Övning 2

Lite repetition från föreläsning 6 gällande konfidensintervall, se föreläsningsfilerna för detaljer.

Läsinstruktion kapitel 11 i IMS (1)

- Läs till och med 11.1.1 om följande begrepp och beteckningar
  - **punktskattning** (point estimate) - baserat på ett urval (stickprov) från populationen har vi en skattning, exv. 166,5 cm för kvinnors medellängd, 38% väljartöd för S, eller (i bokexemplet) en skillnad i befordringsgrad på 29.2%
  - Den grekiska bokstaven μ används för att beteckna populationsmedelvärde, vilket skattas med stickprovsmedelvärdet $\bar{x}$
  - Bokstaven p används för att beteckna en andel (eng: proportion) i populationen, vilket skattas med stickprovsandelen $\hat{p}$ ("p-hatt")*

Punktskattning, variation, standardfel

- För en numerisk variabel - räkna ut standardavvikelsen i stickprovet "som vanligt" (från F3), och sätt in i formeln för standardfel:

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

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

- För en **andel** $\hat{p}$ kan man härleda att standardfelet är

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- där $\hat{p}$ är den skattade andelen

När kan vi använda normalfördelningsapproximationen?

- "Gäller om n är tillräckligt stort och utfallen är ändliga tal"
- **Observationerna ska vara oberoende**
- För **andelar** ska vi verifiera **success-failure condition** (kap 16.2.1):
  - Stickprovsfördelningen för $\hat{p}$, från ett urval av storlek n från en population med sann andel p, är ungefärligen normalfördelad när vi förväntar oss minst 10 lyckade och 10 misslyckade utfall:
  - $np \geq 10$ och $n(1-p) \geq 10$ (i praktiken: använd $\hat{p}$)
- För **numeriska variabler** används **n>30** typiskt som kriterium - vi kan då anta att medelvärdesfördelningen är ungefärligen normalfördelad
  - I statistik (speciellt i större kurser) är fokus mycket på att delaljkontrollera antaganden. "Plotta!".
  - Nyanser diskuteras i kap 19.2.2, exv. om vi har outliers
  - Om data i sig är normalfördelade kan lägre urvalsstorlek vara OK

Konfidensintervall, 95%

- Använd 68/**95**/99.7-regeln

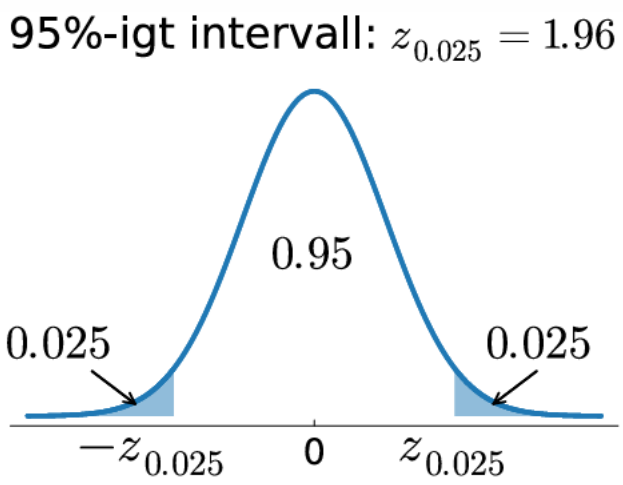
- Då kan vi ta fram ett (ungefärligt) **95%-igt konfidensintervall** för skattningen
  - **Punktskattning** $\pm$ **2**$\times$**standardfelet**, mao
  - $\bar{x} \pm 2\frac{s}{\sqrt{n}}$ - för skattning av medelvärde (med s enligt vanliga formeln)
  - $\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ - för skattning av andel
  - där tvåan kommer från att vi ska vara 2 standardfel "ut från" värdet på skattningen för att 95% av sannolikheten ska täckas (föreg. sida)

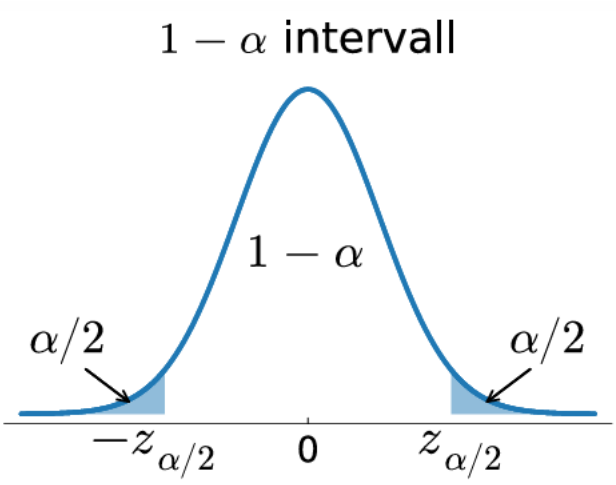
- Det exakta värdet i formeln är inte 2 utan 1.96:
  - $\bar{x} \pm 1.96\frac{s}{\sqrt{n}}$ - för skattning av medelvärde (med s enligt vanliga formeln)
  - $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ - för skattning av andel

Konfidensintervall och konfidensgraden α

- Vi kanske vill ha något annat, exv. ett 90%-igt konfidensintervall
- Allmänt gäller att ett $(1-\alpha)100\%$-igt konfidensintervall (för ett medelvärde) ges av:
  - $\bar{x} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ om σ är känd, annars
  - $\bar{x} \pm Z_{\alpha/2}\frac{s}{\sqrt{n}}$
- Analogt, för andelar, fås ett $(1-\alpha)100\%$-igt konfidensintervall av:
  - $\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$
- Z-värden finns i tabeller över den standardiserade normalfördelningen (och i R), för olika värden på konfidensgraden α

Illustration över hur vi går från vald konfidensnivå (exv. 95%), till det sannolikhetsvärde för vilket vi ska hitta ett Z-värde i standardnormalfördelningstabellen. Ex: 95% (0.95), 5% av observationerna (sannolikheten) ska ligga utanför intervallet, använd 0.025 eller 0.975 (fördelningen är symmetrisk) för att hitta (det absoluta) värdet 1.96 i tabellen – vilket i vår konfidensintervallberäkning innebär 1.96 standardfel ut i respektive riktning, från vår punktskattning.





1. Gå igenom exemplen i pdf-filen Kap13_normalfördelning, från föreläsning 6

2. Använd standardnormalfördelningstabellen (finns på Athena) för att ta fram

- hur stor andel av observationerna (sannolikheten) som ligger under z=-1, dvs mer än en standardavvikelse under medelvärdet

- hur stor andel av observationerna (sannolikheten) som ligger under z=1.5

- hur stor andel av observationerna (sannolikheten) som ligger över z=1.5. Gör uppgiften på två sätt. Vilken räkneregel från övning 1 kan du använda?

- hur stor andel av observationerna (sannolikheten) som ligger mellan z=-1.5 och z=1.5.

3. Kap 16, uppgift 19.

Konfidensintervall, andel

19. **Fireworks on July 4th.** A local news outlet reported that 56% of 600 randomly sampled Kansas residents planned to set off fireworks on July 4^{th}. Determine the margin of error for the 56% point estimate using a 95% confidence level using a mathematical model. (Survey USA 2012)

I tillägg:

- Bekräfta att relevanta antaganden om urvalsstorlek etc. är uppfyllda (success-failure condition)
- För samma uppgift, ta fram ett 92%-igt konfidensintervall.
- För samma uppgift, ta fram ett 99%-igt konfidensintervall.

4. Kap 19, uppgift 3.

3. **Heights of adults.** Researchers studying anthropometry collected body measurements, as well as age, weight, height and gender, for 507 physically active adults. Summary statistics for the distribution of heights (measured in centimeters, cm), along with a histogram, are provided below.[11] (Heinz et al. 2003)

   a. What are the point estimates for the average and median heights of active adults?

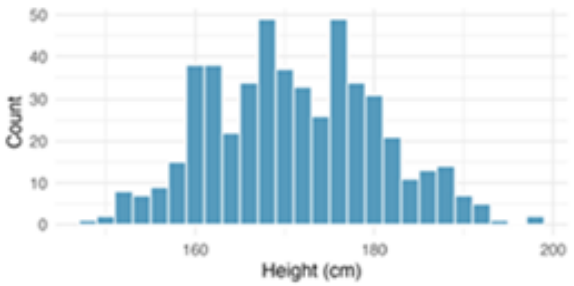
   b. What are the point estimates for the standard deviation and IQR of heights of active adults?

   c. Is a person who is 1m 80cm (180 cm) tall considered unusually tall? And is a person who is 1m 55cm (155cm) considered unusually short? Explain your reasoning.

   d. The researchers take another random sample of physically active adults. Would you expect the mean and the standard deviation of this new sample to be the ones given above? Explain your reasoning.

   e. The sample means obtained are point estimates for the mean height of all active individuals, if the sample of individuals is equivalent to a simple random sample. What measure do we use to quantify the variability of such an estimate? Compute this quantity using the data from the original sample under the condition that the data are a simple random sample.

Min	147.2
Q1	163.8
Median	170.3
Mean	171.1
Q3	177.8
Max	198.1
SD	9.4
IQR	14.0



I tillägg:

- Ta fram ett 95%-igt konfidensintervall för skattningen av populationsmedelvärdet
- Tolka konfidensintervall, generellt

5. Uppgift 11.3 (identifiera hypoteser) (själva bokkapitlet 11 ingår inte, förutom vad som står i läsanvisningen på föreläsning 6)

6. Genomgång av hypotestestexempel från föreläsning 7

7. Genomgång av hypotestestexempel från föreläsning 7 om testet skulle ha varit ensidigt

8. Uppgift 16.23