kungar i en kortlek blir sannolikheten då

$$P(A) = 4/52 \approx 0.08\ .$$

Låt oss vidare anta att vi vet att det dragna kortet är ett bildkort. Vi betecknar $B =$ "bildkort". Vad sannolikheten nu att det dragna kortet är en kung? Eftersom det finns sammanlagt 12 bildkort i en kortlek av vilka 4 är kungar blir sannolikheten

$$P(A|B) = 4/12 \approx 0.33\ .$$

Sannolikheten ovan är den *betingade sannolikheten* för att få en kung givet att vi har dragit ett bildkort.

Multiplikationssatsen

- Sannolikheten för att händelserna A och B inträffar är

$$P(A \cap B) = P(B)P(A|B) = P(A)P(B|A).$$

- Ordet "och" förknippas i sannolikhetskalkyl med *multiplikation*.
- Betydelsen av den betingade sannolikheten i formeln blir tydligare om vi tänker oss att A och B inträffar i följd: först inträffar A och då vi vet att A har inträffat inträffar B.

Exempel. 8. Vi har en skål med 3 vita bollar och 5 röda bollar, dvs totalt 8 bollar. Vi räknar sannolikheten för att då vi slumpmässigt drar två bollar ur skålen är båda röda. Vi har då händelserna

$$A = \text{"vi drar en röd boll"}$$
$$B|A = \text{"vi drar en röd givet att vi redan dragit en röd boll"}$$

som ger sannolikheten

$$P(A)P(B|A) = \frac{5}{8} \cdot \frac{4}{7} = \frac{20}{56} \approx 0.36 \, .$$

Oberoende händelser

- Två händelser A och B är oberoende (betecknas ofta $A \perp\!\!\!\perp B$) om händelserna inte på något sätt påverkar varandra.
- För de oberoende händelserna A och B gäller att

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

  jfr. den allmänna multiplikationssatsen!
- Sannolikheten för B påverkas inte av vår kunskap om händelsen A:

  $$P(B|A) = P(B)$$

  dvs. sannolikheten för B förblir densamma oberoende om vi betingar med A eller inte.

Exempel. 9. Vi har händelserna

$$A = \text{"vi får krona då vi singlar slant"}$$
$$B = \text{"vi får resultatet 4 i tärningskast"}.$$

Det är uppenbart att händelserna är oberoende och sannolikheten för "A och B" blir således

$$P(A \cap B) = P(A)P(B) = \frac{1}{2} \cdot \frac{1}{6} = \frac{1}{12} \approx 0.08 \, .$$

Total sannolikhet och Bayes sats.

- Den totala sannolikheten anger sannolikheten för en händelse som kan inträffa på ett antal alternativa sätt.
- Med hjälp av Bayes sats räknar man ut sannolikheterna för de enskilda alternativa sätten då vi vet att händelsen har inträffat.
- Vi belyser begreppen genom ett exempel:
  `http://web.abo.fi/fak/mnf/mate/kurser/statistik1/TotSann&Bayes.pdf`

5.3 Fördelningar

Slumpvariabler

- En variabel som för varje utfall av ett slumpmässigt försök antar ett reelt tal kallas slumpvariabel.
- Slumpvariabler betecknas ofta med stora bokstäver från slutet av alfabetet: $X, Y, Z, \ldots$

Exempel. 10. Vi singlar två slantar och observerar antalet kronor. I stället för att definiera händelserna $A =$ "vi får 0 kronor", $B =$ "vi får 1 krona", $C =$ "vi får 2 kronor" kan vi definiera en slumpvariabel X som kan anta värdena 0, 1, 2.

- En *diskret* slumpvariabel kan anta ett uppräkneligt antal distinkta värden, medan en *kontinuerlig* slumpvariabel kan anta ett oändligt antal värden i ett givet intervall.

Sannolikhetsfördelning

- Sannolikhetsfördelningen för en slumpvariabel anger med vilken sannolikhet variabeln antar olika värden.

Exempel. 11. Låt den diskreta slumpvariablen X beteckna antalet kronor då vi singlar två slantar. Vi får då följande sannolikhetsfördelning för variablen:

$$\begin{aligned} P(X = 0) &= 0.25 \\ P(X = 1) &= 0.5 \\ P(X = 2) &= 0.25\,. \end{aligned}$$

Väntevärde och varians

- Vi har tidigare sett att empiriska fördelningar av datamaterial kan beskrivas med hjälp av central- och spridningsmått.
- Motsvarande mått används även för att beskriva sannolikhetsfördelningar för slumpavriabler.
- Det genomsnittliga värdet av slumpvariabeln X kallas väntevärde och betecknas $E(X)$ eller μ.
  - Väntevärdet motsvarar medelvärdet av en slumpvariabel då vi upprepar ett slumpmässigt försök oändligt många gånger.
- Variansen, som betecknas $Var(X)$ eller σ^2, är ett mått på hur utspridd fördelningen är kring väntevärdet.
  - Variansen definieras som $Var(X) = E[(X - \mu)^2]$.

Väntevärde

Väntevärdet för en diskret slumpvariabel fås genom att räkna en viktad summa av variabelns alla värden, där vikterna utgörs av sannolikheterna för värdena.

Exempel. 12. Väntevärdet för slumpvariabeln X som betecknar antalet kronor då vi singlar två slantar är

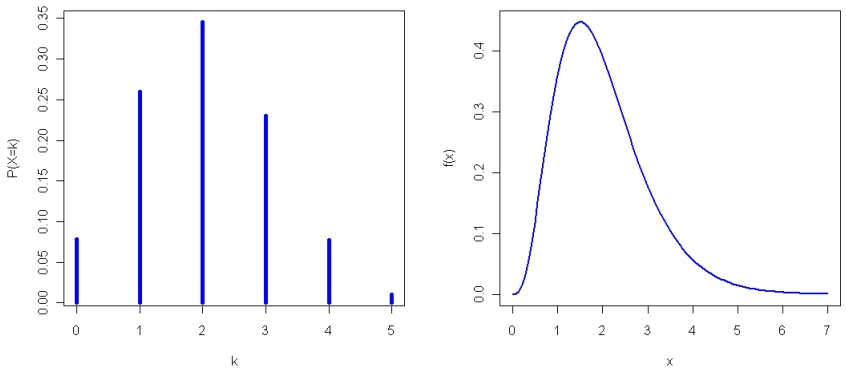
$$\begin{aligned} E(X) &= P(X=0)\cdot 0 + P(X=1)\cdot 1 + P(X=2)\cdot 2 \\ &= 0.25\cdot 0 + 0.5\cdot 1 + 0.25\cdot 2 = 1\ . \end{aligned}$$

Om vi m.a.o. skulle upprepa försöket oändligt många gånger skulle medelvärdet av antalet kronor per försök bli 1.

Sannolikhets- och täthetsfunktion

- I stället för att räkna upp sannolikheter för olika värden av en slumpvariabel är det ofta mera praktiskt att beskriva en fördelning i fom av en *funktion* av slumpvariabeln.
- En sannolikhetsfunktion anger sannolikheten för ett givet värde av en *diskret* slumpvariabel.
- För en *kontinuerlig* slumpvariabel är sannolikheten för enskilda värden 0, varför man i stället använder en sk. täthetsfunktion som konstrueras så att
  - mer sannolika värden får högre täthet (inte samma som sannolikhet!)
  - ytan mellan täthetsfunktionen och x-axeln blir 1.

Figur 13: Sannolikhets- och täthetsfunktion.



Fördelningsfunktion

- En fördelningsfunktion anger för både diskreta och kontinuerliga slumpvariabler sannolikheten att få ett värde som är mindre än eller lika med ett visst värde x

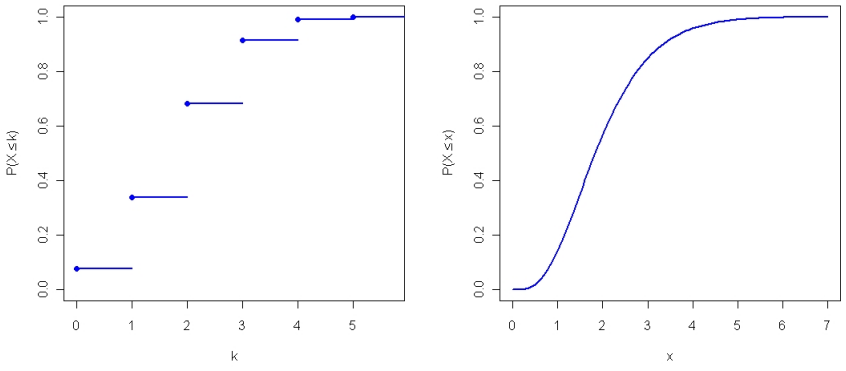
$$F(x) = P(X \leq x)\,.$$

- Sannolikheten för att en slumpvariabel ska anta värden större än a och mindre eller lika med b kan beräknas med:

$$P(a < X \leq b) = F(b) - F(a)\,.$$

Diskret och kontinuerlig fördelningsfunktion.

Figur 14: Diskret och kontinuerlig fördelningsfunktion.



- Med *diskreta* slumpvariabler bör man se upp med att få gränserna rätt då man gör uträkningar med fördelningsfunktionen.

  Exempel. 13. Låt oss anta att X kan få värdena 0, 1, 2, 3. Vi får då

$$\begin{aligned} P(1 < X < 3) &= F(2) - F(1) = P(X = 2) \\ P(1 \leq X < 3) &= F(2) - F(0) \\ P(1 < X \leq 3) &= F(3) - F(1) \\ P(1 \leq X \leq 3) &= F(3) - F(0)\,. \end{aligned}$$

- Då det gäller *kontinuerliga* variabler gör vi ingen skillnad mellan "$<$" och "$\leq$".

- Om alltså X i exemplet ovan hade varit kontinuerlig skulle samtliga sannolikheter ha räknats $F(3) - F(1)$.

Bernoulliförsök och tvåpunktsfördelningen

- Vi utför ett försök som kan resultera i endast två olika utfall A och A^c.
- Försök med endast två möjliga utfall som utgör varandras komplement kallas ofta Bernoulliförsök.
- Vi betecknar framöver $P(A) = p$ och följaktligen $P(A^c) = 1-P(A) = 1-p$.
- *Sannolikhetsfördelningen* av de två möjliga värdena av ett Bernoulliförsök kallas Bernoulli- eller tvåpunktsfördelning.

Binomialfördelningen

Binomialfördelningen kan karakteriseras på följande sätt:

- Ett Bernoulliförsök upprepas så att
  - antalet upprepningar är n
  - sannolikheten $P(A) = p$ hålls konstant över alla upprepningar
  - försöken upprepas oberoende från varandra.
- Vi definierar en slumpvariabel X som anger antalet försök som resulterar i A.
- Slumpvariabeln X följer då en binomialfördelning med parametrarna n och p, vilket betecknas
$$X \sim Bin(n, p) \ .$$
- *Sannolikhetsfunktionen* för en binomialfördelad slumpvariabel definieras som
$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \ ,$$
där k är antalet försök där A inträffar, n är totala antalet försök och p är sannolikheten för att A inträffar i ett försök. Binomialkoefficienten $\binom{n}{k}$ anger på hur många sätt man kan ordna en följd av n försök där A inträffar i k av försöken.
- *Fördelningsfunktionen* för en binomialfördelad slumpvariabel X definieras som
$$F(k) = P(X \leq k) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} \ .$$
- Väntevärdet för binomialfördelningen är $E(X) = np$.
- Variansen för binomialfördelingen är $Var(X) = np(1-p)$.

Exempel. 14. Låt oss anta att andelen ljushåriga i en stad är 30%. Om vi slumpmässigt plockar 5 personer ur befolkningen, vad är sannolikheten att vi får minst 2 och högst 4 ljushåriga i vårt stickprov? Vi låter splumvariabeln X beteckna antalet ljushåriga. Vi får då

$$P(2 \leq X \leq 4) = P(X = 2) + P(X = 3) + P(X = 4) =$$
$$\binom{5}{2} 0.3^2 \cdot 0.7^3 + \binom{5}{3} 0.3^3 \cdot 0.7^2 + \binom{5}{4} 0.3^4 \cdot 0.7 =$$
$$0.3087 + 0.1323 + 0.02835 = 0.46935 .$$

Alternativt kan vi använda oss av fördelningsfunktionen. Från en tabell eller med hjälp av ett statistiskt programpaket kan vi direkt läsa ut värden för $F(4) = P(X \leq 4)$ och $F(1) = P(X \leq 1)$ och får då

$$P(2 \leq X \leq 4) = F(4) - F(1) = 0.99757 - 0.52822 = 0.46935 .$$

Exempel. 15. Väntevärdet för slumpvariabeln X i föregånede exempel är

$$E(X) = n \cdot p = 5 \cdot 0.3 = 1.5 .$$

Vi kan tolka det som att vi i ett mycket stort antal stickprov i medeltal skulle få 1.5 ljushåriga per stickprov.

Variansen för X är

$$Var(X) = n \cdot p \cdot (1 - p) = 5 \cdot 0.3 \cdot 0.7 = 1.05.$$

Normalfördelningen

Normalfördelningen har en mycket central plats inom statistik. Detta är bl.a. för att

- många (men långt ifrån alla!) variabler följer en normal fördelning
- icke-normalfördelade variabler kan ibland transformeras så att de följer en normalfördelning
- många statistiska mått (t.ex. medelvärdet för stora stickprov) följer en normalfördelning.

Normalfördelningen är en kontinuerlig sannolikhetsfördelning med en symmetrisk och "klockformad" täthetsfunktion. Fördelningen bestäms helt av väntevärdet μ och standardavvikelsen σ.

- μ anger var toppen av kurvan befinner sig.
- σ anger hur koncentrerad kurvan är kring μ.

Att en slumpvariabel X följer en normalfördelning med parametrarna μ och σ betecknas

$$X \sim N(\mu, \sigma) .$$

Standardiserad normalfördelning

- Eftersom det för varje tänkbart värdepar (μ, σ) finns en normalfördelning finns det oändligt många normalfördelningar.
- Alla normalfördelningar kan standardiseras till en sk. standardiserad normalfördelning som har väntevärdet 0 och standardavvikelsen 1, dvs. $N(0,1)$.
- Om $X \sim N(\mu, \sigma)$ får vi genom standardisering en ny variabel

$$Z = \frac{(X - \mu)}{\sigma} \sim N(0,1) \; .$$

- Fördelingsfunktionen för en slumpvariabel Z som följer en standardiserad normalfördelning betecknas $\Phi(z) = P(Z \leq z)$.

Exempel. 16. Vid användning av en viss mätmetod antas de erhållna värdena vara normlafördelade med väntevärdet 28.0 och standardavikelsen 0.25, dvs. $N(28.0, 0.25)$. Vi frågar oss nu vad sannolikheten är att ett mätvärde ligger mellan 27.5 och 28.5. Uträkningen blir som följande:

$$\begin{aligned} P(27.5 < X \leq 28.5) &= P(\frac{27.5 - 28.0}{0.25} < Z \leq \frac{28.5 - 28.5}{0.25}) \\ &= \Phi(2) - \Phi(-2) \\ &= 0.9772 - 0.0228 = 0.954 \; . \end{aligned}$$

Om man använder en tabell där endast icke-negativa värden är tabulerade kan man ersätta $\Phi(-2)$ med $1 - \Phi(2)$.

I statistik ofta använda sannolikheter

- För en standardiserad normalfördelning $N(0,1)$ gäller att
  - 95% av alla värden ligger mellan -1.96 och 1.96, dvs.

$$P(-1.96 < Z \leq 1.96) = 0.95$$

  - 99% av alla värden ligger mellan -2.58 och 2.58, dvs.

$$P(-2.58 < Z \leq 2.58) = 0.99$$

  - 99.9% av alla värden ligger mellan -3.29 och 3.29, dvs.

$$P(-3.29 < Z \leq 3.29) = 0.999 \; .$$

- På motsvarande sätt gäller för en allmän normalfördelning $N(\mu, \sigma)$ att

$$\begin{aligned} P(\mu - 1.96 \cdot \sigma < X \leq \mu + 1.96 \cdot \sigma) &= 0.95 \\ P(\mu - 2.58 \cdot \sigma < X \leq \mu + 2.58 \cdot \sigma) &= 0.99 \\ P(\mu - 3.29 \cdot \sigma < X \leq \mu + 3.29 \cdot \sigma) &= 0.999 \; . \end{aligned}$$

Andra fördelningar

Andra i statistiska sammanhang ofta förekommande fördelningar är

- t-fördelningen
- χ^2-fördelningen ("*khi i kvadrat*" -fördelningen)
- F-fördelningen .

6 Punktskattning och konfidensintervall

6.1 Population och stickprov

Population

- Vi har en population vars någon mätbar egenskap X vi är intresserade av.
- Den mätbara egenskapen har inom populationen en fördelning som kan beskrivas med någon lämplig parameter, t.ex.:
  - väntevärdet μ
  - standardavvikelsen σ
  - andelen p av individer i populationen som har någon egenskap av intresse.

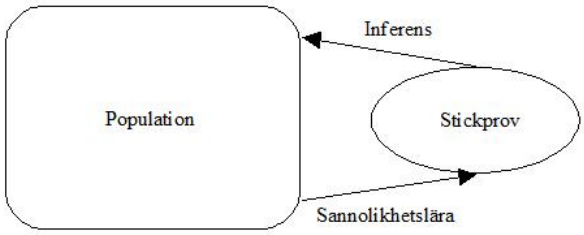
Stickprov

- I praktiken kan man oftast inte göra observationer på en hel population.
- I stället har man ofta tillgång till ett stickprov, dvs. ett slumpmässigt urval av n st individer från populationen.
- Utgående från stickprovet försöker man dra slutsatser om populationsfördelningens okända parametrar. Detta kallas statistisk inferens.

Från sannolikhetslära till inferens

- Med hjälp av sannolikhetslära kan vi dra slutsatser om ett *stickprov* från en population med vissa kända egenskaper.
- I statistisk inferens vänder vi på frågeställningen: vad kan vi säga om en *population* på basen av ett stickprov?
- Bl.a. innefattar statistisk inferens
  - punktskattning
  - intervallskattning (=beräkning av konfidensintervall)
  - hypotesprövning.

Figur 15: Samband mellan sannolikhetslära och inferens.



6.2 Skattning av populationsparametrar

Punktskattning

I punktskattning använder man data från ett stickprov för att räkna ut ett värde som fungerar som en "gissning" på det okända värdet på en populationsparameter av intresse.

- För väntevärdet μ andvänder man som punktskattning oftast stickprovsmedelvärdet $\overline{x}$.
- För *populations*standardavvikelsen σ andvänder man som punktskattning oftast *stickprovs*standardavvikelsen s.
- För proportionen p av individer i en population med en viss egenskap använder man proportionen $\hat{p}$ i stickprovet med samma egenskap.

Stickprovsfördelningen

- Eftersom värdet på en punktskattning avgörs *slumpmässigt* beroende på vilka individer som kommer med i stickprovet kan en skattning betraktas som en *slumpvariabel*.
- Följaktligen har punktskattningen en sannolikhetsfördelning. Vi kallar denna fördelning stickprovsfördelningen (även samplingfördelningen).
- Stickprovsfördelningen talar om för oss hur skattningarna skulle variera om vi hade tillgång till ett mycket stort (oändligt) antal lika stora stickprov?

- Ju större vårt stickprov är, desto mindre spridning har stickprovsfördelningen och desto pålitligare är skattningen.

Konfidensintervall

- Eftersom en punktskattning inte ger en bild av hur pålitlig skattningen är, vill man vanligtvis också bestämma ett sk. konfidensintervall för en parameter.
- Ett konfidensintervall är ett intervall som med en bestämd sannolikhet $1-\alpha$ täcker det sanna parametervärdet. Sannolikheten kallas konfidensgrad.
- Konfidensgraden väljs ofta så att $1-\alpha$ är lika med 0.95 eller 0.99 (kan även anges i procent, t.ex. 95% eller 99%).
- Då man bestämmer ett konfidensintervall för en parameter måste man känna till vissa egenskaper hos stickprovsfördelningen för dess punktskattning.

Medelvärdesfördelningen

- För att kunna härleda ett konfidensintervall för väntevärdet μ måste vi först ta en närmare titt på stickprovsfördelningen för medelvärdet $\overline{X}$, även kallad medelvärdesfördelningen.
- Medelvärdesfördelningen
  - är centrerad kring populationens väntevärde μ.
  - har standardavvikelsen $\sigma/\sqrt{n}$.
  - är approximativt normalfördelad (oavsett hur populationen är fördelad) om stickprovsstorleken n är tillräckligt stor.

Sammanfattningsvis: Tar vi ett stort antal stickprov på n observationer från en godtyckligt fördelad population och räknar ett medelvärde $\overline{x}$ för varje stickprov, kommer medelvärdena att vara ungefär fördelade enligt $N(\mu, \sigma/\sqrt{n})$.

- För att själv experimentera med medelvärdesfördelningar dragna ur olika populationer, se `http://ccl.northwestern.edu/curriculum/ProbLab/CentralLimitTheorem.html`.

Konfidensintervall för väntevärde

Vi ska i följande härleda oss till ett 95% konfidensintervall för väntevärdet μ:

- Vi vet från sannolikhetslära att 95% av värdena i en allmän normalfördelning $N(\mu, \sigma)$ ligger mellan $\mu - 1.96 \cdot \sigma$ och $\mu + 1.96 \cdot \sigma$, dvs.

$$P(\mu - 1.96 \cdot \sigma \leq X \leq \mu + 1.96 \cdot \sigma) = 0.95 \, .$$

- På motsvarande sätt gäller för medelvärdesfördelningen $N(\mu, \sigma/\sqrt{n})$ att

$$P(\mu - 1.96 \cdot \sigma/\sqrt{n} \leq \overline{X} \leq \mu + 1.96 \cdot \sigma/\sqrt{n}) = 0.95 \, .$$

- Vi vet nu att avståndet mellan medelvärdet $\overline{x}$ och väntevärdet μ med 95% sannolikhet är högst $1.96 \cdot \sigma/\sqrt{n}$.

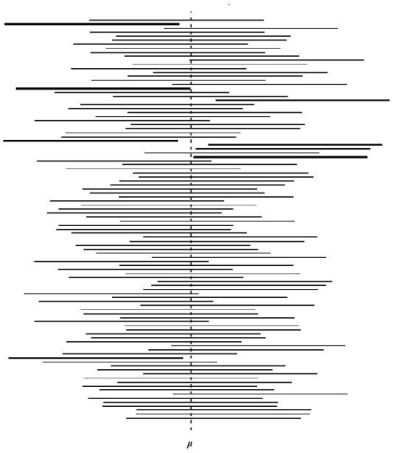
- Likväl kan vi då säga att intervallet

$$\overline{x} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}}$$

  med 95% sannolikhet täcker det okända väntevärdet μ.

- Vi har alltså konstruerat ett 95% konfidensintervall för μ.

Figur 16: Hundra st. 95% konfidensintervall för väntevärdet μ från en och samma population. Vi ser att ca 95% av intervallen täcker det sanna parametervärdet.



- Ett allmänt konfidensintervall för väntevärdet μ med konfidensgraden $1-\alpha$ ges av formeln

$$\overline{x} \pm z \cdot \frac{\sigma}{\sqrt{n}} \, ,$$

  där z är ett sådant värde från den standardiserade normalfördelningen att $P(-z \leq Z \leq z) = 1 - \alpha$.

- Ofta använda konfidensgrader med motsvarande värden för z är

$1-\alpha$	z
0.95	1.96
0.99	2.54
0.999	3.29

Exempel. 17. Vi har en population som följer en normalfördelning med okänt väntevärde μ och standardavvikelsen $\sigma = 3$. Då vi drar ett stickprov om 36 observationer från populationen får vi att medelvärdet är $\overline{x} = 10.0$. Ett 95% konfidensintervall räknas då enligt

$$\begin{aligned}\mu &= 10 \pm 1.96 \cdot \frac{3}{\sqrt{36}} \\ &= 10 \pm 1 \ .\end{aligned}$$

Låt oss nu anta att vi i stället vill bestämma ett 99% konfidensintervall för μ. Vi får då

$$\begin{aligned}\mu &= 10 \pm 2.54 \cdot \frac{3}{\sqrt{36}} \\ &= 10 \pm 1.29 \ .\end{aligned}$$

Genom att öka konfidensgraden blir intervallet bredare!

- Vi har vid beräkning av konfidensintervall hittils antagit att populationens standardavvikelse σ är *känd*.
- I praktiken är detta sällan fallet och vi måste i stället andvända oss av standardavvikelsen från stickprovet, s.
- Då det nu förutom osäkerheten i skattningen $\overline{x}$ har tillkommit en osäkerhet form av skattninen s, kommer konfidensintervallet att bli bredare.

Student's t-fördelning

- Vid härledning av konfidensintervallet för μ utgår man från variablen $\overline{X}$ som efter standardisering följer en standardiserad normalfördelning, dvs.

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1) \ .$$

- Om vi nu ersätter σ med skattningen s följer den motsvarande variabeln

$$\frac{\overline{X} - \mu}{s/\sqrt{n}}$$

  inte längre en normalfördelning, utan i stället en sk. Student's t-fördelning.

Student's t-fördelning (eller bara t-fördelningen):

- liknar den standardiserade normalfördelninen men är något bredare
- är precis som den standardiserade normalfördelningen alltid centrerad kring 0
- bestäms fullt av frihetsgraderna $df = n-1$, där n är antalet observationer i ett stickprov.

Konfidensintervall för väntevärde då σ är okänd

- Ett allmänt konfidensintervall för väntevärdet μ med konfidensgraden $1-\alpha$, då σ är *okänd*, ges av formeln

$$\overline{x} \pm t \cdot \frac{s}{\sqrt{n}} ,$$

  där t är ett sådant värde ur Student's t-fördelning att $P(-t \leq T \leq t) = 1-\alpha$.
- Eftersom t-fördelningens exakta form beror på antalet observationer n måste det aktuella värdet t alltid slås upp i en tabell eller bestämmas med hjälp av ett statistiskt programpaket.
- Användning av t-fördelningen förutsätter att populationen är "någorlunda" normalfördelad, speciellt om vi har ett litet stickprov.

Exempel. 18. Vi har en population som följer en normalfördelning, där väntevärdet μ och standardavvikelsen σ är okända. Då vi drar ett stickprov om 9 observationer från populationen får vi medelvärdet $\overline{x} = 148.0$ och standardavvikelsen $s = 4.87$. Ett 99% konfidensintervall räknas då enligt

$$\mu = 148.0 \pm t \cdot \frac{4.87}{\sqrt{9}} ,$$

där t är ett värde motsvarande konfidensgraden 99% från en t-fördelning med frihetsgraderna $df = 9-1 = 8$. Då $t = 3.36$ blir det sökta konfidensintervallet

$$\begin{aligned} \mu &= 148.0 \pm 3.36 \cdot \frac{4.87}{\sqrt{9}} \\ &= 148.0 \pm 5.5 . \end{aligned}$$

Konfidensintervall för proportioner

- Parametern p anger proportionen eller andelen individer med en viss egenskap i en population.
  - T.ex. kan vi vara intresserade av andelen diabetiker bland den vuxna befolkningen i Finland.
- Vi har tidigare talat om proportionen p i samband med binomialfördelningen.

– T.ex. om vi vet att var tionde vuxen i Finland är diabetiker kan vi med hjälp av binomialfördelningen räkna ut vad sannolikheten är att ett stickprov om $n = 1000$ individer ska innehålla mellan 90 och 110 diabetiker.

Då n är tillräckligt stort och $X \sim Bin(n, p)$ får vi approximativt att

$$X \sim N\big(np, \sqrt{np(1-p)}\big) \ .$$

Om nu X nu står för *antalet* individer med t.ex. diabetes i ett stickprov på n individer, får vi *andelen* av diabetiker i stickprovet med att dividera X med n. Fördelningen för proportionen blir då

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

vilket alltså är stickprovsfördelningen för proportionen $\hat{P} = \frac{X}{n}$. I princip kan vi nu använda oss av samma logik som tidigare för att bilda ett konfidensintervall för p, dvs. $p = \hat{p} \pm z \cdot \sqrt{p(1-p)/n}$, men eftersom p är parametern vi försöker skatta och alltså är okänd för oss, måste vi i kavadratroten ersätta p med den från stickprovet skattade proportionen $\hat{p}$.

- Konfidensintervallet för en okänd proportion p kan nu slutligen skrivas som
$$p = \hat{p} \pm z \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} \ .$$

Exempel. 19. I en liten stad finns 8820 hushåll av vilka man plockar ett stickprov på 200 hushåll. Av dessa visar det sig att 60 har en tavel-TV. Vi beräknar nu ett 95% konfidensintervall för andelen hushåll i *hela staden* med tavel-TV. Då andelen hushåll med tavel-TV i *stickprovet* är $\hat{p} = 60/200 = 0.30$, blir konfidensintervallet

$$\begin{aligned} p &= 0.3 \pm 1.96 \cdot \frac{\sqrt{0.3 \cdot 0.7}}{\sqrt{200}} \\ &= 0.3 \pm 0.06 \ . \end{aligned}$$

Med 95% sannolikhet finns det alltså mellan 2100 och 3200 hushåll med tavel-TV i staden.

Egenskaper hos konfidensintervall

För konfidensintervall gäller allmänt:

- Om vi ökar *konfindensgraden* (från t.ex. 95% till 99%) blir konfidensintervallet *bredare*.
- Om vi ökar *stickprovsstorleken* n blir konfidensintervallet *smalare*.

7 Hypotesprövning

7.1 Allmänna principer och grundbegrepp

Inledande exempel

Exempel. 20. Vi är intresserade av en variabel X om vilken vi kan anta att den är (approximativt) normalfördelad med standardavvikelsen $\sigma = 15$ i den givna populationen. På basen av tidigare erfarenheter är man benägen att tro att populationsmedelvärdet (dvs. väntevärdet μ) ligger kring 100. Det finns dock en misstanke om att det eventuellt skett en ökning i populationsmedelvärdet.

För att undersöka saken drar man ett stickprov om $n = 25$ enheter och räknar medelvärdet på observationerna. Vi får ett stickprovsmedeltal på $\overline{x} = 118.7$, vilket är betydligt större än 100, men kan detta tas som ett bevis på att det faktiskt skett en förändring i populationen?

- Modell för populationen:
$$X \sim N(\mu, 15)$$
- Stickprovsfördelningen för $\overline{X}$ är då $n = 25$:
$$\overline{X} \sim \ N(\mu, \frac{15}{\sqrt{25}}) \ .$$

Om vi nu utgår ifrån att populationsmedelvärdet verkligen är 100, dvs. att ingen förändring har skett, kan vi räkna ut sannolikheten för att t.ex. få ett stickprovsmedelvärde som är större än 115:

Vi har $\mu = 100$ vilket betyder att $\overline{X} \sim N(100, 15/\sqrt{25})$. Sannolikheten för att medelvärdet av 25 observationer är större än 115 är i detta fall

$$P(\overline{X} > 115) = 1 - \Phi\big(\frac{115 - 100}{15/\sqrt{25}}\big) = 0.000000287 \ .$$

Det observerade stickprovsmedelvärdet var 118.7, vilket ju är större än 115. Vi har då två alternativa förklaringar:

1. Något ytterst osannolikt har inträffat.
2. Vårt antagandde om att $\mu = 100$, dvs. att ingen förändring i populationsmedelvärdet har skett, är felaktigt.

Även om den första förklaringen är möjlig (också osannolika händelser kan ibland inträffa!) förefaller den senare förklaringen betydligt rimligare.

- Det antagande vars riktighet vi önskar pröva (testa) kallar vi nollhypotes
  - t.ex. ”populationsmedelvärdet är 100”.
- Ett alternativt antagande som vi har större benägenhet att tro på ifall nollhypotesen verkar orimlig kallar vi mothypotes
  - t.ex. ”populationsmedelvärdet har ökat”.

Översikt av stegen i hypotesprövning

1. Bestäm noll- och mothypotes.
2. Beräkna vad som kan förväntas om nollhypotesen vore sann.
3. Jämför det faktiska utfallet med vad som kan förväntas.
4. Om utfall och förväntan inte stämmer överens förkastar vi nollhypotesen och drar slutsatsen att mothypotesen är mera rimlig.

Hypotesprövning: Steg 1

Formulera en nollhypotes (betecknas H_0) och en mothypotes (betecknas H_1).

- Hypoteserna bör vara så formulerade att de täcker alla tänkbara möjligheter och är varandra uteslutande.
- Till H_0 väljes vanligen det som "gällt tidigare" eller det som man är beredd att hålla fast vid tills man blivit övertygad om att hypotesen är felaktig.
- Vid *jämförelser* är H_0 vanligen den att det inte finns några skillnader mellan de populationer som man jämför.
- Vid studier av *samband* formuleras H_0 vanligen som så att det inte finns något samband.

Hypotesprövning: Steg 2

Drag ett stickprov och beräkna värdet på någon lämplig teststatistika.

- Teststatistikan (kallas även testvariabel) är en sådan funktion av stickprovsobservationerna att man vet dess fördelning om H_0 är sann.
- Detta är nödvändigt för att vi skall kunna bedöma vad som är "sannolika" och "osannolika värden" på teststatistikan om H_0 vore sann.
- Ett exempel på en teststatistika är stickprovsmedelvärdet.

Hypotesprövning: Steg 3

Utgående från fördelningen för teststatistikan, dela in alla dess möjliga värden i sådana som är "sannolika" resp. "osannolika" om H_0 är sann.

- Ofta är teststatistikan sådan att både "väldigt små" och "väldigt stora" värden är osannolika om H_0 är sann. Ett test är då tvåsidigt.
- Om endast "väldigt små" eller "väldigt stora" värden är osannolika är ett test ensidigt.

- Formuleringen av mothypotesen H_1 bestämmer huruvida ett test är två- eller ensidigt.
- De värden i teststatistikans fördelning som kan anses vara "mycket osannolika" om H_0 är sann, utgör det sk. kritiska området.

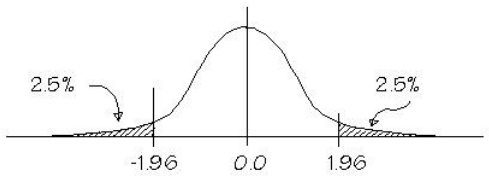
Hypotesprövning: Steg 4

Om värdet på teststatistikan faller inom det kritiska området förkastas H_0.

- Det kritiska området brukar vanligen väljas så att sannolikheten för att få ett värde på teststatistikan som faller inom det kritiska området är bara 0.05 eller 0.01 om H_0 är sann.
- Sannolikheten för att observera ett värde på teststatistikan som faller inom det kritiska området givet att H_0 är sann kallas testets signifikansnivå och betecknas α.
  - Man talar ofta om t.ex. "test på 5%-nivån" eller "test på 1%-nivån".
- Ett värde på teststatistikan som faller inom det kritiska området sägs vara signifikant.

OBS! Att H_0 inte förkastas bevisar ingalunda att nollhypotesen är sann.

Figur 17: Kritiskt område för ett tvåsidigt test på 5%-nivån ($\alpha = 0.05$) då teststatistikan följer en normalfördeling.



Typ I och typ II -fel och testets styrka

- Testets signifikansnivå α kan även tolkas som sannolikheten att förkasta H_0 då den är sann, dvs. sannolikheten för ett typ I -fel.
- I hypotesprövning kan man också begå ett annat slags fel, nämligen att låta bli att förkasta H_0 då H_1 är sann. Sannolikheten för det sk. typ II -felet brukar betecknas β.
- De två felen hänger ihop så att om man minskar risken för den ena kommer risken för den andra att öka. Genom att öka på stickprovsstorleken kan man dock minska båda riskerna samtidigt.

- Testets styrka är dess förmåga att förkasta en felaktig nollhypotes. Då H_1 är sann definieras styrkan som $1 - \beta$.

Tabellen nedan sammanfattar de olika besluten samt deras konsekvenser vid hypotesprövning:

Verklighet	Beslut	
	H_0 **förkastas**	H_0 **förkastas ej**
H_0 **sann**	Typ I -fel (α-fel)	Korrekt
H_1 **sann**	Korrekt	Typ II -fel (β-fel)

p-värden

- Vi påminner oss om att ju mera extremt ("väldigt stort" och/eller "väldigt litet") värde på teststatistikan vi observerar, desto mindre stöd får H_0.
- p-värdet är sannolikheten för att erhålla ett värde på teststatistikan som är minst lika extremt som det som man faktiskt observerat givet att H_0 är sann.
- T.ex är då
  - ett p-värde som är ≤ 0.05 är signifikant på 5%-nivån
  - ett p-värde som är ≤ 0.01 är signifikant på 1%-nivån.

Signifikans och tolkning av p-värden

- Ett statistiskt signifikant resultat behöver inte nödvändigtvis ha någon praktisk relevans.
  - T.ex. då man jämför värdet på en variabel hos två mycket stora grupper kan man lätt upptäcka en statistiskt signifikant skillnad som dock är så liten att den i praktiken inte har någon relevans.
- Då uppdelningen i "signifikanta" och "icke-signifikanta" resultat är något konstgjord, är det i många fall mera meningsfullt att endast rapportera det erhållna p-värdet och sedan verbalt tolka huruvida resultatet talar emot H_0.
- En vanlig missuppfattning är att tro att p-värdet är anger sannolikheten för att H_0 är sann. **Detta stämmer inte.**
- I stället anger alltså p-värdet hur trovärdigt det observerade datamaterialet är *ifall* H_0 vore sann. Föga trovärdiga data (litet p-värde) talar då alltså emot H_0.
- Mer om p-värden, se
  - `http://en.wikipedia.org/wiki/P_value`

samt speciellt den i Wikipedia-artikeln ovan citerade artikeln

- `http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=11159626`.

7.2 En del vanliga test

Test av väntevärde då σ är känt

Vi har ett stickprov från en population vars väntevärde μ är okänt medan standardavvikelsen σ antas vara *känd*, och vi önskar testa hypotesen

$$H_0 : \mu = \mu_0 \ .$$

En lämplig testvariabel är då

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1) \ .$$

I ett *tvåsidigt* test lägger vi upp mothypotesen

$$H_1 : \mu \neq \mu_0 \ .$$

I ett test på signifikansnivå α förkastar vi H_0 om $z < -z_{\alpha/2}$ eller $z > z_{\alpha/2}$, där några vanligt använda värden för $z_{\alpha/2}$ är:

$z_{\alpha/2}$	1.96	2.54	3.29
α	0.05	0.01	0.001

Har vi i stället ett *ensidigt* test formuleras mothypotesen antingen som

$$H_1 : \mu < \mu_0 \ .$$

eller som

$$H_1 : \mu > \mu_0 \ .$$

Om $H_1 : \mu < \mu_0$ förkastas H_0 ifall $z < -z_\alpha$ och om $H_1 : \mu > \mu_0$ förkastas H_0 ifall $z > z_\alpha$. Några vanligt använda värden för z_α är:

z_α	1.64	2.33	3.09
α	0.05	0.01	0.001

Test av väntevärde då σ är okänt

Vi har ett stickprov från en *normalfördelad* population vars väntevärde μ och standardavvikelse σ är *okända*, och vi önskar testa hypotesen

$$H_0 : \mu = \mu_0 \ .$$

En lämplig testvariabel är då

$$T = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \ ,$$

vilken följer en t-fördelning med frihetsgraderna $df = n-1$. De två- och ensidiga mothypoteserna lägger vi upp som tidigare.

- I ett tvåsidigt test på signifikansnivå α förkastas H_0 om $t < -t_{\alpha/2}$ eller $t > t_{\alpha/2}$, där $t_{\alpha/2}$ är ett sådant värde ur en t-fördelning att

$$P(T < -t_{\alpha/2}) + P(T > t_{\alpha/2}) = \alpha \ .$$

- I ett ensidigt test med mothypotesen $H_1 : \mu < \mu_0$ och signifikansnivån α förkastas H_0 om $t < -t_\alpha$.

- I ett ensidigt test med mothypotesen $H_1 : \mu > \mu_0$ och signifikansnivån α förkastas H_0 om $t > t_\alpha$.

- t_α är ett sådant värde ur en t-fördelning att

$$P(T > t_\alpha) = \alpha \ .$$

Det ovan beskrivna testet kallas ofta för t-test, eftersom testvariabeln följer en t-fördelning.

Exempel. 21. Vi antar att variablen X i en population är normalfördelad med μ och σ okända. Vi drar ett stickprov om $n = 60$ enheter och räknar stickprovsmedelvärdet och -standardavvikelsen, vilka blir $\overline{x} = 16.51$ resp. $s = 1.23$. Låt oss nu pröva hypotesen

$$H_0 : \mu = 17.0 \ \text{ mot}$$
$$H_1 : \mu \neq 17.0 \ .$$

Vi har då ett tvåsidigt test och bör alltså förkasta H_0 om $\overline{x}$ är antingen tillräckligt stort eller tillräckligt litet.Vi räknar nu värdet på testvariabeln

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{16.51 - 17.0}{1.23/\sqrt{60}} = -3.09 \ .$$

Det erhållna värdet jämför vi med det kritika värdet för signifikansnivån $\alpha = 0.05$ i en t-fördelning med frihetsgraderna $df = 60 - 1$. T.ex. med hjälp av en tabell finner vi att det kritiska värdet i detta fall är $t_{\alpha/2} = 2.00$. Eftersom teststatistikans värde är $t = -3.09 < -t_{\alpha/2} = -2.00$ förkastar vi H_0. Vi kunde alternativt ha baserat beslutet på att det tvåsidiga p-värdet 0.003 är mindre än 0.05.

Jämförelse av två väntevärden då σ är okänt

Vi har två *normalfördelade* populationer och drar ett stickprov från vardera. Låt storleken på stickproven vara n_1 respektive n_2. Vi antar att populationernas

standardavikelser σ_1 resp. σ_2 är okända men har orsak att tro att de är (nästan) lika, dvs. $\sigma_1 = \sigma_2 = \sigma$. Vi önskar testa hypotesen

$$H_0 : \mu_1 = \mu_2 \ ,$$

vilken även kan formuleras som $H_0 : \mu_1 - \mu_2 = 0$. En lämplig testvariabel är då

$$T = \frac{\overline{X}_1 - \overline{X}_2}{s_p / \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \ ,$$

där s_p är den "poolade" standardavvikelsen. T följer då en t-fördelning med frihetsgraderna $df = n_1 + n_2 - 2$. Mothypotesen är någon av följande:

- $H_1 : \mu_1 \neq \mu_2$
- $H_1 : \mu_1 < \mu_2$
- $H_1 : \mu_1 > \mu_2$.

H_0 förkastas enligt samma principer som i det nyss introcuerade t-testet (även detta är ett t-test).

Jämförelse av proportioner

Vi har två populationer. I vardera populationen har en okänd andel av enheterna en viss egenskap. Vi betecknar andelarna p_1 och p_2. Vi drar var sitt stickprov av storlek n_1 resp. n_2 från de två populationerna. Låt X_1 beteckna antalet enheter i det första stickprovet med den sökta egenskapen och låt X_2 beteckna motsvarande antal i det andra stickprovet. Andelen enheter med egenskapen i vardera stickprov ges då av $\frac{\overline{X}_1}{n_1}$ resp. $\frac{\overline{X}_2}{n_2}$.

- Vi vill nu undersöka om andelarna p_1 och p_2, vilket motsvaras av nollhypotesen

  $$H_0 : p_1 = p_2 \ ,$$

  vilken även kan formuleras som $H_0 : p_1 - p_2 = 0$.

- En lämplig testvariabel är då

  $$Z = \frac{\frac{\overline{X}_1}{n_1} - \frac{\overline{X}_2}{n_2}}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}} \ ,$$

  där $\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$. Variabeln Z följer en $N(0,1)$-fördelning.

- Mothypotesen är någon av följande:
  - $H_1 : p_1 \neq p_2$
  - $H_1 : p_1 < p_2$

– $H_1 : p_1 > p_2$.

Eftersom testvariabeln följer en standardiserad normalfördelning, förkastas H_0 enligt samma principer som t.ex. vid test av väntevärde då σ är känt.

Exempel. 22. I ett förberedande skede av en genetisk studie vill man ta reda på om andelen brunögda invånare i två städer A och B är lika. De okända proportionerna betecknas p_A resp. p_B. Ett stickprov dras ur populationerna i vardera städer. I stad A omfattar stickprovet $n_A = 150$ individer, av vilka antalet brunögda är $x_A = 103$. Motsvarande tal för stad B är $n_B = 120$ och $x_B = 77$

Vi lägger upp hypoteserna

$$H_0 : p_A = p_B$$
$$H_1 : p_A \neq p_B \ .$$

Testvariabeln räknar vi enligt den angivna formeln:

$$z = \frac{\frac{103}{150} - \frac{77}{120}}{\sqrt{\frac{2}{3} \cdot \frac{1}{3} \cdot (\frac{1}{150} + \frac{1}{120})}} \approx 0.78 \ .$$

Testar vi hypotesen på signifikansnivån $\alpha = 0.05$ får vi att $z = 0.78 < z_\alpha = 1.96$, alltså låter vi bli att förkasta H_0.

Test av regressionskoefficient

Vi har i början av kursen bekantat oss med regressionsmodeller där man är intresserad av ett funktionellt samband mellan en beroende variabel och en eller flera förklarande variabler.

- Låt oss nu anta att två variabler X och Y i en population har följande linjära samband

$$Y = \alpha + \beta \cdot X + (\text{slumpmässigt fel}) \ .$$

- α och β är här okända parametrar vars värden skattas från ett stickprov. I praktiken görs detta oftast med hjälp av ett statistiskt programpaket.

- I utskriften av en dylik regressionsanalys brukar det utöver det skattade värdet på regressionskoefficienten även anges ett p-värde, vilket hör ihop med ett test av hypoteserna

$$H_0 : \beta = 0$$
$$H_1 : \beta \neq 0 \ ,$$

  där H_0 innebär att det inte finns något samband mellan variablerna X och Y medan H_1 innebär att det finns ett samband.

Testet är baserat på en t-fördelning och har den bekanta tolkningen av att ju mindre p-värdet är desto mer talar det emot H_0.

Förteckning över vanliga test

En förteckning över vanliga test hittas på sidan
`http://en.wikipedia.org/wiki/Hypothesis_testing#Common_test_statistics`

7.3 Några icke-parametriska test

Icke-parametriska test

- Då vi inte testar värdet på en parameter utan är är t.ex. intresserade av formen på en fördelning talar vi om ett icke-parametriskt test.
- Icke-parametriska test används ofta då man inte kan göra antaganden om normalfördeling eller då data är mätta på nominal- eller ordinalskala.
- Ett par vanliga icke-parametriska test är
  - Mann-Whitney-Wilcoxon testet
    `http://en.wikipedia.org/wiki/Mann-Whitney-Wilcoxon`
  - Kolmogorom-Smirnov testet
    `http://en.wikipedia.org/wiki/Kolmogorov-Smirnov_test`
- Ett av de mest använda icke-parametriska testen är det sk. χ^2-testet, vilket det finns flera olika varianter av.

Test av fördelning med χ^2-test

För att testa hur väl frekvensfördelningen på ett klassificerat datamaterial stämmer överens med en förväntad fördelning kan man använda ett χ^2-test. Hypoteserna i ett dylikt test läggs upp som följande:

- H_0 : Den förväntade och observerade fördelningen är lika.
- H_1 : Den förväntade och observerade fördelningen är olika.

Värdet på testvariabeln beräknas enligt formeln

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \ ,$$

där O_i = observerad frekvens och E_i = förväntad frekvens under antagande att H_0 är sann. Summeringen sker över alla klasser i fördelningen.

- Då H_0 är sann, följer testvariabeln en χ^2-fördelning med frihetsgraderna $df = k - r - 1$, där k är antalet klasser och r är antalet parmametrar som har skattats från datamaterialet.
- I det allra enklaste fallet jämför vi den observerade fördelningen med en likformig fördelning, vilket inte kräver skattning av parametrar. Frihetsgraderna blir då $df = k - 1$.

- H_0 förkastas om testvariabelns värde χ^2 överskrider det signifikansnivån α motsvarande kritiska värdet χ^2_α i en χ^2-fördelning med frihetsgraderna df.

Exempel. 23. Ledningen för ett företag misstänker att en del anställda har tagit för vana att förlänga veckoslutet genom att sjukanmäla sig för fredag och/eller måndag. Saken följdes upp under fyra veckors tid. Ifall ledingens misstanke är obefogad, skulle vi förvänta oss att sjukfrånvarona fördelar sig jämnt över alla arbetsdagar. Resultatet blev följande:

Veckodag	må	ti	on	to	fre
Observerade frånvaron	49	35	32	39	45
Förväntade frånvaron	40	40	40	40	40

Vi lägger upp följande hypoteser:

H_0 : Frånvarona fördelar sig jämnt över alla veckodagar.

H_1 : Frånvarona fördelar sig inte jämnt över alla veckodagar.

Enligt den tidigare angivna formeln blir värdet på testvariabeln

$$\chi^2 = \frac{(49-40)^2}{40} + \frac{(35-40)^2}{40} + \cdots + \frac{(45-40)^2}{40} = 4.9\,.$$

Eftersom den förväntade fördelningen är likformig behöver inga parametrar skattas. Vi jämför således testvariabelns värde med en χ^2-fördelning med frihetsgraderna $df = 5-1 = 4$. Väljer vi signifikansnivån $\alpha = 0.05$ får vi det motsvarande kritska värdet 9.488 t.ex ur en tabell. Eftersom testvariabelns värde inte överskrider det kritiska värdet förkastar vi inte H_0.

8 Analys av korstabeller

Korstabeller

- Vi har tidigare under kursen redan bekantat oss med korstabeller.
- I en korstabell redovisar man fördelningen på två eller flere, vanligen kvalitativa variabler.
- Också om man har att göra med i princip kvantitativa variabler kan det ibland vara skäl att övergå till att studera enbart fördelningen på olika klasser.

Test av oberoende med χ^2-test

Ett liknande χ^2-test som användes för test av fördelning kan även användas för att testa *oberoende* av två korstabulerade variabler.

Tabell 6: Exempel på en korstabell.

Rökvanor	Socioekonomisk status			Summa
	Hög	Medel	Låg	
Röker	51	22	43	116
Rökt tidigare	92	21	28	141
Aldrig rökt	68	9	22	99
Summa	211	52	93	356

- Värdet på testvariabeln beräknas enligt formeln

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \ ,$$

  där O_{ij} = observerad cellfrekvens och E_{ij} = förväntad cellfrekvens under antagande att H_0 är sann. Summeringen sker över de k raderna och l kolumnerna i korstabellen.

- Då H_0 är sann, följer testvariabeln en χ^2-fördelning med frihetsgraderna $df = (k-1)(l-1)$.

- H_0 förkastas enligt samma principer som i det tidigare introducerade χ^2-testet.

Hypoteserna i ett χ^2-oberoendetest är

H_0 : Inget samband, dvs. variablerna oberoende av varandra.
H_1 : Det finns ett samband.

Vi låter p_{ij} beteckna sannolikheten för att en observation tillhör en viss cell i korstabellen. Om variablerna är *oberoende* blir sannolikheten för att en observation tillhör en viss cell

$$\begin{aligned} p_{ij} &= P(\text{observationen tillhör rad } i \textit{ och} \text{ kolumn } j) \\ &= P(\text{observationen tillhör rad } i) \cdot P(\text{observationen tillhör kolumn } j) \ . \end{aligned}$$

De *förväntade* cellfrekvenserna räknas alltså

$$E_{ij} = \frac{(\text{summan av rad } i) \cdot (\text{summan av kolumn } j)}{n} \ .$$

Exempel. 24. För att undersöka effekten av ett nytt vaccin på en sjukdom ges 70 frivilliga försökspersoner vaccinet. I undersökningen ingår även en lika stor kontrollgrupp. De sammanlagt 140 personerna följs upp under en viss tid och man erhåller följande resultat:

	Har insjuknat	Har ej insjuknat	Tot.
Har vaccinerats	20	50	70
Har ej vaccinerats	40	30	70
Tot.	60	80	140

Följande hypoteser formuleras:

$$H_0 : \text{Vaccinet har ingen effekt.}$$
$$H_1 : \text{Vaccinet förebygger sjukdomen.}$$

Om vaccin och sjukdom är oberoende skulle vi förvänta oss följande:

	Har insjuknat	Har ej insjuknat	Tot.
Har vaccinerats	$\frac{70\cdot 60}{140} = 30$	$\frac{70\cdot 80}{140} = 40$	70
Har ej vaccinerats	$\frac{70\cdot 60}{140} = 30$	$\frac{70\cdot 80}{140} = 40$	70
Tot.	60	80	140

Från de två korstabellerna räknar vi sedan värdet på testvariablen

$$\chi^2 = \frac{(20-30)^2}{30} + \frac{(40-30)^2}{30} + \frac{(50-40)^2}{40} + \frac{(30-40)^2}{40} = 11.7\,.$$

Med signifikansnivån $\alpha = 0.01$ och frihetsgraderna $df = (2-1)(2-1) = 1$ förkastar vi H_0 eftersom $\chi^2 = 11.7 > \chi^2_\alpha = 6.635$.

Betingat oberoende och Simpson's paradox

Två variabler som till synes verkar beroende kan vara oberoende om man tar hänsyn till en tredje variabel. Detta kallas betingat oberoende.

- I följande introduktion till analys av korstabeller behandlas även betingat oberoende:
  `http://web.abo.fi/fak/mnf/mate/kurser/statistik1/AnalysAvKorstabeller.pdf`.
- Texten tar även upp det sk. Simpson's paradoxet. För definition och flera exempel, se
  `http://en.wikipedia.org/wiki/Simpson's_paradox`.
- Allmänt om betingat oberoende:
  `http://web.abo.fi/fak/mnf/mate/kurser/statistik1/MarginelltBetingat.pdf`

Oddskvot och relativ risk

Ett mått som ibland används för att beskriva graden av ett samband mellan två korstabulerade *dikotoma* (=av typen ja/nej) variabler är oddskvoten (förkortas ofta OR från *odds ratio*).

- Vi definierar först oddset för händelsen A:

$$\frac{P(A \text{ inträffar})}{P(A \text{ inträffar ej})} = \frac{P(A)}{P(A^c)} = \frac{P(A)}{1 - P(A)} .$$

  Om det är lika sannolikt att händelsen A inträffar som att den inte inträffar får oddset värdet 1.

- Odds kan även beräknas för betingade sannolikheter. T.ex. räknas oddset för att insjukna i ev viss sjukdom givet att man har vaccinerats:

$$\frac{P(\text{Sjuk}|\text{Har vaccinerats})}{P(\text{Ej sjuk}|\text{Har vaccinerats})} = \frac{P(\text{Sjuk}|\text{Har vaccinerats})}{1 - P(\text{Sjuk}|\text{Har vaccinerats})} .$$

- Oddskvoten definieras som kvoten mellan två odds. Vi kan då t.ex. räkna

$$OR = \frac{\frac{P(\text{Sjuk}|\text{Har vaccinerats})}{P(\textbf{Ej sjuk}|\text{Har vaccinerats})}}{\frac{P(\text{Sjuk}|\text{Har } \textbf{ej} \text{ vaccinerats})}{P(\textbf{Ej sjuk}|\text{Har } \textbf{ej} \text{ vaccinerats})}} ,$$

  vilket talar om för oss hur stort oddset är för att insjukna då man blivit vaccinerad i förhållande till motsvarande odds då man ej blivit vaccinerad. Resultatet tolkas på följande sätt

  - OR < 1: vaccinering *minskar* oddset för att insjukna
  - OR $= 1$: vaccinering *påverkar inte* oddset för att insjukna
  - OR > 1: vaccinering *ökar* oddset för att insjukna.

- Ett mått som är ganska likt oddskvoten är den sk. relativa risken (förkortas ofta RR), vilken i fallet ovan räknas som

$$RR = \frac{P(\text{Sjuk}|\text{Har vaccinerats})}{P(\text{Sjuk}|\text{Har } \textbf{ej} \text{ vaccinerats})} .$$

- För små värden är $OR \approx RR$.

- Fastän RR är något enklare och mera intuitiv än OR är den senare ofta mera användbar i statistiska analyser.

- Mera om oddskvoten och dess egenskaper:
  `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1127651`

9 Variansanalys och försöksplanering

Variansanalys

Vi har tidigare i form av ett t-test bekantat oss med jämförelse av väntevärden från två normalfördelade populationer med lika standardavvikelse. Variansanalys (förkortas ofta ANOVA från *analysis of variance*) är en generalisering av det

ovannämnda testet för två eller flera väntevärden. Namnet kan te sig något vilseledande då det ju är väntevärden man testar. Det har dock sitt ursprung i att den sk. F-testvariabeln som metoden bygger på kan tolkas som en kvot av varianser bildade på två olika sätt.

- I variansanalys antar vi alltså att vi har k oberoende (möjligtvis olika stora) stickprov från lika många *normalfördelade* populationer med lika standardavvikelse, dvs. $\sigma_1 = \sigma_2 = \ldots = \sigma_k = \sigma$.
- Hypoteserna kan formuleras på följande sätt:

  $H_0 : \mu_1 = \mu_2 = \ldots = \mu_k$

  H_1 : Åtminstone ett av väntevärdena skiljer sig från de andra.

- F-testvariablen är en kvot av spridningen *mellan* och spridningen *inom* de k grupperna. Om H_0 är sann följer variabeln en F-fördelning med frihetsgraderna $df = (k-1, n-k)$, där n är det totala antalet observationer.
- H_0 förkastas på signifikansnivån α om testvariabelns värde överskrider det kritiska värdet f_α i en F-fördelning med frihetsgraderna $(k-1, n-k)$.

Exempel. 25. Vi har fyra stickprov från normalt fördelade populationer. Av tabellen nedan framgår storleken, medelvärdet och variansen för respektive stickprov:

n_i	7	5	6	6
$\overline{x}_i$	2.4	3.3	3.4	2.6
s_i^2	0.25	0.34	0.10	0.05

Enligt H_0 har alla populationer samma väntevärde. Om det nu visar sig att spridningen mellan stickproven är stor jämfört med spridningen inom stickproven har vi orsak att tro att H_0 är falsk. Spridningen mellan stickproven är 1.53 medan den genomsnittliga spridningen inom stickproven är 0.18 (vi hoppar här över uträkningarna). F-testvariabeln får då värdet $1.53/0.18 = 8.66$. Det kritiska värdet för en F-fördelning med frihetsgraderna $(4-1, 24-4)$ på signifikansnivån $\alpha = 0.05$ är 3.10. Eftersom $8.66 > 3.10$ förkastar vi H_0.

Experimentella försök

- Variansanalysen är en metod som primärt utvecklats för att analysera resultat av experimentella försök.
- I ett typiskt experimentellt försök jämför man effekten av olika behandlingar, vilket här ska uppfattas som en allmän benämning på något som man utsätter försöksenheter för.
- Till skillnad från ett icke-experimentellt försök kan man i ett experimentellt försök påverka vilka enheter som får vilken behandling.

- Ett väl utfört experimentellt försök medger säkrare slutsatser än ett icke-experimentellt försök.

Exempel. 26.

- För att jämföra två medicinska preparat A och B ger en läkare det ena till en patientgrupp och det andra till en annan patientgrupp och jämför resultaten.
- Ett företag överväger att ersätta nuvarande tillverkningsmetod A med en ny metod B. För att undersöka om den nya metoden är bättre än den gamla tillverkar man ett antal enheter enligt vardera metoden och jämför resultatet.
- För att jämföra tre olika undervisningsmetoder delas eleverna i en årskurs i början av terminen slumpmässigt in i tre grupper. I slutet av terminen jämför man den genomsnittliga utvecklingen bland eleverna i de tre grupperna.

Flervägs variansanalys

- Den typ av variansanalys vi ovan har diskuterat kallas för envägs variansanalys eftersom indelningen i grupper sker enligt *en* typ av behandling (t.ex. medicinskt preparat, tillverkningsmetod, undervisningsmetod...).
- Analyserar man kombinationer av *flera* typers behadlingar använder man sk. flervägs variansanalys.
- Mera om variansanalys, se
  `http://en.wikipedia.org/wiki/Analysis_of_variance`

Försöksplanering

- Med försöksplanering strävar man efter att kontrollera effekten av faktorer som kan påverka tillförlitligheten av statistiska analyser i samband med ett experimentellt försök.
- Några för försöksplanering centrala begrepp och tekniker är:
  - Randomisering, dvs. slumpmässig allokering av försöksenheter i olika behandlingsgrupper för att minska på inverkan av okända systematiska fel på slutsatserna.
  - Replikering, vilket betyder att man gör upprepade mätningar av samma enhet för att få en uppfattning av mätfelet.
  - Indelning av liknande enheter i block för uppnå bättre precision ifall det finns stor variation bland enheterna.
- Ett par vanliga försöksplaner är:
  - Fullständigt randomiserat experiment (CRD, *Completely Randomized Design*), se t.ex.
    `http://courses.ncssm.edu/math/Stat_Inst/PDFS/RanDesgn.pdf`.

– Randomiserat blockexperiment (RBD, *Randomized Block Design*), se t.ex.
`http://en.wikipedia.org/wiki/Randomized_block_design`.