Föreläsningsanteckningar

Grundläggande statistik

732G01/732G40

LINKÖPING UNIVERSITY

Kapitel 2 – sid 11-46
Populationer, stickprov och variabler

Beskrivande mått

- Stickprovsmedelvärde

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

- Populationsmedelvärde

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

- Ex: Längden på fem slumpmässigt utvalda personer ur en population: 165, 188, 159, 170, 198

$$\bar{x} = \frac{1}{5}\sum_{i=1}^{5} x_i = \frac{1}{5} \cdot (165 + 188 + 159 + 170 + 198) = 176\ cm$$

Beskrivande mått - Frekvenstabell

- Stickprov

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

- Population

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

där g är antalet klasser i tabellen

Antal dagar (x)	Antal (f)	Andel (%)
0	84	38
1	41	19
2	51	23
3	22	10
4	8	4
5	6	3
6	5	2
7	3	1
Totalt	220	100%

Beskrivande mått

- Stickprovsstandardavvikelse (beräkningsformel)

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

- Population

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

Beskrivande mått - Frekvenstabell

- Stickprov

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{g} f_i \cdot (x_i - \bar{x})^2} = \sqrt{\frac{\sum_{i=1}^{g} f_i \cdot x_i^2 - \frac{\left(\sum_{i=1}^{g} f_i \cdot x_i\right)^2}{n}}{n-1}}$$

- Population

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{g} f_i \cdot (x_i - \mu)^2} = \sqrt{\frac{\sum_{i=1}^{g} f_i \cdot x_i^2 - \frac{\left(\sum_{i=1}^{g} f_i \cdot x_i\right)^2}{N}}{N}}$$

Beskrivande mått - Andel

- Stickprovsandel

$$p = \frac{antal\ enheter\ i\ stickprovet\ med\ studerad\ egenskap}{n}$$

- Population

$$\pi = \frac{antal\ enheter\ i\ populationen\ med\ studerad\ egenskap}{N}$$

- Ex: Bland 550 anställda i ett stickprov uppgav 187 att de röker. Vilken andel röker bland urvalet?

$$p = \frac{187}{550} = 0.34 = 34\%$$

Beskrivande mått - Median

- Medianen är alltid det mittersta värdet i ett storleksordnat material
  - Om n (eller N) är udda: mittersta värdet
  - Om n (eller N) är jämnt: medelvärdet av de två mittersta värdena

- Ex: Längder på fem personer
  159, 165, 170 ,188, 198 Medianen = 170 cm

- Ex: Vikten av fyra personer
  53, 62, 70, 85 Medianen = $\frac{62+70}{2} = 66kg$

Beskrivande mått - Kvartiler

- Första kvartilen (q_1) är mitten av första halvan av materialet
- Andra kvartilen (q_2) är medianen
- Tredje kvartilen (q_3) är mitten av andra halvan av materialet

Beskrivande mått - Typvärde

- Det vanligaste förekommande värdet i en fördelning

- Ex: Vid en tentamen har studenterna följande betyg U, U, G, G, G, VG, VG

  Typvärdet = G

När bör vi använda de olika måtten?

Kvalitativ variabel	Kvantitativ variabel
Typvärde	Median
Median	Kvartiler
Kvartiler	Medelvärde
Andelar	Standardavvikelse
	Andelar

Kapitel 3, sid 47-78
Sannolikhetsteori

LINKÖPING UNIVERSITY

Agenda

- Mängdlära
- Kombinatorik
- Sannolikhetslära

Mängdlära

- Används för att hantera sannolikheter
- Viktig byggsten inom matematik och logik
- Utfallsrummet, S, är samtliga möjliga utfall vid ett experiment
- Ex: Kasta en tärning:

$$S = \{1, 2, 3, 4, 5, 6\}$$

Mängdlära

- Varje del av S kallas för element

- Ex: Låt

$$A = händelsen\ udda\ ögon\ på\ tärningen$$
$$B = händelsen\ högst\ 3\ ögon\ på\ tärningen$$

$$A = \{1, 3, 5\}$$
$$B = \{1, 2, 3\}$$

- Om mängden A ingår i S är A en delmängd av S, vilket betecknas $A \in S$

Snitt och union

- Låt A och B vara två delmängder av S
- Snitt:
  - Är de element som tillhör både A och B

$$A \cap B$$

- Union:
  - Är de element som tillhör A eller B (eller båda)

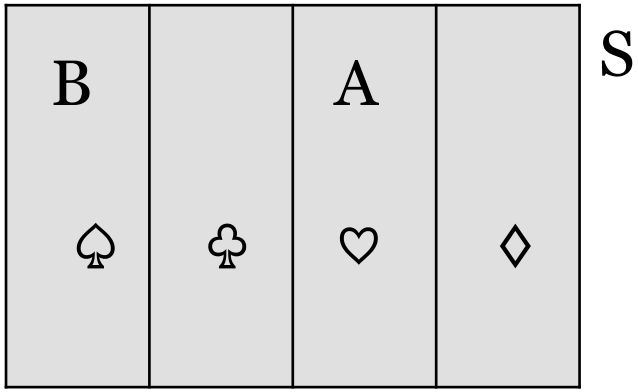
$$A \cup B$$

Disjunkta händelser

- Händelser som inte har ett snitt

- Ex: Dra kort ur en kortlek. Låt:

$$A = kortet \text{ är } en\ hjärter$$
$$B = korter \text{ är } en\ spader$$



Oberoende händelser

- Definieras som att sannolikheten för en händelse inte påverkas av att en annan händelse redan inträffat eller inte

- Går inte att visualisera i Venndiagram

- Ex: Kasta en tärning. Låt:

$$A = 6 \text{ ö}gon\ upp\ \text{på}\ f\text{örsta}\ kastet$$
$$B = 6 \text{ ö}gon\ upp\ \text{på}\ andra\ kastet$$

Händelserna är oberoende för de påverkar inte varandra

- Disjunkta händelser är inte oberoende!

Kombinatorik

- Gren inom matematiken som handlar om att *beräkna på hur många sätt ett givet antal element kan ordnas i mängder*
- Olika metoder
  - Multiplikationsprincipen
  - Permutationer när alla element är olika
  - Permutationer när vissa element är lika
  - Kombinationer utan upprepning
  - Kombinationer vid upprepning

Multiplikationsprincipen

- Ex: Framför oss har vi fyra kulor i olika färger. En röd, en **svart**, en blå och en grön. Vi väljer en, markerar färgen på den, lägger tillbaka den (kallas **med återläggning**). Detta upprepas k gånger.
- På hur många sätt kan detta experiment utföras?

$$n_1 \cdot n_2 \cdot \ldots \cdot n_k = n^k$$

- Visualiseras ofta i ett träddiagram

Permutationer när alla element är olika

- Ex: Samma fyra kulor ligger framför oss. Vi väljer **utan återläggning** två kulor. På hur många sätt kan detta göras, *om ordningen spelar roll*?

Hur många sätt kan man välja k element från n stycken

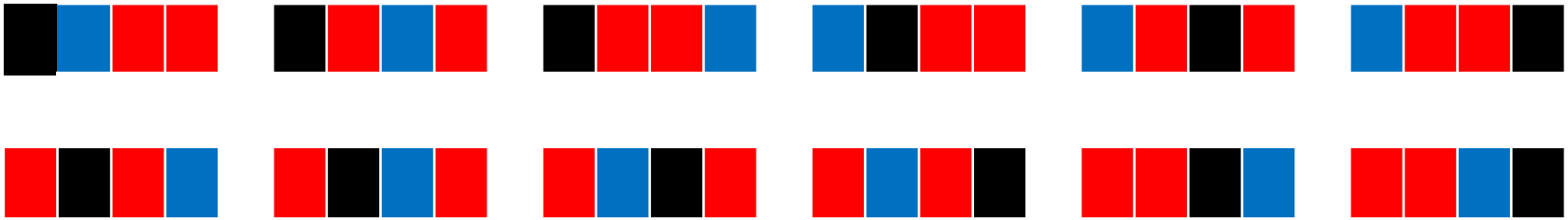
$$P_n^k = \frac{n!}{(n-k)!}$$

Permutationer när vissa element är lika

- Ex: Fyra kulor ligger framför oss. En **svart**, en blå och två röda. På hur många sätt kan vi **utan återläggning** välja ut alla fyra kulorna, *om ordningen spelar roll*?

- Totalt n element, där k_g är antalet element i grupp g

$$P_n^{k_1,k_2,\ldots} = \frac{n!}{k_1! \cdot k_2! \cdot \ldots}$$



Kombinationer utan återläggning

- Ex: Fyra kulor ligger framför oss. En röd, en **svart**, en blå och en grön. Vi väljer **utan återläggning** två kulor. På hur många sätt kan detta göras, *om ordningen inte spelar roll*?

Hur många sätt kan man välja k element från n stycken

$$C_n^k = \binom{n}{k} = \frac{n!}{k!\,(n-k)!}$$

Kombinationer med återläggning

- Ex: Fyra kulor ligger framför oss. En röd, en **svart**, en blå och en grön. Vi väljer **med återläggning** två kulor. På hur många sätt kan detta göras, *om ordningen inte spelar roll*?

$$C'^{k}_{n} = \binom{n+k-1}{k} = \frac{(n+k-1)!}{k!\,(n-1)!}$$

Permutationer och kombinationer

- *Utan återläggning* och *med hänsyn till ordningen:*
  Permutationer när alla element är olika
- *Utan återläggning* och *med hänsyn till ordningen*, vissa element ej åtskiljbara:
  Permutationer när vissa element är olika
- *Utan återläggning* och *utan att ordningsföljden har betydelse*:
  Kombinationer utan upprepning
- *Med återläggning* och *utan att ordningsföljden har betydelse*:
  Kombinationer vid upprepning

Permutationer används när ordningsföljden har betydelse!

Introduktion till sannolikhetslära

- Område inom statistik där vi studerar experiment som beror på slumpen

- Sannolikhet är ett värde mellan 0 och 1 som säger hur trolig händelsen är

- Ex: Sannolikheten för händelse A

$$\Pr(A)$$

Regler för sannolikheter

1. En sannolikhet ligger **alltid** mellan 0 och 1

2. Sannolikheten för alla disjunkta händelser i S kommer tillsammans summera till 1

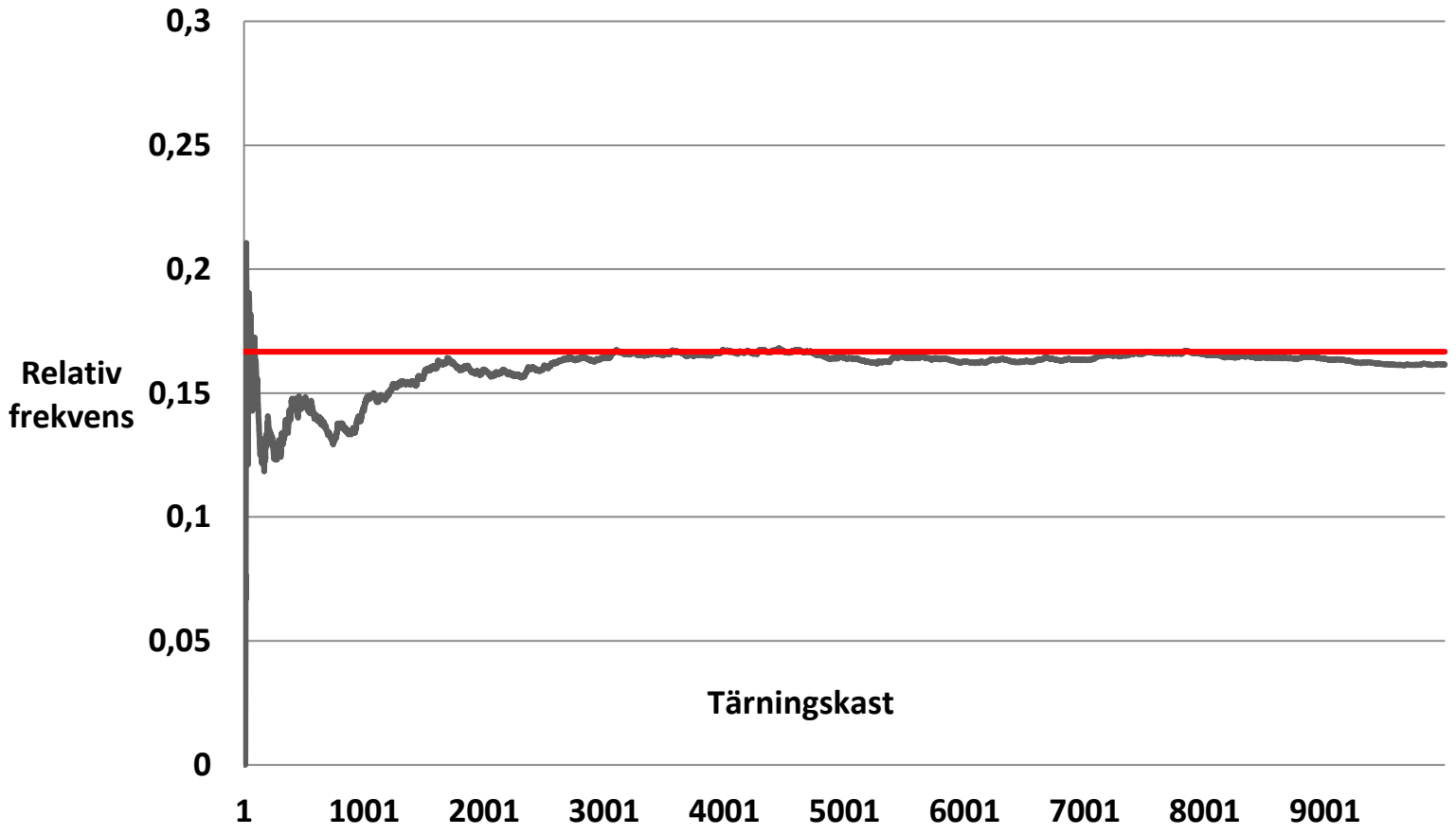
3. Om vi vet sannolikheten för A, $\Pr(A)$, så är sannolikheten att A **inte** inträffar $1 - \Pr(A)$

Den klassiska sannolikhetsdefinitionen

$$\Pr(A) = \frac{antalet\ gynnsamma\ utfall}{totala\ antalet\ utfall}$$

- Ex: Tio kulor i olika färger ligger framför oss, varav en är blå. Vi väljer **utan återläggning** två kulor. Vad är sannolikheten att en av dem är blå?

Relativ frekvens – tolkning av sannolikhet



Additionssatsen för disjunkta händelser

- Om A och B **är disjunkta** gäller

$$\Pr(A \cup B) = \Pr(A) + \Pr(B)$$

- Ex: Dra ett kort ur en kortlek. Vad är sannolikheten att kortet är ett hjärter eller ett spader?

$$\Pr(Hjärter) = \frac{1}{4}$$

$$\Pr(Spader) = \frac{1}{4}$$

$$\Pr(Hjärter\ eller\ Spader) = \frac{1}{4} + \frac{1}{4} = \frac{1}{2}$$

Additionssatsen för icke-disjunkta händelser

- Om A och B **inte är disjunkta** gäller

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

- Ex: Vi drar ett kort ur en kortlek. Vad är sannolikheten att kortet är en hjärter eller en sjua?

$$\Pr(Hjärter) = \frac{1}{4}$$

$$\Pr(Sjua) = \frac{1}{13}$$

$$\Pr(Hjärter\ eller\ Sjua) = \frac{1}{4} + \frac{1}{13} - \frac{1}{52} = \frac{4}{13}$$

Multiplikationssatsen för oberoende händelser

- Om A och B **är oberoende** gäller

$$\Pr(A \cap B) = \Pr(A) * \Pr(B)$$

- Ex: Vi singlar slant två gånger. Vad är sannolikheten för två krona i rad?

$$\Pr(Första\ krona) = \Pr(Andra\ krona) = \frac{1}{2}$$

$$\Pr(Första\ och\ andra\ krona) = \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

Betingad sannolikhet

- Sannolikheten för händelse A givet att händelse B redan inträffat

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

- Ex: Vi drar ett kort och ser att den är röd. Vad är sannolikheten att kortet är ett ess?

$$\Pr(Röd) = \frac{1}{2}$$

$$\Pr(Ess) = \frac{1}{13}$$

Betingad sannolikhet

$$\Pr(Röd \cap Ess) = \frac{1}{26}$$

$$\Pr(Ess|Röd) = \frac{\frac{1}{26}}{\frac{1}{2}} = \frac{1}{13}$$

- Om $\Pr(A|B) = \Pr(A)$ eller $\Pr(B|A) = \Pr(B)$ så är händelserna oberoende

Multiplikationssatsen för beroende händelser

- Om A och B **är beroende** gäller

$$\Pr(A \cap B) = \Pr(A) \cdot \Pr(B|A) = \Pr(B) \cdot \Pr(A|B) = \Pr(B \cap A)$$

- Ex: En skål innehåller 10 röda och 5 blå kulor. Vi väljer slumpmässigt och utan återläggning 2 kulor. Vad är sannolikheten för att bägge är blå?

$$\Pr(Första\ blå) = \frac{5}{15}$$

$$\Pr(Andra\ blå|Första\ blå) = \frac{4}{14}$$

Multiplikationssatsen för beroende händelser

$$\Pr(Första \cap Andra) = \frac{5}{15} \cdot \frac{4}{14} = \frac{2}{21}$$

Lagen om total sannolikhet

- Om $A_1, \ldots, A_g$ är g st parvis disjunkta händelser, vars union bildar hela utfallsrummet, blir:

$$\Pr(B) = \sum_{i=1}^{g} \Pr(A_i) \cdot \Pr(B|A_i)$$

- Ex: Sannolikheten att drabbas as strupcancer är 5% för rökare och 0,1% för icke-rökare. 14% av befolkningen röker. Vad är sannolikheten att en slumpmässigt vald person drabbas av strupcancer?

Bayes sats

- Om $A_1, \ldots, A_g$ är g parvis disjunkta händelser, vars union bildar hela utfallsrummet, blir:

$$\Pr(A_j|B) = \frac{\Pr(A_j) \cdot \Pr(B|A_j)}{\Pr(B)}$$

Där $\Pr(B) = \sum_{i=1}^{g} \Pr(A_i) \cdot \Pr(B|A_i)$

- Ex (fortsättning): Hur stor andel av dem som drabbas av strupcancer är rökare?

Kapitel 4, sid 79-124
Sannolikhetsfördelningar

LINKÖPING UNIVERSITY

Agenda

- Slumpvariabel
- Sannolikhetsfördelning

Slumpvariabel (Stokastisk variabel)

- En variabel som beror av slumpen

- Ex:
  - Tärningskast, längden på en slumpmässigt vald person

- Egenskaper:
  - Väntevärde: $E(X) = \mu = \sum_{i=1}^{g} x_i \cdot p(x_i)$
  - Varians: $Var(X) = \sigma^2 = \sum_{i=1}^{g} p(x_i) \cdot (x_i - \mu)^2 = \sum x_i^2 \cdot p(x_i) - \mu^2$
  - Standardavvikelse: $\sigma = \sqrt{Var(X)}$

Exempel

- Vinstplanen för 16 milj. trisslotter ser ut så här:

Vinst (kr)	Antal (f)	Vinst (kr)	Antal (f)
2500000	8	**750**	1200
1000000	8	**500**	1600
250000	40	**250**	4000
200000	8	**200**	3600
100000	16	**150**	10000
20000	16	**100**	75200
10000	320	**75**	238400
2000	1120	**50**	1672800
1000	1680	**25**	1336000

Exempel (forts.)

Låt $X = vinsten\ på\ lotten$

Utfallsrummet för X och sannolikheten blir då:

X	2500000	1000000	250000	…	50	25
p(x)	$\frac{8}{16000000}$	$\frac{8}{16000000}$	$\frac{40}{16000000}$	…	$\frac{1672800}{16000000}$	$\frac{1336000}{16000000}$

Ex: väntevärdet för vinsten på en slumpmässigt vald lott:

$$E(X) = \sum x_i \cdot p(x_i) =$$

$$2500000 \cdot \frac{8}{16000000} + 1000000 \cdot \frac{8}{16000000} + \cdots + 25 \cdot \frac{1336000}{16000000} =$$

$$12.25kr$$

Linjära variabeltransformationer

- Låt X vara en (slump)variabel med väntevärde $E(X)$ och varians $Var(X)$

$$Y = a + b \cdot X$$

- Då gäller att

$$E(Y) = \mu_y = E(a + b \cdot X) = a + b \cdot E(X)$$
$$Var(Y) = \sigma_y^2 = Var(a + b \cdot X) = b^2 \cdot Var(X)$$

- Ex: Svenska Spel funderar på att höja priset på en Trisslott till 30 kr och samtidigt öka vinsterna med 40 procent. Vad blir den förväntade vinsten efter denna förändring?

Linjära variabeltransformationer

Låt X = vinsten på trisslott **före** ändringen

Vi vet att $E(X) = 12.25$ enligt tidigare beräkning

Låt Y = vinsten minus lottkostnad (nettovinst) efter ändringen

$$E(Y) = a + b \cdot X = -30 + 1.4 \cdot 12.25 = -12.85$$

- Ex: Låt X beteckna nästa veckas värde på Ericsson-aktien. Antag att $E(X) = 50$ och $Var(X) = 25$.
- Vi äger 10 aktier och betecknar nästa veckas värde för dessa med $Y = 10X$.

Sannolikhetsfördelning

- Sammanställning av en slumpvariabels värden och sannolikheten för dessa

- Dessa underlättar komplicerade beräkningar av sannolikheter

- **Diskret** sannolikhetsfördelning:
  - När variabeln endast kan anta heltalsvärden

- **Kontinuerlig** sannolikhetsfördelning:
  - När variabeln kan mätas med flera decimalers noggrannhet

Diskret sannolikhetsfördelning

- Vanlig användning vid ett eller fler delförsök och vid varje delförsök mäts ifall det lyckas eller ej

- Varje delförsök sägs då följa Bernoullifördelningen. Varje delförsök kan bara anta ett av två möjliga värden.

- Ex: Vi definierar händelse $A = sex\ ögon\ upp\ vid\ tärningskast$. Varje delförsök kan antingen lyckas (slå en sexa) eller misslyckas (ej slå en sexa) och är därmed Bernoullifördelad.

Binomialfördelningen

- Beskriver en summa av **oberoende Bernoullifördelade** försök

- Ex: Grobarheten hos en viss typ av frön är 60%. Vi planterar 5 frön under samma förutsättningar och frågar oss: vad är sannolikheten för att två av fröna gror?

  Låt vara $X = antalet\ frön\ som\ gror$. Då gäller:

  $$X \sim bin(n; \pi)$$

  där n är antalet delförsök och π är sannolikheten för ett lyckat utfall

Binomialfördelningen

- Sannolikheten för k lyckade utfall bland n försök beräknas enligt:

$$\Pr(X = k) = \binom{n}{k} \pi^k (1 - \pi)^{n-k}$$

- Kända egenskaper hos Binomialfördelningen:

$$E(X) = n\pi$$

$$Var(X) = n\pi(1 - \pi)$$

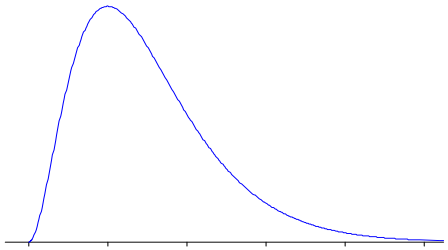
Fler diskreta fördelningar

- Om $X \sim bin(n;\pi)$ och $n > 20$ samt $\pi < 0.05$ kan fördelningen approximeras med Poissonfördelningen.

- Om $X = antalet\ f\ddot{o}rs\ddot{o}k\ tills\ f\ddot{o}rsta\ lyckade\ utfallet$ så används den geometriska fördelningen.

- Om delförsöken är *Bernoullifördelade* och dras utan återläggning och $\frac{n}{N} > 10\%$ används den hypergeometriska fördelningen.

Kontinuerlig fördelning



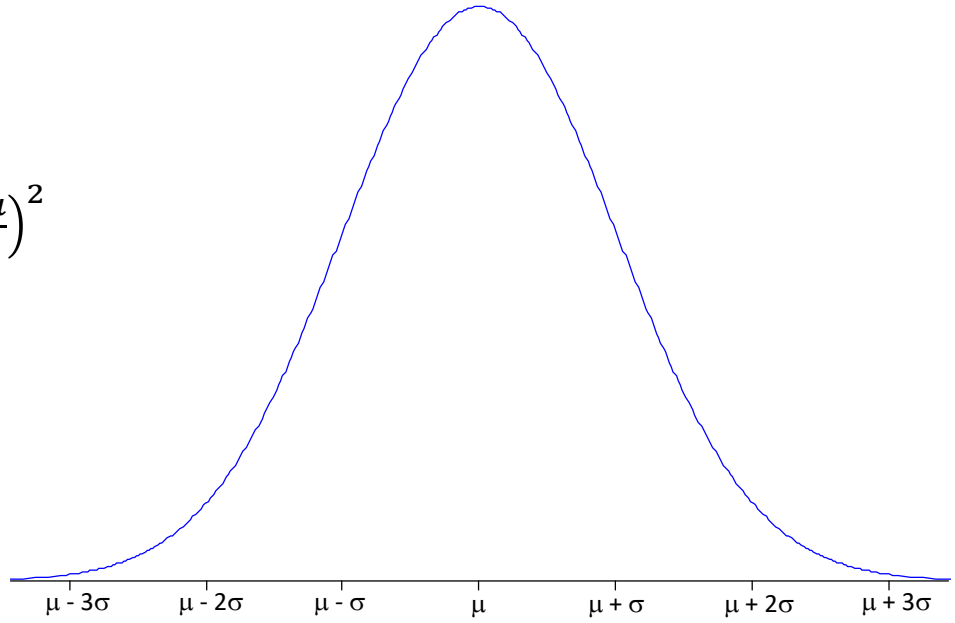
- Fördelningen av en kontinuerlig, kvantitativ variabel visualiseras med ett histogram

- En kurva kan betraktas som ett histogram där varje stapel är oändligt tunn

- Täthetsfunktion är en kurva där arean under kurvan blir 1
  - Detta innebär att vi kan använda den för sannolikheter

Normalfördelningen

- Mycket vanlig och viktig kontinuerlig fördelning
- Fördelningen är symmetrisk kring dess väntevärde
- $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2}$



Normalfördelningen

- Utseendet styrs av två parametrar:
  - Väntevärdet, μ, styr placeringen
  - Standardavvikelsen, σ, styr bredden, **alltid positiv**
- Betecknas $X \sim N(\mu; \sigma)$

- Oavsett parametervärden är arean under kurvan alltid 1

- Ca. 68% av fördelningen ligger inom $\mu \pm \sigma$
- Ca. 95% av fördelningen ligger inom $\mu \pm 2\sigma$

Standardiserad normalfördelning

- Denna fördelning betecknas med att $Z \sim N(\mu = 0; \sigma = 1)$

- Används för att underlätta beräkningar

- Standardiseringsformel

$$z = \frac{x - \mu}{\sigma}$$

där

μ och σ är den normalfördelade variabeln X parametrar och x är det värde vi är intresserade av

Exempel (forts.)

Kan uttrycka problemet som:

$$\Pr(-10 \leq X \leq 10)$$

Standardisering ger:

$$\Pr\left(\frac{-10-2.5}{10} \leq Z \leq \frac{10-2.5}{10}\right)$$

$$\Pr(-1.25 \leq Z \leq 0.75)$$

$$\Pr(Z \leq 0.75) - \Pr(Z \leq -1.25)$$

Från tabell:

$$0.77337 - 0.10565 = 0.66772$$

Att söka x för en given sannolikhet

Ex: Parkeringsgaraget under ett köpcentrum rymmer ett mycket stort antal bilar. Genom inpasseringssystemet vet man att det genomsnittliga antalet bilar som är inne i garaget vid samma tidpunkt är 455, med en standardavvikelse om 60 bilar. Man vet också att antalet bilar i garaget går att betrakta som en normalfördelad slumpvariabel.

Man skulle vilja ta utrymme från garaget för att utöka butiksytan. Hur många platser ska man lämna kvar om man vill att det 95 procent av tiden ska finnas lediga platser?

Normalapproximation av binomialfördelningen

- Låt $X \sim bin(n; \pi)$

- Givet att $n\pi(1-\pi) > 5$ kan X approximeras enligt

$$X \approx N\left(\mu = n\pi; \sigma = \sqrt{n\pi(1-\pi)}\right)$$

- Vi gör detta för att underlätta våra beräkningar

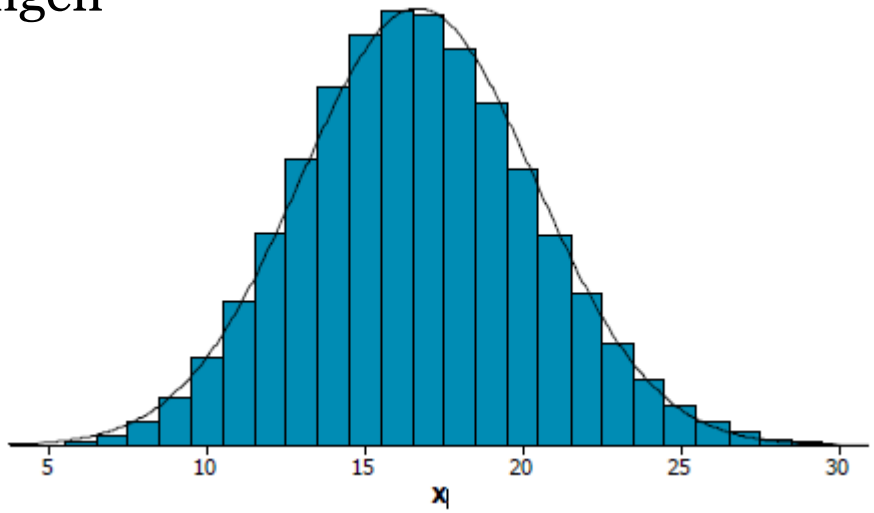
- Ex: Vi kastar en tärning 100 gånger och definierar

$$\mathrm{X} = antalet\ sexor$$

Vad är sannolikheten för att vi ska få sexa 20 gånger eller fler?

Normalapproximation

- Kontinuitetskorrektion:
  - Metod att förbättra approximationen
  - Betraktar tal som intervall och tar med hela intervallet i uträkningen



Kapitel 5, sid 127-152
Stickprovsteori

LINKÖPING UNIVERSITY

Agenda

- Stickprovsteori
- Väntevärdesriktiga skattningar
- Samplingfördelningar
- Stora talens lag, Centrala gränsvärdessatsen

Statistisk inferens

- Population: Den grupp av enheter (ofta individer) vi vill undersöka
- Urvalsram: Förteckning över enheter i populationen
- Urval: De enheter som blivit utvalda i stickprovet

Konsten att dra slutsatser om en population **baserat på ett stickprov** är en av grundpelarna inom statistiken! Det är också vad merparten av denna kurs kommer att handla om.

Obundet slumpmässigt urval (OSU)

- Urvalet är draget på ett sätt att alla enheter i populationen har **samma sannolikhet** att bli valda, nämligen:

$$\frac{n}{N}$$

- Ex: Vår population är alla studenter i ett klassrum, och vi vill undersöka genomsnittsvikten i klassen. Att väga alla skulle ta lång tid, och man vill därför dra ett stickprov om 20 personer.

  Det enklaste sättet att göra ett OSU skulle då vara att skriva ned allas namn på lappar, lägga dem i en låda och dra 20 lappar ur lådan. Då har slumpen valt ut 20 personer åt oss och alla har lika stor chans att bli utvalda.

På-stan urval

- Praktisk tillämpning av OSU:
  - Aktivt söka upp respondenterna
  - Ta hjälp av slumpen!
    - Tillfråga var tionde som passerar
  - Syftet är att göra ett urval bland alla individer inte bara de som ser vänliga ut

Stratifierat urval

- När vi vill dra slutsatser om en **heterogen** population
  - En population som kan delas in i **homogena** undergrupper som vi tror kan påverka undersökningen (t.ex. kön)

- Varje undergrupp kallas för stratum och ett OSU dras ur varje strata

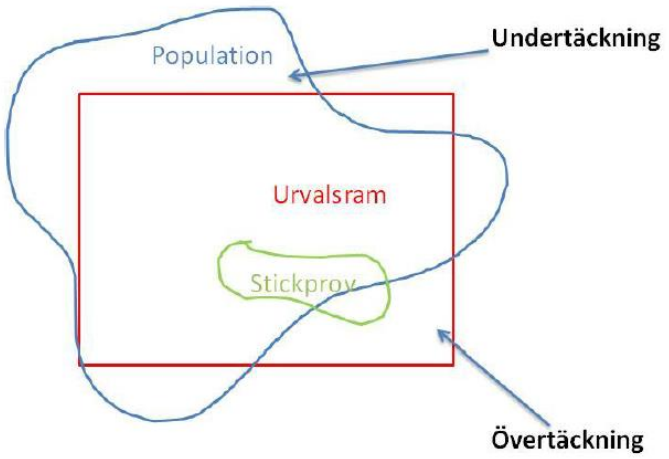
- Stratifierade urval, för en heterogen population, ger normalt mindre standardavvikelse och därmed säkrare slutsatser om populationen

Stratifierat urval

- Ex: Vi delar upp populationen i kvinnor och män, och lägger sedan lapparna med namn i en låda för kvinnor och en för män. Sedan drar vi 10 lappar ur varje låda.

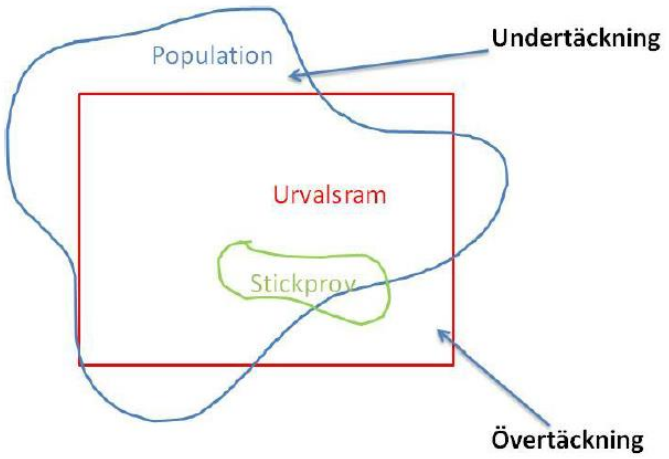
Felkällor vid stickprovsundersökningar

- **Övertäckning**: när det finns enheter i urvalsramen som egentligen inte tillhör målpopulationen
- Ex: Vid studie av vikter bland studenter i ett klassrum används klasslistan som urvalsram. Men vissa studenter har hoppat av utbildningen sedan klasslistan trycktes – de tillhör inte längre målpopulationen utan utgör övertäckning.



Felkällor vid stickprovsundersökningar

- **Undertäckning**: när det finns enheter i målpopulationen som saknas i urvalsramen
- Ex: Vissa studenter har påbörjat sin utbildning sedan klasslistan trycktes. De tillhör därför målpopulationen men har ingen chans att bli utvalda och utgör därför undertäckning.



Felkällor vid stickprovsundersökningar

- **Bortfall**: när enheter inte vill (eller kan) mätas. Skilj på
  - Slumpmässigt bortfall
  - Systematiskt bortfall

- Ex: Socialstyrelsen utsänder en enkät om tobaks- och alkoholvanor. Man kan då tänka sig att nykterister och icke-rökare är mer benägna att besvara enkäten än andra. Slutsatser dragna från enkäten riskerar att bli snedvridna eftersom bortfallet inte är slumpmässigt.

Relation mellan population och stickprov

- Populationsparametrar
  - Okända
  - Vi vill dra slutsatser om
- Stickprovsstatistikor
  - **Skattningar** av parametrarna

	Parameter	Väntevärdesriktig skattning
Medelvärde	$\mu = \frac{\sum x}{N}$	$\bar{x} = \frac{\sum x}{n}$
Varians	σ^2	s^2
Andel	π	p

Väntevärdesriktighet

- Ex: Låt X vara en slumpvariabel med en fördelning. Varje observation i stickprovet, $X_1, \ldots, X_n$, är också slumpvariabler med

$$E(X_i) = \mu$$

$$Var(X_i) = \sigma^2$$

Väntevärdesriktighet

- Genom att utnyttja räkneregler för linjära variabeltransformationer blir då

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E(X_1 + \cdots + X_n) =$$

$$= \frac{1}{n}(\mu + \cdots + \mu) = \frac{1}{n} \cdot n \cdot \mu = \mu$$

- Det förväntade värdet av stickprovsmedelvärdet är populationsmedelvärdet

Väntevärdesriktighet

- Vi visar därmed att skattningen är väntevärdesriktig det vill säga inga systematiska fel görs när stickprovsstatistikan används för att uppskatta populationsparametern.

Medelfel

- En väntevärdesriktig skattning av en parameter har också en osäkerhet

$$Var(\bar{X}) = Var\left(\frac{1}{n}\sum X_i\right) =$$

$$= \left(\frac{1}{n^2}\right) \cdot \left(Var(X_1) + \cdots + Var(X_n)\right) =$$

$$= \frac{1}{n^2} \cdot (\sigma^2 + \cdots + \sigma^2) = \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}$$

Medelfel

- Variansen av stickprovsmedelvärdet påverkas av variansen av slumpvariabeln men också storleken av stickprovet
- Ju större stickprov, desto mindre varians
- Medelfelet $\sigma_{\bar{x}}$ blir då

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

- Medelfelet är en skattning av den genomsnittliga osäkerheten när vi använder en stickprovsstatistika för att skatta en parameter

Egenskaper hos stickprovsstatistikorna

	Lägesmått	Spridning	Medelfel
Medelvärde	$E(\bar{X}) = \mu$	$Var(\bar{X}) = \frac{\sigma^2}{n}$	$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$
Summa	$E(\sum X) = n \cdot \mu$	$Var(\sum X) = n \cdot \sigma^2$	$\sigma_{\sum X} = \sqrt{n} \cdot \sigma$
Andel	$E(P) = \pi$	$Var(P) = \frac{\pi(1-\pi)}{n}$	$\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}$

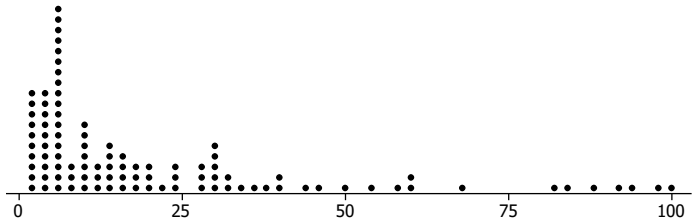
De stora talens lag

Ju större stickprov vi drar, desto mer lika blir stickprovsstatistikorna populationsparametrarna

Samplingfördelning

- Hur ofta kommer stickprovsmedelvärdet att överensstämma med populationsmedelvärdet om vi skulle dra många OSU ur samma population?

1	1	1	1	2	2	2	2	2	2
3	3	3	3	3	3	3	3	4	4
5	5	5	5	5	5	5	5	6	6
6	6	6	6	6	6	6	6	7	8
8	9	9	9	9	9	10	10	11	11
12	13	13	13	14	14	15	16	16	16
17	18	18	19	19	20	22	23	23	24
27	28	28	29	29	29	30	30	32	32
34	36	37	40	40	44	45	50	54	57
59	59	68	81	83	87	91	94	97	100



$\mu = 21.5$

$M = 11.5$

Samplingfördelning

- Från populationen vet vi att

$$\mu = 21.5$$
$$M = 11.5$$

- Vi drar ett stickprov om $n = 10$

1	3	3	5	5	13	14	22	40	81

$$\bar{x} = 18.7$$

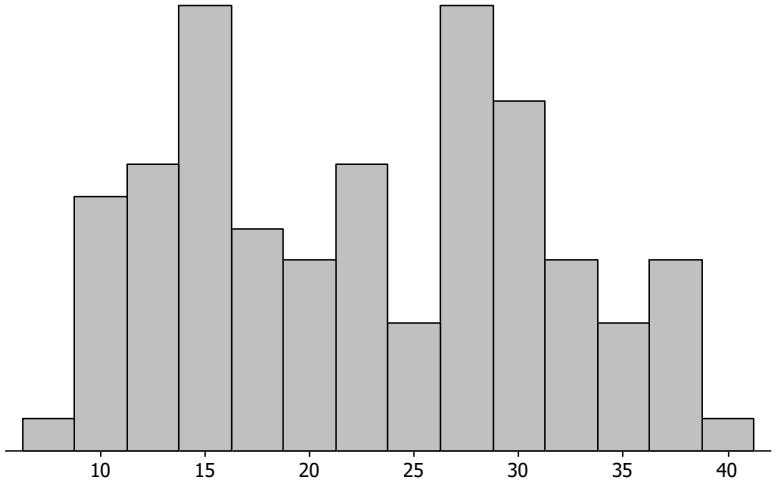
- Vi drar ett till stickprov

2	3	3	3	16	19	22	30	50	100

$$\bar{x} = 24.8$$

Samplingfördelning

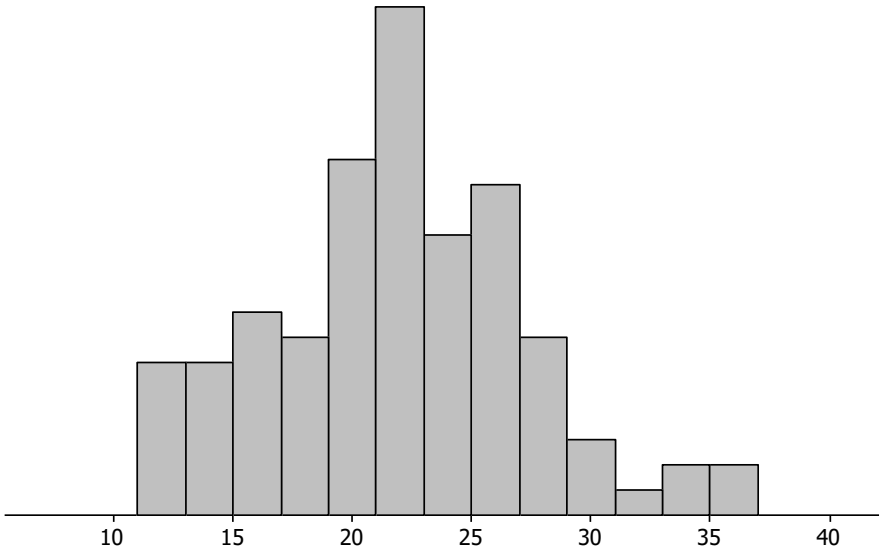
- Om vi drar 100 oberoende stickprov om storleken $n = 10$, beräknar de 100 stickprovsmedelvärdena och visualiserar mätningarna i ett histogram fås följande diagram



$$\bar{\bar{x}} = 22.7$$

Samplingfördelning

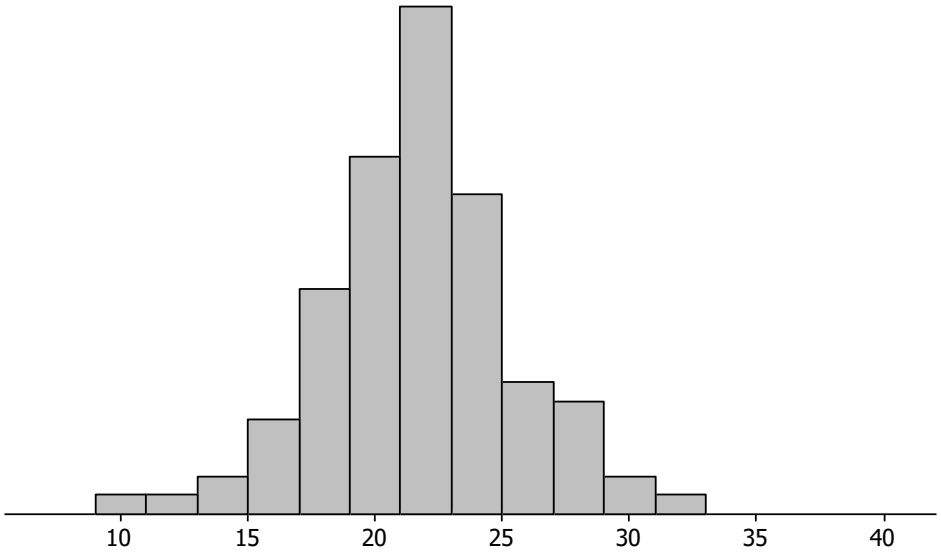
- Experimentet upprepas för 100 oberoende stickprov om storlek $n = 20$



$\bar{\bar{x}} = 22.0$

Samplingfördelning

- Slutligen upprepas experimentet för 100 oberoende stickprov om storlek $n = 30$



$\bar{\bar{x}} = 21.7$

Samplingfördelning

- Stickprovsmedelvärdena följer en fördelning
- Vi kan betrakta denna fördelning som en uppskattning av den fördelning som skulle fås om vi åskådliggjorde stickprovsmedelvärdena för samtliga möjliga stickprov av en viss storlek ur populationen, vilket kallas för en samplingfördelning.

Centrala gränsvärdessatsen

Samplingfördelningen för summor eller medelvärden av n oberoende slumpvariabler med samma fördelning är approximativt normalfördelad om n är tillräckligt stort

Centrala gränsvärdessatsen

- Samplingfördelningen blir mer och mer lik (konvergerar) mot normalfördelningen när stickprovsstorleken ökar
  - Detta gäller även om populationen stickproven dras ifrån inte är normalfördelad
- Vanlig tumregel är $n \geq 30$

Exempel

- Ett flygbolag räknar med att medelvikten på en passagerare är 80kg med en standardavvikelse om 5kg. Vikten för en passagerare är dock inte normalfördelad. En viss flygplanstyp rymmer 290 passagerare.

Linjära variabeltransformationer

Linjära variabeltransformationer av normalfördelade slumpvariabler är alltid normalfördelade

Linjära variabeltransformationer

- Innebörden blir att medelvärden, summor och andelar beräknade på normalfördelade observationer, genom att de dragits ur en population som är normalfördelad, är också normalfördelade oavsett stickprovets storlek
- Ex: Felet hos hastighetsmätaren på en slumpmässigt vald bil av ett visst märke kan ses som normalfördelat och överskattar i medel den sanna hastigheten med 3km/h, med en standardavvikelse på 2km/h. Beskriv fördelningen för hur långt bilen hinner köra på 5 timmar om mätaren visar 100km/h.

Stickprovsstatistikors fördelning

- Om $n \geq 30$ gäller, p.g.a. centrala gränsvärdessatsen, att

$$\bar{X} \approx N\left(\mu_{\bar{X}} = \mu; \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\right)$$

$$\sum X \approx N\left(\mu_{\sum x} = n \cdot \mu; \sigma_{\sum x} = \sqrt{n} \cdot \sigma\right)$$

- Om $n < 30$ krävs att populationen som stickprovet dragits ur är normalfördelad.

Stickprovsandelens fördelning

- Om $np(1-p) > 5$ gäller:

$$P \approx N\left(\mu_P = \pi; \sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}\right)$$

- Detta p.g.a. normallapproximation om n är tillräckligt stort

Exempel

- Vikten av en jordgubbe har väntevärde 13 gram och standardavvikelse 5 gram.

  En låda innehåller 35 jordgubbar. Vad är sannolikheten för att den sammanlagda vikten av lådan överstiger 500 gram om lådan själv väger 50 gram?

Kapitel 6, sid 153-185
Inferens om en population

LINKÖPING UNIVERSITY

Agenda

- Statistisk inferens om populationsmedelvärde
- Statistisk inferens om populationsandel
- Punktskattning
- Konfidensintervall
- Hypotesprövning

Statistisk inferens om populationsmedelvärde

Punktskattning

- Att använda en stickprovsstatistika som en uppskattning av motsvarande parameter

- Stickprovsstatistikor är slumpvariabler och antar olika värden för varje stickprov

- Hur ska vi hantera osäkerheten?

Krav för konfidensintervall för medelvärde

1. Stickprovet är draget som ett OSU
   - Garanterar oberoende mellan observationerna

2. Samplingfördelningen för stickprovsmedelvärdet kan betraktas som normalfördelad
   - Antingen genom centrala gränsvärdessatsen, $n \geq 30$
   - Populationen kan betraktas som normalfördelad

- Om kraven är uppfyllda kan vi skatta osäkerheten genom ett så kallat konfidensintervall

Punktskattning och intervallskattning

- Om kraven är uppfyllda kan vi bilda ett konfidensintervall för populationsmedelvärdet: vi lägger ett osäkerhetsintervall kring punktskattningen vilket tillåter oss att med en viss säkerhet säga att den okända populationsparametern täcks av intervallet.

Dubbelsidigt konfidensintervall för populationsmedelvärde när σ är okänd

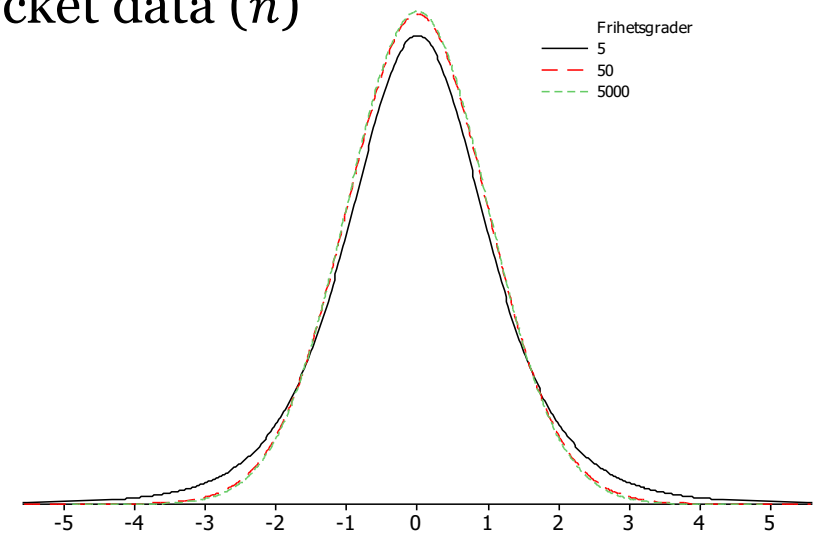
- Givet de två antaganden bildas ett dubbelsidigt konfidensintervall för μ enligt:

$$\bar{x} \pm t_{n-1;1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

- Värdet på t hämtas från t-fördelningen, där $n-1$ är frihetsgrader och α är signifikansnivån

t-fördelningen

- Fördelningen liknar normalfördelningen men beror på frihetsgrader
- Används om stickprovet är litet och om σ är okänd
- Frihetsgrader bestäms av hur mycket data (n) man har
- Fördelningen konvergerar mot normalfördelningen
- Approximativt normal då $n \geq 30$



Exempel

- Ett slumpmässigt urval om 40 studenter vid Linköpings universitet ger medelåldern 21.2 år och standardavvikelsen 4.4 år.

  Bestäm ett intervall som med 95 procents säkerhet täcker den sanna medelåldern bland studerande vid Linköpings universitet.

Enkelsidiga konfidensintervall (KI)

- Nedåt begränsat KI: $\mu > \bar{x} - t_{n-1;1-\alpha} \cdot \frac{s}{\sqrt{n}}$

- Uppåt begränsat KI: $\mu < \bar{x} + t_{n-1;1-\alpha} \cdot \frac{s}{\sqrt{n}}$

Exempelfråga

- Styrelsen i en bostadsrättsförening får in klagomål på att golvvärmen i badrummen är för låg. Man drar ett OSU om 30 badrum bland de omkring 400 badrum som finns i föreningens fastigheter och mäter golvvärmen där. Medeltemperaturen beräknas till 21 grader och standardavvikelsen till 1.6 grader.

  Energimyndigheten rekommenderar att golvvärmen ska ligga på minst 20 grader för att man ska undkomma problem med fuktskador. Föreligger risk för fuktskador i föreningens badrum?

Hypotesprövning för populationsmedelvärde när σ är okänd

Förutsätts åter att

1. stickprovet är draget som ett OSU
2. samplingfördelningen för stickprovsstatistikan kan betraktas som normalfördelad

- En **hypotesprövning** testar om ett påstående (en hypotes) är förenlig med ett observationerna i ett stickprov

- En hypotesprövning kan utföras i 4 steg

Hypotesprövning

- Steg 1: Formulera hypoteser och välj signifikansnivå

$$H_0: \mu = \mu_0$$

$$H_a: \mu > \mu_0$$
$$H_a: \mu < \mu_0$$
$$H_a: \mu \neq \mu_0$$

  - Valet av mothypotes, H_a, bestäms av frågeställningen
  - α = signifikansnivån = risken att förkasta H_0 trots att H_0 är sann
    - Vanliga värden är 1%, 5%, 10%

Hypotesprövning

- Steg 2: Bestäm testvariabeln

$$t_{test} = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

Hypotesprövning

- Steg 3: Beräkna det kritiska området
  - Om H_a: $<$, till vänster om
    $$t_{krit} = -t_{n-1;1-\alpha}$$
  - Om H_a: $>$, till höger om
    $$t_{krit} = t_{n-1;1-\alpha}$$
  - Om H_a: $\neq$, på båda sidor om
    $$t_{krit} = \pm t_{n-1;1-\frac{\alpha}{2}}$$
  - Visualisera detta med ett diagram!

Hypotesprövning

- Steg 4: Dra slutsatser och tolka
  - Besluta att förkasta/inte förkasta H_0
  - Besvara frågeställningen i ord

Exempel

- I ett OSU omfattande 40 personer bland medlemmarna i ett politiskt parti i en region är medelåldern 42.3 år och standardavvikelsen 7.1 år.

  Testa på 5 procents signifikansnivå om medelåldern bland medlemmarna i partiet understiger 45 år.

Om σ är känd

- Väldigt ovanligt att σ är känd

- I detta fallet byts t-fördelningen ut mot normalfördelningen (ex. byts $t_{n-1;1-\alpha}$ ut mot $z_{1-\alpha}$) i formlerna. s byts ut mot σ.

Inferens om populationsandel

Konfidensintervall för populationsandel

- Givet att:
  1. Stickprovet är draget som ett OSU
  2. $np(1-p) > 5$

  bildas ett dubbelsidigt KI för π enligt:

$$p \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n}}$$

- Där värdet på z hämtas ur normalfördelningstabellen

Enkelsidiga intervall

- Nedåt begränsat KI: $\pi > p - z_{1-\alpha}\sqrt{\frac{p(1-p)}{n}}$

- Uppåt begränsat KI: $\pi < p + z_{1-\alpha}\sqrt{\frac{p(1-p)}{n}}$

Exempel

- I en hälsoenkät tillfrågades 100 slumpmässigt utvalda anställda vid ett stort företag om huruvida man regelbundet motionerar eller ej.

  Svar erhölls från 84 anställda och av dessa svarade 65 ja.

  Bestäm ett 95-procentigt konfidensintervall för andelen av de anställda vid det stora företaget som regelbundet motionerar.

Hypotesprövning

- Förutsätter att kraven för andelar är uppfyllda

- Steg 1:

$$H_0: \pi = \pi_0$$

$$H_a: \pi > \pi_0$$
$$H_a: \pi < \pi_0$$
$$H_a: \pi \neq \pi_0$$

Hypotesprövning

- Steg 2:

$$z_{test} = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

- Steg 3:

$$z_{krit} = -z_{1-\alpha}; z_{1-\alpha}; z_{1-\frac{\alpha}{2}}$$

  - Beror på H_a

Hypotesprövning

- Steg 4:
  - Dra slutsats och tolka

Exempel

- I en hälsoenkät tillfrågades 100 slumpmässigt utvalda anställda vid ett stort företag om huruvida man regelbundet motionerar eller ej.

  Svar erhölls från 84 anställda och av dessa svarade 65 ja.

  Undersök om det på 5 procents signifikansnivå finns belägg för påståendet att andelen regelbundna motionärer bland de anställda vid företaget understiger 85 procent.

Ytterligare om inferens

p-värdesmetoden

- Som ett alternativt steg 3 och 4 i en hypotesprövning

- p-värdet är sannolikheten att testvariabeln ska anta ett värde som det vi observerat eller ännu längre ifrån μ_0 sett i den riktning som mothypotesen pekar

- Om p-värdet är litet är H_0 osannolik

- Beslutsregel: om p-värdet < signifikansnivån α så förkastas H_0
  - Vid dubbelsidig mothypotes beräknas p-värdet multiplicerat med 2

Exempel

- I en hälsoenkät tillfrågades 100 slumpmässigt utvalda anställda vid ett stort företag om huruvida man regelbundet motionerar eller ej.

  Svar erhölls från 84 anställda och av dessa svarade 65 ja.

  Undersök om det på 5 procents signifikansnivå finns belägg för påståendet att andelen regelbundna motionärer bland de anställda vid företaget understiger 85 procent genom att beräkna testets p-värde.

Relation mellan hypotesprövning och KI

- Om μ_0 el. π_0 täcks av intervallet kan H_0 inte förkastas
- Vid $H_a: <$ beräknas ett uppåt begränsat KI
- Vid $H_a: >$ beräknas ett nedåt begränsat KI
- Vid $H_a: \neq$ beräknas ett dubbelsidigt KI
- http://rpsychologist.com/d3/CI/

Feltyper och styrka

- Typ I-fel: att förkasta H_0 fastän den är sann
- Typ II-fel: att inte förkasta H_0 fastän den är falsk

- α är sannolikheten (risken) för typ I-fel

Feltyper och styrka

	Sanning om populationen		
Beslut baserat på stickprov		H_0 sann	H_a sann
	Förkasta H_0	Typ I-fel	Korrekt beslut
	Inte förkasta H_0	Korrekt beslut	Typ II-fel

- Det råder ett motsatsförhållande mellan risken för typ I-fel och risken för typ II-fel

- Inom samhällsvetenskaperna brukar man ange $\alpha = 0.05; 0.01; 0.10$ som ger en bra avvägning mellan typerna av fel

Kapitel 7, sid 186-209
Jämförelse av två populationer

LINKÖPING UNIVERSITY

Agenda

- Jämförelse av medelvärden för två populationer
- Jämförelse av populationsandelar för två populationer
- Konfidensintervall och hypotesprövning
- Parvisa observationer

Konfidensintervall (KI) för jämförelse av populationsmedelvärde $(\mu_1 - \mu_2)$

- Krav
  - de två stickproven är dragna som ett OSU
  - Samplingfördelningarna kan betraktas som normalfördelade

$$(\bar{x}_1 - \bar{x}_2) \pm t_{n^*-1;1-\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- där n^* är den minsta av n_1 och n_2

Ensidigt KI för jämförelse av μ

- Nedåt begränsat: $\mu_1 - \mu_2 > (\bar{x}_1 - \bar{x}_2) - t_{n^*-1;1-\alpha}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

- Uppåt begränsat: $\mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{n^*-1;1-\alpha}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Exempel

- I ett medicinskt experiment sammankallade man 80 friska medelålders personer, som under tre månader fick pröva ett nytt medicinskt preparat. Syftet med studien var att utreda om preparatet ger förhöjt blodtryck som en biverkning. 40 av personerna fick preparatet, medan 40 fick placebo (ett verkningslöst preparat). Varken patient eller försöksledare visste under studietiden vem som fick vilket preparat (en så kallad *dubbelblind* studie). Varje person fick varje dag mäta sitt blodtryck, och efter tre månader sammanställdes informationen och räknades om till genomsnittligt blodtryck och standardavvikelse i respektive grupp.
- Går det, på 95 procents konfidensnivå, att påvisa några skillnader i genomsnittligt blodtryck mellan personer som fick aktivt preparat och de som fick placebo?

Grupp	n_i	$\overline{x}_i$	s_i
1 – Aktivt preparat	40	142.5	14.8
2 – Placebo	40	135.9	21.4

KI för jämförelse av andelar i två populationer $(\pi_1 - \pi_2)$

- Krav
  - de två stickproven är dragna som ett OSU
  - $np(1-p) > 5$ för båda stickproven

$$(p_1 - p_2) \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Ensidigt KI för jämförande av π

- Nedåt begränsat: $\pi_1 - \pi_2 > (p_1 - p_2) - z_{1-\alpha}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

- Uppåt begränsat: $\pi_1 - \pi_2 < (p_1 - p_2) + z_{1-\alpha}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$

Exempel

- I utvärderingen av det nya preparatet (se tidigare exempel) så undersökte man även förekomsten av sömnsvårigheter.

  Bland de 40 personerna som fått den aktiva substansen (grupp 1) uppgav 9 att de haft regelbundna sömnsvårigheter under studieperioden. Bland personerna i placebogruppen (grupp 2) var motsvarande siffra 6 personer.

  Går det på 99% konfidensnivå att påvisa skillnad i andelen personer med sömnsvårigheter mellan grupperna?

Hypotesprövning för jämförelser av μ

- Samma krav som för konfidensintervall

- Steg 1: Välj signifikansnivå och formulera hypoteser

$$H_0: \mu_1 - \mu_2 = d_0$$

  - där d_0 är den differens vi testar för, oftast 0

$$H_a: \mu_1 - \mu_2 < d_0$$
$$H_a: \mu_1 - \mu_2 > d_0$$
$$H_a: \mu_1 - \mu_2 \neq d_0$$

  - valet av mothypotes bestäms av frågeställningen

Hypotesprövning för jämförelser av μ

- Steg 2: Bestäm testvariabeln

$$t_{test} = \frac{(\bar{x}_1 - \bar{x}_2) - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Hypotesprövning för jämförelser av μ

- Steg 3: Bestäm det kritiska området
  - Om H_a: $\mu_1 - \mu_2 < d_0$ ligger det kritiska området till vänster om det kritiska värdet $-t_{n^*-1;1-\alpha}$
  - Om H_a: $\mu_1 - \mu_2 > d_0$ ligger det kritiska området till höger om det kritiska värdet $t_{n^*-1;1-\alpha}$
  - Om H_a: $\mu_1 - \mu_2 \neq d_0$ har vi kritiska områden både till vänster och höger om de kritiska värdena $\pm t_{n^*-1;1-\frac{\alpha}{2}}$ i respektive svans
  - Kom ihåg att n^* är den minsta av n_1 och n_2
  - Tips: Rita upp fördelningen och det kritiska området

Hypotesprövning för jämförelser av μ

- Steg 4: Ta beslut och tolka
  - Placera t_{test} på den fördelning som ritats
  - Om t_{test} hamnar i kritiska området ⇒ förkasta H_0

Exempel

Grupp	Antal personer	Genomsnittligt blodtryck	Standardavvikelse
1 – Aktivt preparat	40	142.5	14.8
2 – Placebo	40	135.9	21.4

- Går det, på 5% signifikansnivå, att påvisa att det genomsnittliga blodtrycket är högre i gruppen som fått aktivt preparat?

Hypotesprövning för jämförelser av π

- Krav: stickprovet draget som OSU och $np(1-p) > 5$

- Steg 1: Hypoteser

$$H_0: \pi_1 - \pi_2 = d_0$$
$$H_a: \pi_1 - \pi_2 < d_0$$
$$H_a: \pi_1 - \pi_2 > d_0$$
$$H_a: \pi_1 - \pi_2 \neq d_0$$

Hypotesprövning för jämförelser av π

- Steg 2: Bestäm testvariabeln

$$z_{test} = \frac{(p_1 - p_2) - d_0}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

Hypotesprövning för jämförelser av π

- Steg 3: Bestäm det kritiska området
  - Om $H_a: \pi_1 - \pi_2 < d_0$ ligger det kritiska området till vänster om det kritiska värdet $z_\alpha = -z_{1-\alpha}$
  - Om $H_a: \pi_1 - \pi_2 > d_0$ ligger det kritiska området till höger om det kritiska värdet $z_{1-\alpha}$
  - Om $H_a: \pi_1 - \pi_2 \neq d_0$ har vi kritiska områden både till vänster och höger om de kritiska värdena $\pm z_{1-\frac{\alpha}{2}}$ i respektive svans
  - Tips: Rita upp fördelningen och kritiska området

Hypotesprövning för jämförelser av π

- Steg 4: Ta beslut och tolka
  - Om z_{test} hamnar i kritiska området ⇒ förkasta H_0

Exempel

- I utvärderingen av det nya preparatet (se tidigare exempel) så undersökte man även förekomsten av sömnsvårigheter.

  Bland de 40 personerna som fått den aktiva substansen (grupp 1) uppgav 9 att de haft regelbundna sömnsvårigheter under studieperioden. Bland personerna i placebogruppen (grupp 2) var motsvarande siffra 6 personer.

  Går det på 5% signifikansnivå att påvisa att andelen personer med sömnsvårigheter är större i gruppen som fått den aktiva substansen?

Alternativt

- Steg 3 och 4 kan bytas ut med p-värdesmetoden

  – Om p-värdet $< \alpha \Rightarrow$ förkasta H_0

Exempel

Sämre prognos för män med bröstcancer
Bland män som insjuknar i bröstcancer är överlevnaden betydligt lägre än för kvinnor, enligt en studie vid Akademiska sjukhuset i Uppsala. 99 män med bröstcancer följdes under 15 år och jämfördes med 369 kvinnliga bröstcancerpatienter. Fem år efter diagnosen levde 55 procent av kvinnorna men bara 41 procent av männen.

Östgöta Correspondenten, torsdag 27 oktober 2011

- På vilken signifikansnivå har forskarna kunnat dra denna slutsats?

Parvisa observationer

- Om samma enhet undersöks vi två olika tillfällen uppfylls inte kravet på oberoende mellan stickproven (upprepad mätning) eller något annat beroende finns mellan observationerna.
- Beräkna differensen (d) och sedan utförs hypotesprövning på samma vis som för en population men med följande beteckningar:

$$t = \frac{\bar{d} - \mu_d}{s_d}$$

Exempel

- Vattenplaning är en stor trafikfara, och av stor betydelse är bildäckens förmåga att pressa undan vatten. För att undersöka **vid vilken hastighet vattenplaning uppnås** vid ett kontrollerat experiment på en vattenfylld bana provades två däcktyper: en med traditionellt däckmönster och en med ett nyutvecklat mönster skapat just för att tränga undan vatten. Varje typ av däck provades på 10 typer av bilar eftersom bilens tyngd och aerodynamiska egenskaper också kan påverka vid vilken hastighet vattenplaning uppnås. Följande resultat erhölls.

Deltagare	1	2	3	4	5	6	7	8	9	10
Traditionellt	59	98	62	102	61	115	77	95	74	83
Nytt	64	103	77	99	59	115	79	89	68	85

Exempel

- Är det nya mönstret bättre, sett till vid vilken hastighet vattenplaning uppnås (det är givetvis önskvärt att man ska kunna köra så fort som möjligt utan att få vattenplaning), jämfört med det traditionella mönstret? Utred frågeställningen på 5% signifikansnivå. Vilka antaganden måste göras för att metodiken ska vara tillämpbar?

Exempel

- Skapa en ny variabel som visar differensen!

Deltagare	1	2	3	4	5	6	7	8	9	10
Traditionellt	59	98	62	102	61	115	77	95	74	83
Nytt	64	103	77	99	59	115	79	89	68	85
DIFFERENS	-5	-5	-5	3	2	0	-2	6	6	-2

- Betrakta den nya variabeln som en grupp och använd metoder för inferens om en population

Relation mellan hypotesprövning och KI

- Om d_0 ingår i intervallet som skattats kan H_0 ej förkastas
- Vid $H_a: <$ beräknas ett uppåt begränsat KI
- Vid $H_a: >$ beräknas ett nedåt begränsat KI
- Vid $H_a: \neq$ beräknas ett dubbelsidigt KI

Kapitel 8 – sid 210-229 Inferens om en ändlig population

Läses på egen hand

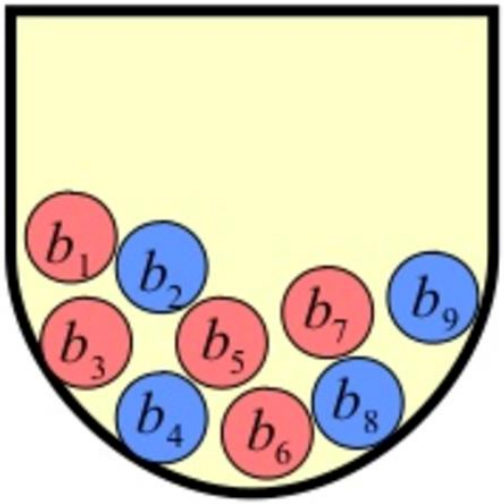
LINKÖPING UNIVERSITY

Agenda

- Statistisk inferens vid ändlig population
- Populationsmedelvärde och totalmängd
- Populationsandel och totalt antal

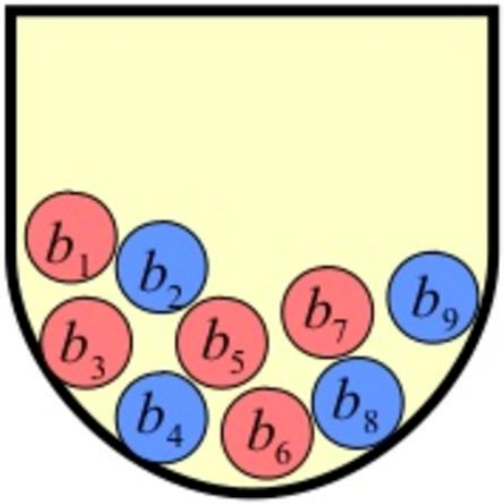
Ändlig population: dragning utan återläggning

- En skål med 5 röda och 4 blå kulor
- Första dragningen: $\frac{5}{9} \approx 55{,}56\%$ sannolikhet att en röd kula dras
- Andra dragningen: $\frac{4}{8} = 50\%$ eller $\frac{5}{8} = 62{,}5\%$ att en röd kula dras, beroende på första dragningen
- Sannolikheten för röd kula förändras efter varje dragning!



Ändlig population: dragning utan återläggning

- Jämför med (approximativt) oändlig population: 5000 röda och 4000 blå kulor.
- Första: $\frac{5000}{9000} \approx 55{,}56\%$ sannolikhet för röd kula
- Andra: $\frac{4999}{8999} \approx 55{,}55\%$ eller $\frac{5000}{8999} \approx 55{,}56\%$ sannolikhet för röd kula
- Vid ändlig population måste vi korrigera för att sannolikheten ändras!



Dubbelsidigt konfidensintervall för populationsmedelvärde vid ändlig population

- Om $\frac{n}{N} > 10\%$ betraktas populationen som ändlig

- Givet att:
  - Stickprovet är draget som ett OSU
  - Samplingfördelningen kan betraktas som normalfördelad

$$\bar{x} \pm t_{n-1;1-\frac{\alpha}{2}} \sqrt{\frac{s^2}{n}\underbrace{\left(1-\frac{n}{N}\right)}_{\text{ändlighetskorrektion}}}$$

Konfidensintervall för totalmängd

- Vi har information om populationsstorleken

- Sanna totalmängden:

$$N \cdot \mu$$

- Punktskattning beräknas enligt:

$$N \cdot \bar{x}$$

- Med intervallet

$$N \cdot \bar{x} \pm t_{n-1;1-\frac{\alpha}{2}} \cdot N \cdot \sqrt{\frac{s^2}{n}\left(1 - \frac{n}{N}\right)}$$

Konfidensintervall för populationsandel π

- Om $\frac{n}{N} > 10\%$ och följande krav har uppfyllts:
  - OSU
  - $np(1-p) > 5$

$$p \pm z_{1-\frac{\alpha}{2}} \sqrt{\frac{p(1-p)}{n-1}\left(1-\frac{n}{N}\right)}$$

Konfidensintervall för totalt antal

- Vi har information om populationsstorleken

- Sanna totalmängden:

$$N \cdot \pi$$

- Punktskattning beräknas enligt:

$$N \cdot p$$

- Med intervallet

$$N \cdot p \pm z_{1-\frac{\alpha}{2}} \cdot N \cdot \sqrt{\frac{p(1-p)}{n-1}\left(1-\frac{n}{N}\right)}$$

Kapitel 9 och 10 – sid 230-284
Samband mellan kvalitativa och kvantitativa variabler

LINKÖPING UNIVERSITY

Agenda

- Samband mellan kvalitativa variabler
- Chitvåtest för analys av frekvenstabell och korstabell
- Samband mellan kvantitativa variabler

Samband mellan kvalitativa variabler

Chitvåtest (χ^2) för analys av frekvenstabell

- Krav:
  - det råder oberoende mellan grupperna i tabellen. Innebörden i detta är att samma element (person) endast får ingå i en grupp.
  - max 20% av de förväntade frekvenserna är mindre än 5
  - alla förväntade frekvenser är större än 1.

- Steg 1: Välj signifikansnivå och formulera hypoteser

$$H_0: Det\ finns\ inga\ skillnader\ i\ frekvens\ mellan\ grupperna$$
$$H_a: Det\ finns\ skillnader\ i\ frekvens\ mellan\ grupperna$$

χ^2-test för analys av frekvenstabell

- Steg 2: Testvariabeln

$$\chi^2_{test} = \sum_{i=1}^{V} \frac{(O_i - E_i)^2}{E_i}$$

  - V är antalet grupper
  - O_i är de **observerade** frekvenserna
  - E_i är de **förväntade** frekvenserna (kraven gäller för dessa), och beräknas genom:

$$E_i = \frac{\sum O_i}{V}$$

χ^2-test för analys av frekvenstabell

- Steg 3: Beräkna kritiska området
  - Ett χ^2-test har alltid kritiskt område till höger av fördelningen

$$\chi^2_{krit} = \chi^2_{V-1;\alpha}$$

- Steg 4: Dra slutsats och tolka
  - Hamnar χ^2_{test} i det kritiska området $\Rightarrow$ förkasta H_0.

Exempel

- En bilförsäljare har sålt bilar av ett visst märke som tillverkas i tre färger: röd, svart och silver. Försäljningen av bilar i respektive färg en viss månad presenteras i följande tabell.

Färg	Antal sålda bilar (f)
Röd	13
Svart	21
Silver	17
Totalt	**51**

- Finns det någon skillnad i popularitet mellan färgerna på 5% signifikansnivå?

χ^2-test för analys av korstabell

- Krav:
  - det råder oberoende mellan cellerna. Innebörden är att samma element (person) inte får förekomma i flera celler i tabellen
  - max 20% av de förväntade frekvenserna är mindre än 5
  - alla förväntade frekvenser är större än 1

- Steg 1: Hypoteser
  H_0: *Det finns inga skillnader i fördelning mellan grupperna*
  (*alternativt: Det finns inget samband mellan grupperna*)

  H_a: *Det finns skillnader i fördelning mellan grupperna*
  (*alternativt: Det finns samband mellan grupperna*)

χ^2-test för analys av korstabell

- Steg 2: Testvariabeln

$$\chi^2_{test} = \sum_{i=1}^{W} \frac{(O_i - E_i)^2}{E_i}$$

  - W är antalet celler i korstabellen
  - O_i är de **observerade** frekvenserna
  - E_i är de **förväntade** frekvenserna (kraven gäller för dessa), och beräknas genom:

$$E_i = \frac{radtot \cdot koltot}{tottot}$$

χ^2-test för analys av korstabell

- Steg 3: Kritiskt område

$$\chi^2_{krit} = \chi^2_{(r-1)(c-1);\alpha}$$

  - r och c är antalet rader respektive kolumner

- Steg 4: Dra slutsats och tolka
  - Hamnar χ^2_{test} i det kritiska området ⇒ förkasta H_0.

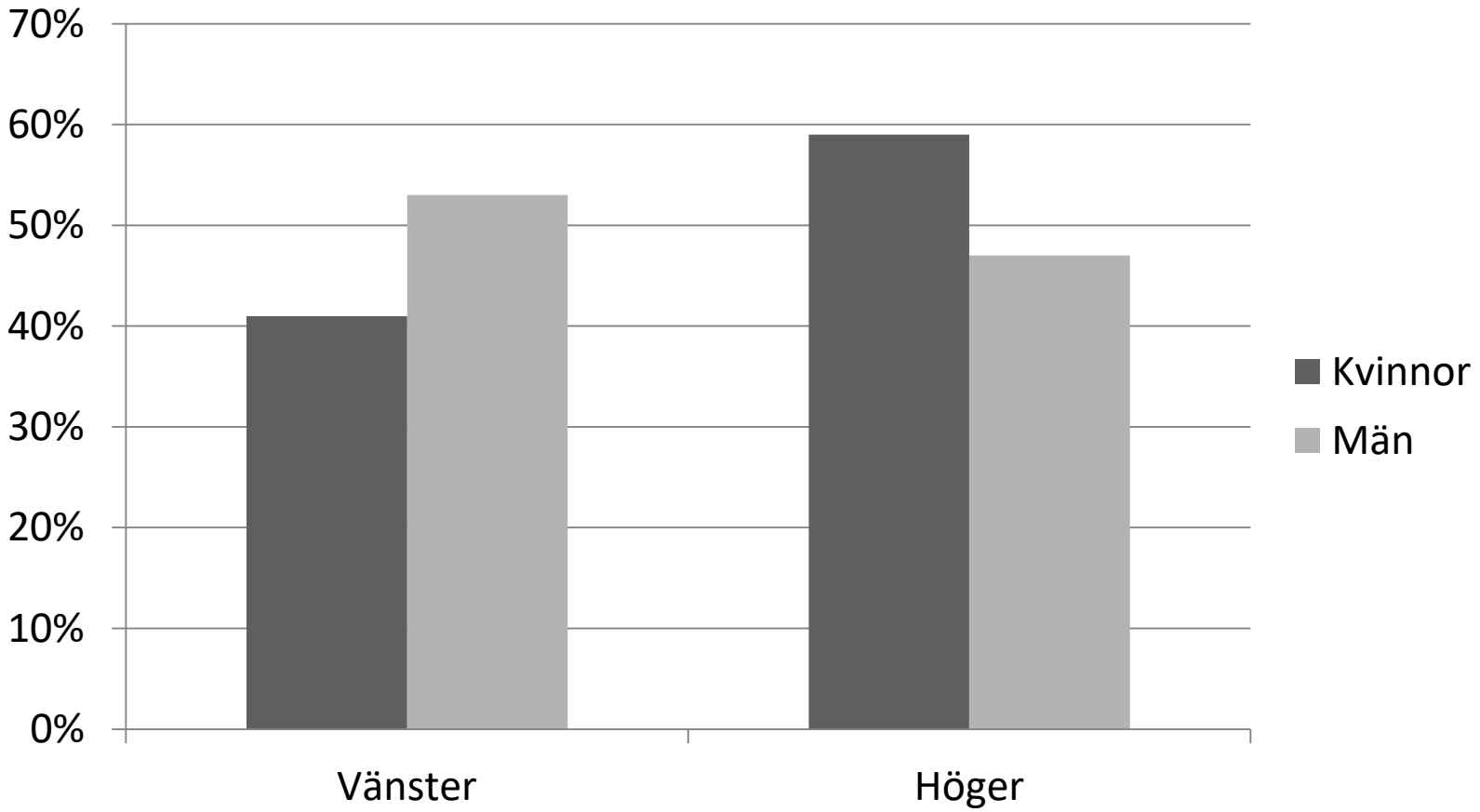
Exempel

- Man drar ett OSU om medlemmar ur en stor politiskt oberoende organisation, och frågar dels om kön, dels om politisk tillhörighet (vänster eller höger). Följande resultat erhålles.

	Vänster	Höger
Kvinna	98	141
Man	67	59

Går det att på 5 procents signifikansnivå påvisa några skillnader mellan kvinnor och män som är medlemmar i organisationen i fråga om politisk tillhörighet?

Exempel (forts.)



Sammanslagningar av variabler vid χ^2-test

- Om inte kraven uppfylls måste ibland sammanslagningar av alternativ genomföras
- Dessa måste ske på logiskt vis, t.ex. måste ordningen bibehållas
- Vid nominalskala är det enklast att skapa ett nytt alternativ som kallas ”Övrigt”

Exempel

- På ett företag angav chefer och övriga hur många dagar i veckan de motionerade

Antal dagar	Chefer	Övriga
0	6	38
1	8	19
2	5	23
3	5	10
4	0	4
5	1	3
6	1	2
7	0	1

Samband mellan kvantitativa variabler

Att studera i ett spridningsdiagram

- Är sambandet linjärt?
  - Undersök om punkterna faller längs en tänkt rät linje

- Lutar punktsvärmen?
  - Lutar punkterna uppåt är det ett positivt samband, nedåt ett negativt samband

Att studera i ett spridningsdiagram

- Hur starkt är sambandet?
  - Undersök hur tätt observationerna ligger längs den tänkta räta linjen. Om de är utspridda är sambandet svagt, om de ligger nära linjen är sambandet starkt

- Finns det några observationer som kraftigt avviker från de övriga?
  - Dessa observationer kallas för uteliggare och kan (men inte behöver) bero på felmätning eller felinmatning

Korrelationskoefficienten

- Matematiskt mått för styrkan av ett linjärt samband mellan två variabler

$$r = \frac{SS_{XY}}{\sqrt{SS_{XX} \cdot SS_{YY}}} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- Korrelationskoefficienten antar värden mellan –1 och +1.
  - Ju närmare –1 desto starkare negativt linjärt samband
  - Ju närmare +1 desto starkare positivt linjärt samband
  - Om korrelationskoefficienten är nära 0 finns inget linjärt samband

Tolkning av korrelationskoefficienten

- Vi tolkar *absolutvärdet* av korrelationskoefficienten (betecknas $|r|$):

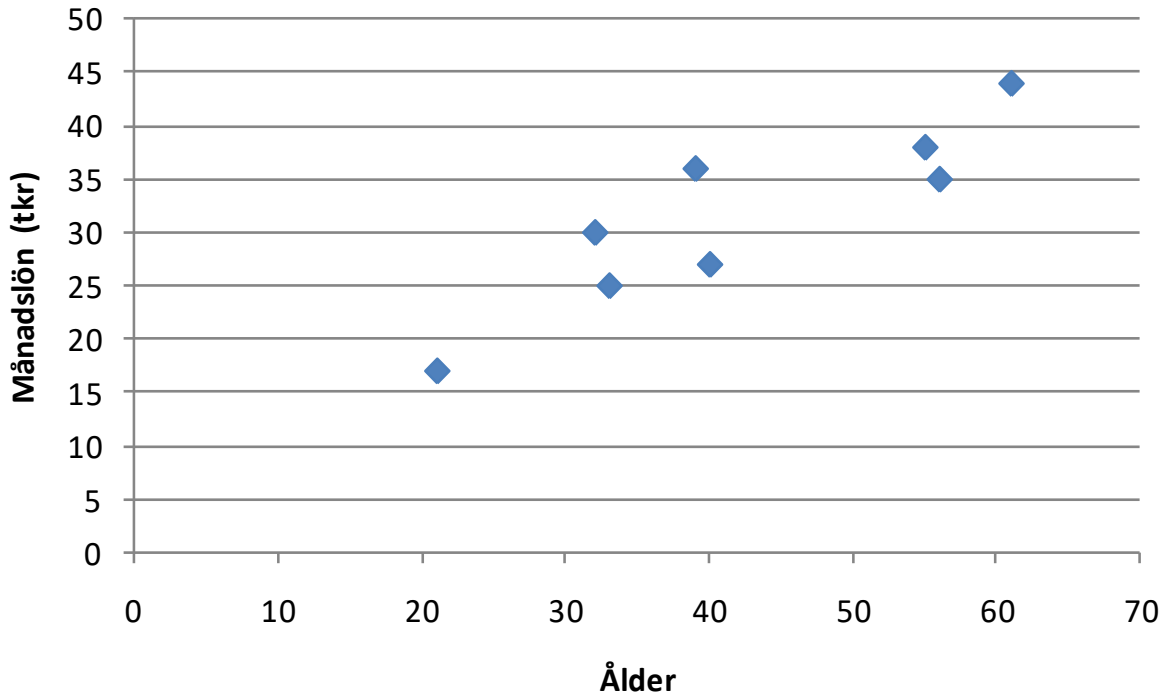
$\|r\|$	Samband
> 0.85	Mycket starkt
0.65 – 0.85	Starkt
0.35 – 0.65	Måttligt
0.20 – 0.35	Svagt
< 0.20	Mycket svagt

Exempel

- Ett företag har ett större antal säljare anställda. Vi har dragit ett OSU om 8 av dessa och för varje utvald person undersökt ålder och månadslön (i tusentals kronor). Föreligger det något samband mellan ålder och månadslön för säljare vid företaget?

Säljare	Månadslön	Ålder
1	17	21
2	30	32
3	27	40
4	35	56
5	44	61
6	38	55
7	36	39
8	25	33

Exempel


Månadslön (tkr)
50
45
40
35
30
25
20
15
10
5
0
0
10
20
30
40
50
60
70
Ålder

Korrelation ≠ Kausalitet