Statistik för Biologer
F1: Introduktion till Statistik och Sannolikhet

Shaobo Jin

Matematiska institutionen

Välkomna till Statistikdelen av Kursen!

Lärare:

1. Shaobo Jin: shaobo.jin@math.uu.se
2. Martin Andersson: martin.andersson@math.uu.se

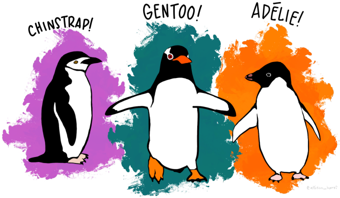
Undervisning i statistikdelen

1. 10 föreläsningar (Shaobo)
2. 5 obligatoriska datorövningar med R
3. 5 lektioner (Martin)
4. En frågestund inför tentan

Vi Börjar Med Pingvin!

Artwork by @allison_horst

Det finns tre arter av pingviner i vår dataset.

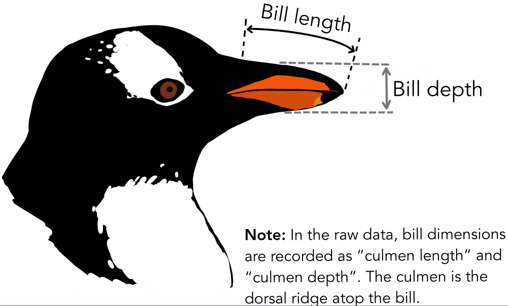



```
library(palmerpenguins)
data(penguins, package =
'palmerpenguins')
```

Mätningar

Forskarna har mätt bland annat

1. Näbbens längd (mm)
2. Näbbens djup (mm)
3. Vingens längd (mm)
4. Vikt (g)
5. Kön (hona/hane)
6. Art (tre arter)

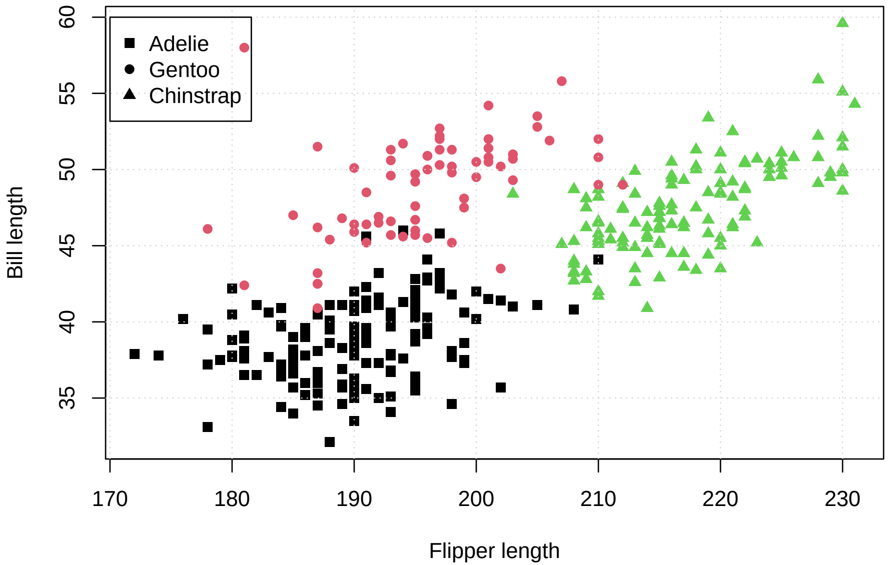


Visualisering: Spridningsdiagram (Scatter Plot)

Alternativ 1: funktionen plot()

```r
pch <- as.numeric(penguins$species)
pch[pch == 1] <- 15
pch[pch == 2] <- 16
pch[pch == 3] <- 17
plot(penguins$flipper_length_mm, penguins$bill_length_mm,
     col = penguins$species, pch = pch,
     xlab = "Flipper length", ylab = "Bill length")
grid()
legend(170, 60, legend = c("Adelie", "Gentoo", "Chinstrap"),
       pch = c(15, 16, 17))
```

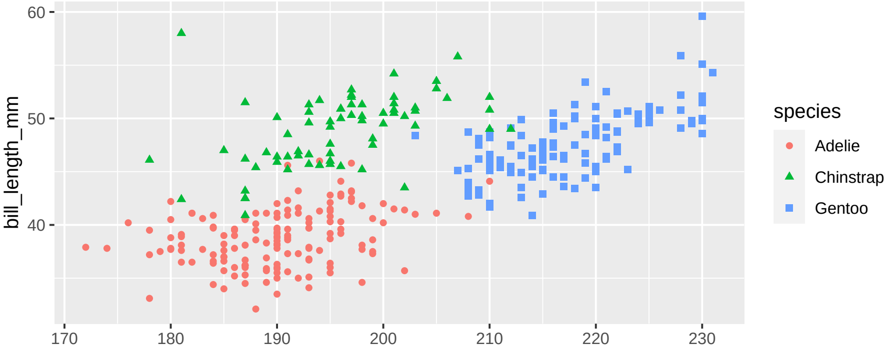
Visualisering: Spridningsdiagram (Scatter Plot)



Visualisering: Spridningsdiagram (Scatter Plot)

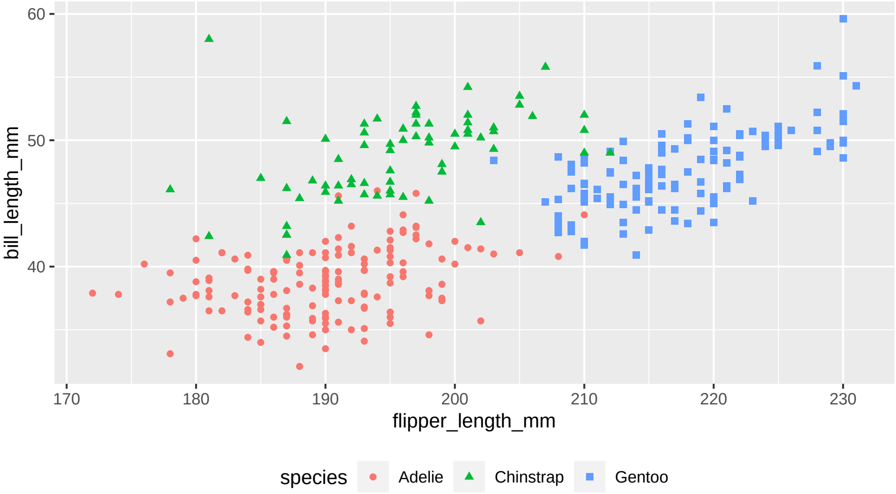
Alternativ 2: funktionen ggplot() av paketet ggplot2()

```
library(ggplot2)
ggplot(penguins, aes(flipper_length_mm, bill_length_mm,
                     color = species, shape = species)) +
geom_point()
```



Forskningsfråga

Finns storleksskillnader mellan olika arter?



Medelvärden

Ett sätt att jämföra storleken för olika grupper är att räkna ut medelvärden (sample mean or mean) för respektiv grupp. Det ger en bild av hur den genomsnittlig längden i gruppen ser ut.

Om x_1, x_2, ..., x_n är våra n mätvärden ges medelvärdet av

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Räkna ut medelvärdet

Våra näbbens längder är 39.1, 39.5, 40.3, 36.7, 39.3. Medelvärdet är

$$\bar{x} = \frac{1}{5}(39.1 + 39.5 + 40.3 + 36.7 + 39.3) = 38.98.$$

Medelvärden av Arter: Näbbens Längd

```
aggregate(bill_length_mm ~ species,
          data = penguins, # Namn av dataset
          FUN = mean) # Rakna ut mean

##     species bill_length_mm
## 1    Adelie       38.79139
## 2 Chinstrap       48.83382
## 3    Gentoo       47.50488
```

Medelvärden av Arter och Kön: Näbbens Längd

```
aggregate(bill_length_mm ~ species + sex,
          data = penguins, # Namn av dataset
          FUN = mean) # Rakna ut mean

##     species    sex bill_length_mm
## 1    Adelie female       37.25753
## 2 Chinstrap female       46.57353
## 3    Gentoo female       45.56379
## 4    Adelie   male       40.39041
## 5 Chinstrap   male       51.09412
## 6    Gentoo   male       49.47377
```

Median

Medianen är det mittersta värdet då observationerna sorteras i storleksordning.

- Om antalet observationer är jämnt är medianen medelvärdet av de två mittersta observationerna.

Räkna ut medianen

Våra näbbens längder är 39.1, 39.5, 40.3, 36.7, 39.3.

1. Längderna sorteras i storleksordning: 36.7, 39.1, **39.3**, 39.5, 40.3.
2. Median är 39.3.
3. Medelvärdet behöver inte vara samma som medianen!
    1. Medelvärdet är 38.98.

Median av Arter och Kön: Näbbens Längd

```r
aggregate(bill_length_mm ~ species + sex,
          data = penguins, # Namn av dataset
          FUN = median) # Rakna ut median
```

```
##     species    sex bill_length_mm
## 1    Adelie female          37.00
## 2 Chinstrap female          46.30
## 3    Gentoo female          45.50
## 4    Adelie   male          40.60
## 5 Chinstrap   male          50.95
## 6    Gentoo   male          49.50
```

Spridningen

Varken medelvärdet eller medianen ger en fullständig bild. Det är också intressant att ha mått på hur stor spridningen är:

1. Variationsbredd (range): det största värdet minus det minsta värdet:
$$\max(x_1, x_2, ..., x_n) - \min(x_1, x_2, ..., x_n).$$
2. Varians (variance): ett mått på hur mycket mätdata avviker från medelvärdet $\bar{x}$,
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$
3. Standardavvikelse (standard deviation, sd): kvadratroten ur variansen,
$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Variationsbredder av Arter och Kön: Näbbens Längd

```
aggregate(bill_length_mm ~ species + sex,
          data = penguins, # Namn av dataset
          FUN = range) # Rakna ut range

##     species    sex bill_length_mm.1 bill_length_mm.2
## 1    Adelie female             32.1             42.2
## 2 Chinstrap female             40.9             58.0
## 3    Gentoo female             40.9             50.5
## 4    Adelie   male             34.6             46.0
## 5 Chinstrap   male             48.5             55.8
## 6    Gentoo   male             44.4             59.6
```

Varianser av Arter och Kön: Näbbens Längd

```
aggregate(bill_length_mm ~ species + sex,
          data = penguins, # Namn av dataset
          FUN = var) # Rakna ut variance

##     species    sex bill_length_mm
## 1    Adelie female       4.116366
## 2 Chinstrap female       9.663824
## 3    Gentoo female       4.207613
## 4    Adelie   male       5.185323
## 5 Chinstrap   male       2.447843
## 6    Gentoo   male       7.401634
```

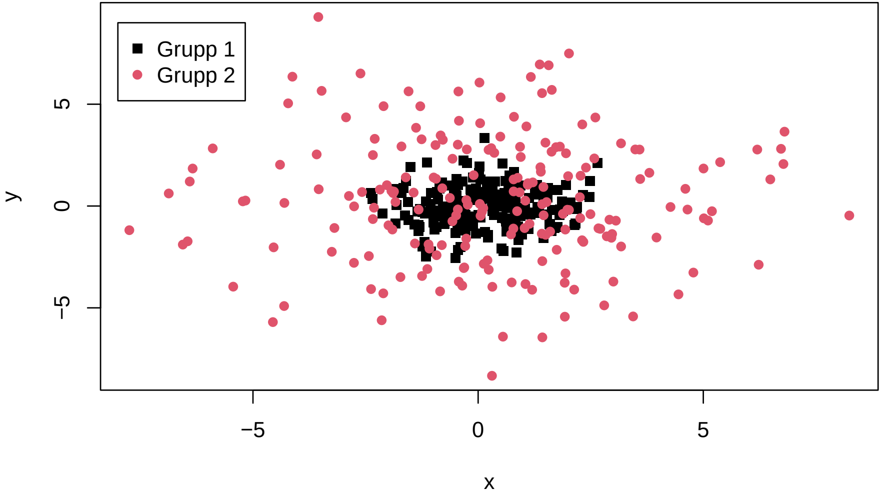
Standardavvikelser av Arter och Kön: Näbbens Längd

```
aggregate(bill_length_mm ~ species + sex,
          data = penguins, # Namn av dataset
          FUN = sd) # Rakna ut standard deviation

##     species    sex bill_length_mm
## 1    Adelie female       2.028883
## 2 Chinstrap female       3.108669
## 3    Gentoo female       2.051247
## 4    Adelie   male       2.277131
## 5 Chinstrap   male       1.564558
## 6    Gentoo   male       2.720594
```

Varför Spridningen?

Lika medelvärden men olika varianser.



Kvartil

1. Den undre kvartilen är medianen i den undre halvan av det ordnade materialet (inklusive medianen vid udda antal observationer).
2. Den övre kvartilen är medianen i den övre halvan av det ordnade materialet (inklusive medianen vid udda antal observationer).

Exempel med 8 observationer

Observationer: 1, 3, 7, 4, 5, 2, 6, 8

1. Sortera alla observationer i stigande storleksordning:

$$1, 2, 3, 4, 5, 6, 7, 8$$

2. Den undre halvan är 1, 2, 3, 4. Den undre kvartilen är $2.5 = (2+3)/2$.
3. Den övre halvan är 5, 6, 7, 8. Den övre kvartilen är $6.5 = (6+7)/2$.

Kvartil

1. Den undre kvartilen är medianen i den undre halvan av det ordnade materialet (inklusive medianen vid udda antal observationer).
2. Den övre kvartilen är medianen i den ovre halvan av det ordnade materialet (inklusive medianen vid udda antal observationer).

Exempel med 9 observationer

Observationer: 1, 3, 7, 4, 5, 2, 9, 6, 8

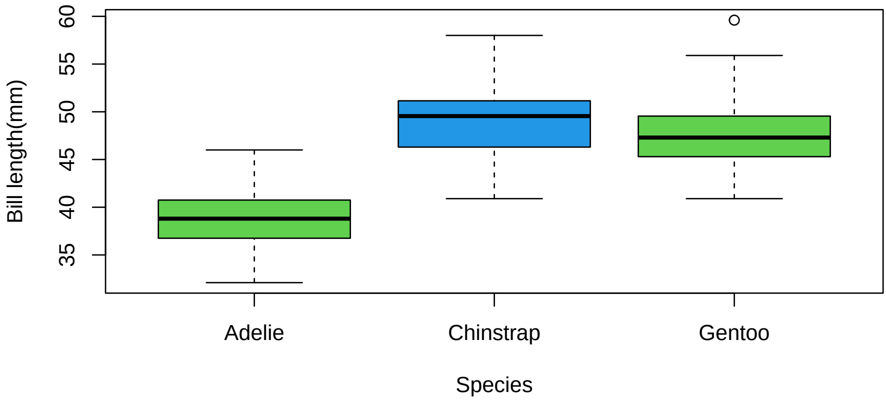
1. Sortera alla observationer i stigande storleksordning:

$$1, 2, 3, 4, 5, 6, 7, 8, 9$$

2. Den undre halvan är 1, 2, 3, 4, 5. Den undre kvartilen är 3.
3. Den övre halvan är 5, 6, 7, 8, 9. Den övre kvartilen är 7.

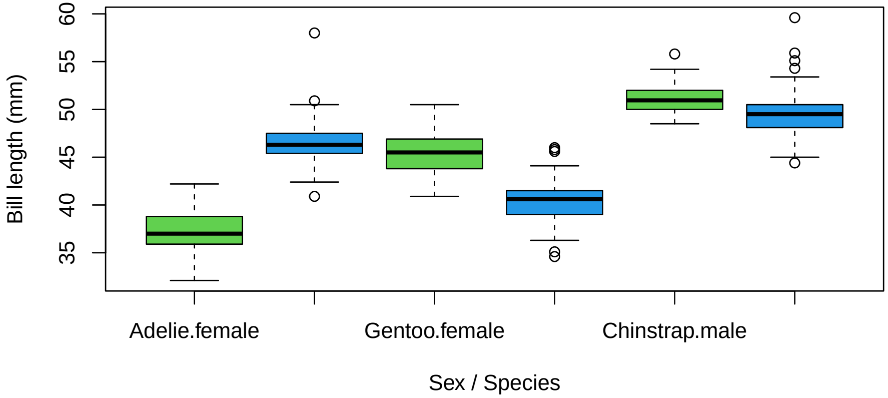
Visualisering: Lådagram (Boxplot)

```r
boxplot(bill_length_mm ~ species, data = penguins,
        xlab = "Species", ylab = "Bill length(mm)",
        col = rep(c(3, 4), 3)) # Color
```



Visualisering: Lådagram (Boxplot)

```r
boxplot(bill_length_mm ~ species + sex, data = penguins,
        xlab = "Sex / Species",  ylab = "Bill length (mm)",
        col = rep(c(3, 4), 3))
```



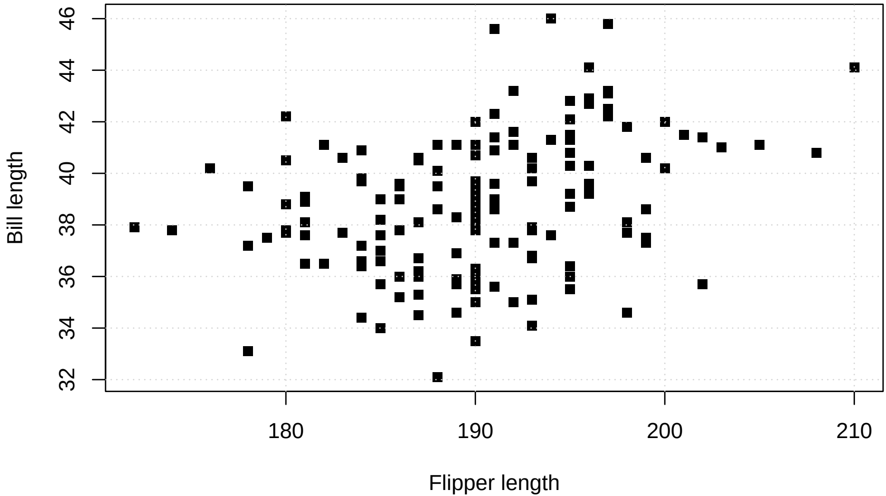
Slumpen

Finns skillnad i längd mellan olika arter?

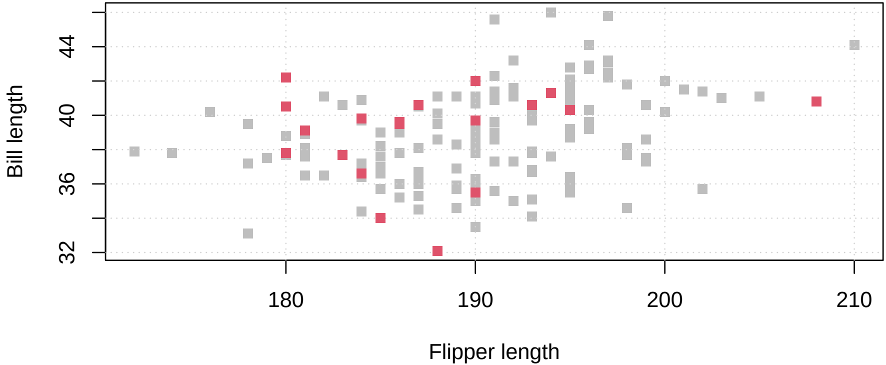
- Medelvärden och fina figurer är bra, men de som vetenskapliga bevis räcker inte!
- De skillnader vi ser skulle kunna bero på andra orsaker:
  1. Skillnaderna kan beror på störande faktorer
     - Mätningarna gjordes olika år.
     - Mätningarna gjordes vid olika boplatser.
  2. Slumpen kan ha påverkat resultatet. Forskarna råkade kanske välja tyngre hanar av ren slump.

Kunde det blivit annorlunda?

Låt oss anta att det finns totalt 152 Adeliepingviner. Men vi bara studerade 20 individer. Vad skulle medelvärdet bli?

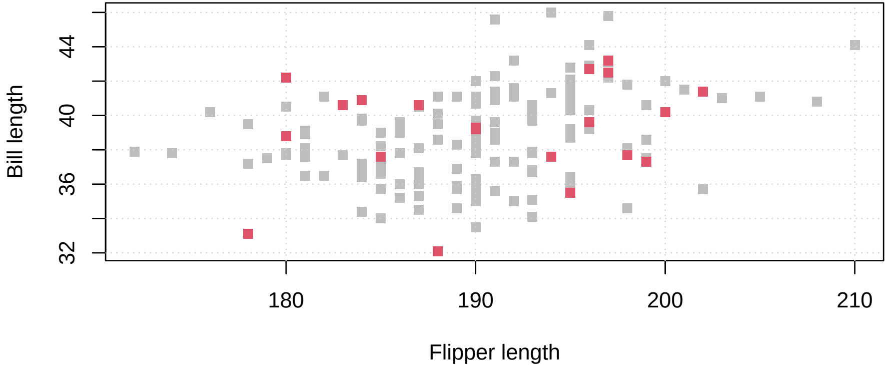


Första Urvalet



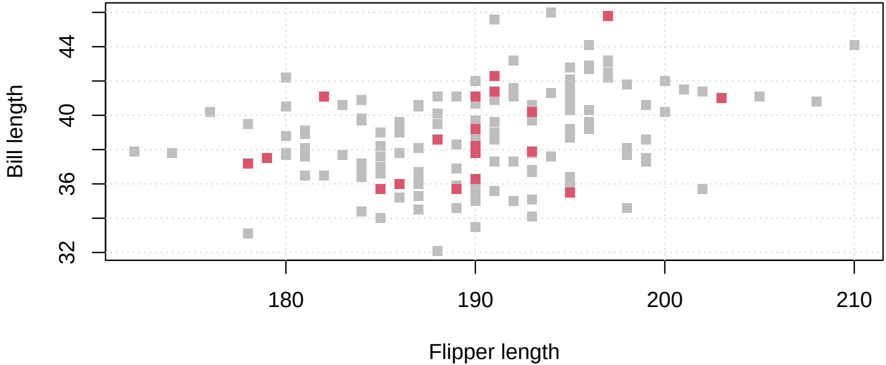

```
##   species    sex bill_length_mm
## 1  Adelie female       37.50000
## 2  Adelie   male       40.22222
```

Andra Urvalet




```
##   species    sex bill_length_mm
## 1  Adelie female       36.68333
## 2  Adelie   male       40.14286
```

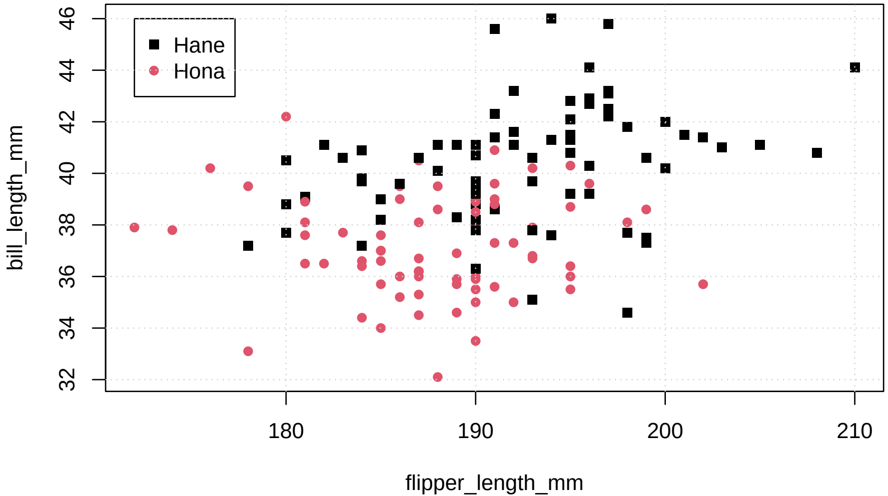
Tredje Urvalet



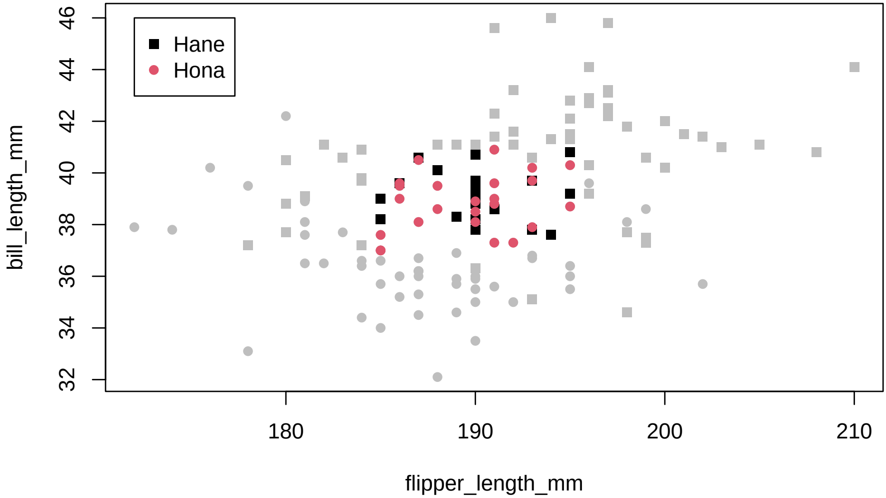

```
##   species    sex bill_length_mm
## 1  Adelie female       37.08571
## 2  Adelie   male       40.12727
```

Kunde det blivit annorlunda?

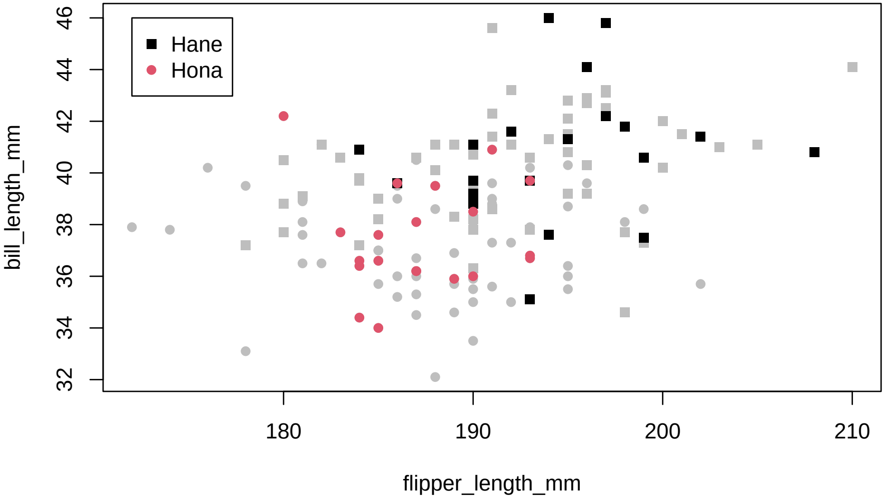
Nu studerar vi bara 20 hanar och 20 honor. Finns storleksskillnader mellan olika kön?



Första Urvalet



Andra Urvalet



Vad Ska Vi Göra?

Hur ska vi kunna känna oss säkra på att den skillnad beror på biologi och inte på slumpen?

1. Idé för att beskriva hur stor skillnaden är konfidensintervall: ange inte bara en punktskattning ("skillnaden är 51 mm") utan ett intervall som visar osäkerheten i skattningen ("skillnaden är 32-70 mm")
2. Idé för att få statistiskt säkerställda resultat är hypotesprövning

Vi kommer att studera båda!

Varför Sannolikhetslära?

Många fenomen är slumpmässiga till sin natur:

1. Antal deletioner i en DNA-sekvens under replikation
2. Vilka kromosomer ett barn ärver från sina föräldrar
3. Var pollen som sprids med vinden hamnar
4. Antal ögon när vi kaster en tärning.

Slumpens Matematik

Låt A vara en händelse. Sannolikheten för A skrivs $P(A)$

- $P()$ betyder probability.

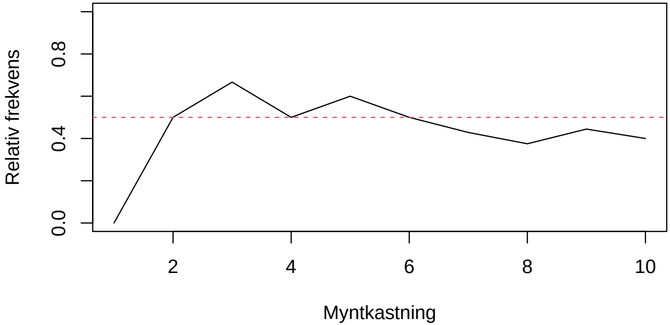
Händelser kan vara nästan vad som helst! Några exempel som vi ska studera i kusen:

1. Man slår en sexa med en tärning
2. Vid vägning av 20 pingviner blir genomsnittet mer än 4000 g

I många situationer vill man veta hur stor sannolikheten för en viss händelse är. Men sannolikheter är också användbara för att beskriva hur (o)säkra vi är på något.

Definition av Sannolikhet

När vi säger att "sannolikheten att man kastar en krona med ett mynt är 50%" menar vi att "den relativa frekvensen av kronor när man kastar ett mynt oändligt många gånger är 50%"



Kasta!

UNIVERSITEIT VAN AMSTERDAM

Zoek...

Vergelijk EN

Onderwijs Onderzoek Nieuws & Agenda Over de UvA Bibliotheek

Heads or Tails: Pure Chance?

18 oktober 2023

When you flip a coin, will it land more often on the same side it started? A well-known physics model suggests it will. Now, for the first time, scientists have gathered robust data to back up this hypothesis. They collected data from 350,757 coin tosses, including 12-hour coin-toss marathons. If you start with the head side up, the coin also more frequently ends up with the head side up.

Kolmogorovs Axiom

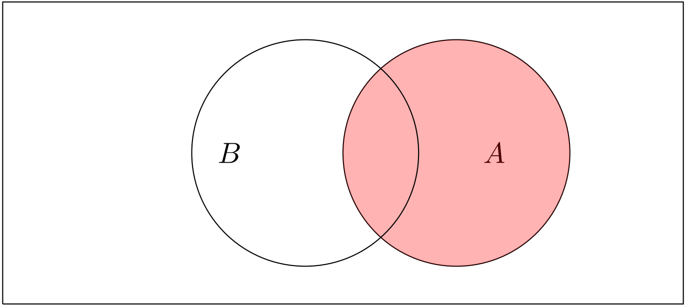
Sannolikheterna för två händelser A och B måste följa några grundläggande regler:

1. $P(A) \geq 0$: sannolikheter kan aldrig vara negativa.
2. $P(\text{minst en av alla möjliga händelser inträffar}) = 1$: något händer alltid!
3. Om A och B är oförenliga händelser, dvs. inte kan inträffa samtidigt, så gäller att

$$P(\text{minst en av A och B inträffar}) = P(A) + P(B).$$

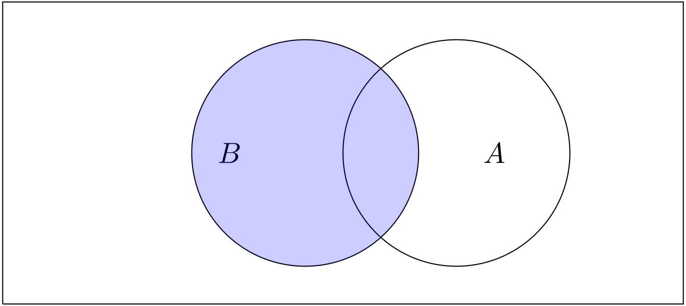
Venndiagram För Händelser

$P(A)$ är den röda ytan.



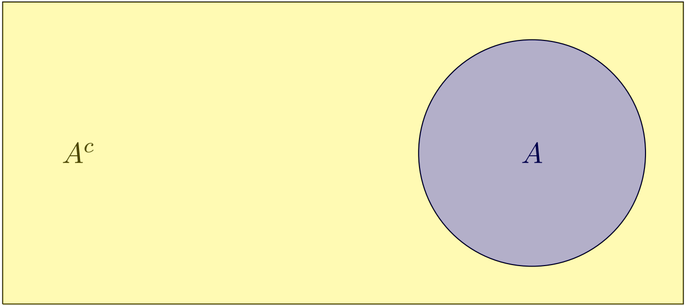
Venndiagram För Händelser

$P(B)$ är den blåa ytan.



Räkneregler för Sannolikheter: Komplementhändelse

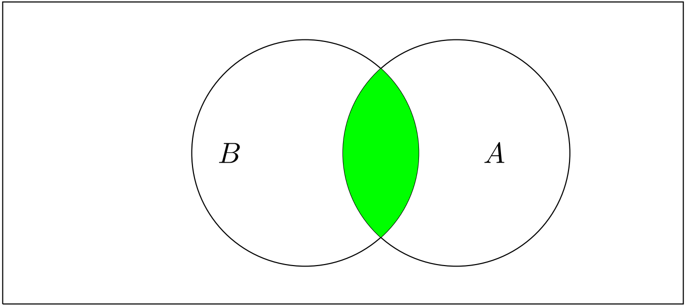
B är en komplementhändelse till A om B inträffar när A inte gör det. I så fall skrivs B som A^c (c =complement). I figuren är den gula ytan A^c. Då gäller att $P(A^c) = 1 - P(A)$.



Om $A = \{\text{Krona}\}$ vid myttkastning med $P(A) = 0.4$ har vi $P(A^c) = P(\text{Klave}) = 1 - P(\text{Krona}) = 1 - 0.4 = 0.6$.

Venndiagram För Händelser: Snitt

$P(A \cap B) = P(\text{både A och B inträffar})$ är den gröna ytan.



Räkneregler för Sannolikheter: Oberoende Händelser

A och B är oberoende om de inte påverkar varandra (om B har inträffat så ger det ingen information om A, och vice versa).

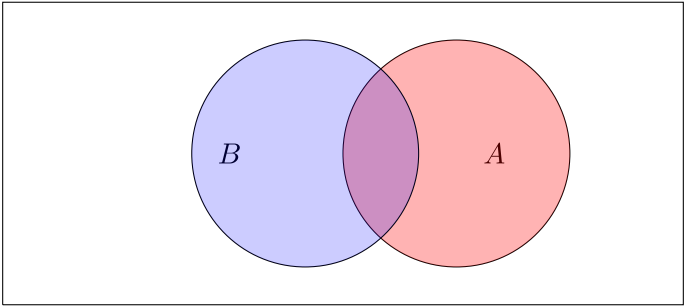
För oberoende händelser gäller att:

$$P(A \bigcap B) = P(\text{både A och B inträffar}) = P(A) \cdot P(B).$$

1. Exempel 1: A = "fadern får en flicka", B = "det kommer att snöa imorgon". Händelserna är oberoende och $P(A \bigcap B) = P(A) \cdot P(B)$.
2. Exempel 2: A = "den färgblinde fadern får en flicka", B = "barnet är färgblind". Händelserna är inte oberoende, och $P(A \bigcap B) \neq P(A) \cdot P(B)$!

Venndiagram För Händelser: Unioner

$P(A \cup B) = P(\text{minst en av A och B inträffar})$ är den färgade ytan.



Allmänt gäller för A och B (inte nödvändigtvis oförenliga) att

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Sammanfattning

1. Många fenomen (inte bara inom biologin) påverkas av slumpen.
2. Slump påverkar alla mätningar inom biologin.
   1. Händelse
   2. Frekvensbaserade sannolikheter
   3. Räkneregler: Unioner, snitt och komplement