Statistik och epidemiologi T5

Anna Axmon
Biostatistiker
Yrkes- och miljömedicin

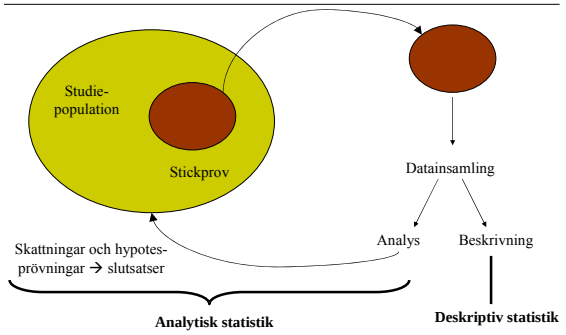
Biostatistik – kursmål

- Dra slutsatser utifrån basala statistiska begrepp och analyser och själva kunna använda sådana metoder.
  - Centralvärden och spridningsmått (deskriptiv statistik)
  - Sambandsanalys och differensanalys (analytisk statistik)
    - Parametriska tekniker
    - Icke-parametriska tekniker

Biostatistik – upplägg

- Föreläsning 1
  - Lägesmått och spridningsmått
  - Punktskattning och tillhörande osäkerhet
  - Introduktion till hypotesprövning
- Föreläsning 2
  - Fördjupning av hypotesprövning
  - Korrelation och linjär regression
  - Lite fler statistiska begrepp

Statistik – en överblick



Datatyper

- Kontinuerliga data – mäts på en skala
  - Exempel: Vikt, blodtryck
- Diskreta data – kontinuerliga data som bara kan anta vissa värden
  - Exempel: Antal
- Värdena är ”sanna” värden
  - 2-1 = 3-2
  - 4 är dubbelt så mycket som 2

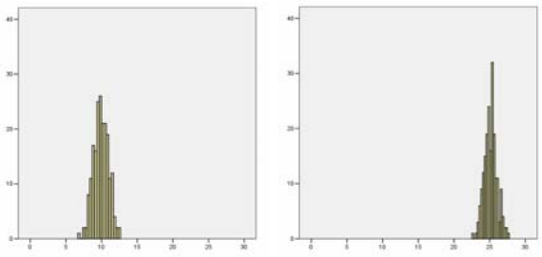
Datatyper

- Ordinaldata – klassdata med rangordning
  - Exempel: VAS-skala
  - 1 < 2 < 3
  - Ej säkert att 2-1 = 3-2
  - Ej säkert att 4 är dubbelt så mycket som 2
- Nominaldata – klassdata
  - Exempel: Bostadsort, kön
  - Ingen rangordning!

Deskriptiv statistik

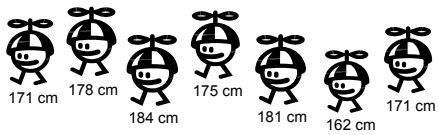
- Beskriva materialet utan att ge alla siffror
- Grafiskt
- Numeriskt
- Viktiga frågor:
  - Var ligger tyngdpunkten?
  - Hur stor är spridningen?

Var ligger tyngdpunkten?



Att ange tyngdpunkten

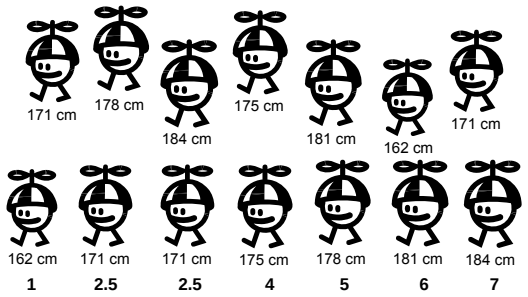
- **Medelvärde** – summan av observationerna delat med antal observationer



$$\frac{\sum x_i}{n} = \frac{171+178+184+175+181+162+171}{7} = 174{,}6$$

Summa

Rangordning



Att ange tyngdpunkten

- **Median** – det mittersta värdet när man sorterat observationerna i storleksordning
- **Typvärde** – det mest förekommande värdet

Längd	162	171	171	175	178	181	184
Rang	1	2,5	2,5	4	5	6	7

Beräkna median

- Har man få observationer är det lätt att hitta mitten
- Har man många observationer kan man använda formeln (n+1)/2 för att få rangen på medianen

Beräkna median – exempel

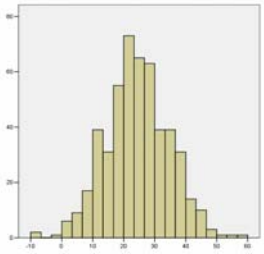
- (n+1)/2 = (7+1)/2 = 4
- Medianen är observationen med rang 4
- Denna observation har värde 175
- Medianen är alltså 175

Längd	162	171	171	175	178	181	184
Rang	1	2,5	2,5	4	5	6	7

Lite mer vetenskapligt…

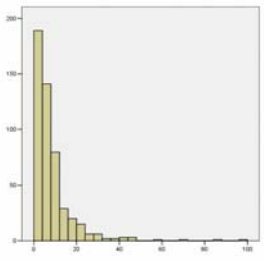
- Tyngdpunkten brukar refereras till som **centralvärde** eller **lägesmått**
- Valet görs utifrån hur data ser ut
  - Symmetriska kontinuerliga data
  - Asymmetriska kontinuerliga data
  - Ordinaldata
  - Nominaldata

Symmetriska kontinuerliga data



- Tyngdpunkten ligger ”mitt i”
- Medel = median
- Exempel: IQ, BMI
- **Använd medel!**
- I bilden: Medel = 24, median = 24

Asymmetriska kontinuerliga data

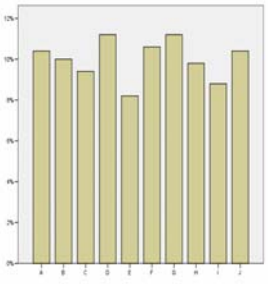


- Data förskjutet
- Medel < median
  ELLER
  medel > median
- Exempel: Många biologiska prover
- **Använd median!**
- I bilden: Medel = 8, median = 5

Varför inte alltid medelvärde?

- En studie visar att det undersökta läkemedlet har sänkt blodtrycket i snitt 10 mm/Hg
- Slutsats: Effektivt läkemedel
- En studie visar att det undersökta läkemedlet har sänkt blodtrycket i snitt 0 mm/Hg
- Slutsats: Inte effektivt läkemedel
- Fem individer hade värdena -1, -1, 0, 1 respektive 50
  - Medelvärde = 9.8
  - Median = 0

Ordinaldata



- Siffrorna är bara ”etiketter” → ej meningsfullt att räkna på
- **Använd median!**
- I bilden: Median = F

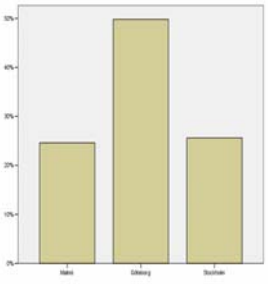
Varför inte medelvärde?

- Man vill undersöka effekten av ett smärtstillande medel
- 140 individer deltar i studien
- De kan ange smärtintensitet som
  - Inte ont alls
  - Lite ont
  - Ganska mycket ont
  - Väldigt ont
- Medelvärdet är 2.64 – hur tolkar man det?

Varför inte medelvärde?

- Värdena anpassades till en skala från 1 till 10 (t.ex. VAS-skala)
  - Inte ont alls = 1 poäng (n=85)
  - Lite ont = 2 poäng (n=15)
  - Ganska mycket ont = 6 poäng (n=15)
  - Väldigt ont = 10 poäng (n=25)
- Medelvärdet är beroende av hur man kodar variabeln!

Nominaldata

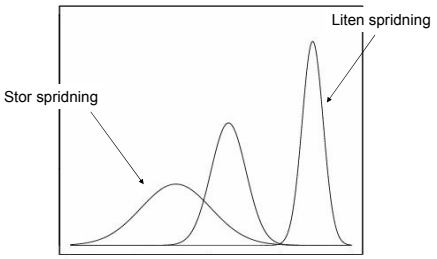


- Lägesmått ej meningsfullt
- Ange exempelvis andelar
- I bilden:
  Malmö = 24%
  Göteborg = 50%
  Stockholm = 26%

Sammanfattning

	Lägesmått	
Symmetriska data	Medel	
Asymmetriska data	Median	
Ordinaldata	Median	
Nominaldata	---	

Spridning



Spridningsmått

- Beskriver hur pass koncentrerade data är kring centralvärdet
- Är ej beroende av var tyngdpunkten ligger
- Precis som för centralvärde används olika mått för symmetriska och asymmetriska data
  - Symmetri – spridningsmåttet baseras på medelvärdet
  - Asymmetri – spridningsmåttet baseras *inte* på medelvärdet

Hur beskriva spridning?

- Genomsnittlig avvikelse från medelvärdet?
- Denna summa skulle dock bli noll!
- Genom att kvadrera varje term slipper man detta problem
- För att få en bättre skattning använder man dock n-1
- Detta kallas för *variansen*

$$\frac{\sum (x_i - \bar{x})}{n}$$

$$\frac{\sum (x_i - \bar{x})^2}{n}$$

$$\frac{\sum (x_i - \bar{x})^2}{n-1}$$

Hur beskriva spridning?

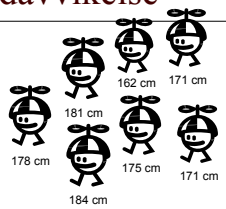
- Genom att ta roten ur variansen får man *standardavvikelsen (standarddeviationen)*

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$$

- Detta spridningsmått har samma enhet som det man mäter

Varians och standardavvikelse

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
171	-3.6	12.96
162	-12.6	158.76
181	6.4	40.96
175	0.4	0.16
184	9.4	88.36
178	3.4	11.56
171	-3.6	12.96



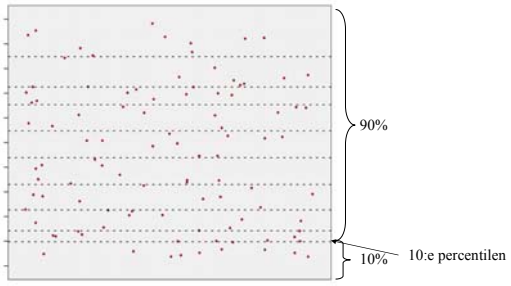
$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1} = 54{,}3$$

$$s = \sqrt{54{,}3} = 7{,}37$$

Percentiler

- Beskriver hur stor andel av observationerna som ligger under värdet
  - 10% ligger under 10:e percentilen
  - 20% ligger under 20:e percentilen
  - Etc…

Percentiler – illustration



Percentiler – beräkna

- Har man få observationer är det lätt att räkna fram percentilerna
- Har man många observationer kan man använda formeln q*(n+1)/100 för att få den q:te percentilen

Percentiler – exempel

- Beräkna 75:e percentilen!
- 0,75*(n+1) = 0,75*(7+1) = 6
- 75:e percentilen är observationen med rang 6
- Denna observation har värde 181
- 75:e percentilen är alltså 181

Längd	162	171	171	175	178	181	184
Rang	1	2,5	2,5	4	5	6	7

Kvartiler

- Undre kvartilen – 25:e percentilen
- Övre kvartilen – 75:e percentilen
- Medianen och kvartilerna delar materialet i fyra lika stora delar
- Interkvartilintervall = kvartilavstånd = skillnaden mellan övre och undre kvartilen

Kvartiler – exempel

BMI	19	20	21	21	22	23	24	24
Rang	1	2	3,5	3,5	5	6	7	8

25:e percentilen
Undre kvartilen
20,5

Median
21,5

75:e percentilen
Övre kvartilen
23,5

Variationsvidd

- Avståndet mellan lägsta och högsta värdet kallas *variationsvidd*
- Kan användas för symmetriska data, asymmetriska data
- För de sju gubbarna är variationsvidden 184-162=22 cm

Hur välja spridningsmått?

- Beräkning av varians utgår från medelvärdet
- Är medelvärde OK är alltså varians OK!
- Beräkning av percentiler utgår EJ från medelvärdet
- Är medelvärdet ej OK kan man ändå beräkna percentiler

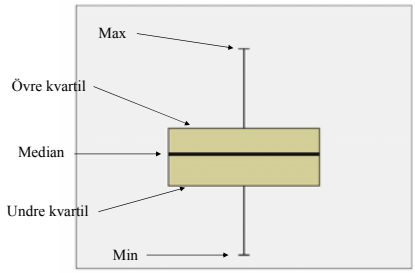
Sammanfattning

	Lägesmått	**Spridningsmått**
Symmetriska data	Medel	Varians *eller* standardavvikelse
Asymmetriska data	Median	Percentiler
Ordinaldata	Median	Percentiler
Nominaldata	---	---

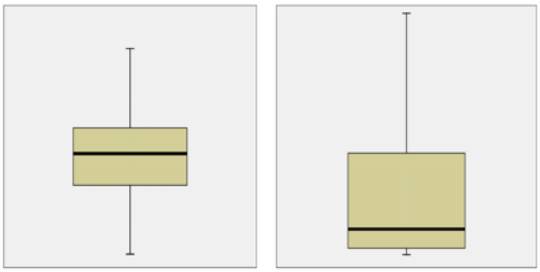
Symmetriskt?

- Median och medelvärde skall vara lika
- Avståndet mellan median och symmetriska percentiler skall vara lika stora
  - Exempel: Medianen → undre kvartilen = övre kvartilen → medianen
  - Kan t.ex. undersökas med Box-Plot

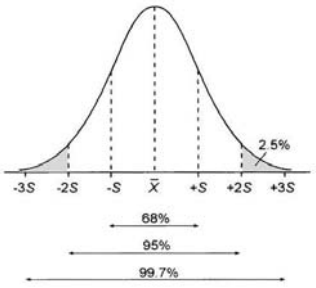
Box-Plot



Symmetriskt eller asymmetriskt?



Normalfördelningen

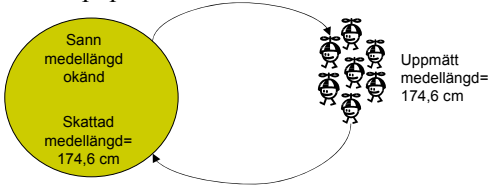


Stickprov vs studiepopulation

- Stickprovet…
  - … är de individer man mäter på
  - … kan man ta reda på allting om
  - … behöver man inte ”gissa” något om
- Studiepopulationen…
  - … är de individer man inte kan mäta (+ stickprovet)
  - … vet man ingenting om
  - … vill man kunna dra slutsatser om

Skattningar

- Data från stickprovet används till att göra gissningar – *skattningar* – angående studiepopulationen



Skattningar – standardfel

- Varje skattning har förstås en osäkerhet
- Denna kan mätas med *standardfelet (standard error* eller *standard error of the mean – SE* eller *SEM)*

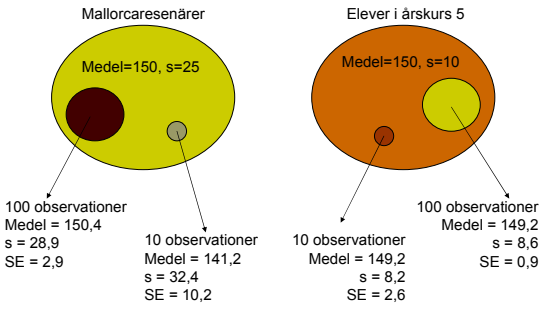
$$SE = \sqrt{\frac{s^2}{n}}$$

- SE beror av…
  - … spridningen i data
  - … antal observationer
- Bland de sju gubbarna är SE = 2,78 cm

Standardfel – exempel

- Population med stor spridning
  - Resenärer på flyg till Mallorca
  - I *studiepopulationen* är medelvärde = 150, standardavvikelse = 25
- Population med liten spridning
  - Barn i årskurs 5
  - I *studiepopulationen* är medelvärde = 150, standardavvikelse = 10

Standardfel – exempel

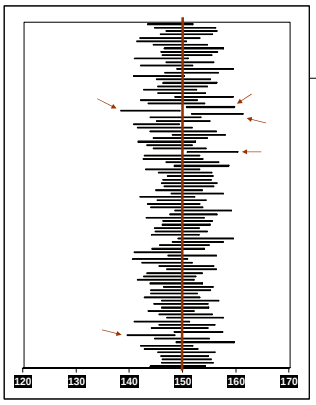


Sammanfattning

- Punktskattningar
  - Stickprovet används för att ”gissa” värden i studiepopulationen
  - Man kan räkna punktskattningar av t.ex. medelvärde och standardavvikelse
- Osäkerhet
  - Standardfel är ett mått på osäkerheten i punktskattningen
  - Ju mindre SE, desto säkrare punktskattning

Skattningar – konfidensintervall

- SE kan användas för att beräkna ett *konfidensintervall* (KI)
- Med en viss sannolikhet täcker konfidensintervallet det ”sanna” värdet
- Konfidensintervallets bredd beror av
  - Storleken på SE (och därmed antalet individer i stickprovet samt spridningen)
  - Konfidensgraden – hur säker man vill vara



- 100 stickprov från Mallorca-populationen
- Ett KI för varje stickprov
- Vissa täcker det sanna medelvärdet, andra inte

Konfidensintervall – definition

- Konfidensgrad 95%
  - Definition: Om man drar 100 stickprov kommer 95 av konfidensintervallen att täcka det sanna medelvärdet
  - ”Slarvig” tolkning: Konfidensintervallet täcker det sanna medelvärdet med 95% sannolikhet
- Motsvarande gäller för andra konfidensgrader, t.ex. 90% och 99%

Konfidensintervall – beräkna

- Generell formel för konfidensintervall: *skattning* $\pm$ *konstant* $\cdot$ *SE*
- Konstanten beror på konfidensgraden
  - 90% → 1,64
  - 95% → 1,96 (eller 2, om huvudräkning…)
  - 99% → 2,56
- Ju högre konfidensgrad, desto bredare konfidensintervall

Konfidensintervall – exempel

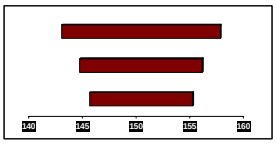
- Stickprov om 100 individer från Mallorcapopulationen

$$\bar{x} \pm c \cdot SE = 150{,}4 \pm 1{,}96 \cdot 2{,}9 = 144{,}7 - 156{,}1$$

- I studiepopulationen ligger medellängden med 95% sannolikhet mellan 144,7 cm och 156,1 cm
- **Det ”sanna” medelvärdet ligger med 95% säkerhet i intervallet medelvärdet ± 2SE!**

Konfidensintervall

- 90% KI = 145,6 – 155,2
- 95% KI = 144,7 – 156,1
- 99% KI = 143,0 – 157,8
- Ju högre konfidensgrad, desto säkrare att man täcker det sanna värdet, och desto bredare KI



Spridning i studiepopulationen

- Skattningar, SE och konfidensintervall säger något om *lägesmåttet* i studiepopulationen
- Ett **referensintervall** säger något om *spridningen* i studiepopulationen

$$\bar{x} \pm 1{,}96 \cdot s$$

- Formeln liknar den för KI, men i stället för SE använder man s

Referensintervall

- Stickprov om 100 individer från Mallorcapopulationen

$$x \pm c \cdot s = 150{,}4 \pm 1{,}96 \cdot 28{,}9 = 93{,}8\text{–}207{,}0$$

- 95% av studiepopulationen bör vara mellan 93,8 och 207,0 cm
- **Intervallet medelvärde ± 2 standardavvikelser täcker 95% av studiepopulationen**

Referensintervall

- 90% referensintervall = 103,0 – 197,8
- 95% referensintervall = 93,8 – 207,0
- 99% referensintervall = 76,4 – 224,4
- Ju större del av populationen man vill täcka, desto större intervall

Viktigt!

- Konfidensintervall och referensintervall är beräknade på data från *stickprovet* men drar slutsatser om *studiepopulationen*!
- Förväxla inte referensintervall med KI!
  - **Konfidensintervall**: Medelvärdet i studiepopulationen ligger med 95% sannolikhet inom gränserna
  - **Referensintervall**: 95% av studiepopulationen har ett värde inom gränserna

Konfidensintervall för en andel

- Ibland är det vi mäter inte en kontinuerlig variabel, utan en *andel*
- Exempel: Vi vill veta hur många som föredrar huvudvärkstablett A före huvudvärkstablett B

Konfidensintervall för en andel

Ind.	Resultat	Ind.	Resultat
1	A	6	A
2	B	7	A
3	B	8	A
4	A	9	B
5	A	10	A

- Punktskattningen är 7/10, d.v.s. 70%
- Hur stor är osäkerheten?

Konfidensintervall för en andel

- Antag att q = punktskattningen
- Det innebär att q är andelen i stickprovet
- Exempel: q = 0,25 → 25% i stickprovet föredrar A
- Konfidensintervall för andelar beräknas då

$$q \pm c \cdot \sqrt{\frac{q(1-q)}{n}}$$

där c är samma konstant som i tidigare beräkningar

Konfidensintervall för en andel

- q = andel som föredrar A = 0,7
- Konfidensgrad = 95% → c = 1,96
- 95% KI

$$0,7 \pm 1,96 \cdot \sqrt{\frac{0,7(1-0,7)}{10}} = 0,7 \pm 0,28 = 0,42 - 0,98$$

- Med 95% sannolikhet ligger den ”sanna” andelen som föredrar A mellan 42% och 98%

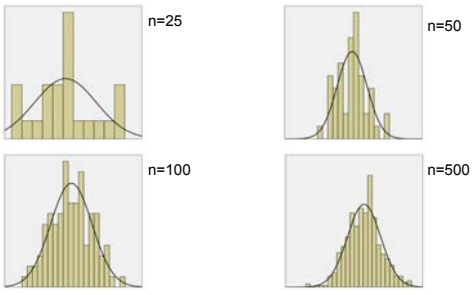
Förutsättningar

- För att konfidens- och referensintervall skall gälla måste
  - Stickprovet vara representativt för studiepopulationen
  - Kontinuerliga data måste vara normalfördelade
    - Stickprovet är normalfördelat
    - Studiepopulationen är normalfördelad
    - Stickprovet är stort

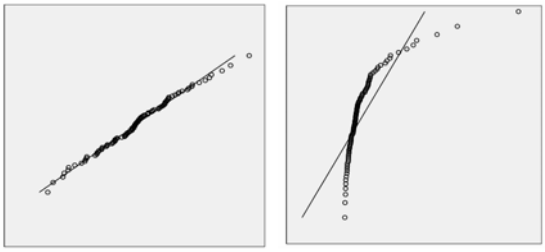
Hur veta om normalfördelat?

- Test av symmetri enligt tidigare
- Histogram
  - Kan vara svårt att avgöra om stickprovet är litet
- Normalfördelningsplot

Histogram



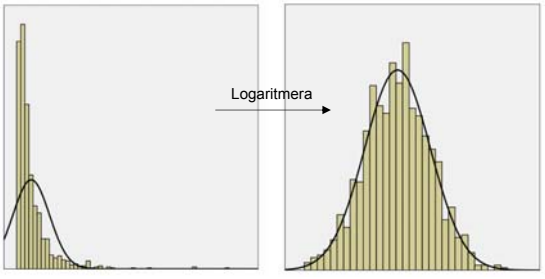
Normalfördelningsplot



Kryphål…

- Om data är snedfördelade kan det hjälpa att logaritmera
- Att logaritmera data är inte att fuska!
- Man ”förflyttar” data över skalan för att få normalfördelning
- Data behåller sina inbördes förhållande (ranger) även efter logaritmering

Logaritmerade data



Baserat på logaritmerade värden

- När man räknar statistik på logaritmerade värden gäller resultaten för de logaritmerade värdena!
- För att få statistik som gäller för originaldata måste man anti-logaritmera
- Då ligger inte längre punktskattningen mitt i konfidensintervallet!

Sammanfattning hittills (1)

- Man mäter egenskaper hos stickprovet för att skatta motsvarande egenskaper i studiepopulationen
- Om data är symmetriska:
  - Lägesmått = medelvärde
  - Spridningsmått = varians/standardavvikelse
- Om data är asymmetriska:
  - Lägesmått = median
  - Spridningsmått = percentiler

Sammanfattning hittills (2)

- Punktskattning: Värdet i stickprovet som används som ”gissning” för värdet i studiepopulationen
- Konfidensintervall och SE: Anger osäkerheten in punktskattningen
- Referensintervall: Ger en uppfattning om spridningen i studiepopulationen
- Konfidens- och referensintervall kan bara beräknas för normalfördelade data

Hypotesprövning

- Man väljer ett stickprov för att dra slutsatser om studiepopulationen
- Man kan dock aldrig *bevisa* något om en studiepopulation
- Däremot kan man avfärda något som mer eller mindre troligt
- Detta gör man genom s.k. hypotesprövning

Hypotesprövning

- Det man vill avfärda som mindre troligt formulerar man som en nollhypotes (H_0)
- Nollhypotesen är vanligtvis hypotesen om ingen effekt
- Det man vill ”ha kvar” formulerar man som en alternativhypotes (H_1)

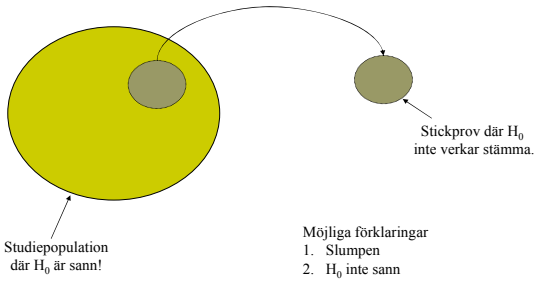
Hypoteser – exempel

- Studie om blodtryckssänkande medel
  - H_0: Ingen effekt
  - H_1: Positiv effekt – enkelsidig alternativhypotes
  - H_1: Positiv eller negativ effekt – dubbelsidig alternativhypotes

Att uttrycka hypoteser

- Hypoteser kan uttryckas på många olika vis
  - Behandlingen har ingen effekt
  - Behandlade och kontroller har samma resultat
- Bäst är dock så numeriskt som möjligt
  - Medelvärdet för behandlade = medelvärdet för kontroller
  - Skillnaden i medelvärde = 0
  - $\bar{x}_B - \bar{x}_K = 0$

Hypotesprövning – p-värde



Möjliga förklaringar
1. Slumpen
2. H_0 inte sann

Hypotesprövning med p-värde

- p-värdet är sannolikheten att man får det resultat man fick (eller ännu mer extremt) om H_0 är sann
- Med ”mer extremt” menas ännu större skillnad
- Om denna sannolikhet är tillräckligt liten förkastas nollhypotesen - ”tillräckligt liten” kan vara t.ex. 1%, 5% eller 10%

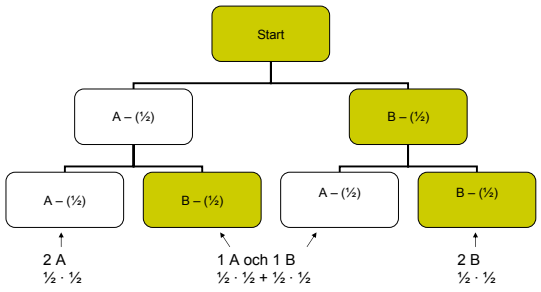
Hypotesprövning med p-värde

- Denna gräns kallas signifikansgräns, och bestäms innan analyserna utförs
- Signifikansgräns + konfidensgrad = 1
- Lite ”slarvigt” uttryckt kan man säga att p-värdet är sannolikheten att resultatet beror på slumpen
- P-värde kan beräknas även om data inte är normalfördelade (fast på olika sätt)!

Hypotesprövning – exempel

- Vi vill testa om huvudvärkstablett A är lika populär som tablett B
- H_0: Ingen skillnad, d.v.s. andelen som föredrar A = 0,5 eller q = 0,5
- H_1: Det finns en skillnad, d.v.s. q ≠ 0,5

Slh för olika utfall om H_0 sann



Hypotesprövning – exempel

Antal A	Slh	Antal A	Slh
0	0,0010	6	0,2051
1	0,0097	7	0,1172
2	0,0440	8	0,0440
3	0,1172	9	0,0097
4	0,2051	10	0,0010
5	0,2460	S:a	1,0000

- Om det inte finns någon skillnad i studiepopulationen:
- Sannolikheten att 7 eller fler (d.v.s. det vi fått eller mer extremt) föredrar A är 0,17, d.v.s. 17%
- Eftersom 17% > 5% förkastas inte H_0

Hypotesprövning med KI

- Hypotesprövning kan också göras med KI
- Det sanna värdet ligger med 95% sannolikhet i KI
- Om H_0 ligger i KI kan H_0 vara det sanna värdet → förkasta inte H_0
- Om H_0 ligger utanför KI är sannolikheten låg att H_0 är det sanna värdet → förkasta H_0
- Test med 95% KI = test med 5% signifikansgräns

Hypotesprövning – exempel

- KI för andel som föredrog huvudvärkstablett A var 42%-98%
- Nollhypotesen (50%) ligger inom detta intervall, d.v.s. nollhypotesen kan vara det ”sanna” värdet
- Alltså kan man *ej* förkasta nollhypotesen
- Hypotesprövning med konfidensintervall och p-värde ger alltid samma resultat!

Hypotesprövning – kontinuerliga data

- En Gato Negro bag-in-box skall, enligt uppgift på lådan, innehålla 3000 ml vin
- I SDS 28/8 2003 mättes innehållet i 10 Gato Negro bag-in-box
- Samtliga lådor innehöll <3000 ml
- Beror detta på slumpen?

Hypotesprövning – kontinuerliga data

H_0 = Gato Negro bag-in-box innehåller 3000 ml vin

H_1 = Gato Negro bag-in-box innehåller **inte** 3000 ml vin

Hypotesprövning – kontinuerliga data

Låda nr	Innehåll (ml)	Avvikelse	Kvadrerad avvikelse
1	2983	42,1	1772,41
2	2959	18,1	3276,61
3	2972	31,1	967,21
4	2935	-5,9	34,81
5	2807	-133,9	17929,21
6	2941	0,1	0,01
7	2943	2,1	4,41
8	2948	7,1	50,41
9	2951	10,1	102,01
10	2970	29,1	846,81
Summa	**29409**	**0,0**	**22034,90**

$$\bar{x} = \frac{29409}{10} = 2940{,}9$$

$$s^2 = \frac{22034{,}90}{10-1} = 2448{,}32$$

$$s = \sqrt{2448{,}32} = 49{,}48$$

$$SE = \frac{49{,}48}{\sqrt{10}} = 15{,}65$$

Hypotesprövning – kontinuerliga data

- 95% konfidensgrad (=5% signifikansgräns) → $c = 1{,}96$

$$95\%KI = 2940{,}9 \pm 1{,}96 \cdot 15{,}65 = 2920{,}2 - 2971{,}6$$

- H_0 (3000 ml) ligger utanför KI → förkasta nollhypotesen
- Statistikprogram ger p=0,004

Hypotesprövning – kontinuerliga data

- Vi har nu testat medelvärdet i ett stickprov mot ett förutbestämt värde
- I sådana lägen använder man ett (*one-sample) t-test*
- Kallas ibland *student's t-test*

Kommer ni ihåg…

- För att man skall få använda medelvärde och…
- …för att konfidens- och referensintervall skall gälla måste
  - Stickprovet vara normalfördelat *eller*
  - Studiepopulationen vara normalfördelad *eller*
  - Stickprovet vara stort
- Det samma gäller för att kunna använda t-test!

Parametriska vs icke-parametriska test

- T-testet är ett s.k. *parametriskt test*
- Namnet kommer från att det bygger på användandet av specifika *parametrar*…
- …nämligen normalfördelningens parametrar.
- Normalfördelningens parametrar är det som definierar fördelningen…
- …alltså medelvärdet och variansen.

Parametriska vs icke-parametriska test

- Test som *inte* bygger på parametrar kallas *icke-parametriska* test eller *fördelningsfria* test
- Dessa använder observationernas ranger i stället för värdena
- Ett icke-parametriskt test för att testa en median mot ett förutbestämt värde är t.ex. *Wilcoxons teckenrangtest*
- (Kommer mer om detta nästa föreläsning…)

Parametriskt vs icke-parametriskt

	Parametriskt	Icke-parametriskt
Utförs på	Värden	Ranger
Kräver normalfördelning	Ja	Nej
Skattar effekt med KI	Ja	Nej
P-värde	Ja	Ja

Detta skall ni kunna!

- Lägesmått och spridningsmått
  - Vilka finns?
  - Hur beräknar man dem?
  - När använder man vilket?
- Analytisk statistik
  - Sätta upp hypoteser
  - Välja test baserat på hur data fördelar sig
  - När lämpligt, beräkna konfidensintervall och referensintervall för **ett** stickprov
  - Tolka konfidensintervall, referensintervall och p-värde

Maila mig!

- Var något extra svårt?
- Vill ni veta mer om något?
- anna.axmon@med.lu.se