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

Se pdf-filen. Flera exempel syftar till att träna på att använda standardnormalfördelningstabellen (finns på Athena).

2. Använd standardnormalfördelningstabellen för att ta fram

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

*Sannolikheten är 0.15866. Vi är på (avrundat) 16e **percentilen** – 16% av Z-värdena är lägre.*

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

0.93319. Percentil 93.

- 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?

Metod 1: Komplementregeln för sannolikheter, från SDM kap 12 och övning 1. Vi vet att arean under kurvan = 1.

Mao:

Sannolikheten att en observation ligger under z=1.5 + Sannolikheten att en observation ligger över z=1.5 = 1

Vi får då:

Sannolikheten att en observation ligger över z=1.5 = 1 – 0.93319 = 0.06881

Metod 2: Använd att vi vet att fördelningen är symmetrisk kring noll. Mao:

Sannolikheten att en observation ligger över z=1.5 = Sannolikheten att en observation ligger under z = - 1.5

Läs av sannolikhetsvärdet vid z = - 1.5

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

Från sannolikheten till vänster om z=1.5 måste vi subtrahera sannolikheten till vänster om z=-1.5, dvs.vi ska ta fram:

Sannolikheten att en observation ligger under z=1.5 - Sannolikheten att en observation ligger under z = - 1.5

Vi får: 0.93319 - 0.06881 = 0.86638

Om du vill - kolla dina svar med pnorm()-kommandot i R (datorlabb 3). Fler exempel finns i boken kap. 13.2.

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.

Bokens svar (finns på s. 488):

19. With a random sample, independence is satisfied. The success-failure condition is also satisfied. $ME = z^\star\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96\sqrt{\frac{0.56\times 0.44}{600}} = 0.0397 \approx 4\%$.

*Vi är intresserade av **andelen** Kansasinvånare (vår population) som ska ha fyrverkerier (kalla denna **populationsandel** p). Vi har $\hat{p}$=0.56, en **punktskattning av andelen**, från vårt stickprov.*

*Vi har n = 600, och såväl $n\hat{p}$ som $n(1-\hat{p})$ är långt över 10, dvs "**success-failure condition**" är uppfyllt. Centrala gränsvärdessatsen ger att vi kan använda normalfördelningsapproximationen (F5, F6).*

*Ta fram **standardfelet** och **felmarginalen**, med formlerna från F6 och enligt facit ovan. **Notera att felmarginalen beror av vald konfidensnivå**, här, 95% konfidens (vi har 1-α = 0.95, α=0.05, $\frac{\alpha}{2}$ = 0.025), vilket ger Z-värdet 1.96.*

Standardfelet = $\sqrt{\frac{\hat{\boldsymbol{p}}(\mathbf{1}-\hat{\boldsymbol{p}})}{n}} = \sqrt{\frac{\mathbf{0.56(1-0.56)}}{600}} \approx 0.02065$

Konfidensintervallet blir $\hat{\boldsymbol{p}}$ ± felmarginalen = $\hat{\boldsymbol{p}}$ ± 1.96 × standardfelet = $\hat{\boldsymbol{p}}$ ± 1.96 × 0.02065 = 0.56 ± 0.0397, vilket ger intervallet (0.5203, 0.5997)

Om vi ska ha ett 92%-igt konfidensintervall (dvs 1-α = 0.92, α=0.08, $\frac{\alpha}{2}$ = 0.04) ska vi hitta de punkter (Z-värden) på standardnormalfördelningen som har 4% av sannolikheten längre ut i svansarna.

Använd standardnormalfördelningstabellen, vi får $Z_{\alpha/2} = Z_{0.04}$ = - 1.75 (och, motsvarande, +1.75 i högersvansen av fördelningen).

Konfidensintervallet blir $\hat{\boldsymbol{p}}$ ± 1.75 × standardfelet = 0.56 ± 1.75 × 0.02065, vilket ger konfidensintervallet (0.523, 0.596)

Om vi ska ha ett 99%-igt konfidensintervall (dvs 1-α = 0.99, α=0.01, $\frac{\alpha}{2}$ = 0.005) ska vi hitta de punkter (Z-värden) på standardnormalfördelningen som har 0.5% av sannolikheten längre ut i svansarna.

Använd standardnormalfördelningstabellen, vi får $Z_{\alpha/2} = Z_{0.005} = -2.58$ (och, motsvarande, +2.58 i högersvansen av fördelningen).

Konfidensintervallet blir $\hat{p} \pm 2.58 \times$ standardfelet = $0.56 \pm 2.58 \times 0.02065$, vilket ger konfidensintervallet (0.507, 0.613)

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?

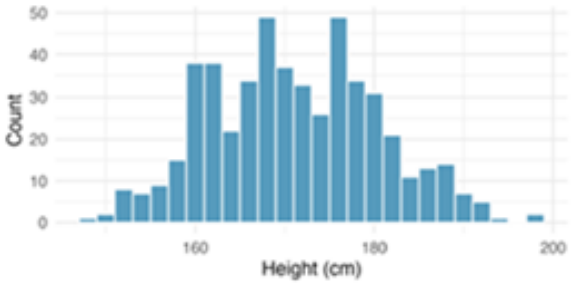
   b. What are the point estimates for the standard deviation and IQR of heights of active adults?

   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

Vi gjorde största delen av denna övning gemensamt på övningen. Här är bokens svar, från s. 491-492:

3. (a) Use the sample mean to estimate the population mean: 171.1. Likewise, use the sample median to estimate the population median: 170.3. (b) Use the sample standard deviation (9.4) and sample IQR ($177.8 - 163.8 = 14$). (c) $Z_{180} = 0.95$ and $Z_{155} = -1.71$. Neither of these observations is more than two standard deviations away from the mean, so neither would be considered unusual. (d) No, sample point estimates only estimate the population parameter, and they vary from one sample to another. Therefore we cannot expect to get the same mean and standard deviation with each random sample. (e) We use the standard error of the mean to measure the variability in means of random samples of same size taken from a population. The variability in the means of random samples is quantified by the standard error. Based on this sample, $SE_{\bar{x}} = \frac{9.4}{\sqrt{507}} = 0.417$.

I tillägg:

*(a, b). Vi vill veta något om **populationen** "active adults" och längd. Vi har inte data på alla i populationen. Vi har data på ett **urval** från populationen, dessa data använder vi för att ta fram **punktskattningar** av populationsvärdena (som är okända, bla. populationsmedel-längden som vi betecknar med grekiska bokstaven μ).*

(c). Frågan är lite vag eftersom "unusal" inte är 100% definierat men om vi går på bokens beskrivning i kapitel 5, kan vi resonera om värden som ligger mer än ett visst avstånd under Q1 eller ett visst avstånd över Q3. Vi skulle också kunna gå exv. 2 standardavvikelser ut från medelvärdet åt respektive håll, och betrakta observationer utanför ett sådant intervall som ovanliga.

(e). Vi får informationen att data har samlats in genom ett slumpmässigt urval från populationen. Vi kan då använda variationen i våra insamlade data, tillsammans med urvalsstorleken, för att ta fram ett mått på hur mycket punktskattningar av populationsmedellängden (μ) skulle variera i olika slumpmässiga urval. Vi har följande formel (F6, hopklipp av två bilder):

- Vi har sett att hur olika punktskattningar är fördelade kan approximeras med normalfördelningen
- **Standardfel** (standard error) är ett mått på variationen i skattningarna:

- **Standardfel** (standard error):

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

- där s är standardavvikelsen i stickprovet

*Standardfelet är alltså ett mått på **variationen i skattningen av medellängden**, om vi skulle dra många olika slumpmässiga urval (av samma storlek som vårt urval). Vi ser i facit ovan att standarfelet är strax över 0.4 cm, alltså långt mycket mindre än variationen i data i sig. Ju större urval vi har, desto mer precist skattar vi medellängden (vilket syns i formeln för standardfelet, där $\sqrt{n}$ finns i nämnaren).*

***Tillägg:** Konfidensintervall*

*Vi har **punktskattningen** $\bar{x}$ = 171.1 (medelvärdet i vårt urval) av **den okända populationsparametern** μ, och standardfelet enligt ovan. Vi ska beräkna ett 95%-igt konfidensintervall (se exv. F6):*

- 95%-igt konfidensintervall ($\alpha/2 = 0.025$) ges av:

$$\bar{x} \pm Z_{\alpha/2}\frac{s}{\sqrt{n}} = \bar{x} \pm Z_{0.025}\frac{s}{\sqrt{n}} = \bar{x} \pm 1.96\frac{s}{\sqrt{n}}$$

(och i vårt fall har vi redan standardfelet, $s/\sqrt{n}$, från ovan.)

Vi får det 95%-iga konfidensintervallet $\bar{x}$ ± 1.96 × standardfelet = 171.1 ± 1.96 × 0.417, vilket avrundat till två decimaler ger intervallet (170.28, 171.92) cm. Glöm inte enhet!

Tillägg: *Tolkning av konfidensintervall (se F6, s. 29)*

Om vi skulle ta många upprepade slumpmässiga urval av samma storlek, och räkna ut konfidensintervall på samma sätt, skulle ungefär 95% av dessa intervall täcka det sanna (och för oss okända) värdet på populationens medellängd (μ).

VI kan också säga att vi har 95% konfidens att vårt intervall, täcker populationsmedelvärdet (μ).

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

Se följande från F7:

- Nollhypotesen (H_0) är typiskt det skeptiska perspektivet, hypotesen att ingen skillnad från ett visst värde eller antagande existerar, exv. vi tror inte att mäns medelvilopuls förändrats från 60 slag/minut

  - *Hypotes om ingen förändring, inget samband, ett skeptiskt perspektiv, är vad som typiskt "kodas som" (utgör) nollhypotesen, medan*

  - *en förändring, ett samband, istället kodas som (utgör) alternativhypotesen.*

Facit finns på s. 485.

6. Genomgång av hypotestestexempel från föreläsning 7

Vi har gått igenom på föreläsning (exemplet med internetanvändning, s. 24-26).

Mer om tvåsidigt test (kommentar till s. 24-26):

Vårt påstående gäller att medelanvändandet i populationen inte är 3.5 h/dag. ***Det kan vara högre eller lägre.*** *Signifikansnivån vi väljer (i vårt fall 0.05) är den risk/sannolikhet vi accepterar att dra fel slutsats,* ***om nollhypotesen är sann****, dvs att vi med bas i vårt stickprov felaktigt drar slutsatsen att användandet inte skulle vara 3.5 h/dag. Eftersom vi ska ta hänsyn till att vi kan få ett högt, såväl som ett lågt värde på testvariabeln, ska vi hitta Z-värden som har sammanlagt 5% av sannolikheten i de båda svansarna (2.5% i respektive svans). Det kritiska värdet blir då ± 1.96 (F7, s. 25).*

7. Hypotestestexempel från föreläsning 7 om testet skulle ha varit ensidigt

Jämför med föreläsning 7, s. 24-26. Vi tänker oss nu istället följande påstående / hypotes, gällande en ökning:

- Antag mobilsurfande var 3.5h/dag/individ, fastslaget av mobiloperatörer
- Vi hävdar att värdet bör ha **ökat** och intervjuar 100 slumpvis utvalda individer, vars användande vi kan anta är oberoende

Vi har mao. hypoteserna:

$H_0 : \mu = 3.5$

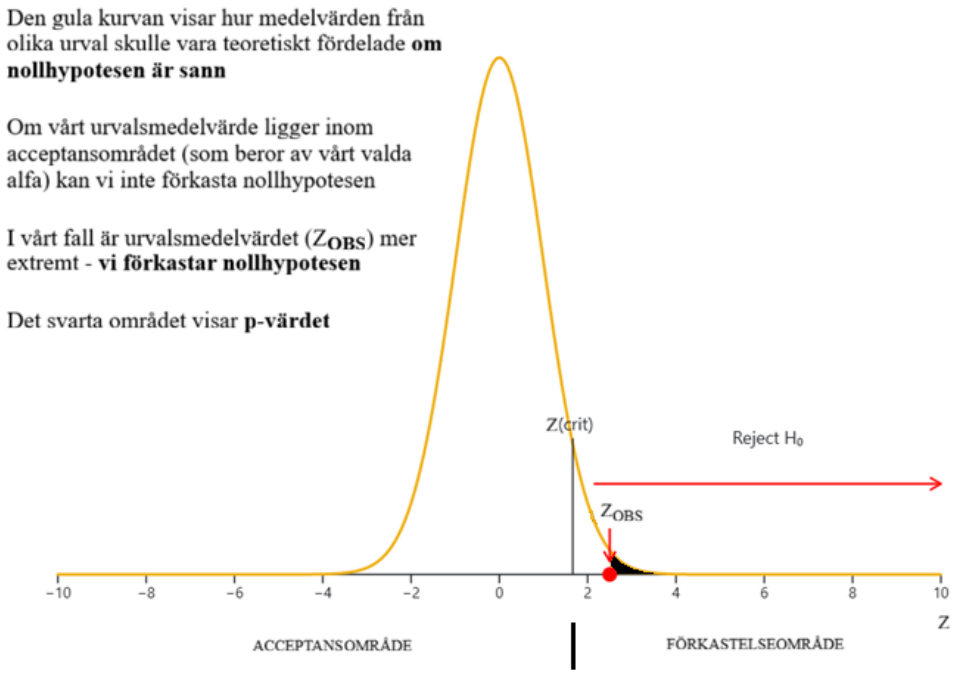
$H_A : \mu > 3.5$

Vi använder samma signifikansvärde, α=0.05

Mer om ensidigt test:

Vårt påstående gäller att användandet är **större än** ett visst värde, vi ska därför bara ta hänsyn till sannolikheten i högra svansen av standardnormalfördelningskurvan, när vi utvärderar påståendet. **Om nollhypotesen är sann** – dvs medelvärdet är 3.5 – är sannolikheten fem procent att observera ett Z-värde över +1.65 (från standardnormalfördelningstabellen). Det vill säga, sannolikheten för ett felbeslut, med detta värde som kritisk gräns, skulle vara fem procent. Vi accepterar denna risk. Om det observerade Z-värdet är + 1.65 eller högre förkastar vi därför nollhypotesen (att medelvärdet i populationen skulle vara 3.5) och tar istället det höga Z-värdet som evidens för alternativhypotesen.

Vi får följande bild:



Eftersom vårt observerade Z-värde är 2.5 (F7, s. 24) och det kritiska Z-värdet är 1.65 förkastar vi nollhypotesen och har istället funnit evidens/stöd för alternativhypotesen (vi säger dock aldrig att vi bevisat en hypotes, men vi har funnit evidens/stöd för alternativhypotesen).

8. Uppgift 16.23

23. **National Health Plan, mathematical inference.** A Kaiser Family Foundation poll for a random sample of US adults in 2019 found that 79% of Democrats, 55% of Independents, and 24% of Republicans supported a generic "National Health Plan". There were 347 Democrats, 298 Republicans, and 617 Independents surveyed. (Kaiser Family Foundation 2019)

    a. A political pundit on TV claims that a majority of Independents support a National Health Plan. Do these data provide strong evidence to support this type of statement? Your response should use a mathematical model.

    b. Would you expect a confidence interval for the proportion of Independents who oppose the public option plan to include 0.5? Explain.

I facit (s. 489) finns följande:

23. (a) We want to check for a majority (or minority), so we use the following hypotheses: H_0 : $p = 0.5$ and H_A : $p \neq 0.5$. We have a sample proportion of $\hat{p} = 0.55$ and a sample size of $n = 617$ independents. Since this is a random sample, independence is satisfied. The success-failure condition is also satisfied: 617×0.5 and $617 \times (1 - 0.5)$ are both at least 10 (we use the null proportion $p_0 = 0.5$ for this check in a one-proportion hypothesis test). Therefore, we can model $\hat{p}$ using a normal distribution with a standard error of $SE = \sqrt{\frac{p(1-p)}{n}} = 0.02$. (We use the null proportion $p_0 = 0.5$ to compute the standard error for a one-proportion hypothesis test.) Next, we compute the test statistic: $Z = \frac{0.55-0.5}{0.02} = 2.5$. This yields a one-tail area of 0.0062, and a p-value of $2 \times 0.0062 = 0.0124$. Because the p-value is smaller than 0.05, we reject the null hypothesis. We have strong evidence that the support is different from 0.5, and since the data provide a point estimate above 0.5, we have strong evidence to support this claim by the TV pundit. (b) No. Generally we expect a hypothesis test and a confidence interval to align, so we would expect the confidence interval to show a range of plausible values entirely above 0.5. However, if the confidence level is misaligned (e.g., a 99% confidence level and a $\alpha = 0.05$ discernibility level), then this is no longer generally true.

Först, gällande fråga (b):

- Fråga (b) handlar om hur konfidensintervall och hypotestest hänger ihop. Vi tänker oss samma konfidensnivå, säg 0.95 (α = 0.05). Om vi skulle göra ett 95%-igt konfidensintervall runt vårt skattade värde 0.55, skulle detta intervall inte innehålla 0.5 (testa gärna), vilket är ett annat sätt att uttrycka vad det tvåsidiga hypotestestet i lösningen ovan visar: att vi finner evidens för att stödet bland independents skiljer sig från 0.5.

  För att vi ska kunna jämföra på detta sätt måste vi arbeta med samma värde på α, vilket är vad som uttrycks i sista meningen i lösningen ovan.

Tillägg – vi tänker oss istället att testet är ensidigt, och illustrerar med en bild.

- Problemställningen gäller populationsandelen (p) av "Independents" som stöder ett förslag på hälsoområdet. Boken har valt att hantera hypotestestet som tvåsidigt (intresset är att studera om andelen skiljer sig från $p_0 = 0.5$) medan vi här istället väljer som alternativhypotes att andelen är > 0.5, dvs ett ensidigt test. Vi har mao:

$$H_0 : p = p_0$$
$$H_A : p > p_0$$

- Från ett slumpvis urval av 617 individer ur gruppen får vi en skattning av populationsandelen: $\hat{p}$=0.55

- Success-failure condition (F6, s. 20) kollas (se lösningen ovan).

- Standardfelet (standard error, SE) beräknas, värdet blir 0.02. (Här skulle författarna kunna ha valt att använda det icke avrundade värdet, men väljer att avrunda 0.02002834 direkt till 0.02).

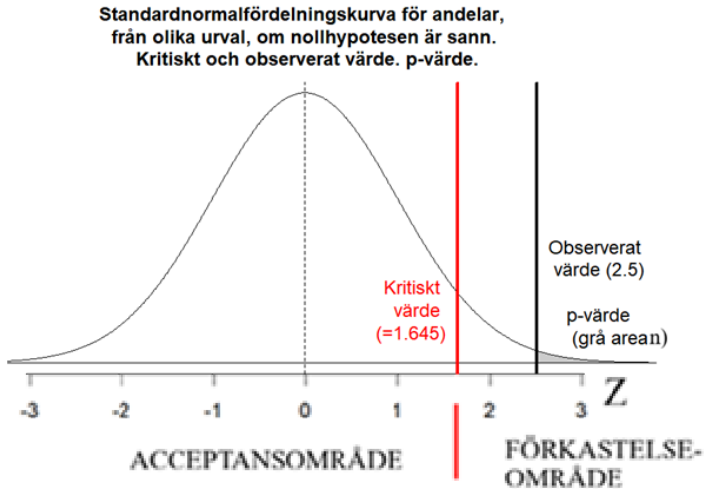
- Testvariabeln (F7, s.27):

  - **Ta fram värdet på testvariabeln:**

$$Z_{\text{obs}} = \frac{\hat{p} - p_0}{\sqrt{\hat{p}(1-\hat{p})/n}}$$

- Vi beräknar testvariabeln **under antagandet att nollhypotesen är sann**, dvs. att populationsandelen är p_0=0.5. Värdet blir (avrundat) 2.5.

- Vi kan rita en bild (med papper och penna, eller med R om vi vill (labb 3)):



- Vi observerar ett mer extremt Z-värde än det kritiska värdet och förkastar därför nollhypotesen, och finner istället evidens för alternativhypotesen (att stödet bland independents är > 0.5).

  - **p-värde:** Sannolikheten att observera ett värde som är minst lika extremt som det observerade, givet att nollhypotesen är sann

- p-värdet ges av sannolikheten att observera ett Z-värde på 2.5 eller högre, vi får värdet från standardnormalfördelningstabellen (övningsuppgift 1 ovan), vi får (avrundat till två decimaler): 0.62%.