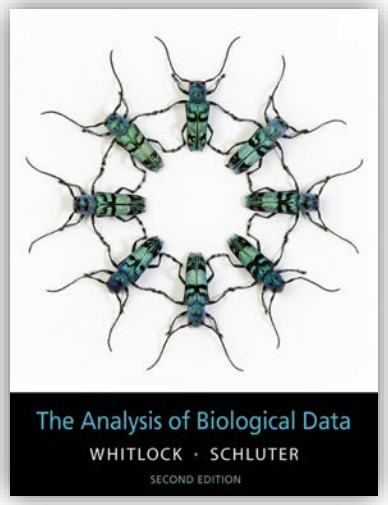
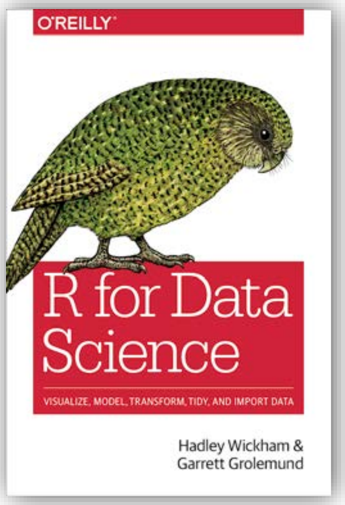
Data Science in Bioinformatics

week.02.remember.your.statistics.class

Palle Villesen & Thomas Bataillon





Outline for week 02

- Any questions from last week ?
- Basic statistics
  - Most data are samples
  - Describing data
    - Descriptive statistics of a sample
    - Displaying data
  - Sampling with uncertainty
    - Sampling distributions

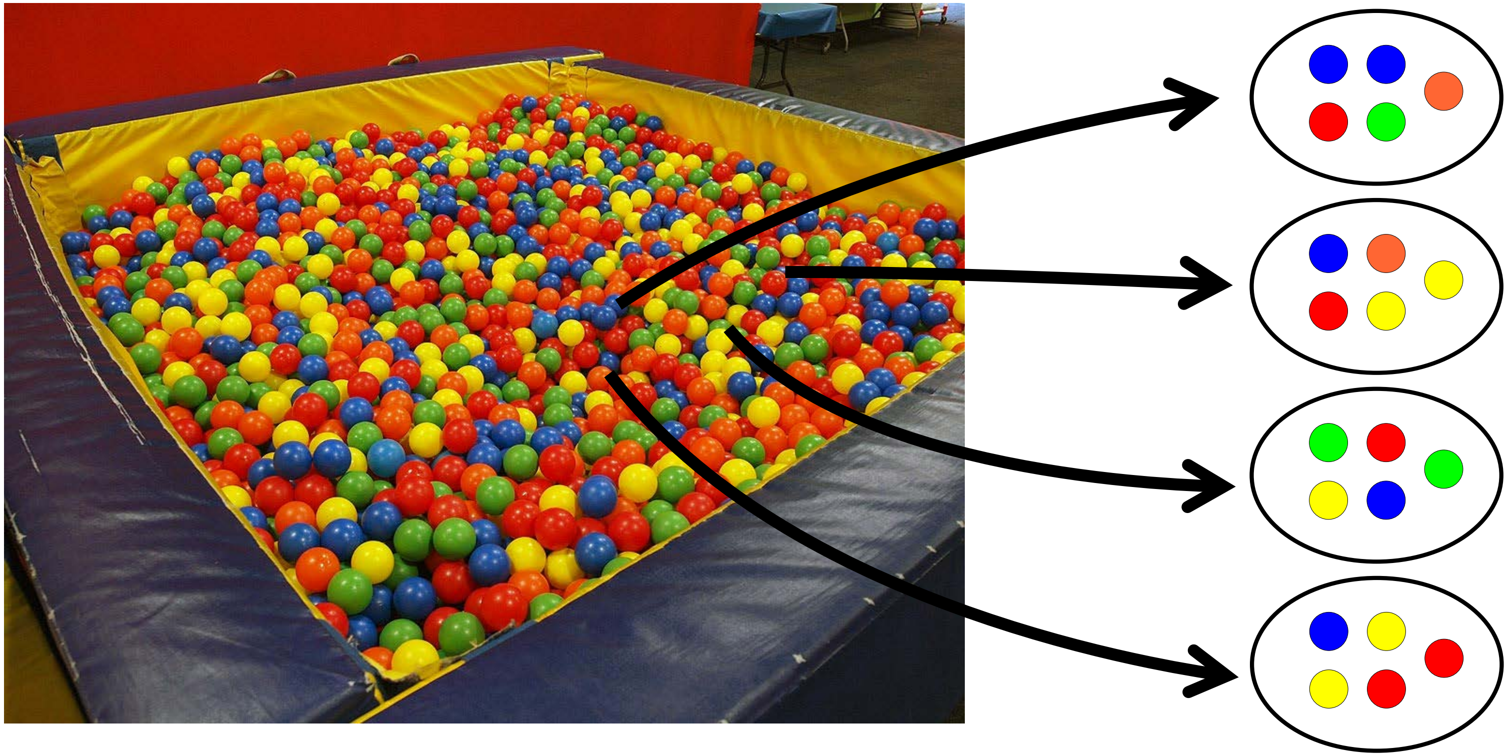
Population VS sample

UNKNOWN

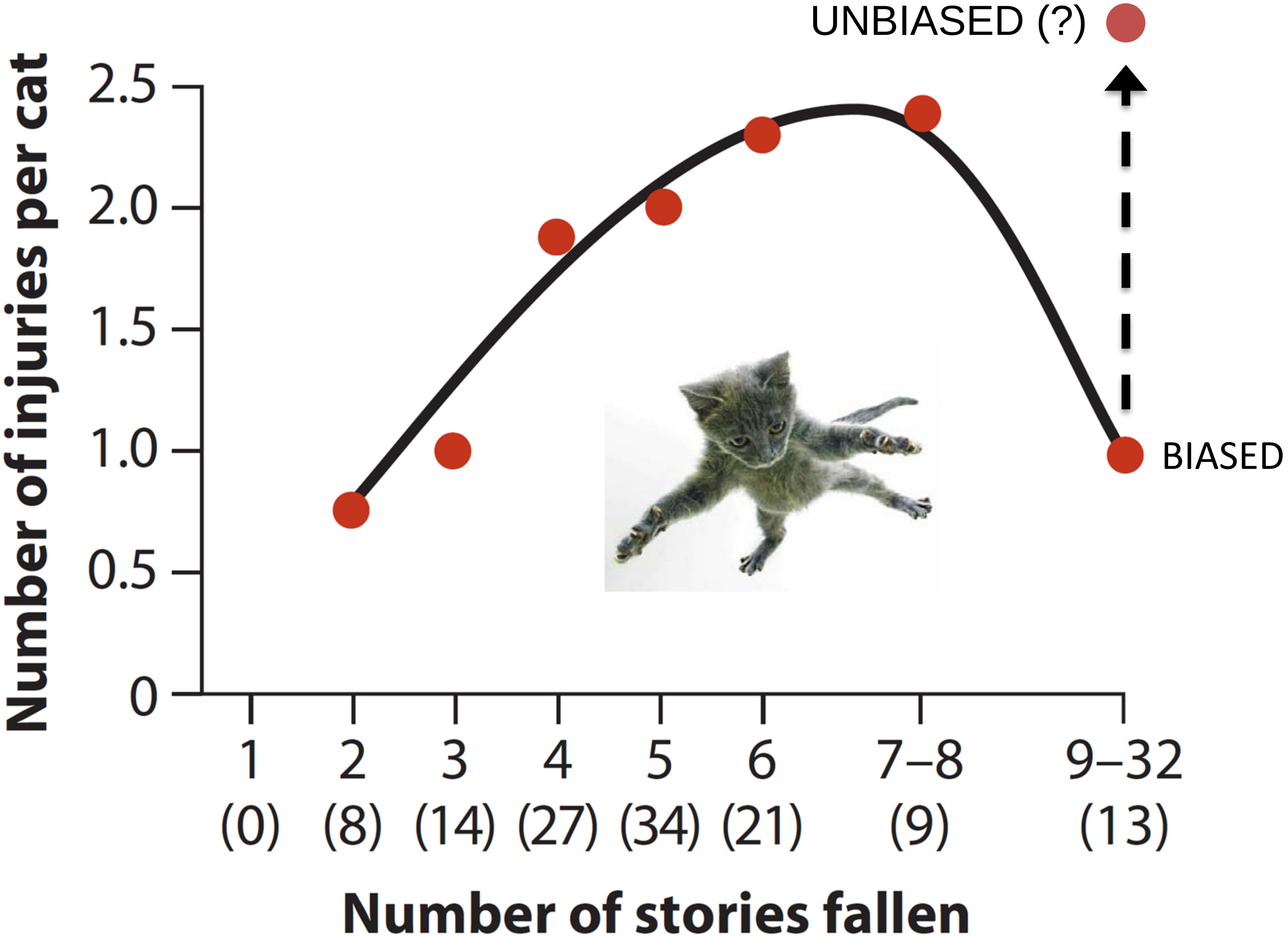
- Virtually infinite
- Parameters
- Probability distribution

KNOWN

- n obs: $x_1, x_2, \ldots x_n$
- Parameter estimates
- Sampling distribution

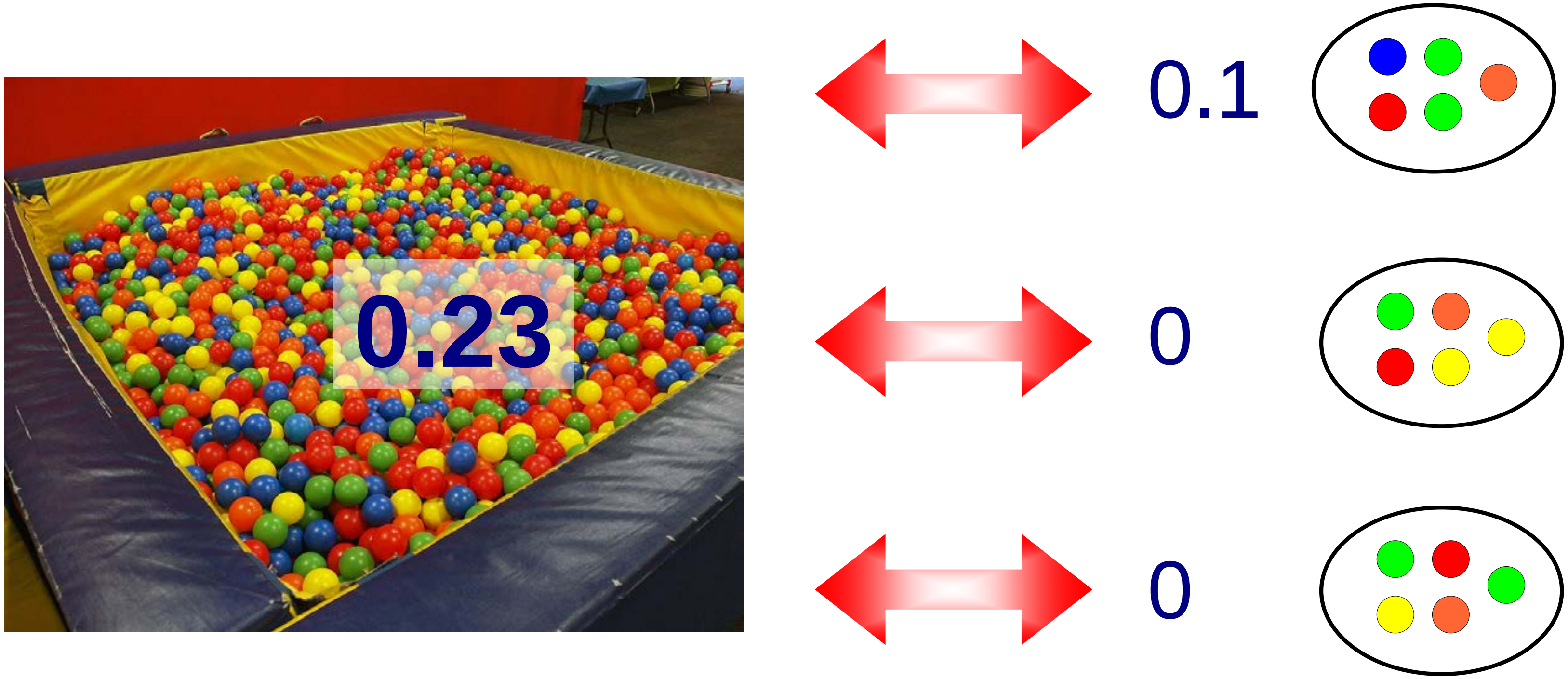


Biased sample: “High-rise syndrome”



Biased sample

- What if the blue balls are more slippery than the others? ...



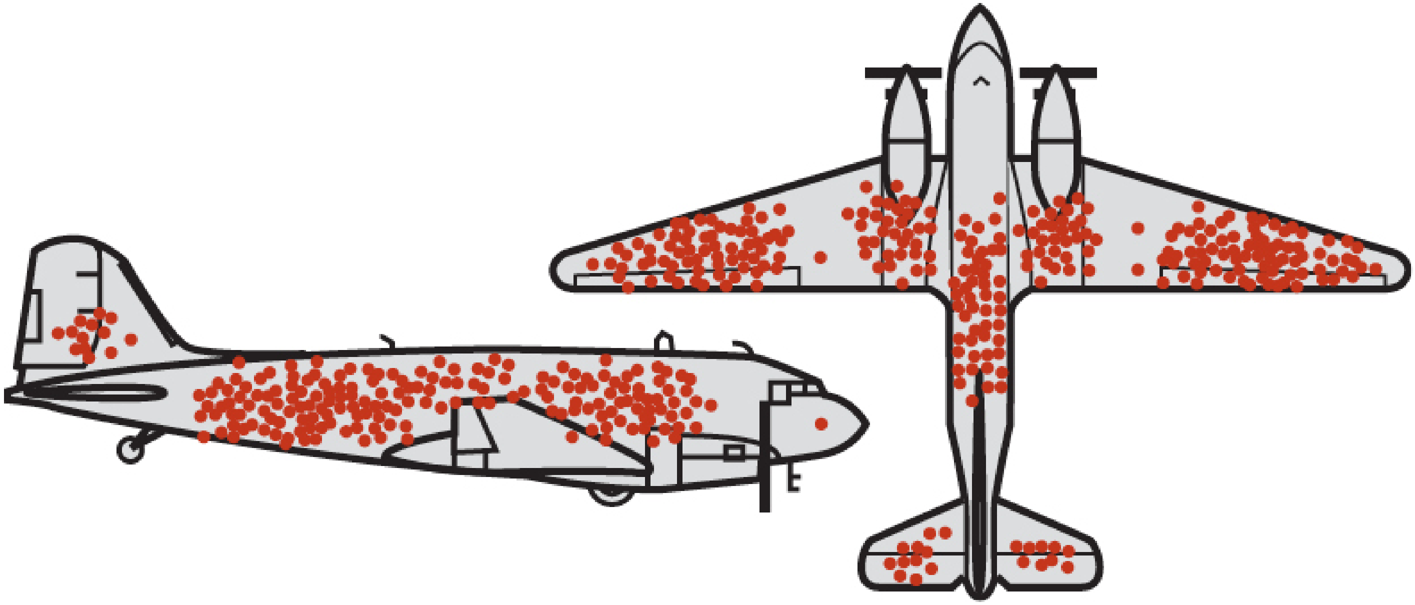
0.23

0.1

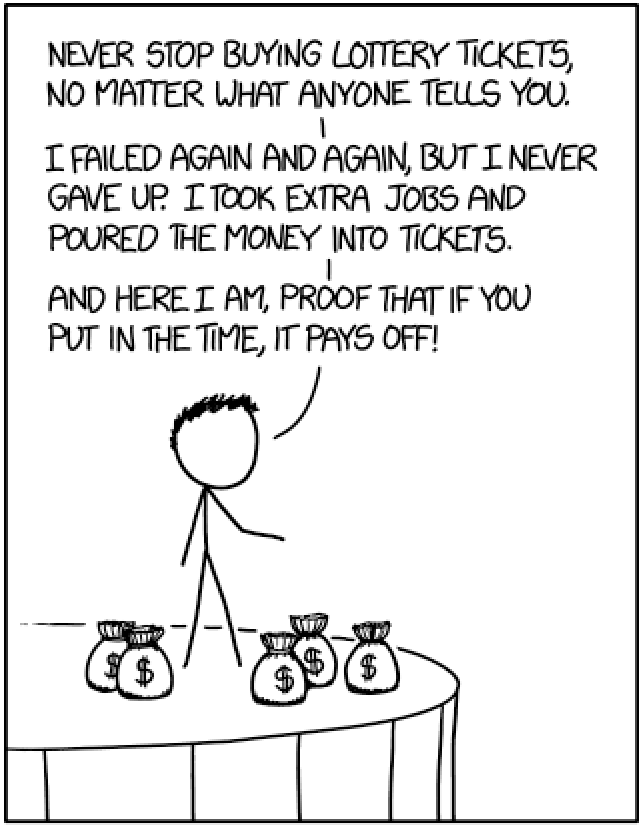
0

0

Bullet holes in returning planes in WWII



How would you enforce the planes?

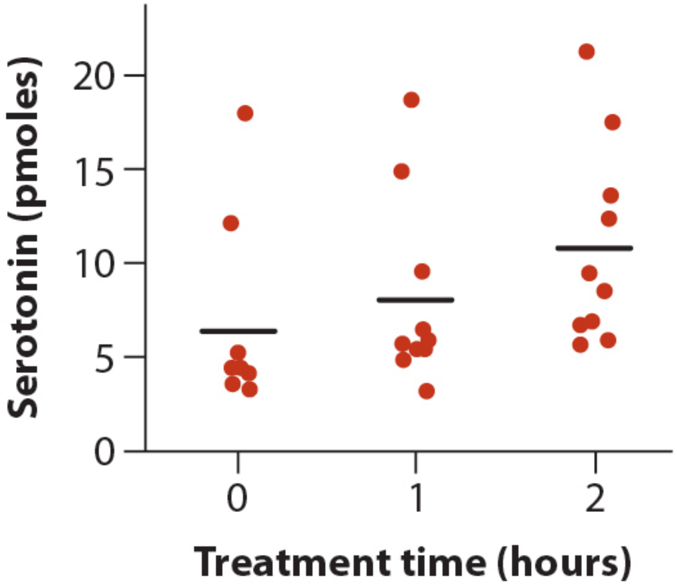

NEVER STOP BUYING LOTTERY TICKETS, NO MATTER WHAT ANYONE TELLS YOU.
I FAILED AGAIN AND AGAIN, BUT I NEVER GAVE UP. I TOOK EXTRA JOBS AND POURED THE MONEY INTO TICKETS.
AND HERE I AM, PROOF THAT IF YOU PUT IN THE TIME, IT PAYS OFF!
$
$
$
$
$

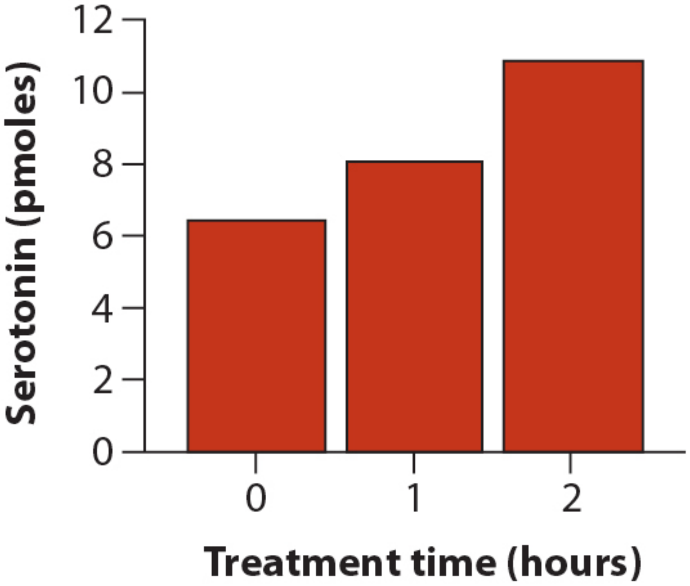
EVERY INSPIRATIONAL SPEECH BY SOMEONE SUCCESSFUL SHOULD HAVE TO START WITH A DISCLAIMER ABOUT SURVIVORSHIP BIAS.

https://xkcd.com/1827/

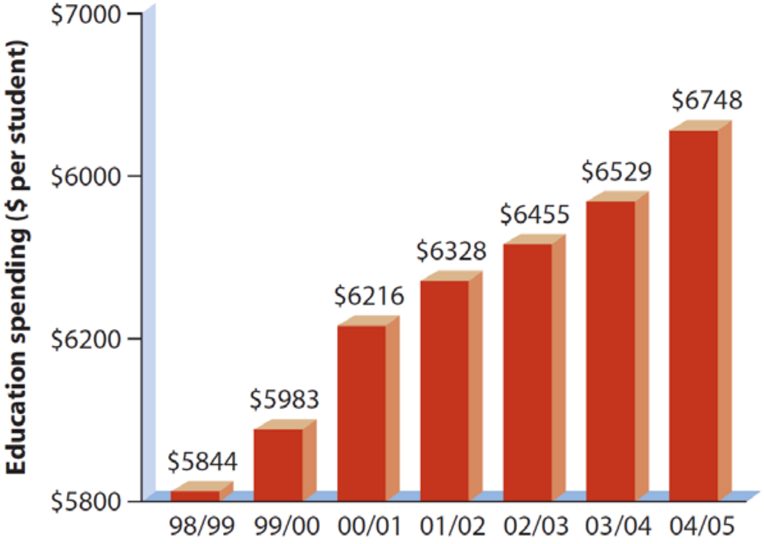
Displaying data

What is best?

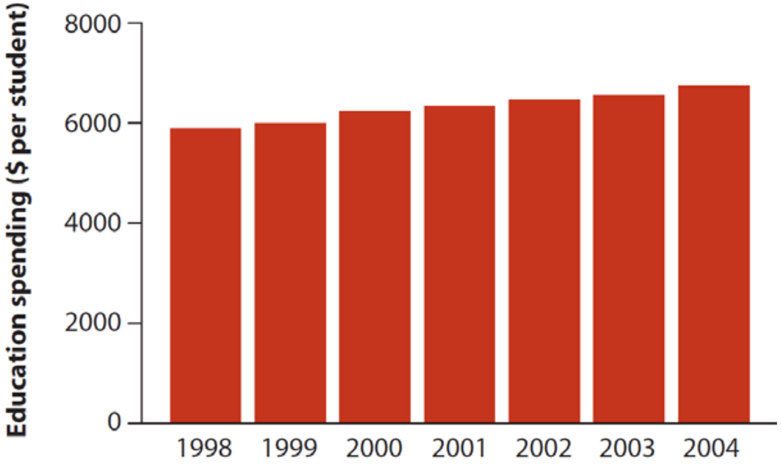

Serotonin (pmoles)
0
5
10
15
20
0
1
2
Treatment time (hours)


Serotonin (pmoles)
0
2
4
6
8
10
12
0
1
2
Treatment time (hours)

What is best?

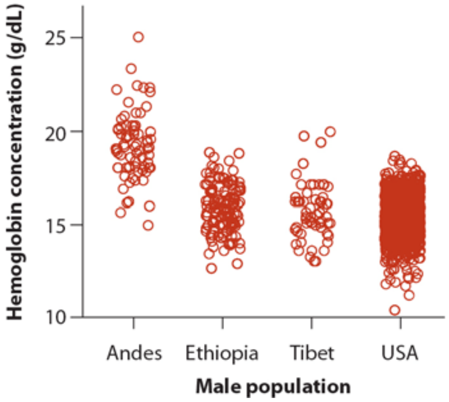


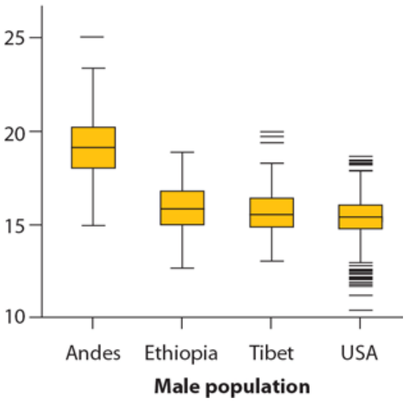
Education spending ($ per student)
$7000
$6000
$6200
$5800
$5844
$5983
$6216
$6328
$6455
$6529
$6748
98/99
99/00
00/01
01/02
02/03
03/04
04/05

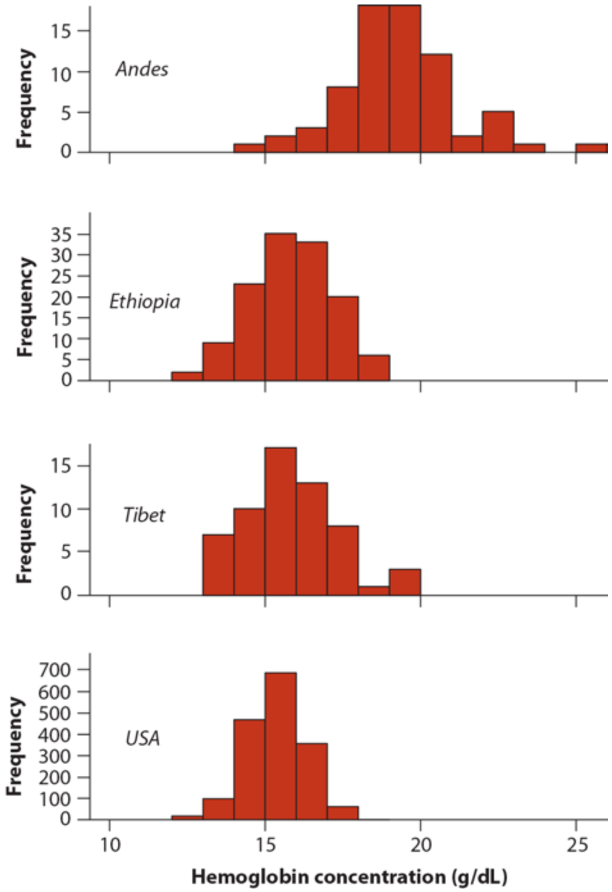


Education spending ($ per student)
8000
6000
4000
2000
0
1998
1999
2000
2001
2002
2003
2004

What is best?

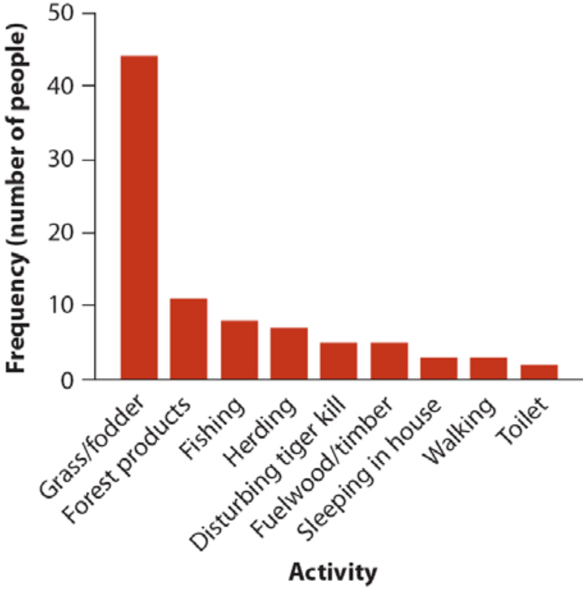

Hemoglobin concentration (g/dL)
25
20
15
10
Andes
Ethiopia
Tibet
USA
Male population

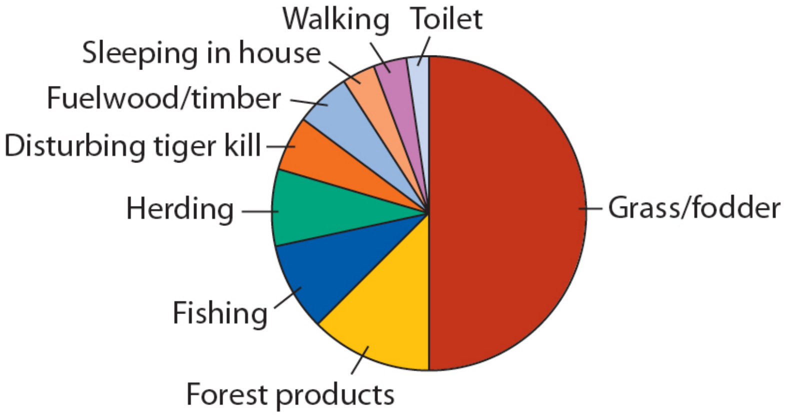

25
20
15
10
Andes
Ethiopia
Tibet
USA
Male population


Frequency
15
10
5
0
Andes
Frequency
35
30
25
20
15
10
5
0
Ethiopia
Frequency
15
10
5
0
Tibet
Frequency
700
600
500
400
300
200
100
0
USA
10
15
20
25
Hemoglobin concentration (g/dL)

Displaying data

- If is really difficult in ggplot, then it is probably a bad idea



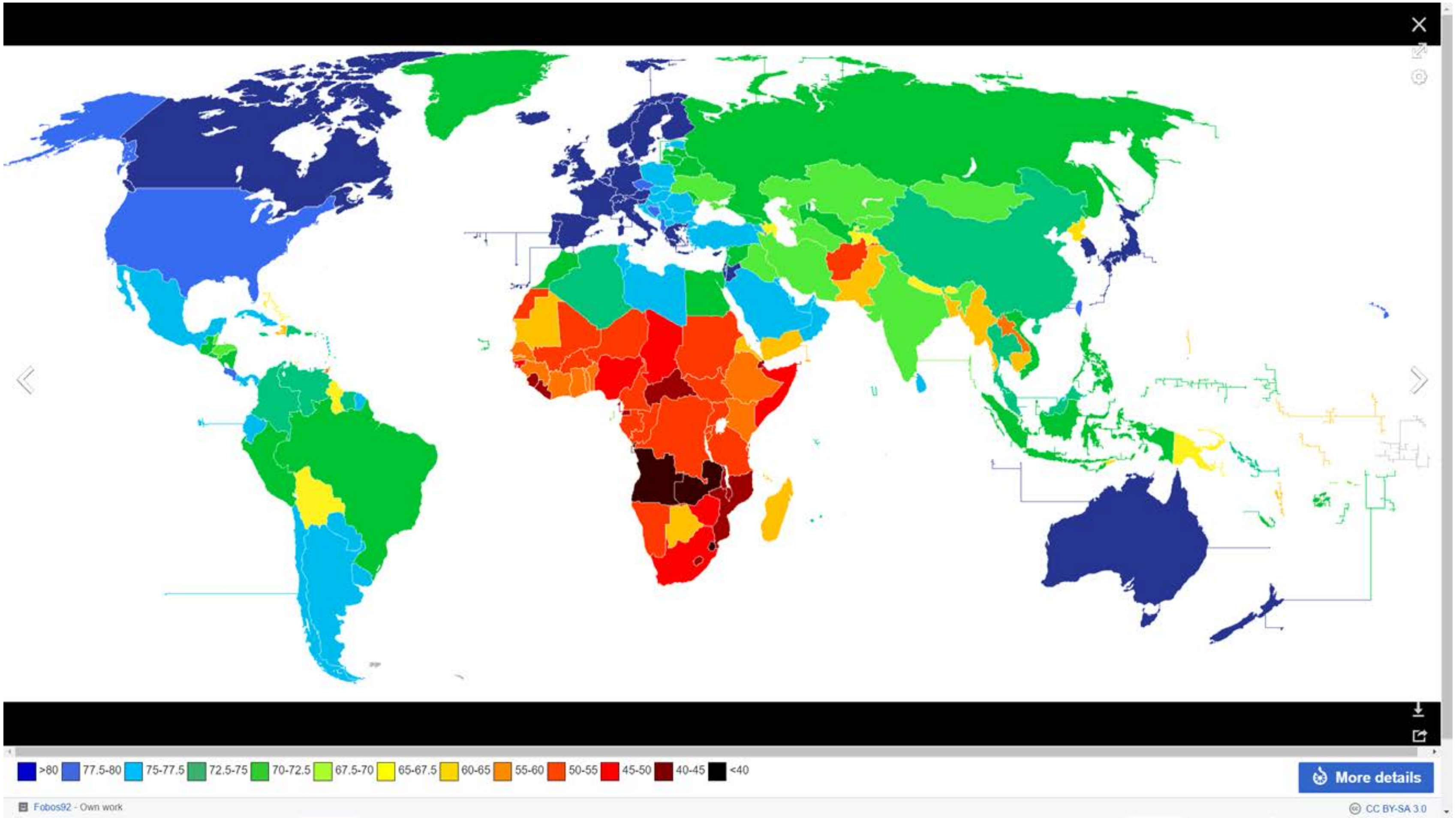


Describing data

- Arithmetic mean = average
- Standard deviation (compares all observations with mean)
- Median = 50% quantile
- IQR (75% quantile – 25% quantile)

- Present a case where one is really bad

Life expectancy is a mean



Fobos92 - Own work

Standard deviation

$$s = \sqrt{\frac{\sum(Y_i - \bar{Y})^2}{n-1}}$$

Median absolute deviation

$$\mathrm{MAD} = \mathrm{median}(\ \left|X_i - \mathrm{median}(X)\right|\ ),$$

Exercise

- Make a small dataset with 3 outliers:

```
x = c(70:90,1000,1100,1200)
```

- Calculate mean and median

```
mean(x)
```

```
median(x)
```

- Calculate sd, iqr and mad

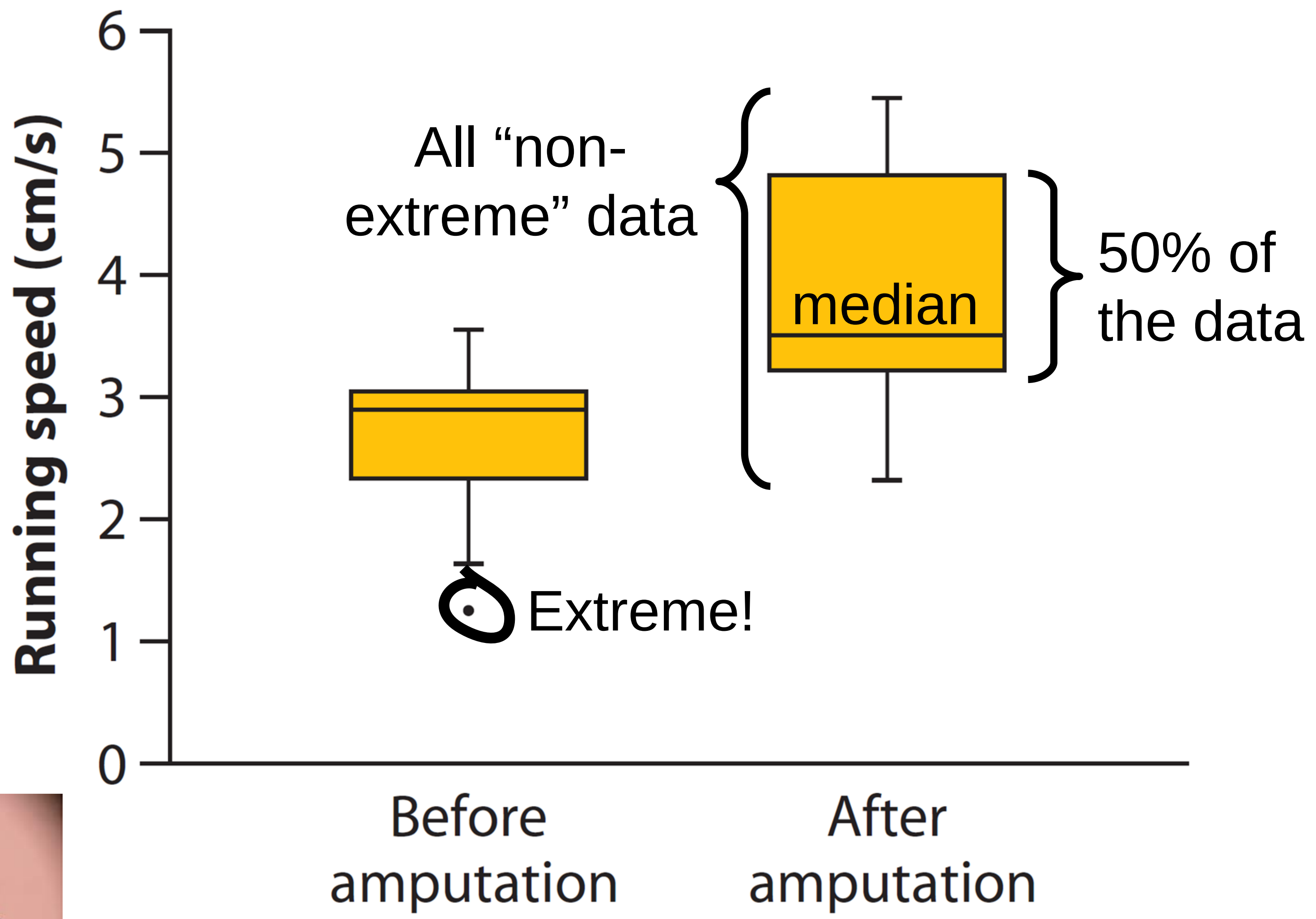
```
sd(x)
```

```
quantile(x)
```

```
mad(x)
```

- Conclusions?

Distribution of the data: boxplots



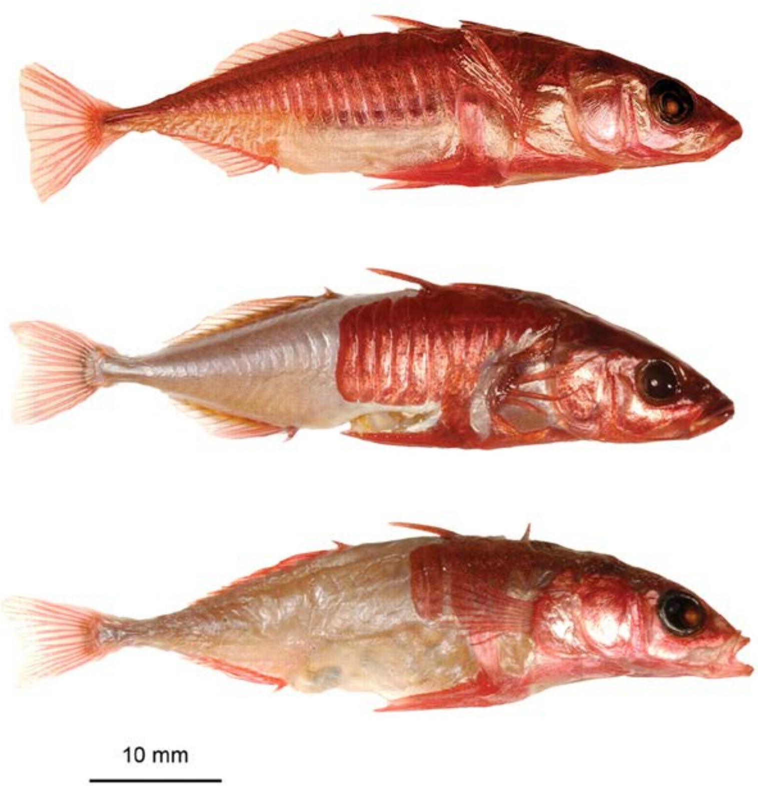
All “non-extreme” data

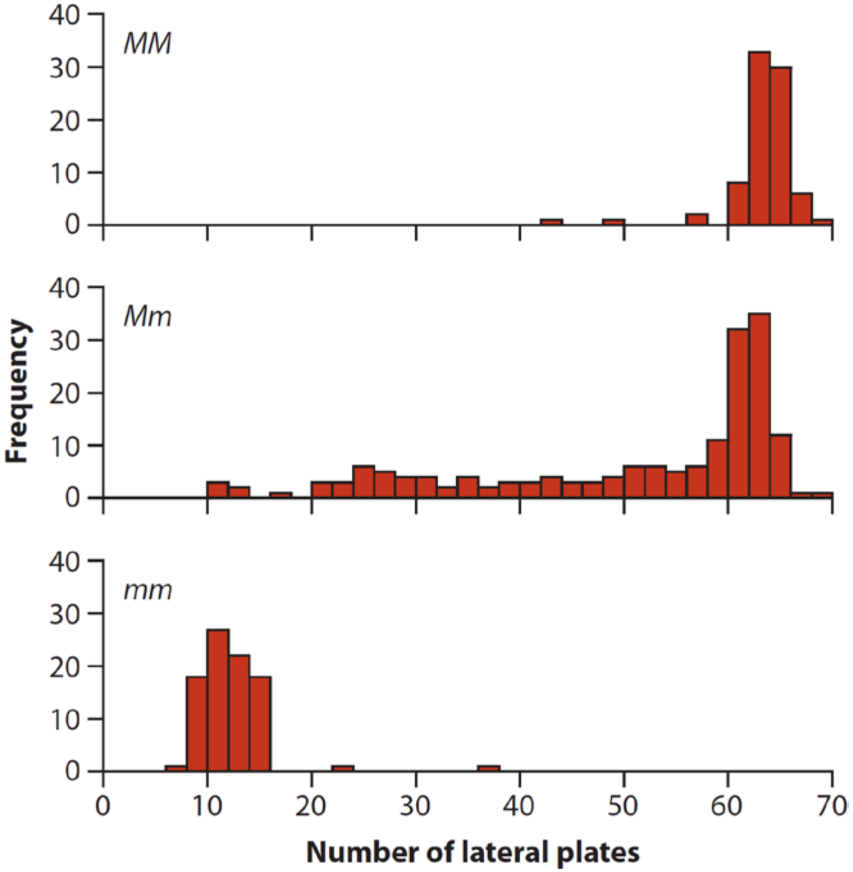
median

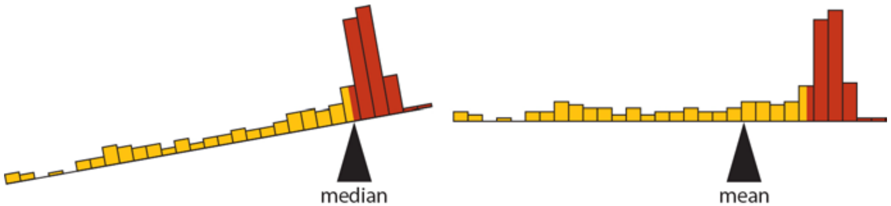
50% of the data

Extreme!

Distribution of the data: histograms





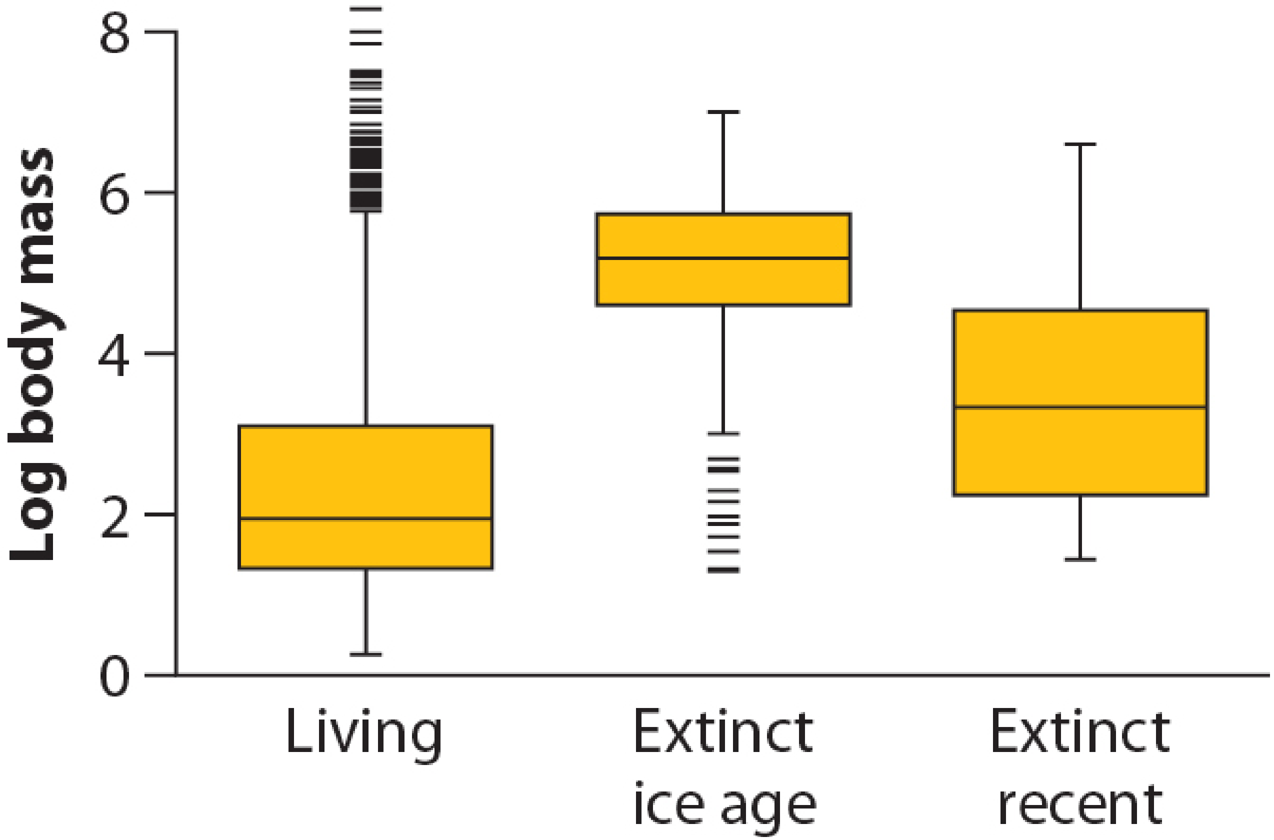


Outliers

- Outliers:
  - Influence mean and sd

- Median and iqr are more robust
  - Median is a single data point
  - iqr only use half of the data

Bodymass of mammals
Discuss and present



Estimating with uncertainty

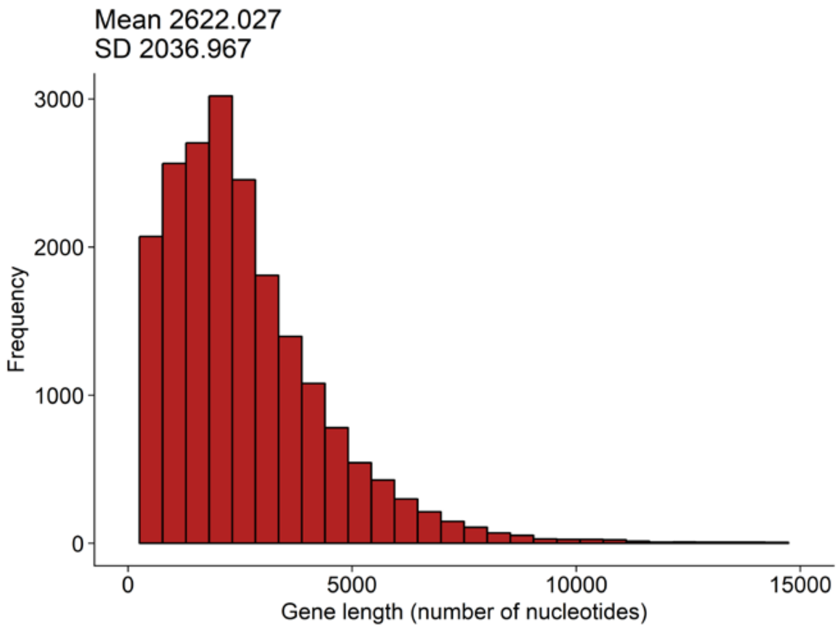
Keywords

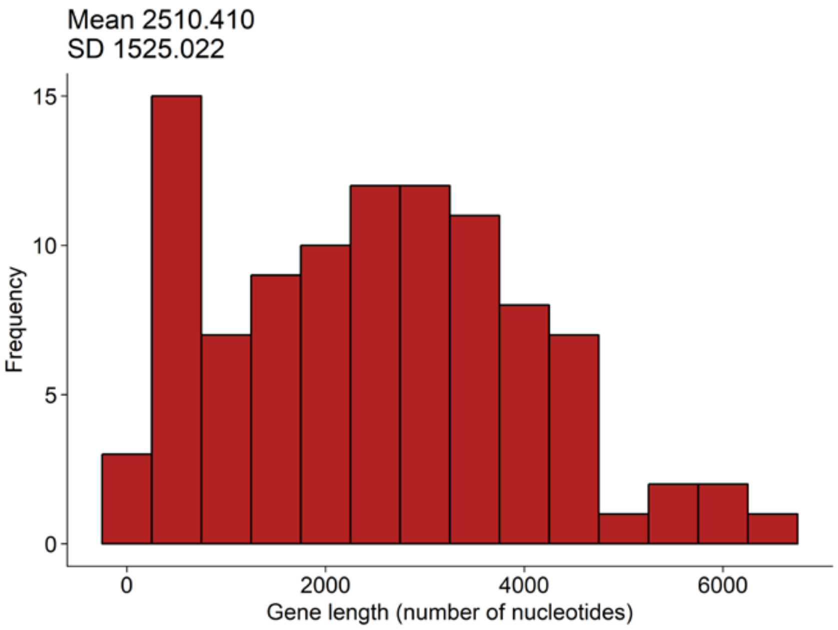
- We want to say something about the **population**
- But we only have a **sample**
- So the **sample estimate** is different from the true value because of **sampling error**
- **The sampling distribution** is the distribution of estimate from different samples
- **The standard error** is the standard deviation of the sampling distribution
- **Confidence intervals on the estimate**
  - The 2SE rule of thumb
  - Bootstrap (chapter 19)

All genes

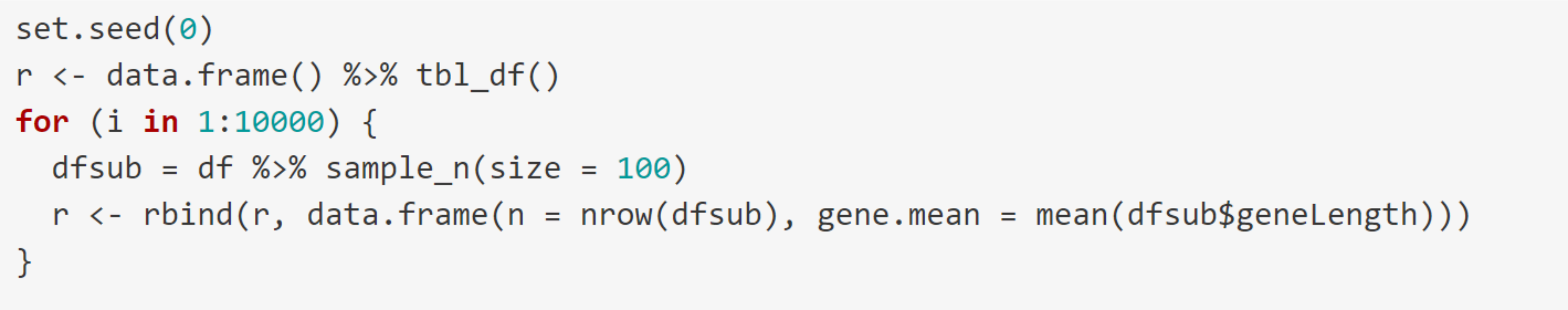
```
df <- read_csv(file =
"chap04e1HumanGeneLengths.csv")
```

```
set.seed(0)
dfsub <- df %>% sample_n(size = 100)
```



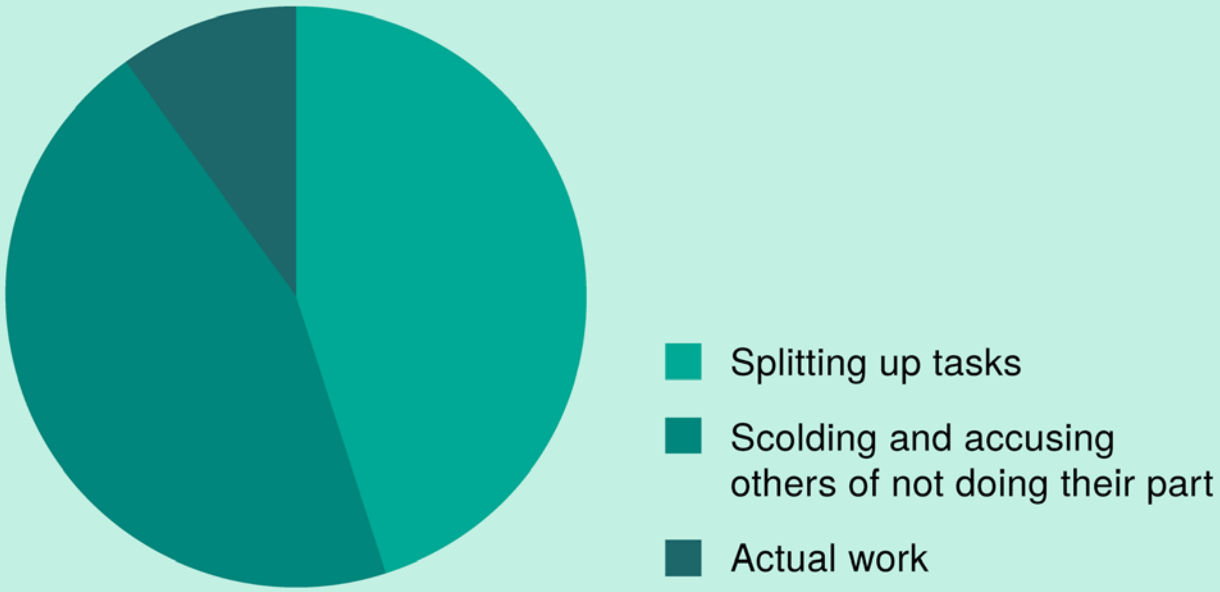


- `df <- read_csv(file="chap04e1HumanGeneLengths.csv")`
- Replicate figure 4.1-3
  - – the sampling distribution of the mean for n=100
- Replicate figure 4.1-4
  - – the sampling distribution of the mean for n=20, n=100 and n=500
- Calculate standard error from your samples (n=20, n=100, n=500)
- Compare with table 4.2-1



```
set.seed(0)
r <- data.frame() %>% tbl_df()
for (i in 1:10000) {
  dfsub = df %>% sample_n(size = 100)
  r <- rbind(r, data.frame(n = nrow(dfsub), gene.mean = mean(dfsub$geneLength)))
}
```

GROUP WORK


🟩 Splitting up tasks
🟩 Scolding and accusing others of not doing their part
🟩 Actual work