

R-anvisningar till *Grundläggande statistik*

2024-10-24

Innehåll

Introduktion	5
Installation av R och RStudio	7
0.1 Inledning	7
0.2 Installation av R	7
0.3 Installation av RStudio	7
0.4 Gränssnittet i RStudio	8
0.5 Paket i R	8
1 Datahantering och grafer	9
1.1 Uppstart och orientering	9
1.2 <i>Packages</i> från CRAN	10
1.3 Objekt och funktioner	11
1.4 Sekvenser av funktioner	13
1.5 Datainskrivning	14
1.6 Urval ur en tabell med select och filter	16
1.7 Grafer med ggplot2	17
1.8 AI-uppgift I	19
2 Beskrivande statistik	21
2.1 Repetition från datorövning 1	21
2.2 Import av data från en Excelfil	22
2.3 Ändra och skapa nya kolumner med mutate	23
2.4 Sammanfattande lägesmått	24
2.5 Sammanfattande spridningsmått	25
2.6 Ordna upp beskrivande statistik och exportera	27
2.7 Kumulativt medelvärde	28
2.8 Tredimensionella grafer med plotly	29
2.9 AI-uppgift II	30
3 Ett stickprov av normalfördelad data	33
3.1 Repetition från datorövning 2	33
3.2 Test av medelvärde för normalfördelad data	34

3.3	Konfidensintervall för normalfördelad data	37
3.4	Normalfördelad data	38
3.5	Centrala gränsvärdesatsen	39
3.6	AI-uppgift III	40
3.7	Bonus. Simuleringar för t-test	41
4	Ett stickprov av icke-normalfördelad data	43
4.1	Repetition av datorövning 3	43
4.2	Proportioner från binär data	44
4.3	Konfidensintervall för proportioner	47
4.4	Chi-två-test för goodness-of-fit	48
4.5	Bonus. Interaktiva kartor med leaflet	50
5	Test för två stickprov	53
5.1	Repetition av datorövning 4	53
5.2	Två stickprov och normalfördelad data	54
5.3	z-test och konfidensintervall för två proportioner	57
5.4	Chi-två-test för korstabeller	58
6	Variationsanalys	61
6.1	Repetition av datorövning 5	61
6.2	Allmänt	63
6.3	Variationsanalys. En faktor	64
6.4	Variationsanalys. En faktor med block	66
6.5	Bonus. Två faktorer med block	68
7	Regression och korrelation	71
7.1	Repetition av datorövning 6	71
7.2	Regression	72
7.3	Korrelation	75

Introduktion

Detta dokument ger en introduktion till R för en kurs i grundläggande statistik.

Installation av R och RStudio

0.1 Inledning

För att köra R-kod på sin dator krävs en installation av programspråket R. För att effektivt arbeta i R används ofta en utvecklingsmiljö (ett tilläggsprogram som på flera sätt förenklar arbetet) och här ges anvisningar till RStudio - den vanligaste utvecklingsmiljön för R. För att komma ingång måste man alltså installera R och RStudio.

0.2 Installation av R

Programspråket R kan laddas ner från <https://www.r-project.org/> med följande steg:

1. Klicka på *CRAN* längst upp till vänster.
2. Klicka på den översta länken under 0-Cloud.
3. Välj en nedladdning beroende på operativsystem.
4. För Windows, välj *base*. För macOS, välj den senaste tillgängliga versionen.
5. Installera R från den nedladdade filen. Installation sker som för andra nedladdade program.

0.3 Installation av RStudio

RStudio kan laddas ner från <https://www.posit.co/> med följande steg:

1. Klicka på *Download RStudio* uppe till höger.
2. Scrolla nedåt och välj *Download* under *Install RStudio*.
3. Installera RStudio från den nedladdade filen. Installation sker som för andra nedladdade program.

0.4 Gränssnittet i RStudio

När man nu öppnar RStudio ser man att fönstret är uppdelat i fyra delar och att varje del består av en eller flera flikar. De viktigaste är i nuläget

- *Console* där kod körs och resultat skrivs ut,
- *Environment* där man ser skapade objekt,
- *History* där man ser tidigare körd kod,
- *Plots* där man ser skapade grafer, och
- *Help* där man ser hjälpsidor för funktioner.

Ofta skriver man inte sin kod direkt i konsollen, utan i ett separat *skript* - en vanlig textfil som innehåller den kod man vill köra. Genom att organisera sin kod i ett skript kan man lätt strukturera och dokumentera sitt arbete. I RStudio kan man öppna ett nytt skript genom att gå till *File > New File > R Script* eller genom att klicka *Ctrl + Shift + N*. Ett tomt skript öppnar sig då i det övre vänstra delfönstret. Om man skriver

```
a <- 5
```

i skriptet och trycker *Ctrl + Enter* bör man se att koden i skriptet körs i konsollen. Om man tittar i fliken *Environment* ska man också se att det nu skapats ett objekt *a*.

0.5 Paket i R

En av de stora styrkorna med R är att språket kan byggas ut av dess användare. De här tilläggen kan sedan samlas i paket (*packages*) och delas med andra. Rs officiella bibliotek för paket kallas för *CRAN* (*Comprehensive R Archive Network*) och består av drygt 20 000 uppladdade paket som innehåller allt från fritt tillgänglig data till avancerade statistiska modeller.

För att använda ett specifikt paket måste det först installeras. Om man vet namnet på paketet man vill installera kan man köra

```
install.packages("tidyverse")
```

I det här fallet installeras paketet **tidyverse**, vilket innehåller funktioner för hantering av data.

I RStudio kan man också installera paket från *Packages*-fliken.

Paket måste också laddas för varje ny session. Innan man kan använda innehållet i ett paket måste man därför köra

```
library(tidyverse)
```


Kapitel 1

Datahantering och grafer

Datorövning 1 handlar om grunderna till R. Efter övningen ska vi kunna

- Starta RStudio och orientera oss i gränssnittet,
- Installera och ladda tilläggspaket (*Packages*),
- Definera objekt och tillämpa funktioner i R,
- Transformera en tabell med data genom att välja kolumner och filtrera rader,
- Skapa grafer med `ggplot2`.

1.1 Uppstart och orientering

För att arbeta i R måste vi installera språket R och ett gränssnitt för att arbeta i R, vanligen *RStudio*. Både R och RStudio kan laddas ner gratis, från <https://www.r-project.org/> respektive <https://posit.co/>.

Starta RStudio, till exempel genom att gå till Startmenyn och söka på RStudio eller genom att dubbelklicka på en fil som öppnas i RStudio. Gränssnittet i RStudio är uppdelat i fyra delar och varje del består av en eller flera flikar. De viktigaste är i nuläget

- *Console* där kod körs och resultat skrivs ut,
- *Environment* där man ser skapade objekt,
- *History* där man ser tidigare körd kod,
- *Plots* där man ser skapade grafer, och
- *Help* där man ser hjälpsidor för funktioner.

Uppgift 1.1 (Help-fliken). Hitta fliken *Help*, klicka på husikonen under fliken. Finns det en länk med *RStudio Cheat Sheets*? Följ den länken för att hitta guider till R som kan bli nyttiga längre fram. För nu, gå tillbaka till RStudio.

Ofta skriver man inte sin kod direkt i konsollen, utan i ett separat *skript* - en

vanlig textfil som innehåller den kod man vill köra. I RStudio kan man öppna ett nytt skript genom att gå till *File > New File > R Script* eller genom att klicka *Ctrl + Shift + N*.

Uppgift 1.2 (Ett första skript). Öppna ett nytt skript genom File-menyn eller genom *Ctrl + Shift + N*. Skriv

```
a <- 5
```

i skriptet och tryck *Ctrl + Enter*. Titta i flikarna *Console* och *Environment*. Har något hänt? Du bör se att koden i skriptet körts i konsollen och att ett nytt objekt *a* ligger i *Environment*.

1.2 Packages från CRAN

En av de stora styrkorna med R är att språket kan byggas ut av dess användare. De här tilläggen kan sedan samlas i paket (*packages*) och delas med andra. Rs officiella bibliotek för paket kallas för *CRAN* (*Comprehensive R Archive Network*).

För att använda ett specifikt paket måste det först installeras. Om man vet namnet på paketet man vill installera kan man köra

```
install.packages("tidyverse")
```

I det här fallet installeras paketet *tidyverse*, vilket innehåller funktioner för hantering av data.

I RStudio kan man också installera paket från *Packages*-fliken.

Uppgift 1.3 (Installera tidyverse-paketet). Kör raden ovan för att installera *tidyverse*. Du kan antingen köra raden genom att skriva den i *Console* eller genom att skriva i ett skript och köra därifrån genom *Ctrl + Enter*.

Uppgift 1.4 (Installera gapminder-paketet). Paketet *gapminder* innehåller lite intressant data vi kommer använda senare. Installera paketet *gapminder* genom att fylla i den saknade biten och köra raden nedan.

```
install.packages("___")
```

Paket måste också laddas för varje ny session. Innan man kan använda innehållet i ett paket måste man därför köra

```
library(tidyverse)
```

Uppgift 1.5 (Ladda gapminder-paketet). Ladda paketet *gapminder* genom att fylla i och köra raden nedan.

```
library(___)
```

Uppgift 1.6 (Paket som inte finns). Vad händer om man försöker installera ett paket som inte finns på *CRAN*? Testa till exempel

```
install.packages("ThisIsNotTheNameOfAnyPackage")
```

och

```
library(ThisIsNotTheNameOfAnyPackage)
```

1.3 Objekt och funktioner

Ett *objekt* i R är en namngiven informationsmängd. Objekt kan se ut på många olika sätt - under kursens gång används objekt som består av insamlad data (konstruerade som vektorer eller tabeller), objekt som är statistiska modeller, och flera andra former. I R skapar man objekt med *assign*-pilen `<-` (mindre än och bindestreck).

I ett tidigare exempel fanns raden

```
a <- 5
```

Här skapas ett objekt med namnet **a** som innehåller informationen 5. *Assign*-pilen pekar alltså på det namn man vill ge objektet och pekar från objektets innehåll.

Ett lite mer komplicerat exempel på ett objekt ges av

```
b <- c(3, 1, 4, 1, 5, 9)
```

Här skapas ett objekt **b** som innehåller en *serie* numeriska värden (en *vektor*). Värdena i en vektor är ordnade och man kan plocka ut ett specifikt värde med hakparenteser.

```
b[3]           # Det tredje värdet i vektorn b
```

```
## [1] 4
```

```
b[c(3,5)]      # Det tredje och femte värdet i b
```

```
## [1] 4 5
```

Uppgift 1.7 (Skapa en vektor). Skapa ett objekt med namnet **new_vector** som innehåller värden 5, 7 och 10 genom att fylla i följande rad.

```
new_vector <- c(, , )
```

Uppgift 1.8 (Ta ut andra värdet). Använd hakparenteser för att plocka ut det andra värdet ur vektorn **new_vector**.

Objekt kan manipuleras genom att tillämpa *funktioner*. En funktion tar någon ingående data och ger något utgående resultat. Funktioner anges genom

att skriva funktionens namn följt av ingående data inom parenteser, och resultatet kan antingen skrivas ut i konsollen eller sparas som ett nytt objekt. En grundinstallation av R innehåller en mängd färdiga funktioner, t.ex.

```
sum(b)
```

```
## [1] 23
```

vilket ger summan av värdena i vektorn `b`,

```
plot(b)
```

som ger en simpel graf, och

```
sqrt(b)
```

```
## [1] 1.732051 1.000000 2.000000 1.000000 2.236068 3.000000
```

som beräknar kvadratroten för varje element i vektorn.

Uppgift 1.9 (Summera vektorn). Fyll i och kör följande rad för att beräkna summan av vektorn `new_vector`

```
sum(____)
```

Vid konstruktionen av vektorn användes också en grundläggande funktion - funktionen `c()`, som tar en serie värden och skapar en sammanhängande vektor av värden.

Alla R-funktioner har en tillhörande hjälpfil som kan plockas fram genom att skriva frågetecken följt av funktionsnamnet, t.ex. `?sum`. Från hjälpfilen får man att `sum()` tar numeriska vektorer som ingående värde och beräknar summan. Man kan styra funktionens beteende genom att sätta ett argument `na.rm` (vilket här styr hur funktionen hanterar saknade värden). Som illustration kan man titta på

```
b <- c(3, 1, 4, 1, 5, 9, NA) # Läger till ett saknat värde
sum(b)                     # na.rm = FALSE är grundinställning
```

```
## [1] NA
```

```
sum(b, na.rm = TRUE)      # na.rm sätts till TRUE
```

```
## [1] 23
```

Det första försöket `sum(b)` ger utfallet `NA`, men om man sätter `na.rm = TRUE` beräknas summan efter att det saknade värdet plockats bort. Notera också att skript kan kommenteras med `#`.

1.4 Sekvenser av funktioner

Ofta vill man genomföra flera operationer på ett objekt. Man behöver då genomföra en sekvens av funktioner. Säg till exempel att man har värdena

$$(-4, -2, -1, 1, 2, 4)$$

och vill ta absolutvärde (vilket gör negativa tal till motsvarande positiva tal) och sedan summera. Den typen av sekvenser kan genomföras på ett par olika sätt. Ett första sätt är att spara resultatet i varje steg och sedan använda utfallet i nästa steg:

```
c <- c(-4, -2, -1, 1, 2, 4)  # Skapa en vektor av värden
c_absolute <- abs(c)         # Ta absolutvärden. Skapa c_absolut
sum(c_absolute)             # Summera värden i c_absolut
```

```
## [1] 14
```

Här skapas ett objekt `c` som innehåller en vektor där några tal är negativa. I nästa rad används `abs` för att skapa absolutvärden. Slutligen summeras absolutvärdena med `sum`. Notera att det är möjligt att skapa ett objekt med namnet `c` trots att det redan är namnet på en funktion - R förstår ur sammanhanget om objektet eller funktionen ska användas. Det kan dock bli lite oklart för en läsare, så försök som regel att undvika att skapa objekt med vanliga funktionsnamn som `sum` och `mean`.

Uppgift 1.10 (Kvadrat, summa och roten ur). Fyll i och kör följande rader för att ta varje värde i `new_vector` i kvadrat, sedan summera, och sedan ta roten ur.

```
new_vector_squared <- new_vector^2 # Ta kvadraten av varje värde
new_vector_squared_sum <- sum(____) # Summera kvadraterna
sqrt(____)          # Ta kvadratroten ur summan
```

Ett alternativ till att spara utfallet i varje steg är att skriva en senare funktion runt en tidigare funktion. Det fungerar för att R utvärderar funktioner inifrån-ut. Med samma exempel som tidigare får man

```
sum(abs(c(-4, -2, -1, 1, 2, 4)))
# Ta summan av absolutvärden av vektorn
```

medan beräkningen i övningen blir

```
sqrt(sum(new_vector^2)) # Ta roten ur summan av vektorn i kvadrat
```

Den här typen av skrivning kan spara plats men blir snabbt svårläst.

Ett sista alternativ är att använda en så kallad *pipe* (namnet kommer från att en sekvens funktioner kallas en *pipeline*). En pipe skrivs `%>%` och kan i RStudio tas fram med kortkommandon *Ctrl + Shift + M*. Pipen tar utfallet av en funktion till vänster och sänder till en funktion till höger. Den kan utläsas i dagligt tal som *och sen*. Med samma exempel som tidigare kan vi skriva

```
library(tidyverse)

c(-4, -2, -1, 1, 2, 4) %>%      # Skapa en datamängd och sen
  abs() %>%                    # ta absolutvärden, och sen
  sum()                        # beräkna summan.
```

```
## [1] 14
```

Uppgift 1.11 (Kvadrat, summa och rot med pipe). Fyll i de saknade funktionerna och kör följande rader för att ta varje värde i `new_vector` i kvadrat, sedan summera, och sedan ta roten ur, denna gång genom att länka funktionerna med en pipe `%>%`.

```
new_vector^2 %>% # Ta kvadraterna av new_vector, och sen
  ___() %>%      # beräkna summan, och sen
  ___()          # Ta kvadratroten med sqrt()
```

1.5 Datainskrivning

Det första praktiska steget i en statistisk analys är att importera data. I R kan det göras genom att direkt skriva in sin data och spara som ett nytt objekt, men ett bättre och vanligare sätt är att importera sin data från en extern fil eller databas.

I ett tidigare exempel användes funktionen `c()` för att skapa en vektor av data. Ofta ordnas flera vektorer i en tabell där varje kolumn är en vektor och varje rad en observation av någon enhet. En datatabell (en `data.frame` i R) skapas genom funktionen `data.frame()` följt av namngivna vektorer. Exempeldata kan skrivas in genom följande.

```
dat <- data.frame(Vecka = c(7,7,7,7,7,7,
                           11,11,11,11,11,11),
                  Behandling = c("A","A","A","B","B","B",
                                  "A","A","A","B","B","B"),
                  Vikt = c(232,161,148,368,218,257,
                          1633,2213,972,2560,2430,855),
                  N = c(2.63,2.90,2.99,3.54,3.30,2.85,
                       1.53,1.90,NA,2.58,NA,NA))

dat
```

Radbrytningar och blanksteg är oviktiga i R, och används bara för läsbarhet här. Saknade värden skrivs in som `NA` för *not available*. Notera att alla kolumner inte behöver vara av samma datatyp men att värden inom en kolumn måste vara det. Här är *Behandling* text medan övriga kolumner är numeriska.

Uppgift 1.12 (Alea iacta est). Kasta din tärning tio gånger och skriv in resultatet i en datatabell i R med hjälp av grundkoden nedan. Om du saknar en

tärning, fråga lämplig person om du kan få en. Behåll tärningen, den behövs till nästa datorövning (och närhelst man står inför ett avgörande livsbeslut).

```
dat_dice <- data.frame(Kast = c(1,2,3,4,5,6,7,8,9,10),
                      Utfall = c(,_,_,_,_,_,_,_,_,_))
dat_dice
```

Objektet `dat` är av typen `data.frame` - en tabell med rader och kolumner. Man kan ange en specifik kolumn i en `data.frame` med dollartecken följt av kolumnens namn.

```
dat$Vikt
```

```
## [1] 232 161 148 368 218 257 1633 2213 972 2560 2430 855
```

Man kan också plocka ut rader och kolumner med hakparenteser och ordningstal.

```
dat[2,3]      # Andra raden och tredje kolumnen
dat[2, ]      # Tomt värde ger alla värden.
dat[, 3]      # Alla värden i kolumn 3
```

Uppgift 1.13 (Plocka ut en specifik kolumn). I den tidigare övningen skapade du ett objekt `dat_dice`. Använd dollartecken för att plocka ut kolumnen *Utfall* från det objektet.

```
dat_dice$___
```

Genom att plocka ut en kolumn från en `data.frame` kan man beräkna vanlig beskrivande statistik med funktioner som `mean()` (medelvärde) och `sd()` (standardavvikelsen).

```
mean(dat$Vikt)
```

```
## [1] 1003.917
```

```
sd(dat$Vikt)
```

```
## [1] 951.3067
```

Funktionen `plot()` ger en enkel graf.

```
plot(dat$Vecka, dat$Vikt)
```

Uppgift 1.14 (Tärningsgraf). Använd datan i objektet `dat_dice` och skapa ett diagram med kolumnen *kast* på x-axeln och kolumnen *Utfall* på y-axeln.

```
plot(dat_dice$___, dat_dice$___)
```

1.6 Urval ur en tabell med `select` och `filter`

En vanlig operation på en tabell är att göra ett urval - antingen ett urval av rader (t.ex. ett visst land), vilket kallas *filtrering* eller ett urval av variabler (t.ex. land och befolkning), vilket kallas *selektion*. Här tittar vi på hur det kan göras med funktionerna `filter()` och `select()` från paketet `tidyverse`. Vi använder `gapminder`-datan som kan laddas med `library(gapminder)`.

För att filtrera på ett givet land kan man använda pipe-funktionen från datan till en filter-funktion, t.ex.

```
gapminder %>%
  filter(country == "Sweden")      # Ta gapminder-datan och sen
                                   # filtrera för ett land
```

Inom filterfunktionen anges ett logisk villkor `country == Sweden` och utfallet är de rader där villkoret är sant. Notera de dubbla likhetstecknen - de måste användas för ett logisk villkor eftersom enkelt likhetstecken används för att skapa objekt och sätta funktionsargument.

Uppgift 1.15 (Filtrera för land). Vad måste ändras i koden för att istället plocka ut rader där landet är Finland? Hur många rader har det urvalet i datan?

```
gapminder %>%
  filter(country == "Sweden")      # Ta gapminder-datan och sen
                                   # filtrera för ett land
```

Om man vill välja flera länder kan man använda funktionen `%in%` på ett liknande sätt.

```
gapminder %>%
  filter(country %in% c("Sweden", "Finland"))
```

och om man vill ha mer än ett villkor kan man rada dem i filter-funktionen:

```
gapminder %>%
  filter(country %in% c("Sweden", "Finland"), # Ta datan, och sen
         year == 2002)                       # filtrera för land
                                             # och för år
```

För att se fler eller färre rader kan man använda en pipe `%>%` till funktionen `print()`. Följande skriver ut fem rader

```
gapminder %>%
  filter(country %in% c("Sweden", "Finland")) %>%
  print(n = 5)
```

Om man istället vill göra ett urval av kolumner kan man använda `select()`. Som argument anges de kolumner man vill välja, t.ex.

```
gapminder %>%
  select(country, lifeExp)          # Ta datan, och sen
                                   # välj kolumner
```

Uppgift 1.16 (Befolkade länder). Funktionen `arrange()` sorterar data efter

en angiven kolumn. Följande stycke ger oss Europas länder efter befolkning.

```
gapminder %>%
  filter(continent == "Europe", year == 1962) %>%
  select(country, lifeExp, pop) %>%
  arrange(-pop)
```

Gör lämpliga ändringar för att få Asiens länder från 2002 ordnade efter förväntad medellivslängd.

1.7 Grafer med ggplot2

Vi kan nu börja titta på grafer. R har en mängd grundläggande funktioner för grafer. Vi såg tidigare ett exempel på funktionen `plot()`.

```
gm_2002 <- gapminder %>% filter(year == 2002)
```

```
plot(gm_2002$lifeExp, gm_2002$gdpPercap)
```

Ett object `gm_2002` skapas efter att ha filtrerat för året 2002. Tecknet `$` används för att välja kolumnerna `lifeExp` och `gdpPercap` ur objektet `gm_2002`.

För mer avancerade grafer används dock ofta funktioner ur Rs paketbibliotek. Här illustreras det mest populära - `ggplot2`. I `ggplot2` byggs grafer upp med tre grundläggande byggstenar:

- *data*, informationen man vill visualisera,
- *aesthetics*, en koppling mellan data och visuella element såsom grafens axlar, objekts storlek och färg,
- *geometries*, de geometriska former som visas i grafen.

En graf skrivs med en startfunktion `ggplot` som anger namnet på datan och grafens *aesthetics*, och därefter sätts geometriska element genom funktioner som börjar med `geom_`. Ett spridningsdiagram kan t.ex. skapas med `geom_point`.

```
ggplot(gm_2002, aes(x = lifeExp, y = gdpPercap)) +
  geom_point()
```

Grafen kan byggas ut genom att sätta *aesthetics* för färg och storlek. Man kan också dela en graf i småfönster med `facet_wrap` och styra grafens utseende genom att sätta ett tema såsom `theme_bw`.

```
ggplot(gm_2002, aes(x = lifeExp, y = gdpPercap,
                    color = continent, size = pop)) +
  geom_point() +
  facet_wrap(~ continent)
```

Uppgift 1.17 (Livslängd över tid). Vad ska ändras i stycket nedan för att skapa en graf med år på x-axeln, medellivslängd på y-axeln och skilda småfönster för olika kontinenter?

```
ggplot(gm_2002, aes(x = _____, y = _____)) +
  geom_point() +
  facet_wrap(~ continent)
```

Andra graftyper kan skapas med andra `geom_`-funktioner. Stapeldiagram ges av `geom_col` (`col` för *column*). Man kan också använda `geom_bar` om man bara vill räkna antal rader per någon kategori. Följande väljer ut en kontinent och ått och plottar ländernas befolkning stapeldiagram.

```
dat_small <- gapminder %>%
  filter(continent == "Europe", year == "2002")

ggplot(dat_small, aes(pop, country, fill = gdpPercap)) +
  geom_col(color = "black")
```

Argumentet `fill` styr färgen för ytor (här staplarnas ytor) medan `color` i `geom_col()` styr kanten runt varje stapel.

Man kan styra grafiken i en `ggplot` genom funktionen `theme()`. Det är ett ganska komplicerat ämne, men låt oss titta på några grunder. Vi börjar med att skapa en enkel graf: en boxplot över medellivslängd per kontinent.

```
dat_small <- gapminder %>%
  filter(year == 2002)

ggplot(dat_small, aes(x = lifeExp, y = continent)) +
  geom_boxplot()
```

Vi kan ändra utseendet på grafen genom argument inom geometrier och med funktionen `theme()`. I `theme()` sätter man de specifika egenskaper man vill ändra genom att tillskriva dem ett *element*. Valet av element beror på typen av grafiskt objekt - text sätts t.ex. med `element_text()` och ytor med `element_rect()` (för *rectangle*). Vi ger ett exempel med ändrad bakgrund, rut-mönster, och teckenstorlek.

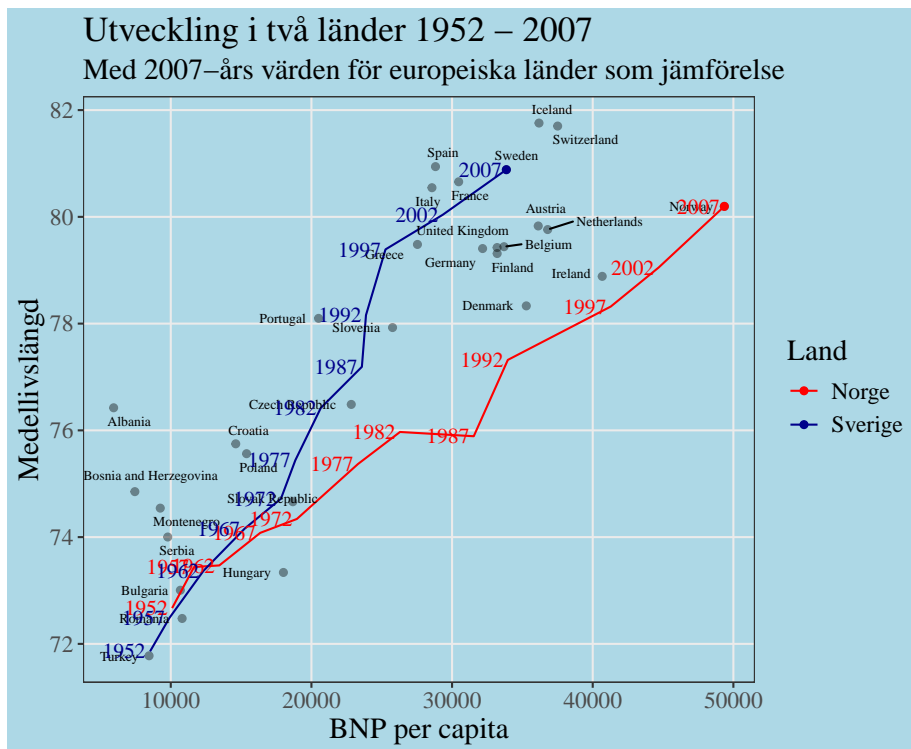
```
ggplot(dat_small, aes(lifeExp, continent)) +
  geom_boxplot(fill = "lightblue") +
  theme(panel.background = element_rect(fill = "red3"),
        text = element_text(size = 15,
                              color = "white", family = "serif"),
        axis.text = element_text(color = "white"),
        plot.background = element_rect(fill = "grey30",
                                         color = "black"),
        panel.grid.major.y = element_blank())
```

Uppgift 1.18 (Temaval 1). Ändra färgvalen i grafen ovan för att skapa snyggast möjliga graf. Funktionen `colors()` ger de färger som finns tillgängliga i R. Man kan också använda hex-koden för färger, t.ex. `fill = "#ffdd00"`.

Uppgift 1.19 (Temaval 2). Ändra färgvalen i grafen ovan för att skapa fulast möjliga graf. Visa de två graferna för någon annan och se om de kan säga vilken som är vilken.

1.8 AI-uppgift I

Nedan har vi en graf gjord med `ggplot2` med flera olika *geom* och ändringar från standardtemat.



Uppgift 1.20 (Dekonstruera grafen). Vilka geografiska former (punkter, linjer ytor) visas i grafen och hur är de kopplade till datans variabler? Vilka ändringar har gjorts från standardtemat (bakgrundsfärg, typsnitt, teckenstorlek)?

Uppgift 1.21 (Rekonstrera grafen). Använd en chatbot (såsom Microsofts copilot eller OpenAIs ChatGPT) för att rekonstrera grafen. Börja med att beskriva grafen i ord och be om R-kod. Klipp ut koden och kör i R. Läs koden du fått. Finns det några delar som är svåra att förstå. Med vissa chatverktyg kan man också klippa in bilden och be om den bakomliggande koden.

Uppgift 1.22 (Steg i Excel). Hur skulle man gå tillväga för att skapa samma graf i Excel? Be en chatbot om en stegvis instruktion.

Kapitel 2

Beskrivande statistik

Datorövning 2 handlar om beräkning av beskrivande statistik i R. Efter övningen ska vi kunna

- Importera data från en excelfil,
- Beräkna lämpliga lägesmått för en variabel,
- Beräkna lämpliga spridningsmått för en variabel,
- Konstruera grafer som jämför två eller flera grupperns läge och spridning.

2.1 Repetition från datorövning 1

När man arbetar i R är det klokt att använda ett avancerat gränssnitt som RStudio och att skriva sin kod i ett separat skript. I RStudio kan man starta ett nytt skript genom *Ctrl + Shift + N*.

Mycket funktionalitet i R ligger i tilläggspaket (*Packages*). Paket måste installeras första gången de används och laddas varje session de används, t.ex.

```
# install.packages("tidyverse") # Installera tidyverse  
library(tidyverse)              # Ladda ett paket
```

Objekt skapas med *assign*-pilen `<-`. Det är ett sätt att importera data till R. Det är dock vanligare importera från en extern fil. Mer om det senare.

Vi arbetar nästan alltid med data på tabellformat där variablerna är kolumner och observationerna rader. I gapminder-datan ges raderna av ett land för ett visst år.

```
library(gapminder)  
gapminder          # Skriv ut objektet gapminder
```

Funktioner agerar på objekt och ger något utfall. Här nedan beräknas medelvärdet av livslängd med funktionen `mean()`. Dollartecknet används för att ange

en specifik kolumn i dataobjektet. Funktioner styrs av möjliga argument - här används `na.rm` för att ange att saknade värden inte ska tas med i beräkningen

```
mean(gapminder$lifeExp, na.rm = T) # Beräkna medel av lifeExp
```

```
## [1] 59.47444
```

Funktionerna `filter()` och `select()` kan användas för att välja kolumner och rader. Funktioner kan länkas samman med en pipe `%>%` för att skapa sekvenser av funktioner. Man kan tänka på pipen som *och sen*.

```
gapminder %>%
  filter(country == "Norway", year > 1972) %>%
  select(country, lifeExp)
```

Slutligen tittade vi på grafer med `ggplot2`-paketet. En `ggplot` byggs upp med tre grundelar: data, geometrier (grafens objekt och former), och *aesthetics* (utseendet och placering av geometrierna). I ett enkelt spridningsdiagram är data två numeriska variabler, geometrierna är punkter, och punkternas placering ges av en x-koordinat och en y-koordinat. Ytterligare *aesthetics* kan vara punkternas färger (`color`) och storlek (`size`).

```
dat_small <- gapminder %>% filter(year == 2007)
ggplot(dat_small, aes(x = gdpPercap, y = lifeExp,
                     size = pop, color = continent)) +
  geom_point()
```

2.2 Import av data från en Excelfil

Inom vetenskapen är Excel det vanligaste filformatet för mindre datamängder. Till den här delen ska vi återigen arbeta med data från *Gapminder*.

Uppgift 2.1 (Excelfil från Canvas). Hitta excel-filen *Gapminder.xlsx* på Canvas och ladda ner den. Hitta mappen som filen laddats ned till.

I R kan man läsa in data från en Excel-fil med funktionen `read_excel()` från paketet `readxl`. Som argument till funktionen sätts filens sökväg - dess placering på hårddisken. Stycket nedan importerar från en excel-fil som ligger på hårddisken *C:* i mappen *Downloads*, under *User_name*, under *Users*.

```
library(readxl) # Ladda readxl
```

```
gapminder <- read_excel("C:/Users/Name/Downloads/Gapminder.xlsx")
# Läs in från en lokal excel-fil
gapminder
```

Uppgift 2.2 (Importera från excel-fil). Var ligger den nedladdade filen *Gapminder.xlsx*? Gör lämplig ändring i koden ovan för att läsa in data från den filen. Notera att R använder högerlutande snedstreck `/`, så om en kopierad sökväg har

vänster-snedstreck måste de ändras. Kontrollera att datan blivit korrekt inläst genom att köra objektnamnet `gapminder`.

En R-session har alltid en grundmapp, ett *Working directory*. Man kan se vilken mapp det är genom att köra

```
getwd() # Ange working directory
```

En filsökväg kan anges antingen som en fullständig sökväg, som ovan, eller relativt *working directory*. Om man till exempel har en fil *Gapminder.xlsx* som ligger i en mapp *Data* som i sin tur ligger i *working directory*, kan man importera data från filen med

```
gapminder <- read_excel("Data/Gapminder.xlsx")
# Läs in från en lokal excel-fil (relativt wd)
gapminder
```

Uppgift 2.3 (Working directory). Identifiera *working directory* för din nuvarande R-session genom att köra `getwd()`.

RStudio har också en inbyggd funktionalitet för att importera data. Man kan hitta den genom att gå till *Environment*-fliken och sedan *Import Dataset*. Det kan vara en bra hjälp, i synnerhet om man vill sätta datatyp för någon specifik kolumn.

Om du inte har tillgång till Canvas kan Gapminder-datan alternativt hämtas från paketet Gapminder.

```
# install.packages("gapminder")
library(gapminder)
gapminder
```

2.3 Ändra och skapa nya kolumner med mutate

Variabler kan omräknas och nya variabler kan skapas med `mutate`-funktionen. I gapminder-datan finns bnp per capita. Om man vill ha nationell BNP kan man skapa en ny kolumn och beräkna den som bnp per capita gånger populationen.

```
gapminder <- gapminder %>%           # Ta datan, och sen
  mutate(gdp = gdpPercap * pop)      # Beräkna bnp
```

Den inledande delen med `gapminder <-` gör så att utfallet av beräkningen sparas i objektet `gapminder`. Vi kan skriva ut objektet och se resultatet av beräkningen:

```
gapminder
```

Om man vill skapa en kolumn med mellanrum i namnet måste man skriva namnet inom *backticks* ‘ för att ange att namnet ska tolkas som en enhet. Jag

rekommenderar att undvika mellanrum i kolumnnamn och istället använda stora bokstäver eller understreck för ett nytt ord (`NationalGDP` eller `National_GDP`).

```
gapminder <- gapminder %>%
  mutate(`National GDP` = gdpPercap * pop)
gapminder
```

2.4 Sammanfattande lägesmått

Den importerade datan anger medellivslängd, populationsstorlek och bnp per capita per land och år. Vi kan börja med att producera en bubbelgraf över datan - en av de presentationer Gapminder ofta använder. En bubbelgraf är ett spridningsdiagram där punktens storlek beror på en tredje variabel.

```
ggplot(gapminder, aes(x = gdpPercap, y = lifeExp,
                      size = pop, color = continent)) +
  geom_point() +
  facet_wrap(~ year)
```

En interaktiv version kan vara bra om man vill identifiera någon specifik punkt.

```
g <- ggplot(gapminder, aes(x = gdpPercap, y = lifeExp, size = pop,
                          color = continent, text = country)) +
  geom_point() +
  facet_wrap(~ year)

# install.packages("plotly") # Kör om ej installerat tidigare
library(plotly)             # Ladda paketet plotly
ggplotly(g)                  # Interaktiv graf g
```

Under föreläsningen såg vi exempel på två lägesmått: medelvärdet (egentligen det *aritmetiska* medelvärdet) och medianen. De har bägge enkla funktioner i R: `mean()` respektive `median()`. Vi plockar ut en variabel ur datan och beräknar bägge.

```
gapminder$gdpPercap      # Vektorn med data
mean(gapminder$gdpPercap) # Beräkna medelvärdet
median(gapminder$gdpPercap) # Beräkna medianen
```

Samma sak kan göras med en pipe `%>%` och `summarise()`.

```
gapminder %>%
  summarise(Mean = mean(gdpPercap),
            Median = median(gdpPercap))
```

Uppgift 2.4 (Lägesmått av livslängd). Gör lämpliga ändringar i exemplet ovan för att beräkna lägesmått för medellivslängd (`lifeExp`).

Den andra lösningen, med en pipe och `summarise()`, kan enkelt utvecklas med

ett `group_by()`-steg för att beräkna medel och median per någon grupp, t.ex. per år.

```
gapminder %>%
  group_by(year) %>%
  summarise(Mean = mean(gdpPercap),
            Median = median(gdpPercap))
```

Ta datan, och sen
gruppera efter år, och
beräkna medelvärde och
medianen av gdpPercap

Uppgift 2.5 (Lägesmått per kontinent). Gör lämpliga ändringar i exemplet ovan för att beräkna lägesmått per kontinent. Vad måste läggas till för att också beräkna maximum och minimum per kontinent (funktionerna `max()` och `min()`)?

Uppgift 2.6 (Upprepade mätningar). Finns det några problem med att beräkna medelvärde per kontinent på den här datan?

I vetenskapliga publikationer redovisas medelvärden ofta med ett stapeldiagram. Som exempel ges staplar för medelvärdet av bnp per kontinent för 2007.

```
dat_gdp_2007 <- gapminder %>%
  filter(year == 2007) %>%
  group_by(continent) %>%
  summarise(Mean = mean(gdpPercap))

ggplot(dat_gdp_2007, aes(continent, Mean)) +
  geom_col()
```

Ta datan, och sen
filtrera för 2007, och sen
gruppera kontinenter, och
beräkna medelvärdet
Illustrera med kolumner (columns)

Uppgift 2.7 (Graf för livslängd). Gör om stapeldiagrammet. Denna gång med livslängd (`lifeExp`) istället för bnp per capita (`gdpPercap`).

2.5 Sammanfattande spridningsmått

Under föreläsningarna såg vi några mått på spridning: varians, standardavvikelse och kvartilavstånd (IQR, *inter-quartile range*). De har alla motsvarande funktioner i R (`var()`, `sd()`, och `IQR()`) som kan användas på samma sätt som funktionerna för lägesmått.

```
gapminder$gdpPercap      # Vektorn med data

var(gapminder$gdpPercap) # Beräkna varians
sd(gapminder$gdpPercap)  # Beräkna standardavvikelse
IQR(gapminder$gdpPercap) # Beräkna kvartilavstånd
```

Alternativt med en pipe och `summarise()`.

```
gapminder %>%
  summarise(Var = var(gdpPercap),
            Sd = sd(gdpPercap),
            IQR = IQR(gdpPercap))
```

Ta datan, och
beräkna varians,
standardavvikelse,
och kvartilavstånd

Lösningen med `pipe` och `summarise()` kan som tidigare utvecklas med `group_by()`.

```
gapminder %>%
  group_by(year) %>%
  summarise(Varians = var(gdpPercap),
            Standardavvikelse = sd(gdpPercap),
            Kvartilavstånd = IQR(gdpPercap))
```

Uppgift 2.8 (Graf för livslängd). Gör lämpliga ändringar i det sista exempel för att istället beräkna spridningsmått för livslängd.

Vi avslutar med tre vanliga illustrationer av vetenskaplig data - ett linjediagram med felstaplar, ett stapeldiagram med felstaplar, och ett låddiagram. För linjediagrammet beräknar vi medelvärdet och spridningsmått för bnp över år och kontinent. Som spridningsmått använder vi standard error (sv. *medelfel*), vilket beräknas som standardavvikelse delat på roten ur antalet observationer. Funktionen `n()` ger antalet observationer per kontinent och år.

```
dat_year_continent <- gapminder %>%
  group_by(year, continent) %>%
  summarise(Mean = mean(gdpPercap),
            SE = sd(gdpPercap) / sqrt(n()))
dat_year_continent
```

Med `ggplot2` kan vi bygga ett linjediagram med `geom_line()` och lägga till felstaplar med `geom_errorbar()`. Den senare behöver `aes()`-argument för `ymin` och `ymax` - nedre och övre del av felstapeln. Vi sätter dem till medelvärdet minus respektive plus ett medelfel.

```
ggplot(dat_year_continent, aes(x = year, y = Mean,
                              color = continent)) +
  geom_line() +
  geom_errorbar(aes(ymin = Mean - SE, ymax = Mean + SE))
```

Uppgift 2.9 (Bredd). Felstaplarna från `geom_errorbar()` har väldigt breda ändar. Använd hjälpsidan för `geom_errorbar`, i synnerhet exemplet längst ned, och se om det går att ändra bredden.

En graf med staplar och felstaplar kan konstrueras på ett liknande sätt. Följande exempel visar staplar över livslängd per kontinent. Felstapeln ges av standardavvikelsen.

```
dat_2007 <- gapminder %>%           # Ta datan, och sen
  filter(year == 2007) %>%         # filtrera på år, och sen
  group_by(continent) %>%         # gruppera efter kontinent,
  summarise(Mean = mean(lifeExp), # summera med medelvärde,
```

```
SD = sd(lifeExp))      # och standardavvikelse
dat_2007
```

Vi bygger en ggplot med `geom_col()` och `geom_errorbar()`. Felstapels konstruktion kan anges i en notis med funktionen `labs()`.

```
ggplot(dat_2007, aes(continent, Mean, fill = continent)) +
  geom_col() +
  geom_errorbar(aes(ymin = Mean - SD, ymax = Mean + SD)) +
  labs(title = "Average life expectancy by continent, 2007",
        caption = "Errorbars given by mean +/- standard deviation.",
        Source: Gapminder")
```

Uppgift 2.10 (Staplar för 1982). Gör lämpliga ändringar i exempel ovan för att konstruera ett stapeldiagram med felstaplar för året 1982 och variabeln medellivslängd.

Ett lådagran anger fördelningen av en variabel genom att illustrera minimum, maximum och kvartiler. Kvartiler är mått som delar en datamängd i fyra lika stora delar (så att en fjärdedel ligger under första kvartilen, en fjärdedel mellan första och andra kvartil, och så vidare). Med `ggplot2` kan vi bygga ett lådagran med `geom_boxplot()`. Exempel ger en låda per år och kontinent.

```
ggplot(gapminder, aes(year, lifeExp, fill = continent,
                      group = year)) +
  geom_boxplot() +
  facet_wrap(~ continent)
```

Uppgift 2.11 (Group-argumentet). I lådagranmet används argumentet `group`. Vad gör det? Vad händer om man tar bort det?

2.6 Ordna upp beskrivande statistik och exportera

Efter att ha beräknat någon beskrivande statistik kan det vara bra att titta på hur resultaten kan snyggas upp och exporteras i något lämpligt format. Ta den tabell med medelvärden vi producerade i ett tidigare exempel.

```
dat_2007 <- gapminder %>%      # Ta datan, och sen
  filter(year == 2007) %>%    # filtrera på år, och sen
  group_by(continent) %>%     # gruppera efter kontinent,
  summarise(Mean = mean(lifeExp), # summera med medelvärde,
            SD = sd(lifeExp))    # och standardavvikelse
dat_2007
```

Objekt kan exporteras från R på liknande som det importeras - med särskilda exportfunktioner (*write*-funktioner) beroende på filtyp. För att exportera till en

csv-fil man man använda `write_csv()`. Ingående argument är det objekt man vill exportera och den fil man vill exportera till. R ger ingen varning om man skriver över en existerande fil, så var lite försiktiga här.

Precis som vid import använder R *working directory* om inget annat anges. Följande exporterar objektet `dat_2007` till en csv-fil *Exporterad data från R.csv* i *working directory*.

```
getwd() # Se nuvarande working directory
write_csv(dat_2007, "Exporterad data från R.csv") # Exportera data
```

Därifrån skulle man kunna öppna filen i något kalkylprogram, snygga till layouten, och sedan klippa in i ett textdokument.

2.7 Kumulativt medelvärde

Om man har data som av någon anledning samlas in i sekvens kan det vara intressant att beräkna och illustrera den med ett *kumulativt medelvärde*. En serie med kumulativa medelvärden beräknas genom att för varje nytt värde ta medelvärden av de värden man hittills samlat in - vid tio värden tar man medelvärdet av de tio, vid elva värden medelvärdet av de elva, och så vidare.

Med de tärningsvärden vi hade innan kan vi beräkna ett kumulativt medelvärde genom att först beräkna summan med `cumsum()` och sedan dela på antalet kast. För att förenkla beräkningen av antalen tar vi fram sekvensen med antal kast i ett `mutate()`-steg.

```
dat_dice <- data.frame(Utfall = c(6,3,2,3,5)) %>%
  mutate(Kast = 1:n())
dat_dice

dat_dice <- dat_dice %>%
  mutate(kumulativ_summa = cumsum(Utfall),
         kumulativt_medelvärde = kumulativ_summa / Kast)
dat_dice
```

Uppgift 2.12 (Kumulativt medelvärde). Vad ska läggas till för att stycket nedan ska ge en linjeförlopp över medelvärdet?

```
ggplot(dat_dice, aes(x = Kast, y = ___)) +
  ___()
```

Uppgift 2.13 (Fler tärningskast). Kasta din tärning ytterligare några gånger, gärna på en mjuk yta. Fyll i dina utfall och gör grafen från föregående uppgift. Kan man se en tendens för medelvärdet att minska i varians vid fler kast? Svänger kurvan in mot något specifikt värde?

```
dat_dice <- data.frame(Utfall = c(___)) %>%
  mutate(Kast = 1:n(),
```

```

    kumulativ_summa = cumsum(Utfall),
    kumulativt_medelvärde = kumulativ_summa / Kast)
dat_dice

```

Uppgift 2.14 (Kumulativ frekvens). Om man vill titta på andelen gånger ett visst utfall inträffat talar man om *kumulativ frekvens* snarare än *kumulativt medelvärde*. Använd stycket nedan för att titta på andelen gånger utfallet varit en etta (ett *positivt* utfall, i begreppets kliniska mening). Om ett inte är ett möjligt utfall på din tärning, ändra ettan till något mer lämpligt.

```

dat_dice <- data.frame(Utfall = c(___)) %>%
  mutate(Kast = 1:n(),
         positivt_utfall = Utfall == 1,
         kumulativt_antal = cumsum(positivt_utfall),
         kumulativ_frekvns = kumulativt_antal / Kast)
dat_dice

ggplot(dat_dice, aes(x = Kast, y = kumulativ_frekvns)) +
  geom_line()

```

2.8 Tredimensionella grafer med plotly

De två avslutande avsnitten är mindre viktiga och kan läsas i mån av tid.

Paketet `plotly` kan användas för att göra interaktiva graf och grafer med tre dimensioner. Börja med att ladda paketet.

```
library(plotly)
```

Vi börjar med ett enkelt exempel på en 3d-graf med lite skapad data.

```

dat_ex <- data.frame(Var1 = c(1,2,3), Var2 = c(3,1,2),
                    Var3 = c(2,3,1), Type = c("A", "B", "C"))
dat_ex

plot_ly(dat_ex, x = ~Var1, y = ~Var2,
        z = ~Var3, color = ~Type) %>%
  add_markers()

```

Om grafen inte kommer upp direkt kan det fungera att trycka på den lilla ikonen med ett fönster och en pil i *Viewer*-fliken. Grafen ska då öppnas i en webbläsare.

Syntaxen till `plot_ly()` är inte helt olik `ggplot()`. Först anges datan, därefter argument för x- y-, och z-koordinater. Notera tilde-tecknet `~` före variabelnamnen. Eventuell färg sätts med `color`. Efter det lägger man till punkter (här *markers*) med en pipe in i `add_markers()`. Vi vill göra en liknande graf med gapminder-datan, men får börja med att filtrera på ett visst år.

Uppgift 2.15 (Filtrera för år). Vad måste läggas till i funktionen nedan för att filtrera för data där året är 2007?

```
dat_2007 <- gapminder %>%
  filter(year == ___)
```

Vi kan nu konstruera en 3d-graf med datan.

Uppgift 2.16 (Gapminder i 3d). Vad måste läggas till i funktionen nedan för en 3d-graf med befolkningsmängd (`pop`) på x-axeln, livslängd (`lifeExp`) på y-axeln, bnp per capita (`gdpPerCap`) på z-axeln, och färg efter kontinent (`continent`)? För att kunna identifiera specifika länder kan man också sätta argumentet `text`.

```
plot_ly(data_2007, x = ~pop, y = ~___, z = ~___,
        color = ~continent, text = ~country) %>%
  add_markers()
```

Uppgift 2.17 (Log-transformationer). Inom statistiken är det vanligt att transformera variabler för att ta bort extremeffekter och visa på specifika dataegenskaper. En vanlig transform är att *logaritmera* ett värde, vilket innebär att man istället för att använda det ursprungliga värdet använder exponenten i någon bas (ofta basen tio). Ta till exempel värdet 10000, dess tio-logaritm är 4, eftersom 10 upphöjt i 4 är 10000. Logaritmer är vanliga vid data med extremvärden.

Grafen i uppgiften ovan präglas mycket av skillnader i bnp och befolkningsstorlek. Testa att tio-logaritmera variablerna och se om det blir en mer eller mindre överskådlig graf. Logaritmen kan göras genom att byta den ursprungliga variabeln mot en variabel transformerad med `log10()`. Fyll i stycket nedan.

```
plot_ly(dat_2007, x = ~log10(pop), y = ~log10(___),
        z = ~___, color = ~___, text = ~country) %>%
  add_markers()
```

Uppgift 2.18 (Följa ett land). Likt en ggplot kan man lägga till graf-element. Här använder man dock en pipe för lägga till ett nytt element. Fyll i kodstycket nedan. Vad, om något, har lagts till i grafen?

```
plot_ly(dat_2007, x = ~log10(pop), y = ~log10(___),
        z = ~___, color = ~continent, text = ~country) %>%
  add_markers() %>%
  add_lines(data = gapminder %>% filter(country == "Costa Rica"))
```

2.9 AI-uppgift II

Uppgift 2.19 (Standardavvikelse från gapminder-datan). Använd en chatbot (såsom Microsofts co-pilot eller OpenAIs ChatGPT) för att få kod som beräknar standardavvikelsen av medellivslängd från gapminder-datan. Fråga också om tolkning av standardavvikelsen och om det finns några specifika tolkningsproblem för just den här gapminder-datan.

Uppgift 2.20 (3d-graf återigen). Använd en chatbot för att få kod som skapar en 3d-graf lik den vi skapade ovan. Se om det går att utveckla grafen ytterligare, t.ex. med en linje som går ner till grafens *botten*, eller linjer som kopplar samma länder till kontinentens medelvärde.

Kapitel 3

Ett stickprov av normalfördelad data

Datorövning 3 handlar om hypotestest och konfidensintervall för ett stickprov av normalfördelad data. Efter övningen ska vi kunna

- genomföra och tolka ett t-test för normalfördelad data,
- beräkna och tolka ett konfidensintervall för normalfördelad data,
- använda simulerad data för att förstå t-testets egenskaper.

3.1 Repetition från datorövning 2

När man startar en ny R-session bör man ladda de paket man vet kommer behövas med `library()`. Om paket inte finns installerade måste man först köra `install.packages()`.

```
# install.packages("tidyverse")  
library(tidyverse)
```

I datorövning 2 tittade vi på hur insamlade variabler kan sammanfattas med lägesmått och spridningsmått. Ett enkelt sätt att ta fram dem är att använda `summarise()` och ange de mått och variabler man vill använda. Vi hade uppe ett exempel på data från Gapminder som vi importerade från en excel-fil. För nu kan vi dock hämta datan från paketet `gapminder`.

```
# install.packages("gapminder")  
library(gapminder)  
  
gapminder %>%  
  filter(year == 2007) %>%  
  group_by(continent) %>%
```

```
summarise(`Livslängd, medel` = mean(lifeExp),
          `Befolkning, median` = median(pop),
          `Bnp per capita, standardavvikelse` = sd(gdpPercap))
```

Beskrivande mått sammanfattas ofta i någon enkel vetenskaplig graf. Två vanliga val är lådagrammet, som illustrerar kvartiler och möjliga extremvärden, och stapeldiagrammet med felstaplar. Vi ger först ett exempel på ett lådagram över livslängd per kontinent uppdelat efter år.

```
ggplot(gapminder, aes(lifeExp, continent, fill = continent)) +
  geom_boxplot() +
  facet_wrap(~ year)
```

Därefter ett exempel på ett stapeldiagram med felstaplar för samma data. Felstapeln ges av standardavvikelsen.

```
dat_sum <- gapminder %>%
  group_by(continent, year) %>%
  summarise(Mean = mean(lifeExp),
            SD = sd(lifeExp))
dat_sum

ggplot(dat_sum, aes(continent, Mean, fill = continent)) +
  geom_col() +
  geom_errorbar(aes(ymin = Mean - SD, ymax = Mean + SD),
               width = 0.3) +
  facet_wrap(~ year)
```

3.2 Test av medelvärde för normalfördelad data

Om man har en normalfördelad variabel och vill testa om populationens medelvärde är skilt från något hypotetiskt värde μ_0 kan man använda ett *t-test för ett stickprov*. Ta som exempel följande data på 8 observationer av havreskörd. Av någon historisk anledning vill man testa om populationsmedelvärdet är skilt från 50.

```
library(tidyverse)

dat <- data.frame(y = c(49.8, 58.4, 49.4, 57.1,
                      52.2, 49.1, 44.6, 55.4))
dat

ggplot(dat, aes(y, 0)) +
  geom_point() +
  geom_vline(xintercept = 50, color = "red")
```

I grafen ser vi att värdena ligger jämt spridda kring 50, så 50 är nog ganska rimligt som medelvärde, men låt oss göra ett formellt test. I R kan ett t-test genomföras med `t.test()`.

```
t.test(dat$y, mu = 50)  # Ett t-test på variabeln y

##
## One Sample t-test
##
## data:  dat$y
## t = 1.2086, df = 7, p-value = 0.266
## alternative hypothesis: true mean is not equal to 50
## 95 percent confidence interval:
##  48.08713 55.91287
## sample estimates:
## mean of x
##          52
```

Utskriften ger ett p-värde från vilket vi kan dra en slutsats. I det här fallet är p-värdet högt (över fem procent) så vi kan inte förkasta nollhypotesen (vilken är att populationsmedelvärdet är lika med 50).

Vi tittar nu på stegen bakom t-testet. Ett t-test bygger, som alla hypotestest, på en serie steg:

1. sätt upp en *nollhypotes* och en *alternativhypotes*,
2. beräkna ett *testvärde* från en testfunktion,
3. identifiera en *testfördelning*,
4. beräkna ett *p-värde*, eller uppskatta ett genom att ställa testvärde mot ett kritiskt värde,
5. dra en klar *slutsats* om statistisk signifikans.

Vi vill testa om medelskörden är skild från 50, så hypoteser ges av

- H_0 : μ lika med 50
- H_1 : μ ej lika med 50

Alternativhypotesen är tvåsidig - vi tittar både på möjligheten att populationsmedelvärdet är större och på möjligheten att det är mindre.

Uppgift 3.1 (Ensidig mothypotes). Hur hade hypoteserna sett ut om vi ville testa om medelvärdet är *större* än 50?

Från datan har vi åtta observation och vi kan ta fram stickprovsmedelvärde och standardavvikelse i R.

```
mean(dat$y)

## [1] 52
```

```
sd(dat$y)
```

```
## [1] 4.680354
```

Vårt mål är att testa ett medelvärde och det är rimligt att anta normalfördelning för den undersökta variabeln. Det lämpliga testet är då ett t-test för ett stickprov och testvärdet kan beräknas av en testfunktion som ges av det observerade stickprovet minus nollhypotesens värde, delat på standardavvikelsen delat på roten ur antalet observationer.

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$$

Uppgift 3.2 (Handräkning). Beräkna t-värdet på miniräknare eller telefon. Finns det beräknade t-värdet i utskriften från `t.test()`?

Nästa steg är att identifiera testfördelning, det vill säga den slumpfördelning testvärdet följer om nollhypotesen är sann. I det här fallet är testfördelningen en t-fördelning med $n - 1$ frihetsgrader. Vi har åtta observationer, så antalet frihetsgrader blir 7. I R kan man ta fram täthetsfunktionen för en t-fördelning med `dt()` och fördelningsfunktionen med `pt()`. Kurvorna kan plottas med `geom_function()` i en `ggplot`.

```
ggplot() +
  geom_function(fun = dt, args = list(df = 7)) +
  xlim(-5,5)
```

```
ggplot() +
  geom_function(fun = pt, args = list(df = 7)) +
  xlim(-5,5)
```

Från testfördelningen kan vi nu ta fram ett p-värde. P-värdet kan illustreras som ytan under t-fördelning *bortom* test-värdet. I ett tvåsidigt test tar vi med bägge svansarna.

```
ggplot() +
  geom_function(fun = dt, args = list(df = 7)) +
  geom_vline(xintercept = c(-1.2086, 1.2086),
             color = "red") +
  xlim(-5,5)
```

Ytorna i *svansarna* motsvarar p-värdet. De anger sannolikheten att få ett större t-värde än det vi fått, under antagandet att nollhypotesen stämmer. Det exakta p-värdet gavs av `t.test()`-funktionen som 0.266. En tolkning av det är om försöket upprepas ett stort antal gånger och nollhypotesen är sann, kommer vi 26.6 procent av gångerna få ett större testvärde än 1.2086. Det vi observerar är ganska sannolikt om nollhypotesen stämmer, vilket tyder på att nollhypotesen är rimlig.

Det avslutande steget är att dra en formell slutsats och ge ett tydligt svar. Det beräknade p-värdet ställs mot en förbestämd signifikansnivå, oftast fem procent. Här är p-värdet över den nivån, så vi kan inte förkasta nollhypotesen. Slutsatsen är att det *inte* finns någon signifikant skillnad från 50 i havreskörd.

Uppgift 3.3 (Kritiskt värde). Om man gör ett t-test för hand kan man inte enkelt ta fram ett p-värde, men kan se om p-värdet är större eller mindre än fem procent genom att ställa testvärdet mot ett kritiskt värde. Använd en tabell för t-fördelning för att hitta det kritiska värdet.

I R kan man ta fram kritiska värden med `qt()`. För fem procent i svansarna har man 0.025 i respektive svans och det kritiska värdet ges av

```
qt(0.975, 7)
```

Uppgift 3.4 (Ensidigt test). Använd `?t.test` för att ta fram funktionens hjälpsida. Försök att utifrån hjälpsidan beräkna ett *ensidigt* test för att se om medelskörden är *större* än 50.

3.3 Konfidsensintervall för normalfördelad data

Utfallet av ett t-test beror på nollhypotesen värde, så om vi ändrar det kommer vi få ett annat p-värde.

```
t.test(dat$y, mu = 50)
t.test(dat$y, mu = 48)
```

Från p-värdena kan man dra slutsatsen att förkasta vid nollhypotesen att μ är 48, men inte förkasta vid nollhypotesen att μ är 50. Värdet 50 är alltså i någon mening ett mer troligt värde på populationens medelvärde än vad 48 är. *Konfidsensintervall* kan ses som en generalisering av den tanken: ett konfidsensintervall ger ett spann av värden där man *inte* förkastar. Intervallet tolkas vanligen som att det täcker det sanna populationsmedelvärdet med en viss konfidens.

För ett stickprov och antagen normalfördelning ges konfidsensintervallet (med konfidensgraf $1 - \alpha$) av

$$\bar{y} \pm t_{(1-\alpha/2, n-1)} \frac{s}{\sqrt{n}}$$

Kvantilen från t-fördelningen kan hämtas från en tabell (samma som det kritiska värdet i testet) eller genom R. Antalet frihetsgrader ges av antalet observationer minus ett. I det här fallet ges delarna av

```
mean(dat$y)
sd(dat$y)
qt(0.975, 7)
```

och konfidsensintervallet ges alltså av

$$52 \pm 2.365 \cdot 4.680/\sqrt{8}$$

Funktionen `t.test()` ger automatiskt ett konfidensintervall, direkt under utfallet av testet. Notera att konfidensintervallet inte beror på nollhypotesen. Konfidensintervall kan beräknas med skilda konfidensnivåer, oftast 95 procent, vilket sätts med argumentet `conf.level`.

Uppgift 3.5 (Konfidensnivå). Gör lämplig ändring i koden nedan för att beräkna ett 99-procentigt konfidensintervall, istället för ett 95-procentigt.

```
t.test(dat$y, conf.level = 0.95)

##
## One Sample t-test
##
## data:  dat$y
## t = 31.425, df = 7, p-value = 8.538e-09
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  48.08713 55.91287
## sample estimates:
## mean of x
##          52
```

Är ett 99-procentigt konfidensintervall bredare eller smalare än ett 95-procentigt?

3.4 Normalfördelad data

Eftersom t-testet bygger på att data är normalfördelad är det förstås bra att kunna undersöka om det antagandet stämmer. Ett sätt är att göra ett histogram över datan - om den underliggande variabeln är normalfördelad bör stickprovet ge den typiska klockformen. Det här kräver dock ganska mycket data. Ta ett histogram för havredatan som exempel

```
ggplot(dat, aes(y)) + geom_histogram(bins = 5)
```

Låt oss titta på histogram över genererad normalfördelad data.

```
n <- 10
ggplot() + geom_histogram(aes(x = rnorm(n)), bins = 30)
```

Uppgift 3.6 (Histogram för normalfördelning). Testa koden ovan för lite olika värden på `n`. Det kan vara nyttigt att sätta antalet staplar `bins` för att få ett bättre histogram. Hur stort måste `n` vara för att ge en karaktäristisk klockform för histogrammet?

Ett annat vanligt alternativ för att grafisk undersöka om data följer en ungefärlig normalfördelning är en QQ-graf (*QQ-plot*). En qq-graf är ett spridningsdiagram med teoretiska kvantiler på en axel och datans kvantiler på den andra axeln. Om data perfekt följer en normalfördelning kommer grafen visa en rak diagonal linje. En QQ-graf kan tas fram med `qqnorm()` eller `geom_qq()` i en ggplot. En diagonal linje för jämförelse kan läggas till med `geom_qq_line()`.

```
ggplot(dat, aes(sample = y)) + geom_qq() + geom_qq_line()
```

Punkterna ligger nära linjen. Vi kan återigen demonstrera med lite genererad data.

```
n <- 100
dat_norm <- data.frame(y = rnorm(n))
ggplot(dat_norm, aes(sample = y)) + geom_qq() + geom_qq_line()
```

Uppgift 3.7 (Histogram för normalfördelning). Funktionen `runif()` ger slumpmässiga värden mellan 0 och 1. Testa att ändra i kodstycket ovan så att slumptal genereras med `runif()` istället för `rnorm()`. Hur påverkar det QQ-grafen?

3.5 Centrala gränsvärdesatsen

Även om data inte är normalfördelad kan t-testet vara ett lämpligt val av test. Detta beror på *centrala gränsvärdesatsen*, som säger att summor (och därmed även medelvärden) av lika slumpvariabler går mot en normalfördelning där antalet observationer ökar. Den tidigare uppgiften gav att `runif()` inte ger normalfördelad data. Vad händer om vi tar medelvärdet av flera observationer från `runif()`? Följande kod beräknar tiotusen medelvärden av två observationer.

```
antal <- 2

dat_sim_unif <- expand_grid(Observation = 1:antal,
                           Upprepning = 1:10000) %>%
  mutate(y = runif(n())) %>%
  group_by(Upprepning) %>%
  summarise(y = mean(y))

ggplot(dat_sim_unif, aes(y)) + geom_histogram(bins = 50)

ggplot(dat_sim_unif, aes(sample = y)) +
  geom_qq() + geom_qq_line()
```

Fördelningen för summan är inte likformig, men inte heller särskilt normalfördelad. Vad händer om vi ökar antal termer i summan?

Uppgift 3.8 (Antalet observationer för normalfördelning). Vad måste ändras i koden ovan för beräkna medelvärdet av tio observationer? Följer de medelvärdena en ungefärlig normalfördelning? Vad är det lägsta antalet observationer som

ger ungefärligen normalfördelade medelvärden?

Det kan också finnas situationer där någon matematisk transformation kan göra icke-normal data till normalfördelad data. Vanliga transformationer är att ta en kvadratrots eller att *logaritmera* datan. Som exempel kan vi återvända till gapminder-datan vi använde i en tidigare datorövning. Paketet **patchwork** kan användas för att placera flera grafer bredvid varandra. Den exakta koden är mindre viktig här.

```
library(gapminder)
gapminder_2007 <- gapminder %>% filter(year == 2007)

g1 <- ggplot(gapminder_2007, aes(pop)) +
  geom_histogram(bins = 30)
g2 <- ggplot(gapminder_2007, aes(sample = pop)) +
  geom_qq() +
  geom_qq_line()
g3 <- ggplot(gapminder_2007, aes(sqrt(pop))) +
  geom_histogram(bins = 30)
g4 <- ggplot(gapminder_2007, aes(sample = sqrt(pop))) +
  geom_qq() +
  geom_qq_line()
g5 <- ggplot(gapminder_2007, aes(log10(pop))) +
  geom_histogram(bins = 30)
g6 <- ggplot(gapminder_2007, aes(sample = log10(pop))) +
  geom_qq() +
  geom_qq_line()

library(patchwork)
g1 + g3 + g5 + g2 + g4 + g6
```

I grafen har vi den ursprungliga variabeln (befolkning per land 2007), den kvadratrots-transformerade variabeln (`sqrt()`) och den log-transformerade variabeln (`log10()`). De två första fallen påverkas kraftigt av extremvärden och är klart icke-normala medan den log-transformerade variabeln ger en ungefärlig normalkurva och följer diagonalen väl i QQ-grafen.

3.6 AI-uppgift III

Det finns många studier som pekar på att p-värdet är svårtolkat. Det är särskilt vanligt att p-värdet tolkas som sannolikheten att nollhypotesen är sann. Den korrekta tolkningen är dock snarare tvärtom: p-värdet är sannolikheten för datautfallet givet att nollhypotesen är sann.

Uppgift 3.9 (Tolkning av p-värdet). Be en AI-chatbot besvara en t-test-fråga t.ex. från en tidigare tenta eller uppgiften i början av den här övningen. Fråga särskilt om tolkningen av p-värdet och om det finns några vanliga feltolkningar.

3.7 Bonus. Simuleringar för t-test

Följande kod simulerar ett dataset om tio observationer från en normalfördelning med medelvärde 7 och standardavvikelse 5, beräknar ett hypotestest med nollhypotesen att populationsmedelvärdet är 7, och beräknar ett konfidensintervall.

```
dat_sim <- data.frame(y = rnorm(10, mean = 7, sd = 5))
t.test(dat_sim$y, mu = 7)
```

Uppgift 3.10 (Upprepad simulering). Kör de två raderna i stycket ovan ett tiotal gånger. Du bör se att man ibland förkastar nollhypotesen trots att den ska stämma. Kan du få en känsla för hur stor andel av gångerna man felaktigt förkastar?

Låt oss upprepa simuleringen tusen gånger. Ett sätt är att upprepa ett steg flera gånger är genom en for-loop.

```
dat_sim <- data.frame()
for(i in 1:1000){
  new_data <- data.frame(y = rnorm(10, mean = 7, sd = 5))
  test <- t.test(new_data$y, mu = 7)
  new_results <- data.frame(t_value = test$statistic,
                           p_value = test$p.value,
                           ci_lower = test$conf.int[1],
                           ci_upper = test$conf.int[2])
  dat_sim <- bind_rows(dat_sim, new_results)
}
```

Enligt statistisk teori ska t-värdet följa en t-fördelning med nio frihetsgrader. Vi kan undersöka det genom ett histogram.

```
ggplot(dat_sim) +
  geom_histogram(aes(t_value, y = after_stat(density)),
                bins = 50,
                fill = "white", color = "black") +
  geom_function(fun = dt, args = list(df = 9))
```

Den teoretiska t-fördelning passar histogrammet nästan perfekt.

Vidare ska p-värdet vara under fem procent fem procent av gångerna.

```
mean(dat_sim$p_value < 0.05)
```

Även det stämmer någorlunda väl. Det här innebär alltså att om man har en signifikansnivå på fem procent kommer man förkasta nollhypotesen fem procent av gångerna även om nollhypotesen stämmer. Det kallas ett *falskt positivt* utfall.

Uppgift 3.11 (Signifikant skillnad). Stycket nedan simulerar data när populationsmedelvärdet är 9 och t-test har nollhypotesen att populationsmedelvärdet

är 7. Här vill vi alltså förkasta nollhypotesen.

```
dat_sim <- data.frame()
for(i in 1:1000){
  new_data <- data.frame(y = rnorm(10, mean = 9, sd = 5))
  test <- t.test(new_data$y, mu = 7)
  new_results <- data.frame(t_value = test$statistic,
                           p_value = test$p.value,
                           ci_lower = test$conf.int[1],
                           ci_upper = test$conf.int[2])
  dat_sim <- bind_rows(dat_sim, new_results)
}
```

Använd kod från den första simuleringen för att undersöka hur väl histogrammet stämmer med den teoretiska fördelningen och för att se hur stor andel av gångerna man förkastar nollhypotesen på signifikansnivån 5 procent.

Uppgift 3.12 (Konfidensintervallets bredd). Ett konfidensintervall blir smalare och smalare ju större stickprovet är. Koden nedan ger medelvärdet för stickprovsbredden i simulerad data med standardavvikelsen 1.

```
n <- 10

dat_sim <- data.frame()
for(i in 1:1000){
  new_data <- data.frame(y = rnorm(n, mean = 0, sd = 1))
  test <- t.test(new_data$y, mu = 7)
  new_results <- data.frame(t_value = test$statistic,
                           p_value = test$p.value,
                           ci_lower = test$conf.int[1],
                           ci_upper = test$conf.int[2])
  dat_sim <- bind_rows(dat_sim, new_results)
}

mean(dat_sim$ci_upper - dat_sim$ci_lower)
```

Ungefär hur många observationer behövs för att konfidensintervallets bredd ska bli under 1, under 0.9, under 0.8, och så vidare ned till 0.1?

Kapitel 4

Ett stickprov av icke-normalfördelad data

Datorövning 4 handlar om hypotestest och konfidensintervall för ett stickprov av icke-normalfördelad data. Efter övningen ska vi kunna

- genomföra och tolka ett z-test för proportioner,
- genomföra och tolka ett chi-två-test för nominal data
- beräkna och tolka ett konfidensintervall för proportioner.

Om det finns tid för en bonussektion kommer vi också titta på interaktiva kartor med `leaflet`.

4.1 Repetition av datorövning 3

När man startar en ny R-session bör man ladda de paket man vet kommer behövas med `library()`. Om paket inte finns installerade måste man först köra `install.packages()`.

```
# install.packages("tidyverse")  
library(tidyverse)
```

I datorövning 3 tittade vi på analys av ett stickprov av normalfördelad data. Det underliggande upplägget är att vi vill säga något om en population genom att titta på ett stickprov. Ett *t-test* kan användas för att testa en given *nollhypotes*. Ett konfidensintervall ringar in populationens medelvärde med en viss konfidensgrad - oftast 95 procent. Ett normalfördelningsantagande kan undersökas med histogram eller QQ-grafer.

Ta som exempel följande data på jord-pH och låt oss anta att det är relevant att testa om populationens medelvärde är 7.

```
dat_pH <- data.frame(pH = c(6.3, 6.55, 6.75, 6.4,
                             7.25, 6.65, 6.8, 7.3, 7.15, 6.7))

ggplot(dat_pH, aes(pH, 0)) +
  geom_point(size = 4) +
  geom_vline(xintercept = 7, size = 5,
             color = "red", alpha = 0.4) +
  theme(axis.text.y = element_blank(),
        axis.title.y = element_blank(),
        axis.ticks = element_blank())
```

Hypoteser ges av

- H_0 : μ lika med 7
- H_1 : μ ej lika med 7

Testet kan genomföras med funktionen `t.test()`.

```
t.test(dat_pH$pH, mu = 7)
```

Ett p-värde över 0.05 ger att vi inte förkastar nollhypotesen - den observerade skillnaden mot 7 är inte statistiskt signifikant. Konfidensintervallet ger att populationens medelvärde ligger mellan 6.54 och 7.03 med 95 procents konfidens.

En QQ-graf kan visa om det finns några avvikelser från normalantagandet. Med små datamängder är det oftast svårt att se några tydliga tecken på icke-normalitet.

```
ggplot(dat_pH, aes(sample = pH)) +
  geom_qq() +
  geom_qq_line()
```

4.2 Proportioner från binär data

Binär data är data där en observation har ett av två utfall, vilka kan kodas som noll och ett. Man talar ibland om utfallet ett som ett *positivt* utfall. Binär data kan sammanfattas med en proportion - antalet positiva utfall delat på det totala antalet upprepningar. En proportion kan testas med ett z-test för proportioner (eller *relativ frekvens*). Testet följer stegen för hypotestest (Hypoteser - Testvärde - Testfördelning - P-värde (eller jämförelse med kritiskt värde) - Slutsats).

Låt oss importera lite exempeldata och beräkna ett exempel. Följande rad importerar matchresultat i fotbollsallsvenskan för damer 2000-2020.

```
library(tidyverse)
dat_allsv <- read_csv("https://raw.githubusercontent.com/adamflr/ST0060/main/Data/Allsv")
```

```
ggplot(dat_all, aes(hemmamal, bortamal)) +
  geom_jitter(size = 0.1)
```

Uppgift 4.1 (En interaktiv målgraf). Kör stycket nedan för en interaktiv målgraf. Vilken match gav det högsta antalet gjorda bortamål?

```
# install.packages("plotly")
library(plotly)
g <- ggplot(dat_all, aes(hemmamal, bortamal,
  text = paste(sasong, hemma, "-", borta))) +
  geom_jitter(size = 0.1)

ggplotly(g)
```

Gammal bollkunskap säger att var tredje match är en bortavinst. Vi kan testa det med ett z-test för proportioner. För att ta fram antalet bortasegrar och totalt antal matcher använder vi `count()` på kolumnen `resultat`. Man kan också använda funktionen `table()` för ett liknande resultat.

```
dat_all %>% count(resultat)
table(dat_all$resultat)
```

Datan har 947 bortasegrar av totalt $947 + 1803$ matcher. Vår skattade proportion p och totala antal n är alltså

```
n <- 947 + 1803
p_est <- 947 / n

p_est
n
```

För att genomföra ett z-test sätter vi upp hypoteser om proportionen bortasegrar.

- Nollhypotes H_0 : p lika med 0.33
- Alternativhypotes H_1 : p ej lika med 0.33

Ett test kan köras i R med `prop.test()`.

```
prop.test(x = 947, n = 2750, p = 0.33, correct = F)

##
## 1-sample proportions test without continuity correction
##
## data: 947 out of 2750, null probability 0.33
## X-squared = 2.5661, df = 1, p-value = 0.1092
## alternative hypothesis: true p is not equal to 0.33
## 95 percent confidence interval:
## 0.3268327 0.3623288
```

```
## sample estimates:
##           p
## 0.3443636
```

P-värdet ställs mot en förbestämd signifikansnivå (vanligen 5 procent). I det här fallet leder det höga p-värdet till att nollhypotesen accepteras.

Om vi tar en närmre titt på testets steg börjar vi med att beräkna ett testvärde.

```
p0 <- 0.33
z_value <- (p_est - p0) / sqrt(p0 * (1 - p0) / n)
z_value
```

```
## [1] 1.601904
```

Därefter kan p-värdet räknas ut som arean under en standardiserad normalfördelning bortom z-värdet. Eftersom vi har en tvåsidig mothypotes adderas de två svansarna.

```
ggplot() +
  geom_function(fun = dnorm) +
  geom_vline(xintercept = c(-z_value, z_value)) +
  xlim(-3,3)
```

Areans yta kan tas fram med normalfördelningens fördelningsfunktion `pnorm()`.

```
2 * pnorm(-z_value)
```

```
## [1] 0.1091769
```

Testets p-värde är ungefär 11 procent. Vår observation är alltså inte ett orimligt utfall om den faktiska sannolikheten för bortaseger är 0.33 och vi kan inte förkasta nollhypotesen på femprocentnivån.

Om man löste uppgiften för hand skulle man istället för att beräkna p-värdet jämföra z-värdet med ett kritisk värde ur en tabell. Det kritiska värdet för femprocentig signifikans är 1.96. Vi kan också ta fram det genom `qnorm(0.975)`.

Vi kan jämföra den beräkning med den direkta R-funktionen.

```
prop.test(x = 947, n = 2750, p = 0.33, correct = F)
```

Den stegvisa beräkningen gav samma utfall som funktionen (`p-value = 0.1092`). Funktionen ger inte z-värdet utan ett chi-två-värde (2.5661). Här är det värdet lika med z-värdet i kvadrat.

```
z_value^2
```

Uppgift 4.2 (Test för proportionen oavgjorda). Samma gamla bollkunskap säger att 20 procent av matcher blir oavgjorda. I datan är 518 av 2750 matcher oavgjorda. Ställ upp hypoteser och fyll i koden nedan för att testa om bollkunskapen stämmer. Beräkna också z-värdet för hand.

```
prop.test(x = ___, n = ___, p = ___, correct = F)
```

Uppgift 4.3 (Test för proportionen hemmasegrar). Slutligen är då resten av matcherna, 1285 av 2750, hemmasegrar. Gammal bollkunskap säger: *47 procent av alla matcher är hemmasegrar*. Genomför ett z-test för att testa det påstående.

```
prop.test(x = ___, n = ___, p = ___, correct = F)
```

Uppgift 4.4 (Population och stickprov). Ett hypotestest bygger på en underliggande tanke med en population (med någon för oss okänd proportion positiva utfall) och ett stickprov (i vilket vi kan observera andelen positiva utfall). Det är inte alltid uppenbart vad som egentligen är populationen. I fallet med fotbollsdatan, vad kan ses som populationen? Hur långt skulle man kunna generalisera de slutsatser man kan dra från datan?

Hypotestestet för proportioner som används här, *z-testet*, bygger på en normalapproximation av en binomialfördelning. Approximation blir bättre när antalet observationer är stort och nollhypotesens värde p_0 ligger nära 0.5. En vanlig tumregel för när approximationen är giltig är att n gånger p_0 gånger $(1 - p_0)$ ska vara större än 10.

4.3 Konfidensintervall för proportioner

Konstruktionen av ett konfidensintervall för en proportion är ganska lik konstruktionen för ett medelvärde. För en skattad proportion p och antal observationer n kan man beräkna p plus/minus ett z -värde från tabell gånger medelfelet, där medelfelet ges av roten ur $p \cdot (1 - p)/n$. För exemplet med bortasegrar i allsvenskan är $p = 0.344$ och $n = 2750$. Tabellvärdet hämtas från en tabell över kvantiler. För ett 95-procentigt konfidensintervall tar vi kvantilen 0.975 (2.5 procent i respektive svans) vilket ger värdet 1.96. Konfidensintervallet ges av

```
n <- 947 + 1803
p <- 947 / n

p - 1.96 * sqrt(p * (1 - p) / n)
p + 1.96 * sqrt(p * (1 - p) / n)
```

Notera att 0.33, det värde som var nollhypotesen i det tidigare testet, *ingår* i intervallet. Om man tittar på utskriften från `prop.test()` kan man se ett konfidensintervall. Det intervallet är dock inte beräknat på samma sätt den formel som förekommer på föreläsningarna. För att få matchande utskrift kan vi använda paketet `binom` och funktionen `binom.asymp()`.

```
#install.packages("binom")
library(binom)
binom.asymp(x = 947, n = 2750)
```

Uppgift 4.5 (99-procentigt konfidensintervall). Gör lämplig ändring i koden nedan för att beräkna ett 99-procentigt konfidensintervall för andelen bortasegrar.

```
binom.asymp(x = 947, n = 2750, conf.level = 0.95)
```

4.4 Chi-två-test för goodness-of-fit

Ett proportionstest kan ses som ett test av en variabel med två möjliga klasser som utfall. Ett *goodness-of-fit*-test utvecklar det till valfritt antal klasser. Testet utförs som ett chi-två-test genom att beräkna ett observerat antal O och ett förväntat antal E för varje klass. Testvärdet ges av att man beräknar $(O - E)^2 / E$ för varje klass och sedan summerar. Testfunktionen är en chi-två-fördelning där antalet frihetsgrader beror på antalet klasser.

Låt oss göra ett exempel baserat på fotbollsdatan. Där hade vi utfallet 947, 518 och 1285 för bortaseger, oavgjort och hemmaseger. Klassisk bollkunskap ger oss sannolikheterna 33, 20 och 47 procent. Testets hypoteser ges av

H_0 : sannolikheterna för de olika utfallet ges av 33, 20 respektive 47 procent

H_1 : minst något utfall har en annan sannolikhet än 33, 20 respektive 47 procent

För att få de förväntade värdena E multipliceras nollhypotesens sannolikheter med det totala antalet matcher.

```
O <- c(947, 518, 1285)
E <- c(0.33, 0.20, 0.47) * 2750
```

Uppgift 4.6 (Granska E). Skriv ut objektet E och jämför med de observerade värdena. Notera att de förväntade värdena inte måste vara heltal, trots att de observerade värdena förstås alltid kommer vara det.

Testvärdet beräknas genom formeln för varje term följt av summan.

```
chisq_value <- sum((O - E)^2 / E)
```

P -värdet beräknas från en chi-två-fördelning. Antalet frihetsgrader ges av antalet klasser minus antalet skattade parametrar minus ett. I det här fallet har inga parametrar skattats från datan så antalet frihetsgrader blir två. Ett chi-två-test beräknas med kvadrater så vi är enbart intresserade av högra svansen.

```
ggplot() +
  geom_function(fun = dchisq, args = list(df = 2)) +
  geom_vline(xintercept = chisq_value) +
  xlim(0, 10)
```

Man kan också beräkna ytan i svansen med `pchisq()`. Ett minus det resultatet ger den övre delen.


```
1 - pchisq(chisq_value, df = 2)
```

P-värdet är alltså 0.16, över den klassiska signifikansnivån på 5 procent, vilket ger att vi inte kan förkasta nollhypotesen. Om man gör ett chi-två-test för hand jämför man det observerade chi-två-värdet med ett tabellvärde över kvantiler. Tabellvärdet kan också hämtas med funktionen `qchisq()`, i det här fallet

```
qchisq(0.95, df = 2)
```

Notera att man tar 0.95 eftersom man alltid tittar på den yttre svansen i ett chi-två-test.

R har en inbyggd funktion för chi-två-test. Dess argument ges av observerade antal och sannolikheter.

```
chisq.test(0, p = c(0.33, 0.2, 0.47))
```

Testet ger samma chi-två-värde och p-värde som beräknats ovan.

Situationen med flera klasser kan som sagt ses som en generalisering av fallet med två klasser. Det är alltså logiskt att chi-två-test kan användas även när man har två klasser. Följande exempel ger samma test som vi sett tidigare av andelen bortasegrar.

```
chisq.test(x = c(947, 1803), p = c(0.33, 0.67), correct = F)
```

Likt `prop.test()` sätter vi `correct` till `FALSE` för att inte göra en korrektion. Notera att `x` här anges som positiva och negativa utfall istället för positiva utfall och totalt antal utfall, vilket var fallet i `prop.test()`.

Chi-två-testet bygger på en underliggande normal-liknande approximation. En vanlig tumregel är att alla förväntade värden ska vara större än 5. R ger en varning om så inte är fallet.

```
chisq.test(c(6,4), p = c(0.51, 0.49))
```

Uppgift 4.7 (Chi-två med lika sannolikheter). En vanlig tillämpning av goodness-of-fit-testet är för att testa om alla klasser är lika sannolika. En jämn fördelning är grundinställning i `chisq.test()` så i det fallet behöver man bara ange de observerade värdena. En datainsamling om M&M-godis gav följande antal.

```
dat_mnm <- data.frame(Color = c("blue", "brown", "green",
                                "orange", "red", "yellow"),
                      Count = c(180, 80, 88, 160, 134, 166))

ggplot(dat_mnm, aes(Color, Count, fill = Color)) +
  geom_col() +
  scale_fill_manual(values = dat_mnm$Color)
```

Använd de observerade värdena i kolumnen `Count` för att testa om alla godisfärger är lika vanliga.

4.5 Bonus. Interaktiva kartor med leaflet

Paketet `leaflet` (<https://rstudio.github.io/leaflet/>) kopplar R till leaflet - ett verktyg för interaktiva kartor som ofta använd för kartor online.

Uppgift 4.8 (Installera leaflet). Installera och ladda `leaflet` genom att fylla i och köra raden nedan.

```
install.packages("leaflet")
library(leaflet)
```

Som en första kontroll kan vi köra den exempelkod som ges på hemsidan länkad till ovan.

```
m <- leaflet() %>%
  addTiles() %>%
  addMarkers(lng = 174.768, lat = -36.852,
             popup="The birthplace of R")
m
```

På canvassidan finns en excel fil med data tillgänglig på Artportalen - *Artportalen, feromoninventering, 2011-2013.xlsx*. Datan kommer från ett inventeringsprojekt vid SLU.

Uppgift 4.9 (Importera datan). Ladda ner excel filen och läs in datan med `read_excel()` från paketet `readxl`.

```
library(readxl)
dat_leaf <- read_excel("___")
```

Kartans utseende kan ändras genom att ange en källa och karttyp. Tillgängliga alternativ kan skrivas ut med `providers`.

```
leaflet() %>%
  addTiles() %>%
  addProviderTiles(providers$Stamen.Toner)
```

Uppgift 4.10 (Baskarta). Skriv ut tillgängliga baskartor med `providers`. Välj ett alternativ slumpmässigt och ändra koden ovan för att se hur det ser ut.

För att lägga till datapunkter kan man använda `addCircleMarkers()`.

```
leaflet() %>%
  addTiles() %>%
  addCircleMarkers(lng = dat_leaf$lng,
                  lat = dat_leaf$lat, radius = 10)
```

Uppgift 4.11 (Cirkelstorlek). Ändra storleken på cirkarna från `addCircleMarkers()` genom argumentet `radius`.

Slutligen kan vi lägga till en etikett för år med argumentet `popup`. Texten kommer upp när man klickar på en punkt.

```
leaflet() %>%  
  addTiles() %>%  
  addCircleMarkers(lng = dat_leaf$lng,  
                   lat = dat_leaf$lat, radius = 10,  
                   popup = dat_leaf$Rödlistade)
```

Uppgift 4.12 (Artnamn). Ändra i koden ovan för att ange artnamn som popup-text istället för rödlistestatus. Funktionen `paste()` kan också vara intressant för att ta med mer information i texten: `paste(dat_leaf$Rödlistade, dat_leaf$Antal)` skulle exempelvis ge både status och antal individer vid den specifika observationen.

Kapitel 5

Test för två stickprov

Datorövning 5 handlar om hypotestest och konfidensintervall för jämförelse av två stickprov. Efter övningen ska vi kunna genomföra och tolka

- t-test för jämförelse av två medelvärden,
- z-test för jämförelse av två proportioner,
- chi-två-test för jämförelse av fördelningar mellan två eller flera grupper,
- konfidensintervall för skillnaden mellan två medelvärden eller två proportioner.

5.1 Repetition av datorövning 4

När man startar en ny R-session bör man ladda de paket man vet kommer behövas med `library()`. Om paket inte finns installerade måste man först köra `install.packages()`.

```
# install.packages("tidyverse")  
library(tidyverse)
```

I datorövning 4 tittade vi på tester som inte bygger på normalfördelning: dels tester för proportioner, där varje observation är något av två möjliga utfall; dels tester för kategoridata, där varje observation har ett utfall i någon av flera kategorier.

Proportioner kan testas med ett z-test. Säg att man testat om en doft har en effekt på en insekt, man skickar 20 insekter i ett y-rör och 16 går mot doften. Hypoteserna ges av

- H_0 : proportionen i populationen är 0.5,
- H_1 : proportionen i populationen är inte 0.5.

I R genomförs ett z-test med `prop.test()`. Funktionen gör en korrektion som ger ett något bättre test än det man ofta beräknar för hand. Korrektionen kan sättas av med argumentet `correct`.

```
prop.test(x = 16, n = 20, p = 0.5, correct = F)
```

Ett lågt p-värde tyder på en signifikant skillnad från 0.5.

För att få samma konfidensintervall som det man beräknar för hand kan man använda `binom.asymp()` från paketet `binom`.

```
library(binom)
binom.asymp(16, 20)
```

Populationens proportion ligger med 95 procents konfidens mellan 0.62 och 0.98.

Om utfallen är mer än två kategorier kan en hypotes om datans fördelning testas med ett goodness-of-fit-test, vilket är ett slags chi-två-test. Testet bygger på att observerade antal (O) ställs mot förväntade antal (E). I R genomförs ett chi-två-test med `chisq.test()`. Säg som exempel att man studerar fågelpopulationer genom en observationsstudie. Observerade antal av fyra fågelarter är 102, 53, 75 och 12 och de förväntade andelarna av arterna är 50, 15, 25 och 10 procent. Testets hypoteser ges av

- H_0 : proportionerna följer den förväntade fördelningen,
- H_1 : proportionerna följer inte den förväntade fördelningen,

och testet ges av

```
chisq.test(c(102, 53, 75, 12), p = c(0.5, 0.15, 0.25, 0.10))
```

Ett lågt p-värde tyder på att de antagna proportionerna inte stämmer. Antalet frihetsgrader ges i ett chi-två-test av antalet klasser minus antalet skattade parametrar minus ett. I det här fallet finns fyra klasser och ingen skattad parameter.

5.2 Två stickprov och normalfördelad data

Vid normalfördelad data från två stickprov eller grupper vill vi nästan alltid testa om populationerna har samma medelvärde. Det kan också ses som att vi testar om differensen mellan medelvärdena är noll. Vi skiljer mellan två fall: *matchade stickprov* - där varje observation i den ena gruppen är *kopplad* till en observation i den andra gruppen; och *oberoende stickprov* - där det inte finns någon sådan koppling mellan stickproven. Typiska exempel på matchade stickprov är när man mäter samma individ för och efter en behandling och syskonstudier där ett syskon får en behandling och den andra en annan behandling.

5.2.1 t-test för två matchade stickprov

Vid matchade stickprov kan varje observation i en behandlingsgrupp paras med en observation i den andra gruppen. Själva testet är ett t-test för *ett* stickprov på differensserien beräknat från varje par. I R kan man antingen beräkna den differensserien eller använda `t.test()` med två dataserier och argumentet för parvisa observationer satt till sant, `paired = T`. Som exempel ges följande data från en studie på äpple, där trädhöjd mätts före och efter en näringsbehandling.

```
dat_apple <- tibble(Tree = 1:4,
  Before = c(48, 43, 30, 47),
  After = c(51, 44, 42, 54))
dat_apple
```

Datan kan illustreras med ett punktdiagram där en linje binder samman paret. För att enkelt skapa grafen i `ggplot2` kan man först omstrukturera datan till lång form genom `pivot_longer`.

```
dat_long <- dat_apple %>%
  pivot_longer(-Tree, names_to = "Time", values_to = "Height")
dat_long
```

Uppgift 5.1 (Äppelgraf). Fyll i kodstycket nedan för en graf av äppeldatan. Axlarna ges av `Time` och `Height`. Två observationer kan kopplas genom att sätta `Tree` som grupp.

```
ggplot(dat_long, aes(____, ____, group = ____)) +
  geom_point() +
  geom_line()
```

För att testa för skillnad före och efter behandling sätter vi upp hypoteser

- H_0 : μ före behandling är lika med μ efter behandling
- H_1 : μ före behandling är skild från μ efter behandling

Testet kan antingen utföras som ett enkelt t-test på differensserien

```
t.test(dat_apple$Before - dat_apple$After)
```

eller som ett t-test för två stickprov där man särskilt anger att datan är parad

```
t.test(dat_apple$Before, dat_apple$After, paired = T)
```

För bägge alternativen måste datan vara ordnad så att de två vektorerna matchar varandra parvis. Ett p-värde på 0.0987 ger att man inte förkastar vid en signifikansnivå på fem procent. Vi drar därmed slutsatsen att det inte finns någon signifikant skillnad före och efter behandling.

Uppgift 5.2 (Ensidigt test). Gör ett tillägg till ett av kodstyckena med `t.test()` för att beräkna ett ensidigt test med mothypotesen att träden ökar

i höjd efter behandling. Hjälpsidan för `t.test()` kan tas fram genom att köra `?t.test()`.

Konfidensintervallet beräknas från differenserna på samma sätt som vid ett stickprov med normalfördelad data. Tolkningen liknar den för ett stickprov: med 95 procents konfidens ligger den sanna skillnaden i medelvärden i intervallet.

5.2.2 t-test för två oberoende stickprov

Ett t-test för två oberoende stickprov testar om två populationsmedelvärden är lika. Ta som exempel följande data på jordgubbsskörd vid två olika näringsbehandlingar (A och B). Här är stickproven inte matchade - det finns **ingen** direkt koppling mellan en observation i den ena behandlingsgruppen till någon observation i den andra.

```
dat_berry <- data.frame(Behandling = c("A", "A", "A", "A",
                                       "B", "B", "B", "B"),
                       Vikt = c(40, 48.2, 39.2, 47.9,
                                57.5, 61.5, 58, 66.5))

dat_berry
```

Datan kan illustreras med ett enkelt punktdiagram.

```
ggplot(dat_berry, aes(Behandling, Vikt)) +
  geom_point()
```

Ett t-test för två oberoende stickprov har nollhypotesen att grupperna har samma populationsmedelvärde och alternativhypotesen att populationsmedelvärdena är skilda (för det tvåsidiga fallet):

- H_0 : μ_A är lika med μ_B
- H_1 : μ_A är ej lika med μ_B

Testet kan utföras i R genom funktionen `t.test()`. Data kan antingen anges som en formel med dess data `Vikt ~ Behandling`, `data = dat_berry` (vilket man kan läsa som *vikt uppdelat efter behandling*) eller som två skilda vektorer. Det förra alternativet är oftast enklare om man har datan på lång form - med en kolumn som anger grupp (i exemplet *Behandling*) och en kolumn som anger utfallsvärdet (i exemplet *Vikt*).

För formen med formel ger det

```
# Formelskrivning
t.test(Vikt ~ Behandling, data = dat_berry, var.equal = T)
```

och för formen med vektorer

```
# Två separata vektorer
## Filtrera ut data där behandling är A
Vikt_A <- dat_berry$Vikt[dat_berry$Behandling == "A"]
```


5.3. Z-TEST OCH KONFIDENSINTERVALL FÖR TVÅ PROPORTIONER 57

```
## Filtrera ut data där behandling är B
Vikt_B <- dat_berry$Vikt[dat_berry$Behandling == "B"]

t.test(Vikt_A, Vikt_B, var.equal = T)
```

Argumentet `var.equal = T` används för att beräkna testet där gruppernas varianser antas vara lika. Grundinställningen är testet där varianser inte antas vara lika, så `t.test(Vikt ~ Behandling, data = dat)` ger ett lite annat resultat.

Uppgift 5.3 (Ej lika varianser). Vilka resultatvärden ändras i utskriften om man sätter `var.equal = F`?

Testet ger ett p-värde på 0.0018, vilket leder till att nollhypotesen förkastas på enprocentsnivå. Detta tyder på att det finns en viktskillnad mellan behandlingarna. Utskriften ger också ett 95-procentigt konfidensintervall på $(-24.898, -9.202)$. Tolkningen är att *skillnaden* mellan populationsmedelvärden ligger i intervallet med 95 procents konfidens. Notera att värdet noll inte ligger i intervallet.

Om man har fler än två grupper kan man vilja göra parvisa t-test - alltså ett t-test för varje par av grupper. Ett exempel på funktionen `pairwise.t.test()` ges nedan. Funktionen bygger på att datan är i *lång* form, med en kolumn som anger det numeriska utfallet och en kolumn som anger behandlingen.

```
pairwise.t.test(dat_berry$Vikt, dat_berry$Behandling,
                p.adjust.method = "none", pool.sd = F)
```

```
##
## Pairwise comparisons using t tests with non-pooled SD
##
## data:  dat_berry$Vikt and dat_berry$Behandling
##
##      A
## B 0.002
##
## P value adjustment method: none
```

Matchade observationer kan också kallas *parade* (eng. paired) så se upp med terminologin. Funktionen `pairwise.t.test()` för *parvisa jämförelse* mellan behandlingar, men testerna är t-test för oberoende stickprov.

5.3 z-test och konfidensintervall för två proportioner

Om man vill jämföra två proportioner kan man använda z-testet för två stickprov. Säg till exempel att man har två sorter av någon planta och vill se hur stor proportion som är infekterad av bladmögel. I den ena gruppen (sort A) är 17 av

50 infektera och i den andra (sort B) är 26 av 60 infekterade. Testets hypoteser är i det tvåsidiga fallet

- H_0 : proportion A är lika med proportion B
- H_1 : proportion A är skild från proportion B

I R kan testet genomföras med `prop.test()`-funktionen. Funktionen första argument är antalen infekterade, som en vektor med två värden, och dess andra argument är totalerna. Likt testet med ett stickprov finns en möjlighet att göra en kontinuitetskorrektion med `correct`-argumentet. För att få samma resultat som räkning för hand anger vi att korrektion inte ska göras med `correct = F`.

```
prop.test(x = c(17, 26), n = c(50, 60), correct = F)
```

Notera att funktionen inte ger ett z-värde utan ett χ^2 -värde (utskrivet **X-squared**). Det beror på att funktionen beräknar z-testet som ett likvärdigt χ^2 -test. Det z-värde man får om man genomför testet som ett z-test är detsamma som roten ur utskriftens χ^2 -värde. Testet ger ett högt p-värde på 0.32 vilket innebär att nollhypotesen inte förkastas: det finns ingen signifikant skillnad i infektionsproportion.

Funktionen `prop.test()` ger också en utskrift av konfidensintervallet. Tolkning är att skillnaden i proportioner mellan populationerna ligger i intervallet med 95 procents konfidens. Notera att nollan ingår i intervallet.

Uppgift 5.4 (Burfågel). Det finns en förvånansvärt stor mängd studier på kopplingen mellan innehav av burfågel och lungcancer. En sådan studie (Kohlmeier et al 1992) ger följande antal för burfågeläggande och lungcancer.

```
dat_bird <- data.frame(Burfågel = c("Burfågel", "Ej_burfågel"),
                      Lungcancer = c(98, 141),
                      Ej_lungcancer = c(101, 328))
dat_bird
```

Datan tyder på att människor med burfågel har en förhöjd risk att drabbas av lungcancer. Genomför ett z-test för att se om andelen burfågelägare än densamma i de två patientgrupperna.

```
prop.test(x = c(____, ____), n = c(____, ____), correct = F)
```

5.4 Chi-två-test för korstabeller

Data med två kategoriska variabler kan presenteras med en korstabell. Ta som (ett något deppigt) exempel överlevnadsdata från Titanic. Datan finns tillgänglig i R som `Titanic`. I detta fall ges överlevnad filtrerad på vuxna män, uppdelat efter klass.

```
dat_titanic <- Titanic %>%
  data.frame() %>%
```

```
filter(Sex == "Male", Age == "Adult")
dat_titanic
```

En korstabell kan konstrueras med `pivot_wider`.

```
dat_wide <- dat_titanic %>%
  pivot_wider(names_from = Survived, values_from = Freq)
dat_wide
```

Datan tyder på att överlevnad är beroende av klass. Vi kan illustrera med uppdelade staplar.

```
ggplot(dat_titanic, aes(Class, Freq, fill = Survived)) +
  geom_col(position = position_fill(), color = "black") +
  scale_fill_manual(values = c("red4", "white"))
```

Argumentet `position` i `geom_bar` används för att skapa proportionella staplar.

Ett chi-två-test på en korstabell har nollhypotesen att det inte finns något samband mellan variabeln för rader och variabeln för kolumner. Chi-två-testet kan tillämpas på korstabeller med godtyckligt antal rader och kolumner. Antal frihetsgrader ges av antal rader minus ett gånger antal kolumner minus ett. Testet kan enkelt göras med `chisq.test()`. Som ingångsvärde kan man plocka ut kolumnerna med numeriska värden genom hakparenteser.

```
dat_wide[, 4:5] # De två numeriska kolumnerna

chisq.test(dat_wide[, 4:5])
```

Utskriften ger teststorheten, antal frihetsgrader, och p-värdet. I det här fallet är p-värdet mycket litet och slutsatsen blir att nollhypotesen förkastas - det finns ett samband mellan klass och överlevnad. Antalet frihetsgrader ges av antalet rader minus ett gånger antalet kolumner minus ett (här $(4 - 1) \cdot (2 - 1) = 3$).

Chi-två-testet är ett asymptotiskt test - dess egenskaper är beroende av *stora* stickprov. Som gräns för storleken används ofta att samtliga förväntade antal ska vara större än 5. Funktionen ger en varning om förväntade värden är små. En möjlig lösning i sådana fall är att slå ihop klasser.

```
test_result <- chisq.test(dat_wide[, 4:5])
test_result$expected # Samtliga förväntade värden över 5
```

Om detta krav inte är uppfyllt skriver funktionen ut en varning.

Uppgift 5.5 (Ogiltig approximation). Ta följande lilla korstabell och kör `chisq.test()` för att få ett felmeddelande.

```
dat <- matrix(c(4,2,5,1), 2)
dat
```

```
##      [,1] [,2]
```

```
## [1,]    4    5
## [2,]    2    1
```

Uppgift 5.6 (Burfågeln återvänder). En svensk studie på koppling mellan burfågel och lungcancer (Modigh et al, 1996) ger följande antal (för män).

```
dat_bird_swe <- data.frame(Burfågel = c("Burfågel", "Ej_burfågel"),
                           Lungcancer = c(108, 144),
                           Ej_lungcancer = c(171, 256))
dat_bird_swe
```

Genomför ett chi-två-test för att se om andelen cancerdrabbade är densamma i de två burfågelsgrupperna. Formulera tydliga hypoteser. För att få utfall som stämmer med en handräkning kan man sätta `correct = F`.

```
dat_bird_swe[, c(2,3)]
chisq.test(___, correct = F)
```

Kapitel 6

Variansanalys

Datorövning 6 handlar om variansanalys. Efter övningen ska vi kunna

- beräkna en anova-modell i R,
- ta fram och tolka en anova-tabell,
- göra lämpliga tester av modellantaganden,
- göra parvisa jämförelser mellan behandlingar.

6.1 Repetition av datorövning 5

När man startar en ny R-session bör man ladda de paket man vet kommer behövas med `library()`. Om paket inte finns installerade måste man först köra `install.packages()`.

```
# install.packages("tidyverse")  
library(tidyverse)
```

I datorövning 5 tittade vi på tester för två stickprov. För normalfördelad data kan man då använda ett t-test för två stickprov (för två matchade stickprov eller för två oberoende stickprov beroende på situation) för data med utfall i två eller flera kategorier kan man använda ett z-test för två stickprov eller ett chi-två-test för en korstabell.

Ett t-test för matchade stickprov används när de två grupper man jämför är matchade så att en observation i den ena gruppen är kopplad till en observation i den andra gruppen. Ett t-test för oberoende stickprov används om man inte har matchade stickprov, det vill säga då det inte finns någon koppling mellan behandlinggrupperna.

Ta som exempel följande fiskefångster för sex båtar från två regioner och två fiskearter.

```
dat_fish <- data.frame(
  Vessel = c("A", "B", "C", "D", "E", "F"),
  Region = c("N", "N", "N", "S", "S", "S"),
  Species1 = c(115.7, 98.5, 82.1, 89.2, 95.7, 99.4),
  Species2 = c(122.8, 105.3, 99.8, 106.8, 114, 102.7)
)
```

Vi vill här testa om det finns en skillnad mellan arter och om det finns skillnad mellan regioner.

```
dat_long <- dat_fish %>%
  pivot_longer(-c(Vessel, Region),
    names_to = "Species", values_to = "Catch")
ggplot(dat_long, aes(Species, Catch, group = Vessel)) +
  geom_point() +
  geom_line() +
  labs(title = "Fångster av två arter",
    subtitle = "Linje kopplar värden från samma fartyg")

ggplot(dat_fish, aes(Species1, Region)) +
  geom_point() +
  labs(title = "Fångster i två regioner")
```

För arterna har vi matchad data - varje observation av den ena arten är kopplad till en observation från den andra arten eftersom den kommer från samma båt - och vi kan testa om medelfångsterna av de två arterna är lika med ett t-test. Hypoteserna ges av

- H_0 : populationsmedelvärdet av fångster för art 1 är lika med det för art 2,
- H_1 : populationsmedelvärdena är ej lika.

I R kan ett test för matchad data genomföras med `t.test()` och argumentet `paired`, eller genom att beräkna differensen per båt och göra ett t-test för ett stickprov.

```
t.test(dat_fish$Species1, dat_fish$Species2, paired = T)

# Alternativt
# t.test(dat_fish$Species1 - dat_fish$Species2)
```

Det beräknade p-värdet ställs mot en signifikansnivå, vanligen fem procent, och om p-värdet är under signifikansnivån förkastar vi nollhypotesen. I det här exemplet tyder på värdet på att nollhypotesen inte stämmer - en art är vanligare än den andra.

För att jämföra regioner kan vi göra ett t-test för två oberoende stickprov. Hypoteser ges av

- H0: populationsmedelvärden av fångster är lika mellan regioner,
- H1: populationsmedelvärden av fångster är ej lika mellan regioner.

Testet kan genomföras med `t.test()` och kan antingen göras med ett antagande om lika varianser (vilket motsvarar det som görs för hand under kursen) eller utan det antagandet. Variabler kan anges med en formel som `Species1 ~ Group`, vilket vi kan tänka på som värden för art 1 uppdelat efter grupp.

```
t.test(Species1 ~ Region, data = dat_fish, var.equal = T)
```

Ett högt p-värde tyder på att det inte finns någon skillnad i fångst mellan regioner.

För data där utfallen är två eller flera kategorier kan ett chi-två-test testa om det finns något samband mellan två variabler. Följande data anger vilka partier ett urval väljare från tre kommuner planerar rösta på i nästa riksdagsval.

```
dat_parti <- data.frame(Kommun = c("Malmö", "Lund", "Kävlinge"),
                       S = c(54, 102, 40),
                       M = c(30, 98, 53),
                       MP = c(7, 50, 5))
dat_parti
```

För att testa om det finns något samband mellan kommun och parti sätter vi upp hypoteserna

- H0: det finns inget samband mellan parti och kommun (ingen skillnad mellan kommuner),
- H1: det finns något samband mellan parti och kommun.

Detta kan testas med ett chi-två-test med funktionen `chisq.test()`. Som argument ges den numeriska delen av korstabellen - vi tar alltså bort den första kolumnen för kommun.

```
chisq.test(dat_parti[, -1])
```

Det låga p-värdet på 0.000037 ger att vi förkastar nollhypotesen och drar slutsatsen att det finns ett samband mellan kommun och parti.

6.2 Allmänt

Variansanalys (eller *anova-modellen*) är en statistisk modell där medelvärdet varierar beroende på en behandling och ett normalfördelat slumpfel. Från en anova-modell kan man beräkna ett F-test, som testar om det finns någon övergripande gruppskillnad, och post-hoc-test, som jämför specifika grupper med varandra.

Den specifika modellen beror på försöksupplägget. Här ges exempel på variansanalys med en faktor, en faktor med block, och två faktorer.

6.3 Variansanalys. En faktor

Vid variansanalys med en faktor har man observationer av en kontinuerlig utfallsvariabel från två eller flera behandlingsgrupper. Som exempel används en datamängd på ett odlingsförsök med tre behandlingar (varav en kontroll). Exemplet finns tillgängligt i R som `PlantGrowth`.

`PlantGrowth`

Datan har 30 observationer av vikt `weight` och varje observation tillhör någon specifik behandling `group`. Datan kan illustreras med ett spridningsdiagram.

```
ggplot(PlantGrowth, aes(group, weight)) +  
  geom_point()
```

Behandling 1 verkar vara något lägre än kontrollen medan behandling 2 verkar vara något högre.

En anova-modell kan i R skattas med funktionen `lm()` (för *linjär modell*). Från modellobjektet kan man sedan plocka fram en anova-tabell (som bland annat anger utfallet av F-testet) och genomföra parvisa jämförelser genom `emmeans`.

```
mod <- lm(weight ~ group, data = PlantGrowth)
```

Modellen anges som en formel `weight ~ group`, vilket kan utläsas *vikt beroende på behandlingsgrupp*. Därefter anges data med argumentet `data`.

För anova-tabellen finns flera alternativ. Här används funktionen `Anova()` från paketet `car`.

```
library(car)  
Anova(mod)
```

Anova-tabellen ger kvadratsummor (`Sum Sq`), frihetsgrader (`Df`) och utfallet av ett F-test. Testets hypoteser ges av

H_0 : alla behandlingsgrupper har samma medelvärde H_1 : alla behandlingsgrupper har inte samma medelvärde

Det låga p-värdet tyder på att nollhypotesen bör förkastas, vilket alltså pekar på att det finns någon eller några skillnader i medelvärde.

Uppgift 6.1 (Anova för hand). Anovatabell från `Anova()` ger kvadratsummor och frihetsgrader. Använd den informationen för att, för hand, beräkna medelkvadratsummor och F-värdet.

Uppgift 6.2 (Tabellvärde för F-fördelningen). Anova-tabellen ger ett p-värde från vilket vi kan dra en direkt slutsats. Om man istället löser uppgiften för hand ställer man det beräknade F-värdet mot ett kritiskt värde från en tabell över F-fördelningen. Se efter om man kan hitta ett lämpligt tabellvärde för det aktuella testet (med 2 och 27 frihetsgrader). Det är möjligt att det inte finns en rad för 27 i en vanlig F-fördelningstabell, använd isåfall värdet på närmast övre

rad (t.ex. 26 eller 25). I R kan kvantiler för F-fördelningen tas fram med `qf()`, t.ex.

```
qf(0.95, 2, 27)
```

En naturlig följdfråga är vilka behandlingsgrupper som skiljer sig åt. För att besvara det krävs *parvisa jämförelser* där behandlingarna jämförs två och två. Parvisa jämförelse kan göras med paketet `emmeans` och funktionen med samma namn. Funktionen tar modellobjektet som första argument och en formel för jämförelsetyp som andra argument (här `pairwise ~ group`, en parvis jämförelse mellan nivåer i `group`).

```
# install.packages("emmeans")
library(emmeans)
emmeans(mod, pairwise ~ group)
```

I den nedre tabellen med jämförelser ges alla parvisa jämförelser. Nollhypotesen är att de två grupper som jämförs har samma medelvärde - ett lågt p-värde tyder alltså på att de två grupperna är signifikant skilda. Notera också att p-värden justeras med tukey-metoden, även känt som Tukeys HSD.

Om man istället vill använda Fishers LSD kan man styra justeringen med argumentet `adjust`.

```
emmeans(mod, pairwise ~ group, adjust = "none")
```

Parvisa jämförelser presenteras ofta med signifikansbokstäver (en. *compact letter display*, `cld`). Dessa kan plockas fram med `multcomp`-paketet och funktionen `cld()`.

```
em <- emmeans(mod, pairwise ~ group)

# install.packages("multcomp")
# install.packages("multcompView")
library(multcomp)
cld(em, Letters = letters)
```

Tolkning av grupperingen till höger är att grupper som delar en bokstav inte är signifikant skilda. I det här fallet är den lägsta nivån skild från de två högsta. I övrigt finns inga signifikanta skillnader. Jämför gärna med p-värdena från tabellen med parvisa jämförelser. Man bör se att parvisa jämförelser med ett p-värde under fem procent motsvaras av att de behandlingarna inte delar någon bokstav i bokstavstabellen.

Uppgift 6.3 (Anova med två behandlingar). Följande kod skapar en datamängd med två behandlingar.

```
dat_two <- PlantGrowth %>% filter(group %in% c("trt1", "trt2"))
```

Använd den datan för att göra ett t-test för två oberoende stickprov med lika

varians, ett t-test för två oberoende stickprov utan antagande om lika varians, och ett F-test (ofullständig exempelkod nedan). Vad kan sägas om p-värdena från de tre testen?

```
t.test(____ ~ group, data = dat_two, var.equal = T)
t.test(weight ~ ____, data = dat_two, var.equal = F)

mod <- lm(weight ~ group, data = dat_two)
Anova(mod)
```

Uppgift 6.4 (Mass-signifikans). Anledning till att vi justerar p-värden är att man vid varje test har en sannolikhet att förkasta. Om man gör ett stort antal tester är man nästan garanterad att få något (falskt) signifikant resultat. Justering höjer p-värdena för att minska den risken. Följande kod simulerar data med 5 grupper och producerar de parvisa jämförelserna.

```
n_groups <- 5
dat_sim <- expand_grid(obs = 1:10, group = letters[1:n_groups]) %>%
  mutate(y = rnorm(n()))
mod <- lm(y ~ group, dat_sim)
emmeans(mod, pairwise ~ group, adjust = "none")
```

Kör koden tio gånger. Hur många gånger av de tio ger de parvisa jämförelserna någon signifikant skillnad (det vill säga något p-värde under 0.05)?

Uppgift 6.5 (Äppelinfektionsimport). En studie har givit ett mått på infektion hos äppelträd. Fyra sorter jämförs med tre replikat per sort. Data finns i filen *Äppelangrepp* i excel-filen *Uppgiftsdata.xlsx* på canvassidan. Fyll i kodstycket nedan för att importera datan.

Uppgift 6.6 (Äppelinfektionsgraf). Fyll i kodstycket nedan för att skapa en graf av äppeldatan.

```
ggplot(____, aes(x = ____, y = ____)) +
  geom_point()
```

Uppgift 6.7 (Äppelinfektionsmodell). Fyll i kodstycket nedan för att skatta en anovamodell och ta fram anovatabellen. Vad är F-testets noll- och alternativhypotes? Vilken slutsats kan man dra från testet?

```
mod <- lm(____ ~ ____, data = dat_apple)
Anova(mod)
```

6.4 Variansanalys. En faktor med block

I en blockdesign delas försöksobjekten (de enheter man ger en behandling och sedan mäter, t.ex. en försöksruta eller en planta) in i grupper av lika objekt (ett *block*). Sedan ger man enheterna inom blocket varsin behandling. Blockförsök är ofta balanserade, så att varje behandling förekommer en gång i varje block.

Som exempel på ett blockförsök kan vi titta på datan `oats` från paketet `MASS`. Datan kommer från ett agrikulturellt försök och blockdesignen sker genom att man delar in ett fält i flera delar (blocken) och sätter varje behandling i varje block. Datan har två faktorer (kväve `N` och sort `V`), men låt oss i den här första delen titta på en specifik sort.

```
library(MASS)
oats_marvel <- oats %>% filter(V == "Marvellous")
oats_marvel
```

En vanlig illustration av ett blockförsök är ett punktdiagram kombinerat med ett linjediagram.

```
ggplot(oats_marvel, aes(N, Y, color = B, group = B)) +
  geom_point(size = 4) +
  geom_line()
```

Färg och linje sammanbinder observationer från samma block. Det finns tecken på en blockeffekt: block I är nästan alltid högst och block V är nästan alltid lägst. Det finns också en tydlig behandlingseffekt i att högre kväve ger högre skörd.

Blockeffekten kan enkelt föras in i modellen genom att lägga till variabeln `B` i `lm`-funktionen. Anova-tabellen och parvisa jämförelser kan göras på samma sätt som tidigare. Resultaten påverkas av att modellen har en blockfaktor; man behöver vanligen inte ange det explicit.

```
mod_b1 <- lm(Y ~ N + B, data = oats_marvel)
Anova(mod_b1)
```

P-värdet från F-testet på variabeln `N` är nu klart mindre än tidigare. Detta beror på att en stor del av variationen kan förklaras med blockeffekten, vilket är tydligt i att blockeffekten också har ett litet p-värde i F-testet.

Det kan vara intressant att jämföra med modellen utan block.

```
mod_wo_block <- lm(Y ~ N, data = oats_marvel)
Anova(mod_wo_block)
```

Det som är residualens kvadratsumma i modellen utan block är i blockmodellen uppdelat i en blockeffekt och en residualterm. Eftersom F-testet bygger på en jämförelse mellan behandlingseffekten och residualtermen leder blockdesignen till starkare signifikans i blockmodellen. Å andra sidan kostar blockfaktorn frihetsgrader vilket ger oss ett svagare test. Effekten av att ta med ett block beror alltså på om det finns en verklig skillnad mellan blocken eller ej.

Vi kan gå vidare med att titta på parvisa jämförelser mellan kvävenivåer. Funktionen `emmeans()` och `cld()` fungerar som tidigare.

```
cld(emmeans(mod_bl, ~ N), Letters = letters)
```

Signifikansbokstäver anger att den lägsta nivån är skild från övriga och att den näst lägsta är skild från den högsta. Även här kan det vara intressant att jämföra med modellen utan block.

```
cld(emmeans(mod_wo_block, ~ N), Letters = letters)
```

Modellen utan block ger samma medelvärden `emmean` men större medelfel `SE` och färre signifikanta skillnader.

Uppgift 6.8 (Block med två behandlingar. Graf). Det minsta möjliga blocket är det med två behandlingar. Vi filtrerar havredatan för att den situationen.

```
dat_small_block <- oats %>%
  filter(V == "Marvellous", N %in% c("0.6cwt", "0.0cwt"))
dat_small_block
```

Fyll i stycket nedan för att skapa en graf med `N` på x-axeln, `Y` på y-axeln och en gruppering som länkar observationer från samma block.

```
ggplot(dat_small_block, aes(x = ___, y = ___, group = ___)) +
  geom_point() +
  geom_line()
```

Uppgift 6.9 (Block med två behandlingar. Test). Eftersom det är ett försök med en förklarande faktor och block kan man modellera det med den tidigare blockmodellen. Men eftersom man bara har två observationer per block kan man också se det som matchade stickprov, vilket kan lösas med ett t-test. Fyll i stycket nedan för att göra de två testen - utfallsvariabeln är skörd `Y` och den förklarande faktorn är kvävenivån `N`. Jämför resultaten.

```
mod <- lm(___ ~ ___ + B, data = dat_small_block)
Anova(mod)

t.test(___ ~ ___, data = dat_small_block, paired = ___)
```

6.5 Bonus. Två faktorer med block

Exempeldata på havre tar med två förklarande faktorer och ett block. Datan kan illustreras med ett punktdiagram där `facet_wrap` delar grafen efter sort.

```
ggplot(oats, aes(N, Y, color = B)) +
  geom_point(size = 4) +
  facet_wrap(~ V)
```

Grafen visar samma kvävesamband som tidigare. Det finns inga tydliga skillnader mellan sorter, möjligen har sorten Victory givit något lägre skörd än övriga.

Det finns fortfarande en tydlig blockeffekt, till exempel har block I höga värden och block V låga värden.

Modellen skattas genom att lägga till variabeln för sort (V för variety) i `lm`-formeln. En modell med två faktorer kan antingen vara med eller utan en *interaktion*. Interaktionstermen fångar påverkan mellan faktorerna. Ett exempel hade varit om någon sort svarat starkare på ökad kväve än någon annan. Standardmodellen är att ta med interaktionen, vilket vi anger genom att sätta `N * V` istället för `N + V`. Blocket tas fortfarande med som en adderad faktor

```
mod_two_fact <- lm(Y ~ N * V + B, data = oats)
```

Anovatabellen kan plockas fram på samma sätt som tidigare.

```
Anova(mod_two_fact)
```

Raden N:V gäller interaktionseffekten mellan kväve och sort. I det här fallet är det ingen signifikant interaktion - vilket tyder på att sorterna svarar på kvävebehandling på liknande sätt. Samtliga huvudeffekter (raderna för N, V och B) är signifikanta. Kvadratsummorna och p-värdena tyder på att kväve förklarar mer av variationen än sort, vilket också är i linje med grafen ovan.

Vid flerfaktoriella försök kan man presentera parvisa jämförelser på flera olika sätt. Man kan ange huvudeffekter för en faktor utan att ange den andra faktorn, man kan ange medelvärden för samtliga kombinationer av två faktorer, och man kan ange medelvärden uppdelat efter nivåer i en annan faktor.

```
emmeans(mod_two_fact, ~ N)
emmeans(mod_two_fact, ~ N + V)
emmeans(mod_two_fact, ~ N | V)
```

Även här kan man göra jämförelser mellan nivåer genom att sätta `pairwise ~ N + V` eller beräkna signifikansbokstäver med `cld`. Följande kod jämför kvävenivåer *inom* sort.

Uppgift 6.10 (Sort uppdelat efter kvävenivå). Gör lämplig ändring i koden ovan för att jämföra sorter *inom* kvävenivå. Finns det några signifikanta skillnader?

Uppgift 6.11 (Interaktion med ett block). I modellen ovan är block en *additiv* faktor - den ingår inte i någon interaktionseffekt. Vad händer med testerna om man skattar modellen där samtliga interaktioner tas med? Varför?

```
mod_two_fact <- lm(Y ~ N * V * B, data = oats)
```


Kapitel 7

Regression och korrelation

Datorövning 7 handlar om regression och korrelation. Efter övningen ska vi kunna

- skatta en regressionsmodell i R,
- testa parametrar i modellen med F-test och t-test,
- göra lämpliga tester av modellantaganden,
- beräkna och tolka korrelationen mellan två variabler.

7.1 Repetition av datorövning 6

När man startar en ny R-session bör man ladda de paket man vet kommer behövas med `library()`. Om paket inte finns installerade måste man först köra `install.packages()`.

```
# install.packages("tidyverse")  
library(tidyverse)
```

I datorövning 6 tittade vi på variansanalys - en metod som gör det möjligt att utvidga t-testet för två grupper till ett godtyckligt antal grupper eller kombinationer av faktorer. I variansanalys skattar man en modell som förklarar ett datautfall. Utifrån modellen sätter man upp en anova-tabell som delar upp den totala variansen i en förklarad del och en kvarvarande residualdel. Anova-tabell ger också ett F-test som testar om det finns några skillnader mellan grupper. Från en skattad modell kan man sedan göra parvisa jämförelser mellan specifika grupper och testa modellantaganden (främst antagande om normalfördelning och lika varians inom grupper).

Ta som exempel följande data på tandtillväxt (`len`) hos marsvin under C-vitaminbehandling i olika doser (`dose`) och två olika metoder (`supp`), tillgängligt i R som objektet `ToothGrowth`.

```

ToothGrowth <- ToothGrowth %>% mutate(dose = as.character(dose))

ggplot(ToothGrowth, aes(len, supp, fill = dose)) +
  geom_boxplot()

```

Ett lådagram visar en klar skillnad mellan doser och en svagare skillnad mellan metoder. Det finns också tecken på att metoderna svarar olika på dos i att metoden *VC* ligger lägre än *OJ* vid de låga doserna men över (eller iallafall lika) vid den höga dosen.

En envägsanova-modell (en modell med en faktor) kan skattas med `lm()` och en anovatabell kan tas fram med `Anova()` från paketet `car`.

```

mod <- lm(len ~ dose, data = ToothGrowth)

library(car)
Anova(mod)

```

F-testets nollhypotes är att alla grupper (här alla doser) har samma populationsmedelvärde. Det låga p-värdet pekar på en klar skillnad mellan doser.

En anovamodell kan användas för parvisa jämförelse, vilket ibland kallas post-hoc-test. Den vanligaste är Tukey-testet, men andra tester kan också förekomma. Testet kan utföras med `emmeans()` från paketet med samma namn. Följande ger en jämförelse mellan doser.

```

library(emmeans)
emmeans(mod, pairwise ~ dose)

```

7.2 Regression

I en regression modelleras en numerisk variabel som en funktion av en annan numerisk variabel. Vid enkel linjär regression finns *en* sådan *förklarande variabel* och förhållandet mellan variablerna antas vara linjärt.

Ta som exempel data på förväntad medellivslängd och bnp per capita. Datan hämtas från `gapminder`-paketet. Paketet `ggrepel` och funktionen `geom_text_repel()` kan användas för att sätta punktetiketter som inte överlappar. För enklare tolkning av modellen transformeras bnp per capita till att vara i tusen dollar, snarare än dollar.

```

library(gapminder)
dat_eu07 <- gapminder %>%
  filter(year == 2007, continent == "Europe") %>%
  mutate(gdpPercap = gdpPercap / 1000)

library(ggrepel)
ggplot(dat_eu07, aes(gdpPercap, lifeExp)) +

```



```
geom_point() +
geom_text_repel(aes(label = country), size = 3)
```

Datan visar ett positivt samband mellan variablerna - högre bnp per capita är kopplat till högre medellivslängd.

Uppgift 7.1 (Data för 1957). Vad måste ändras i stycket nedan för att plocka ut data och göra en graf för Europa 1957?

```
dat_eu57 <- gapminder %>%
  filter(year == 2007, continent == "Europe") %>%
  mutate(gdpPercap = gdpPercap / 1000)

ggplot(dat_eu57, aes(gdpPercap, lifeExp)) +
  geom_point() +
  geom_text_repel(aes(label = country), size = 3)
```

En regressionmodell kan i R skattas med `lm()`-funktionen. Syntaxen är väldigt lik den för anovamodellen, men istället för en faktor som förklarande variabel används nu en kontinuerlig variabel.

```
mod <- lm(lifeExp ~ gdpPercap, data = dat_eu07)
summary(mod)
```

Funktionen `summary` ger en sammanfattning av modellen. Skattningen av modellens konstanta parameter ges som raden (**Intercept**) och dess tolkning är som förväntat värde i medellivslängd om bnp per capita är noll. Det är ofta lutningsparametern som är mer intressant. Skattningen av lutningsparametern ges på den rad som har samma namn som den förklarande variabeln, här `gdpPercap`. Den skattade parametern är 0.2146. Lutningsparametern har den generella tolkning som ökningen i y-variabeln när x-variabeln ökar med 1. I det här fallet ger 0.2146 att ett lands medellivslängd ökar med ungefär 0.2146 år (eller 78 dagar) när bnp per capita ökar med 1000 dollar.

Uppgift 7.2 (Modell för 1957). Skatta samma modell som ovan, denna gång med data från 1957. Tolka lutningsparametern i ord. Är effekten av ökad bnp större 2007 än den var 1957?

Man kan enkelt rita ut regressionlinjen i en graf med `geom_smooth()` och argumentet `method` satt till `lm`.

```
ggplot(dat_eu07, aes(gdpPercap, lifeExp)) +
  geom_point() +
  geom_text_repel(aes(label = country), size = 3) +
  geom_smooth(method = lm)
```

Den blå linjen illustrerar regressionlinjen $72.27 + 0.2146x$. Det grå bandet kring linjen är ett konfidensintervall för skattningen av y-variabeln.

Uppgift 7.3 (Graf för 1957). Använd `geom_smooth(method = lm)` för att lägga till en regressionslinje för data för 1957. Hur mycket påverkar de två avvikande länderna?

Utskriften från `summary` ger också tester av parametrarna (den högra kolumnen `Pr(>|t|)` ger p-värdet för ett test där nollhypotesen är att populationsparametern är noll). I det här fallet är både intercept och lutning skilda från noll. Motsvarande F-test för lutningen kan tas fram med en anova-tabell.

```
library(car)
Anova(mod)
```

Testerna av en regressionsmodell bygger på ett normalfördelningsantagande och ett antagande om *homoskedasticitet* (lika varians i y oavsett position på x-axeln). Antagandena kan undersökas genom att titta på skattningens *residualer* - skillnaden mellan det faktiska y-värdet och modellens värde. Residualerna kan undersökas med ett histogram eller en QQ-plot. En annan vanlig diagnosplot är ett spridningsdiagram med skattade värden på x-axeln och residualerna på y-axeln.

```
dat_eu07 <- dat_eu07 %>%
  mutate(Residualer = residuals(mod),
         Skattade = fitted(mod))

ggplot(dat_eu07, aes(sample = Residualer)) +
  geom_qq() + geom_qq_line()

ggplot(dat_eu07, aes(Skattade, Residualer)) + geom_point()
```

Om data följer en normalfördelning bör histogrammet visa en ungefärlig normalkurva, QQ-plotten bör visa punkter på den diagonala linjen och spridningsdiagrammet bör visa en slumpmässig spridning av punkter. Graferna pekar i det här fallet inte på några tydliga avvikelser från normalfördelningsantagandet, möjligen pekar QQ-plotten på mindre spridning i svansarna än en teoretisk normalfördelning.

Uppgift 7.4 (Icke-linjära samband). Låt oss titta på hela gapminder-datan för 2007.

```
dat_2007 <- gapminder %>% filter(year == 2007)
ggplot(dat_2007, aes(gdpPercap, lifeExp)) + geom_point()
```

Hur ser sambandet mellan bnp och medellivslängd ut? Vad skulle vara problematiskt med simpel linjär regression i det här fallet? När vi tittade på normalfördelningen sa vi att man ofta kan logaritmera en variabeln och få *bättre* egenskaper. Vad ska ändras i koden ovan för att använda logaritmerad `gdpPercap` istället för den ursprungliga variabeln? Är det sambandet mer linjärt?

Uppgift 7.5 (Log-transformerad data). Vad ska ändras i koden nedan för att använda logaritmerad `gdpPercap` istället för den ursprungliga variabeln? Är det sambandet mer linjärt?

```
dat_2007 <- gapminder %>% filter(year == 2007)
ggplot(dat_2007, aes(gdpPercap, lifeExp)) + geom_point()
```

7.3 Korrelation

Korrelation ger ett mått mellan -1 och 1 på hur väl två variabler samvarierar. En korrelation över noll tyder på ett positivt samband mellan variablerna - en observation med ett högt värde i den ena variabeln har också ett högt värde på den andra - medan en korrelation under noll tyder på ett negativt samband. I R kan korrelation beräknas med `cor()` och två variabler som första och andra argument. Funktionen `cor.test()` ger ett test där nollhypotesen är att korrelationen är noll.

```
cor(dat_eu07$lifeExp, dat_eu07$gdpPercap)
cor.test(dat_eu07$lifeExp, dat_eu07$gdpPercap)
```

Medellivslängd och bnp per capita har en stark positiv korrelation på 0.85 och den korrelation är signifikant skild från noll ($p < 0.001$). Notera att p-värdet är detsamma som för lutningsparametern i regressionen.

Uppgift 7.6 (Korrelationsmatris). Om man har fler än två variabler sammanfattas korrelationer ofta med en korrelationsmatris.

```
dat_eu07[, 4:6]
cor(dat_eu07[, 4:6])
```

Vad är korrelationen mellan befolkningsstorlek och bnp per capita?

Uppgift 7.7 (Anscombes data). Den raka regressionslinjen eller det enkla korrelationsmättet säger lite om hur data egentligen ser ut. En vanlig illustration av detta är *Anscombes kvartett*, fyra exempel konstruerade av den brittiske statistikern Francis Anscombe 1973. Datan finns tillgänglig i R som datasetet `anscombe`.

```
anscombe
```

Plotta de fyra graferna (`x1` paras med `y1` och så vidare) i spridningsdiagram och beräkna korrelation för varje par. Ett exempel ges för den första mängden nedan. Kommentera utfallet.

```
ggplot(anscombe, aes(x1, y1)) + geom_point()
cor(anscombe$x1, anscombe$y1)
```

För en modern variant av Anscombes data kan man titta på *The Datasaurus Dozen*. Det är också en serie konstruerade variabler som har liknande egenskaper sett till medelvärden, standardavvikelser och korrelationer, men skiljer sig åt om du ritas i en graf.