Föreläsning 4

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

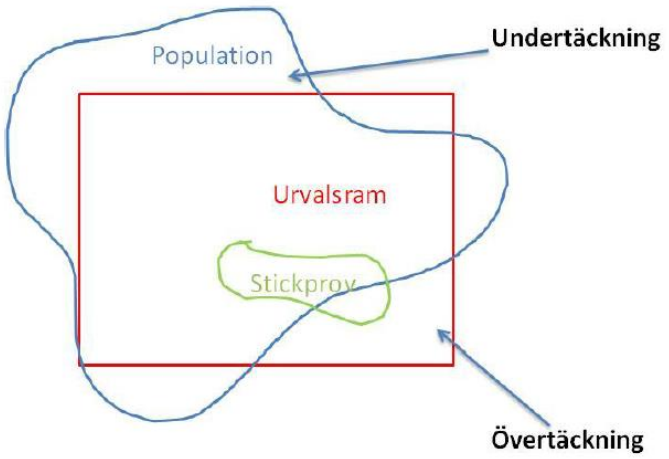
- Stratifierade urval, för en heterogen population, ger normalt mindre standardavvikelse och därmed säkrare slutsatser om populationen

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.

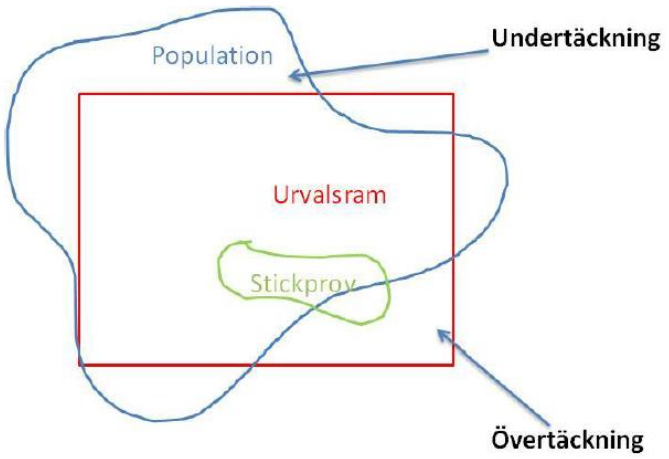
Felkällor vid stickprovsundersökningar

- **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	$E(P) = \pi$	$Var(P) = \frac{\pi(1-\pi)}{n}$	$\sigma_P = \sqrt{\frac{\pi(1-\pi)}{n}}$

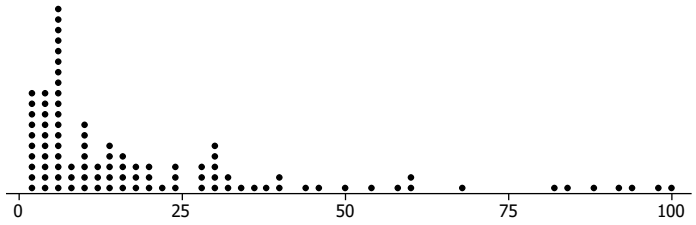
De stora talens lag

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

$$\bar{x} = 18.7$$

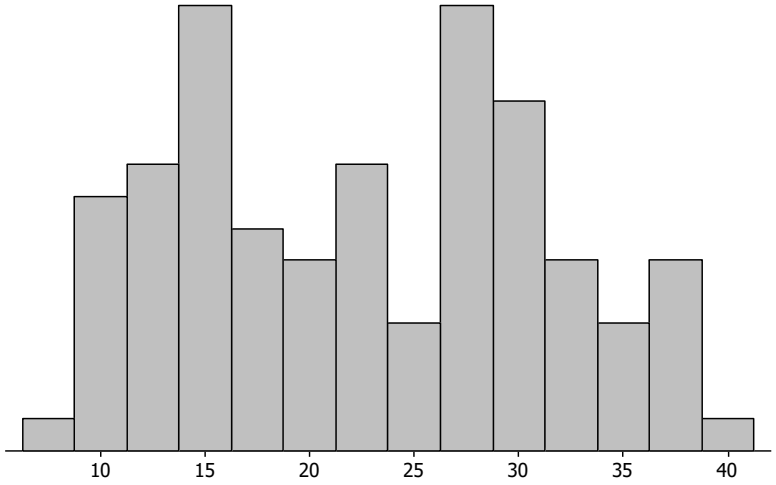
- Vi drar ett till stickprov

2	3	3	3	16	19	22	30	50	100

$$\bar{x} = 24.8$$

Samplingfördelning

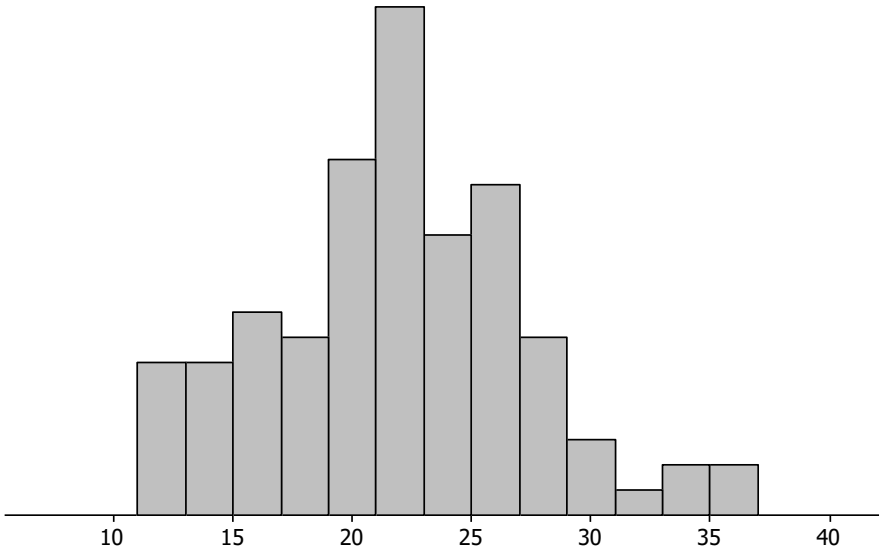
- Om vi drar 100 oberoende stickprov om storleken $n = 10$, beräknar de 100 stickprovsmedelvärdena och visualiserar mätningarna i ett histogram fås följande diagram



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

Samplingfördelning

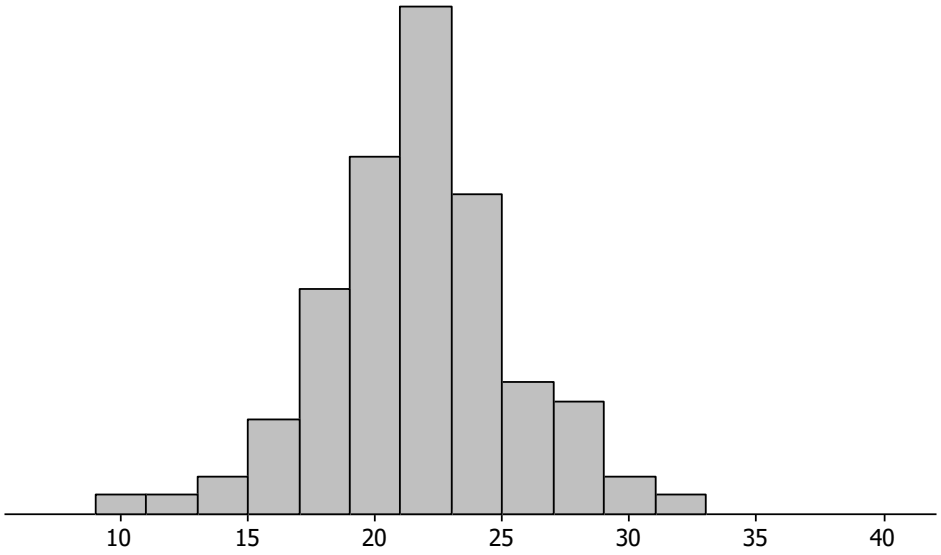
- Experimentet upprepas för 100 oberoende stickprov om storlek $n = 20$



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

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?