Statistik 1 för biologer, logopeder och psykologer

Föreläsningar, del 4

Innehåll

Innehåll

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.

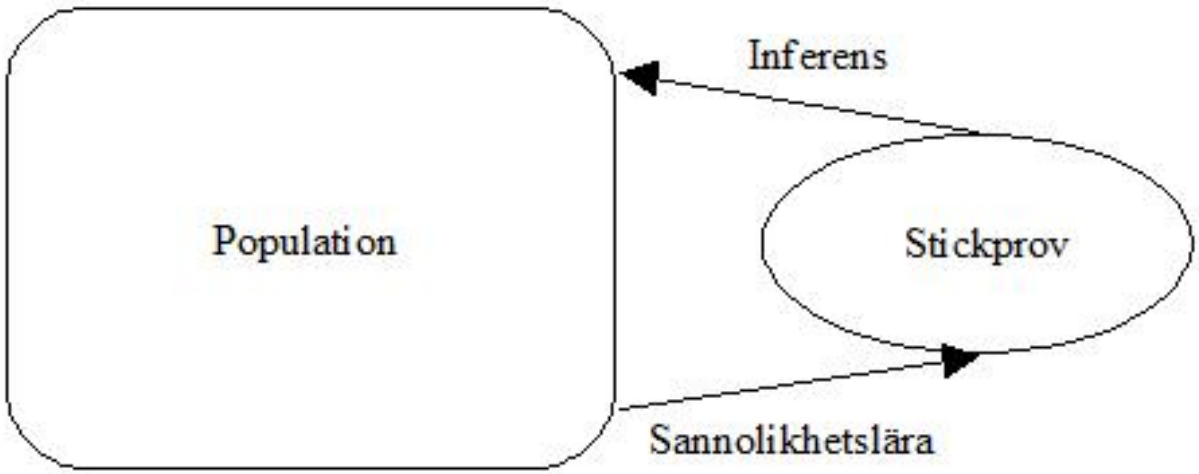
Stickprov

- 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.

Från sannolikhetslära till inferens



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.

Exempel.

- 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 $N(\mu, \sigma/\sqrt{n})$.

Medelvärdesfördelningen

- 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

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

Konfidensintervall för väntevärde

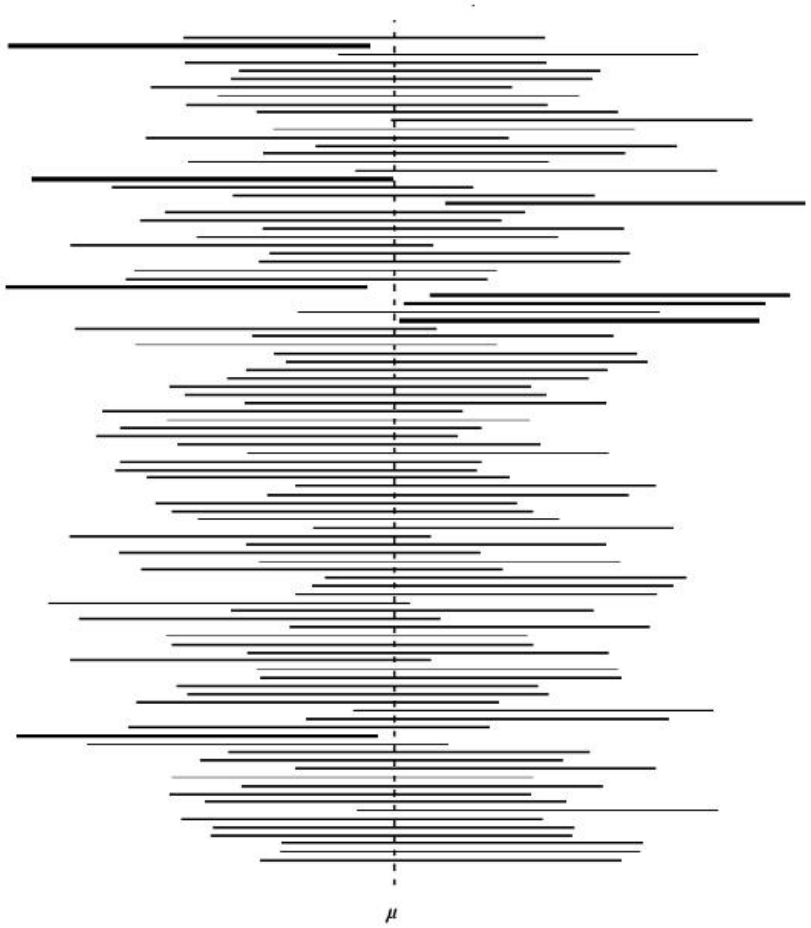
- Vi vet nu att avståndet mellan medelvärdet $\overline{x}$ och väntevärdet μ med 95% sannolikhet är högst $1.96 \cdot \sigma/\sqrt{n}$.
- Likväl kan vi då säga att intervallet

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

  med 95% sannolikhet täcker det okända väntevärdet μ.
- Vi har alltså konstruerat ett 95% konfidensintervall för μ.

Konfidensintervall för väntevärde

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



Konfidensintervall för väntevärde

- 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

Konfidensintervall för väntevärde

Exempel.

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!

Konfidensintervall för väntevärde

- 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

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 $f = n - 1$, 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 från 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.

Konfidensintervall för väntevärde då σ är okänd

Exempel.

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 $f = 9 - 1 = 8$. 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

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

Konfidensintervall för proportioner

- 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 $\hat{P} = \frac{X}{n}$.

Konfidensintervall för proportioner

- 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

$$p = \hat{p} \pm z \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} .$$

Konfidensintervall för proportioner

Exempel.

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*.