

# Introduction to statistics using R

Seminar series - session 1

Chloé Warret Rodrigues



# Session1 - Learning objectives

- Understand what is “statistics”
- Get an idea of application range
- Understand the difference between **descriptive** and **inferential** statistics
- Measure central tendency, variability (R)
- Review data types
- What is a variable

# What is “statistics”?

A collection of methods for:

- collecting,
- organizing,
- summarizing,
- analyzing,
- interpreting,

data

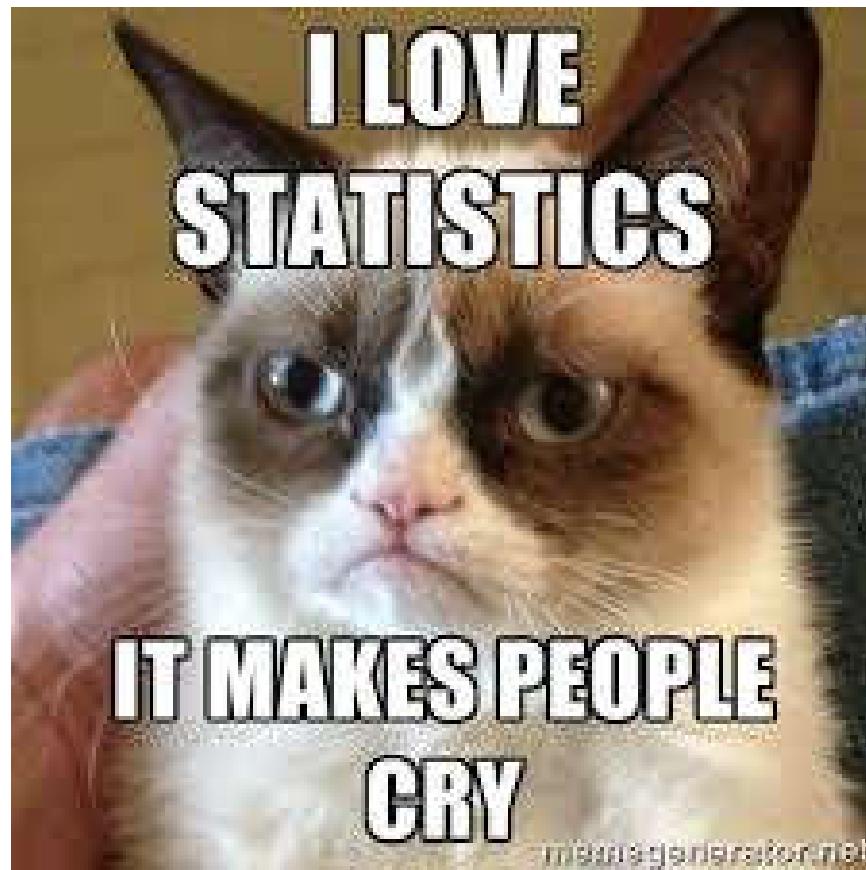
- and make decisions based on data

All steps in red: we can do with R software.

# Why use statistics?

- Answer questions or test hypotheses
- Understand mechanisms and relationships between variables
- Ensure observed patterns are not due to chance
- Quantify phenomena
- Monitoring
- Predict future events
- Make informed decisions

# Some examples of why to use stats in Biology



# Ex.1: test a hypothesis

Background:

- Arctic ecosystems: food is plenty in summer, scarce in winter



# Ex.1: test a hypothesis

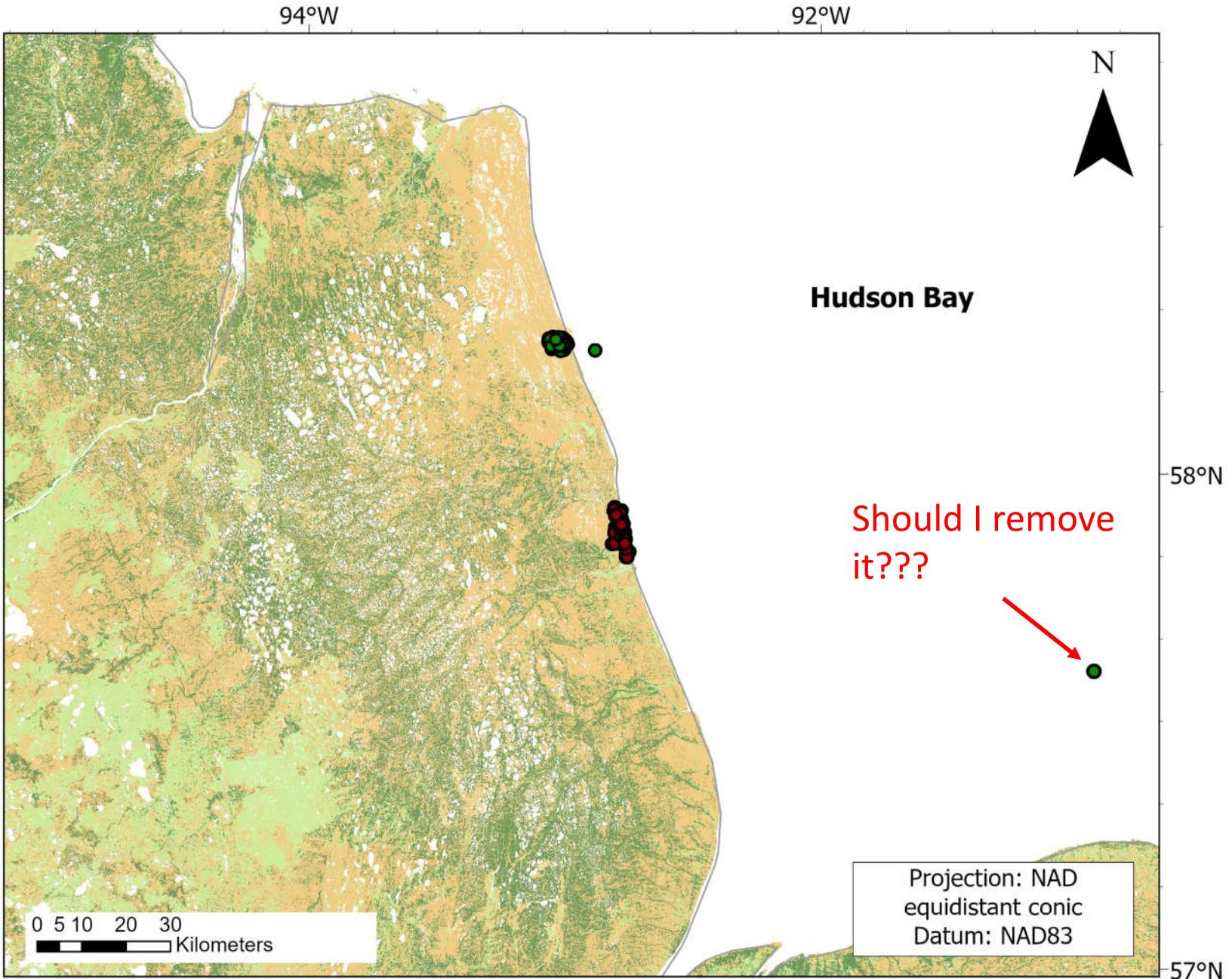


Hypothesis: Fox movement tactic  
changes seasonally



Collect data: Live trapping and satellite telemetry

# Clean and organize the data



# Analyze data

R RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

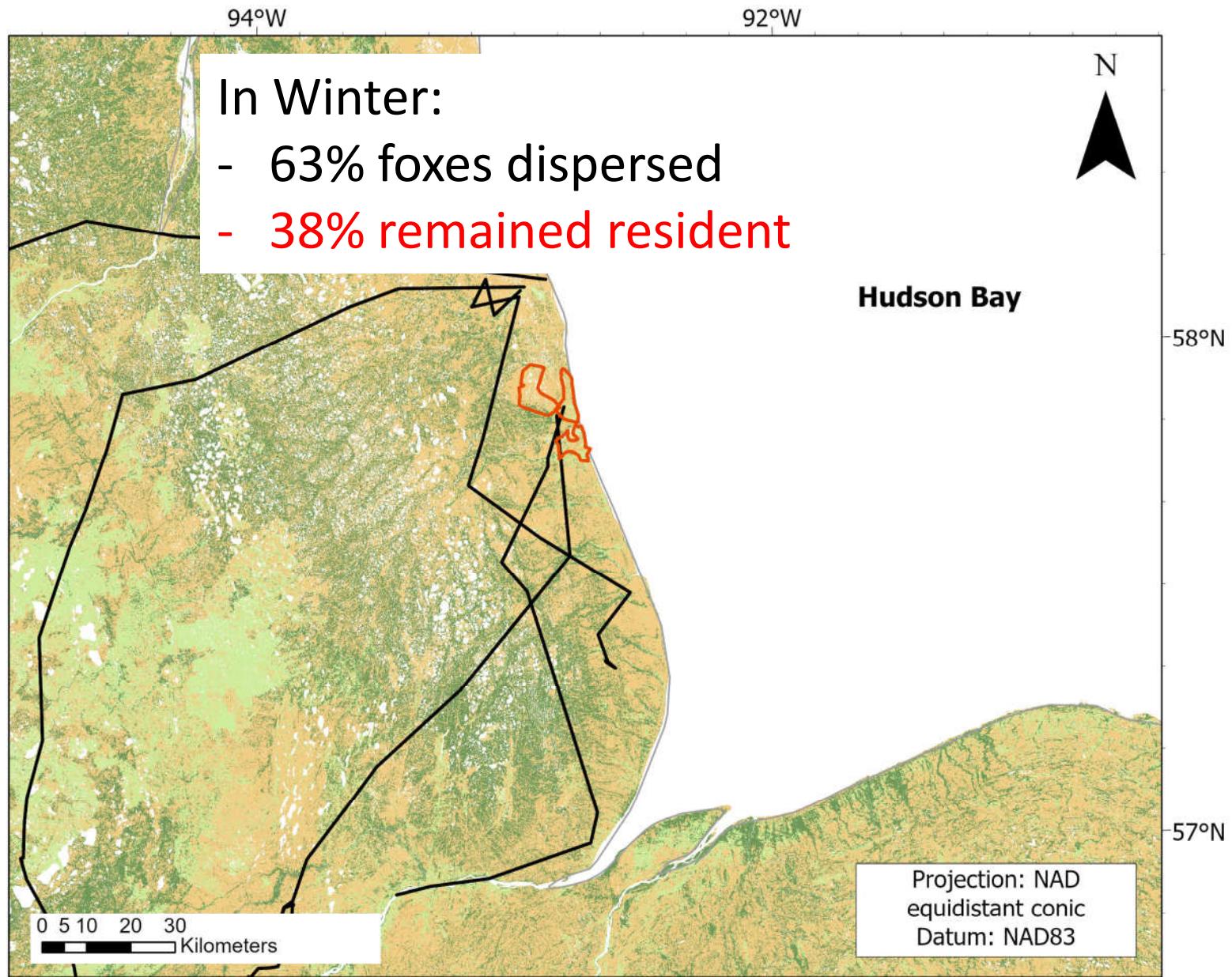
Source on Save Go to file/function Addins

grouse\_gout.R × Parasite\_summary.R × Gout\_in\_SG.md × Cloacal\_prolapsus\_KZ.R × Enjil\_MC\_EAB06.R × LoCoH\_Churchill\_winter\_20182020.R × eab1.c ×

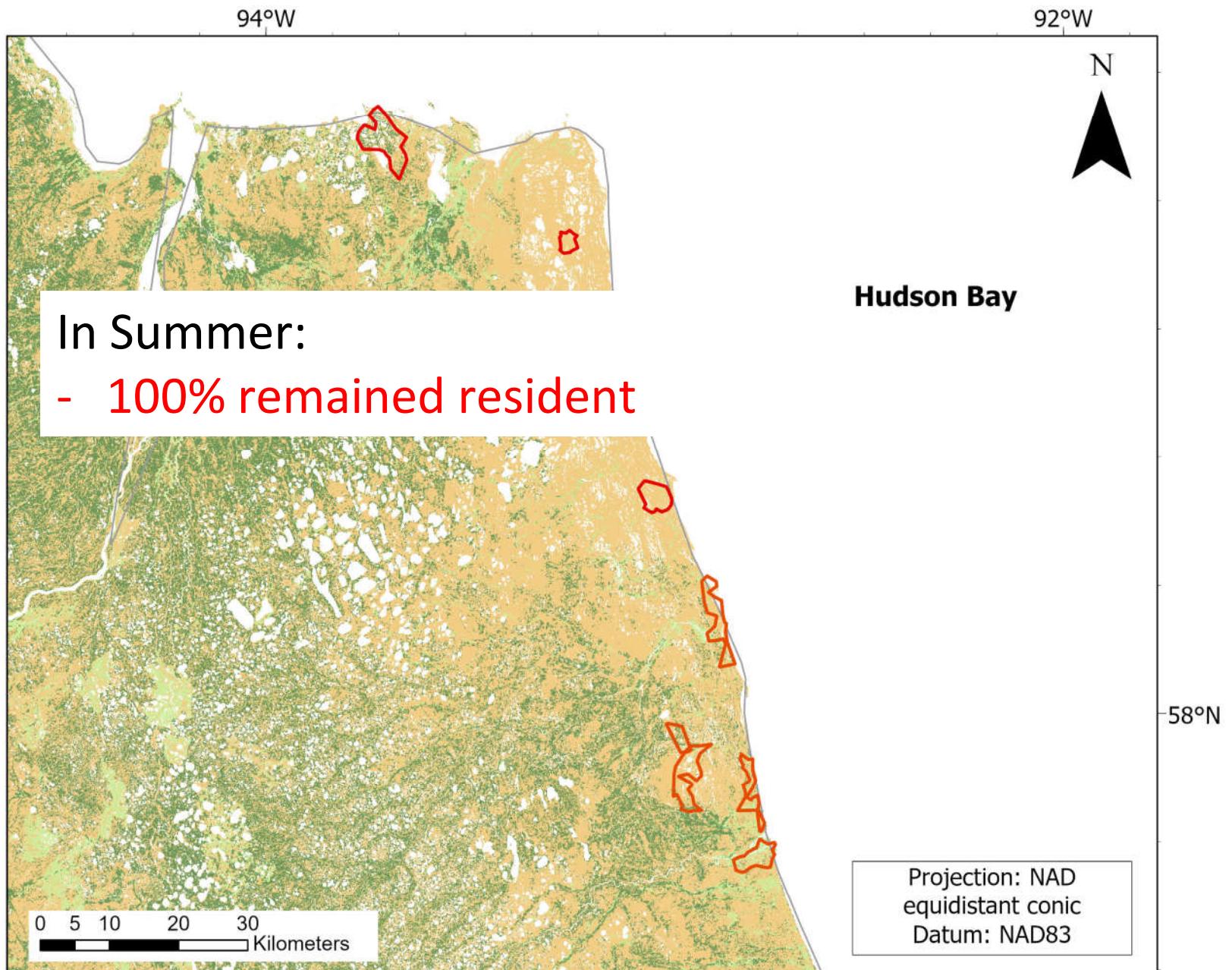
Run Source

```
13 timestamp<-fox$date
14 xy2017<-SpatialPoints(fox[, c("easting","northing")], proj4string=CRS("+proj=utm +zone=15 +datum=NAD83 +ellps=GRS80"))
15 xy2017 <-coordinates(xy2017)
16 head(xy2017)
17 colnames(xy2017) <-c("x","y")
18
19 #merge the timestamp column to the SPDF
20 merge<-data.frame(fox)
21 head(merge)
22 df <- merge[,-(18)]
23 head(df)
24
25 head(as.character(df$date))
26 df.gmt<-as.POSIXct(strptime(merge[, "date"], "%d-%m-%y"))
27 df.gmt[1:3]
28
29
30 ######
31 # create lxy object and investigate it #
32 #####
33
34 #create and check lxy object
35 s.lxy <-xyt.lxy(xy=xy2017, dt=df.gmt, id=fox$Name, proj4string=CRS("+proj=utm +zone=15 +datum=NAD83 +ellps=GRS80"),
36 show.dup.dt=TRUE, show.bad.timestamps=TRUE)
37
38 #####
39 # identify nearest neighbours - a-method #
40 #####
41 ######
42 # ***2018*** #
43 #####
44
45 #subset individuals
46 u.lxy.2018 <-lxy.subset(s.lxy, id = "Uhtred")
47 MM.lxy.2018 <-lxy.subset(s.lxy, id = "MadMax2018")
48 FJ.lxy.2018 <-lxy.subset(s.lxy, id = "FoxyJim")
49 AB.lxy.2018 <-lxy.subset(s.lxy, id = "AnneBoleyn2018")
50 LR.lxy.2018 <-lxy.subset(s.lxy, id = "Little Red2018")
51
52 ##Parameter s will be set to 0 since time sn't relevant here. All hulls will purely be determined spatially
53
54
```

# Quantify observations



# Quantify observations



# Interpret your results



Yes, fox movement tactic changes  
seasonally

# Interpret your results

Fox movement tactic changes seasonally, **but why?**



# Interpret your results

Fox movement tactic changes  
seasonally, **but why?**

And hard to get...



So, not worth staying if you're starving!

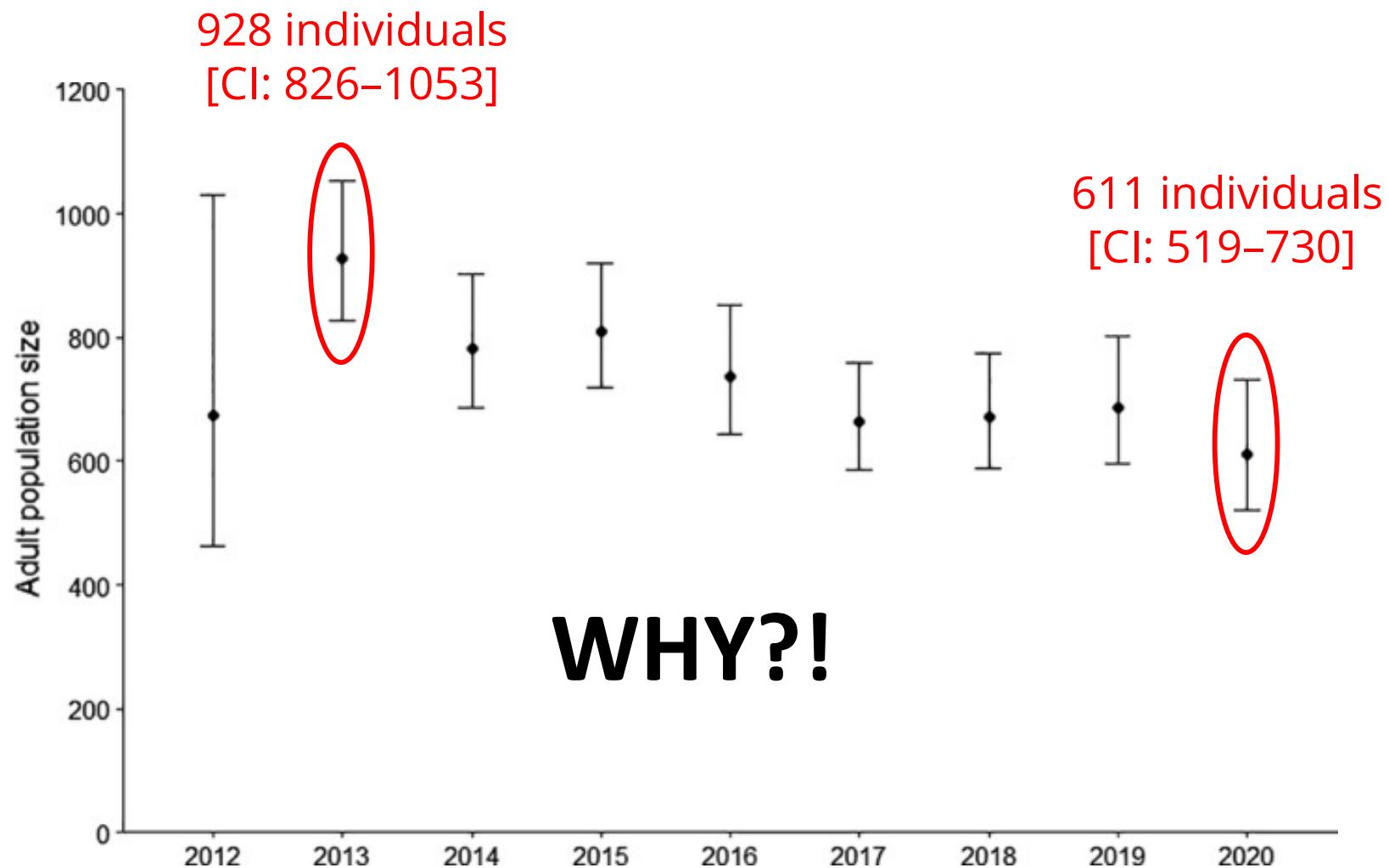


# Population survey

1. Collect data: capture-  
Recapture

# Ex.2: Managing endangered populations

Background: an endangered iguana population is declining



# Ex.2: Understand a population decline and propose conservation measures

Background: an endangered iguana population is declining

Low fecundity?



Adult mortality?



Eggs not hatching?



Juvenile mortality?

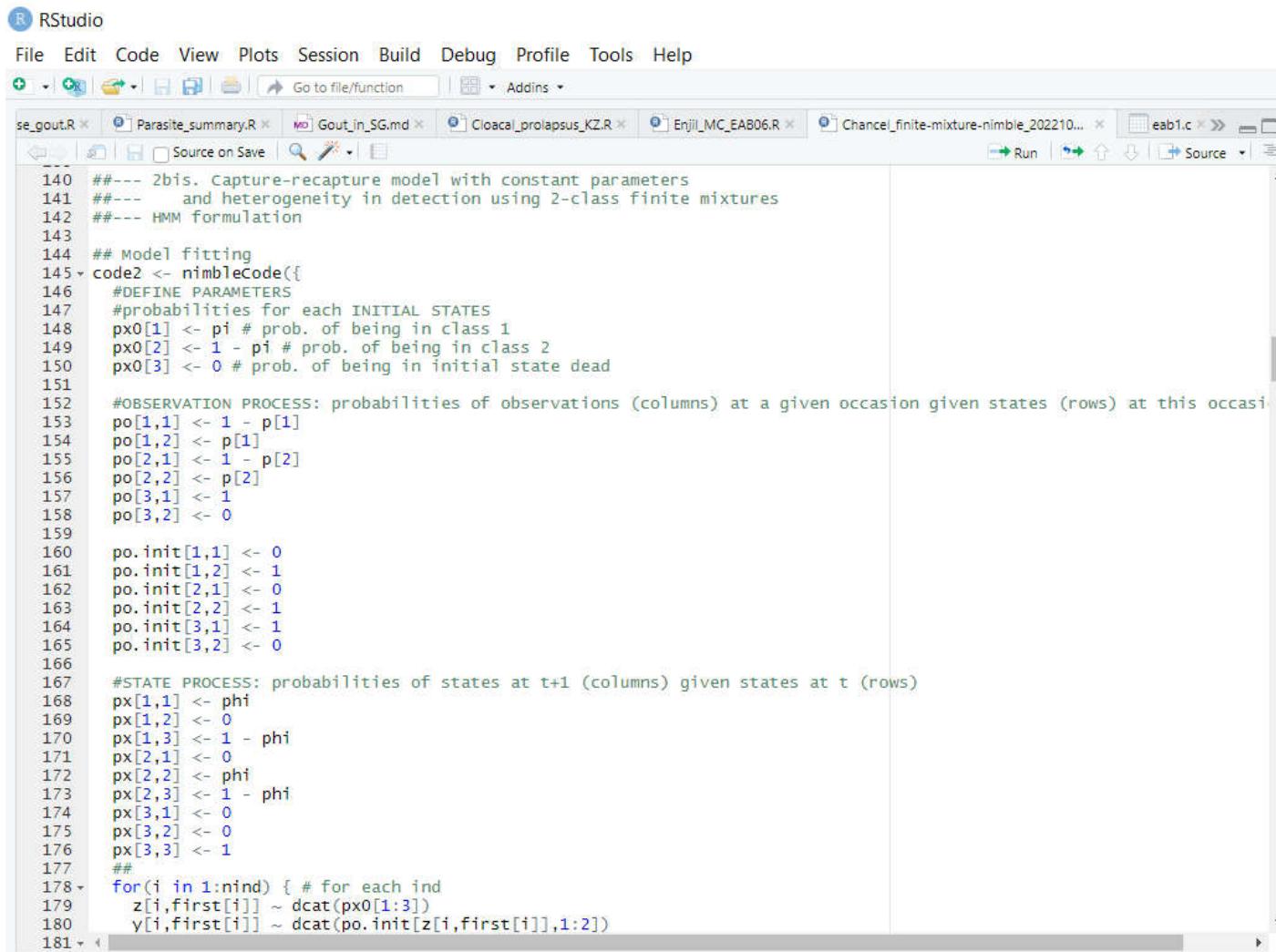
# Demography analyses

1. Collect data: capture-Recapture
2. Organizing: create a capture history database

	<i>Occasion 1</i>	<i>Occasion 2</i>	<i>Occasion 3</i>	<i>Occasion 4</i>	<i>Occasion 5</i>
<i>Animal 1</i>	1	1	0	1	1
<i>Animal 2</i>	0	1	0	1	1
<i>Animal 3</i>	1	1	0	1	0

# Demography analyses

1. Collect data: capture-Recapture
2. Organizing: create a capture history database
3. Analyzing: Demography models



The screenshot shows the RStudio interface with the following details:

- Header:** RStudio, File, Edit, Code, View, Plots, Session, Build, Debug, Profile, Tools, Help.
- Toolbar:** Go to file/function, Addins.
- Script Editor:** A code editor containing an R script titled "se\_gout.R".
- Code Content:** The script is a nimbleCode for a capture-recapture model. It defines parameters for initial states (px0), observation process (po), and state process (px). It also includes a loop for individual records (i) and a dcat distribution for observations.

```
##--- 2bis. Capture-recapture model with constant parameters
##--- and heterogeneity in detection using 2-class finite mixtures
##--- HMM formulation
## Model fitting
code2 <- nimbleCode({
  #DEFINE PARAMETERS
  #probabilities for each INITIAL STATES
  px0[1] <- pi # prob. of being in class 1
  px0[2] <- 1 - pi # prob. of being in class 2
  px0[3] <- 0 # prob. of being in initial state dead
  #OBSERVATION PROCESS: probabilities of observations (columns) at a given occasion given states (rows) at this occasion
  po[1,1] <- 1 - p[1]
  po[1,2] <- p[1]
  po[2,1] <- 1 - p[2]
  po[2,2] <- p[2]
  po[3,1] <- 1
  po[3,2] <- 0
  po.init[1,1] <- 0
  po.init[1,2] <- 1
  po.init[2,1] <- 0
  po.init[2,2] <- 1
  po.init[3,1] <- 1
  po.init[3,2] <- 0
  #STATE PROCESS: probabilities of states at t+1 (columns) given states at t (rows)
  px[1,1] <- phi
  px[1,2] <- 0
  px[1,3] <- 1 - phi
  px[2,1] <- 0
  px[2,2] <- phi
  px[2,3] <- 1 - phi
  px[3,1] <- 0
  px[3,2] <- 0
  px[3,3] <- 1
  ##
  for(i in 1:nind) { # for each ind
    z[i,first[i]] ~ dcat(px0[1:3])
    y[i,first[i]] ~ dcat(po.init[z[i,first[i]],1:2])
  }
})
```

# Demography analyses

## 4. Interpreting results:



Adult survival = 0.85

HIGH

Recruitment = 0.09

LOW



# Demography analyses

## 5. make decisions: improve survival of immatures



control predators

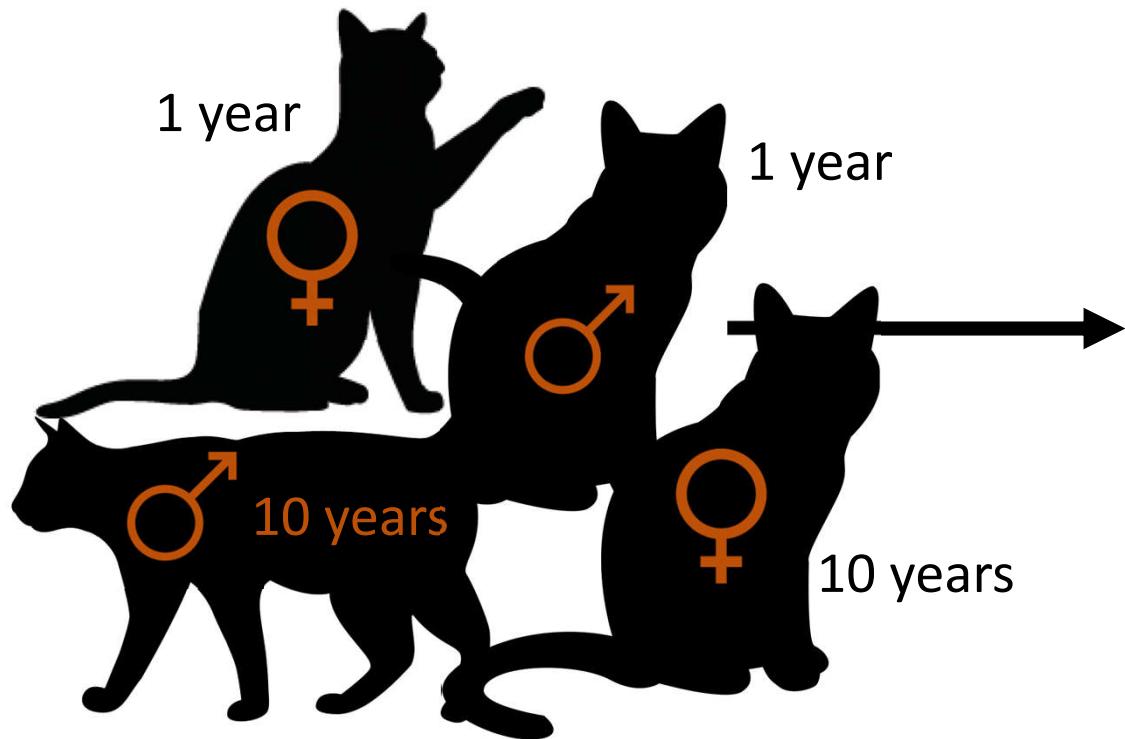
# Ex.3: Monitoring patient's health

Paramètres	Résultat	Unités
ALP	100	U/l
ALT	123	U/l
UREA	0.45	g/l
GLU	1.18	g/l
CRE	13.9	mg/l
TBIL	0.6	mg/l
TP	85	g/l
UR/C	32.38	

???



# Ex.3: Monitoring patient's health



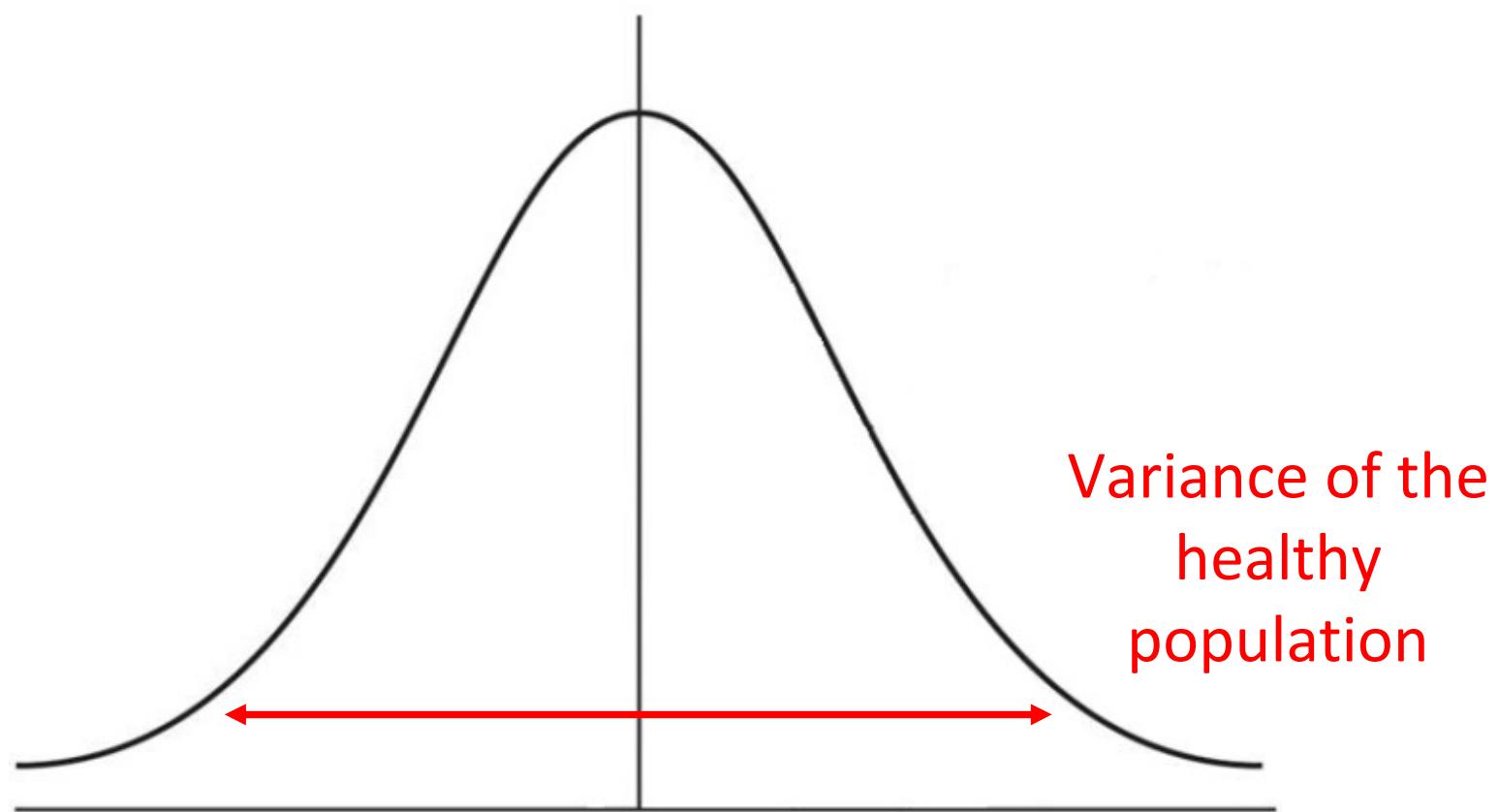
Sample healthy individuals



Collect data

Estimate parameters of a population based on the statistics of the sample

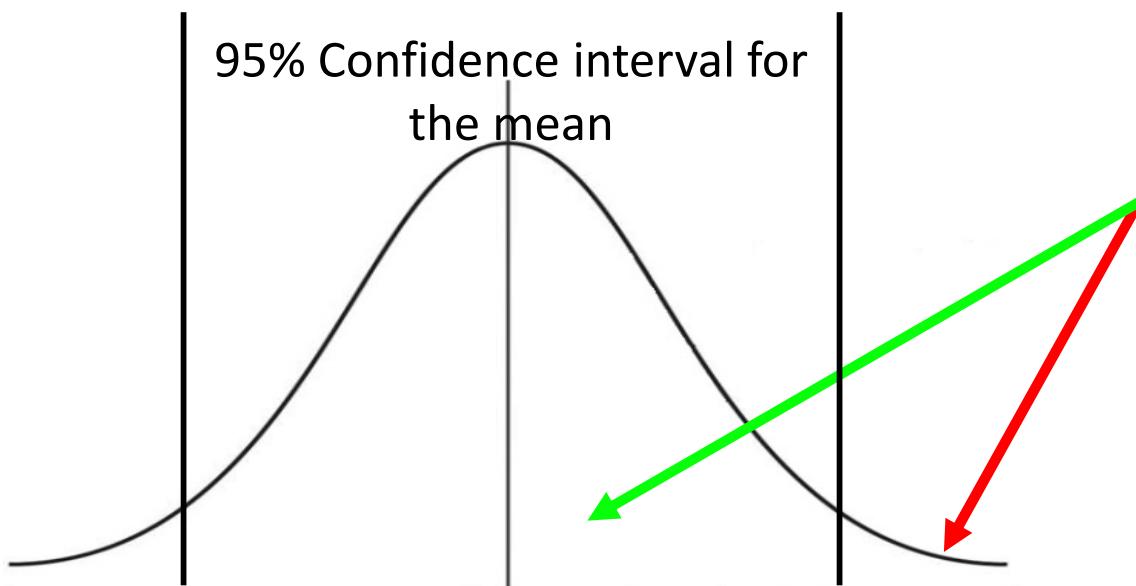
Mean of the healthy population



Summarize data

# Some examples: Monitoring patient's health

Paramètres	Indi	Alrm.	Résultat	Unités	Normalités	<-----N-----+>
ALP			100	U/l	10- 130	
ALT	+		123	U/l	10- 90	
UREA			0.45	g/l	0.19- 0.60	
GLU			1.18	g/l	0.60- 1.50	
CRE			13.9	mg/l	3.1- 14.0	
TBIL			0.6	mg/l	0.0- 9.9	
TP	+		85	g/l	58- 78	
UR/C			32.38		0.00- 0.00	



# The usual steps

1. Everything starts with a QUESTION or HYPOTHESIS  
(or at least it should 😊)
2. Decide WHAT data to collect and HOW to collect them

Collect,  
organize

Collect data  
Prepare data  
set

Think about  
structure!

Enter data  
Clean data

Summarize

Descriptive  
statistics

Plots

Analyze

Compare  
groups

Determine  
relationship  
between variables

# The usual steps

1. Everything starts with a QU  
(or at least it should 😊)
2. Decide WHAT data to collect ↗

**Describe distributions:**

Characterize the variability around central tendency

Collect,  
organize

Collect data  
Prepare data  
set

Think about  
structure!

Enter data  
Clean data

Summarize

Descriptive  
statistics

Plots

Analyze

Compare  
groups

Determine  
relationship  
between variables

# The usual steps

1. Formulate a hypothesis

**Make inferences about populations based on samples:**

Use a mathematical model applied from the theory of probability

Collect, organize

Collect data

Prepare data set

Think about structure!

Enter data

Clean data

2. Analyze

Descriptive statistics

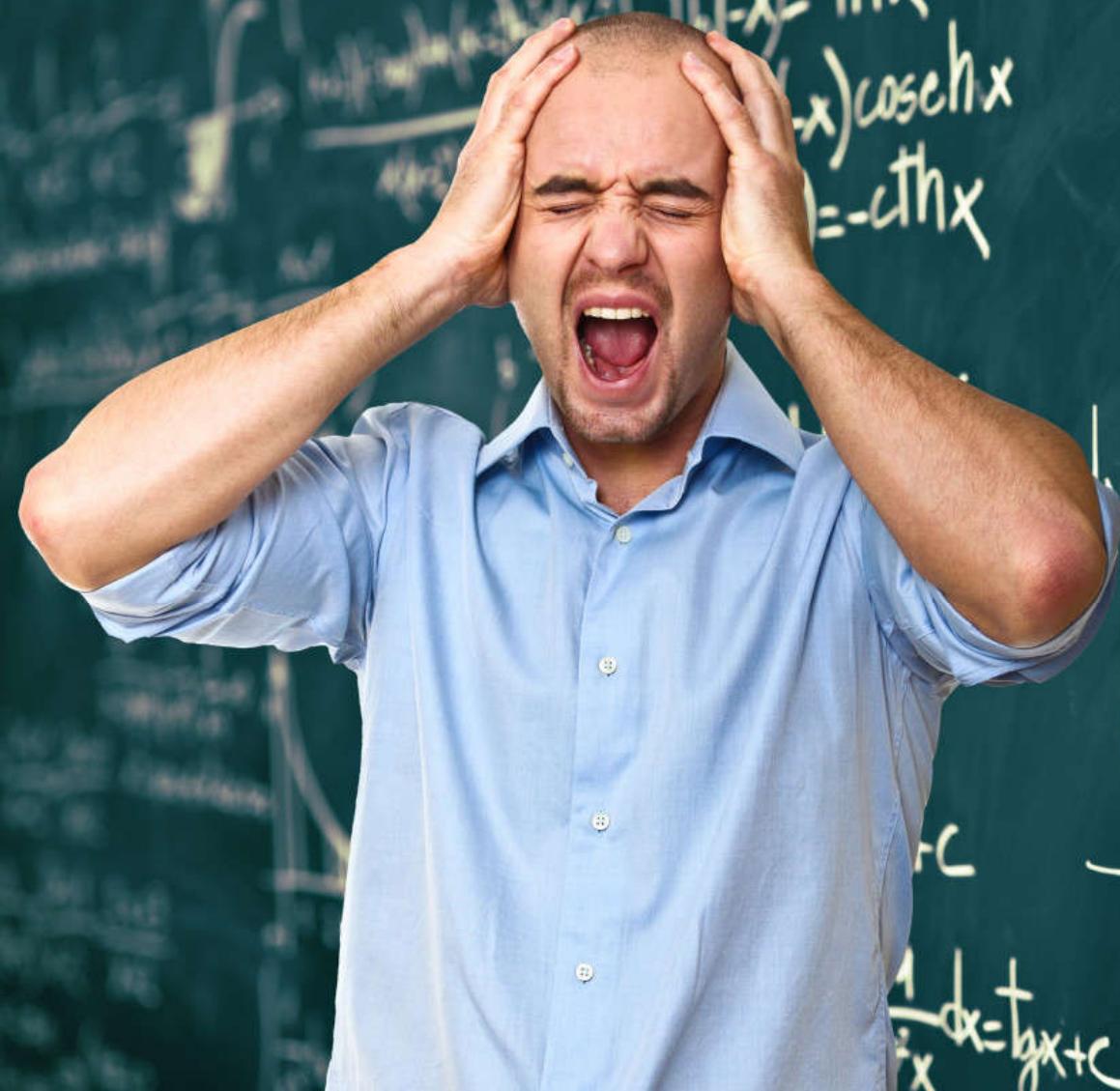
Plots

Analyze

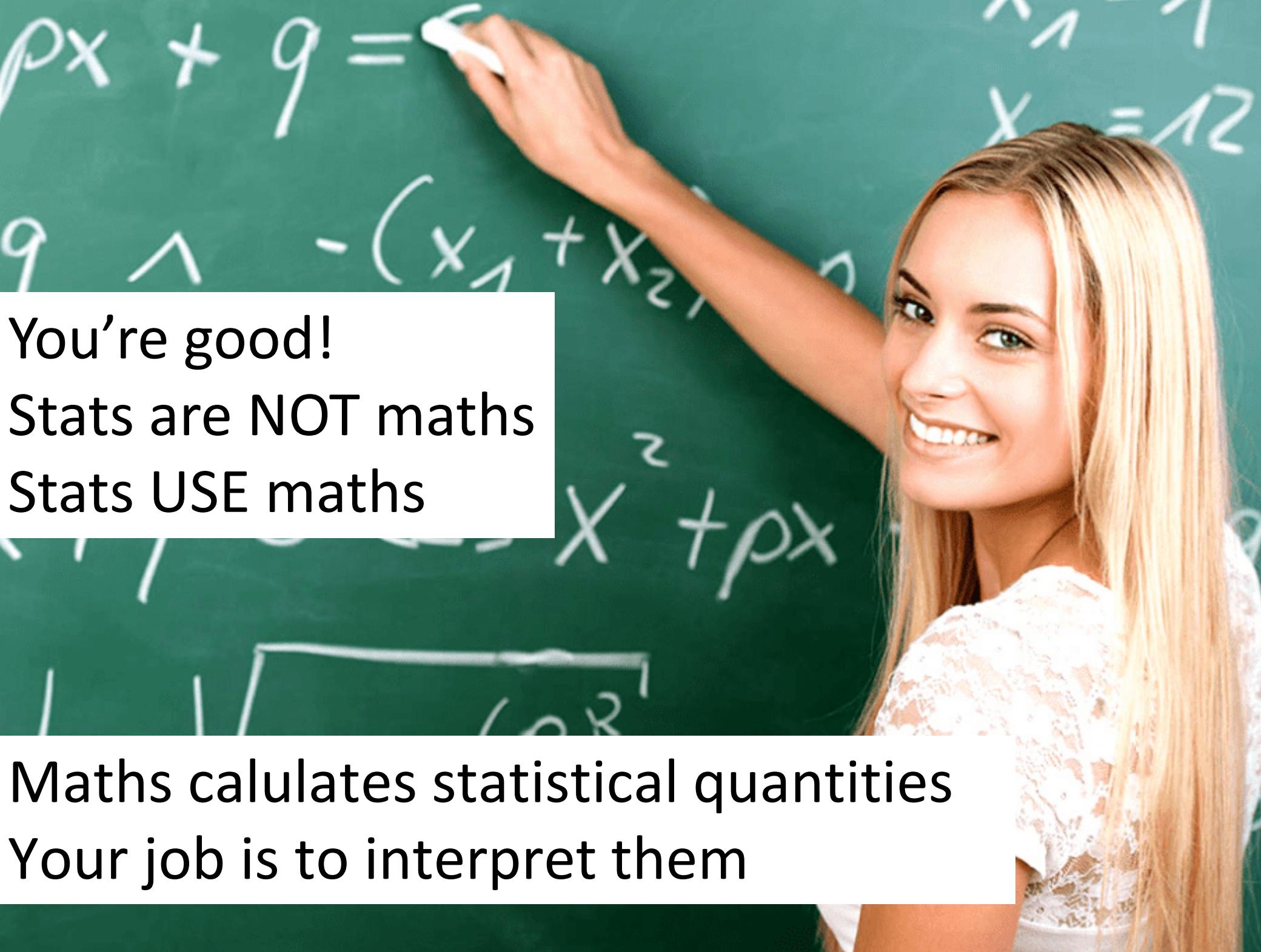
Compare groups

Determine relationship between variables

3. Interpret them



But what if I don't do maths?!



You're good!

Stats are NOT maths

Stats USE maths

Maths calculates statistical quantities

Your job is to interpret them

# Statistics

## Descriptive

Organizing, summarizing,  
and presenting data

## Inferential

Drawing conclusions about  
a population, based on  
data from a sample

# Statistics

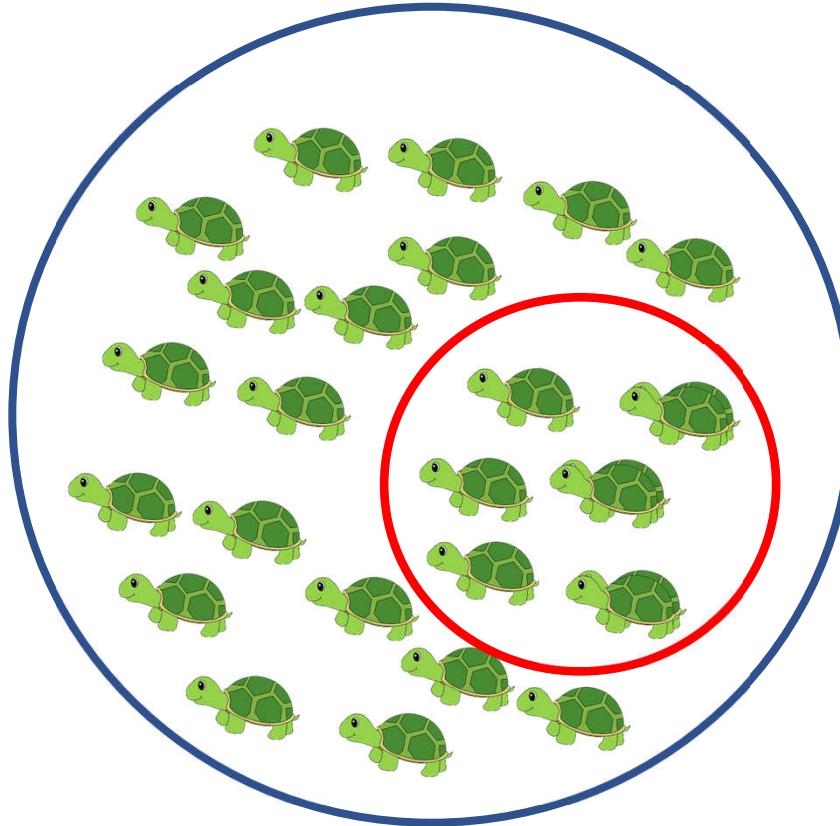
## Descriptive

Organizing, summarizing,  
and presenting data

## Inferential

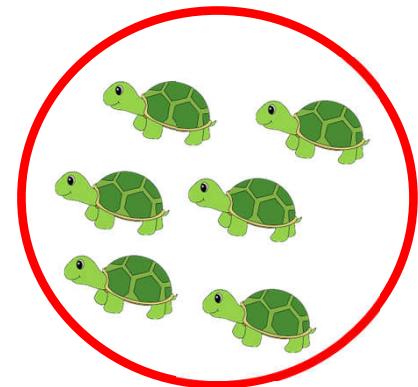
Drawing conclusions  
about a population, based  
on data from a sample

?



Population: impossible to observe all outcomes

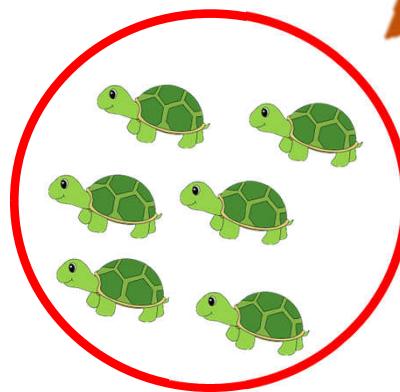
Sample:  $n$  individuals independently obtained and characteristic of your population



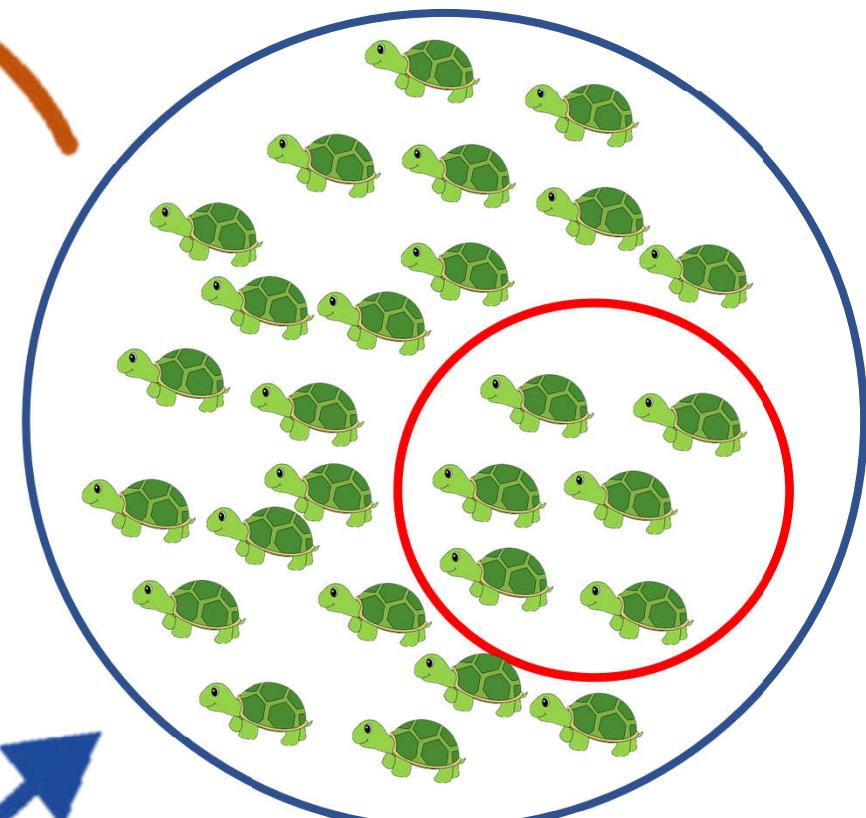
Your data

Population parameter:  
unknown

## Sampling



Calculate quantity in  
sample: “statistic”



Make inferences:  
approximate true population  
values with estimators

# Statistic vs Parameter

Sample		Population
$\bar{x}$	$\leftarrow$ mean	$\rightarrow \mu$
$s$	$\leftarrow$ St. dev.	$\rightarrow \sigma$
$\hat{p}$	$\leftarrow$ proportion	$\rightarrow p$
$n$	$\leftarrow$ Size	$\rightarrow N$

A note on semantics:

- Statistic describes a sample
- Parameter describes a population

A note on  
semantics



# Descriptive statistics

Learning objectives:

- Define descriptive statistics
- Different data types
- Measure **central tendency**
- Measure **variability**
- Define variable

# Descriptive statistics

Measure quantities of statistical populations  
that summarise or describe an aspect of the  
population

**No generalization** beyond your data

Data

Quantitative

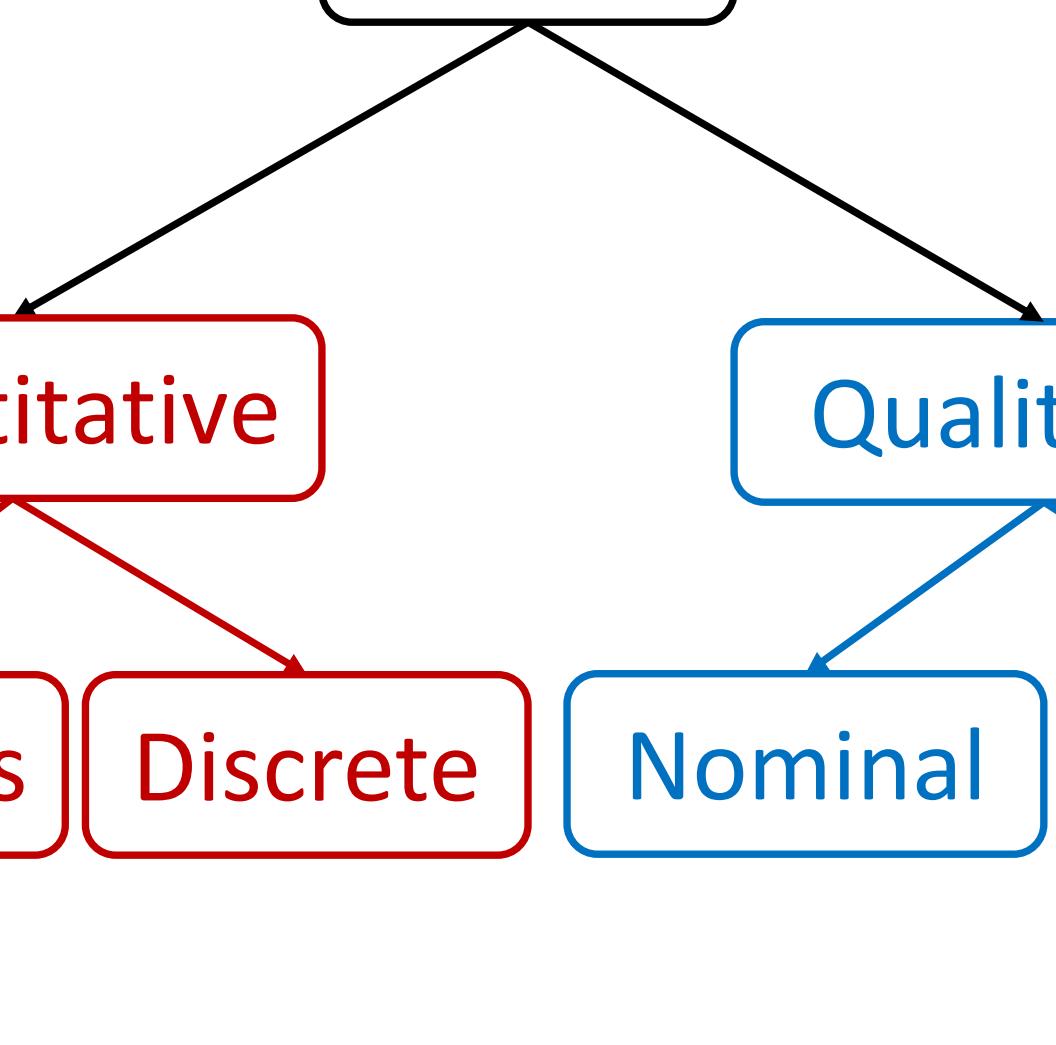
Continuous

Discrete

Qualitative

Nominal

Ordinal



Data

Quantitative

Continuous

- Theoretically: infinite number of values within a range
- In practice: limited by measurement accuracy

Ex.:

- Wingspan
- Body mass
- Animal growth
- Skull length
- precipitation

Data

Quantitative

Discrete

- Limited number of values within a range
- Often but not always integers

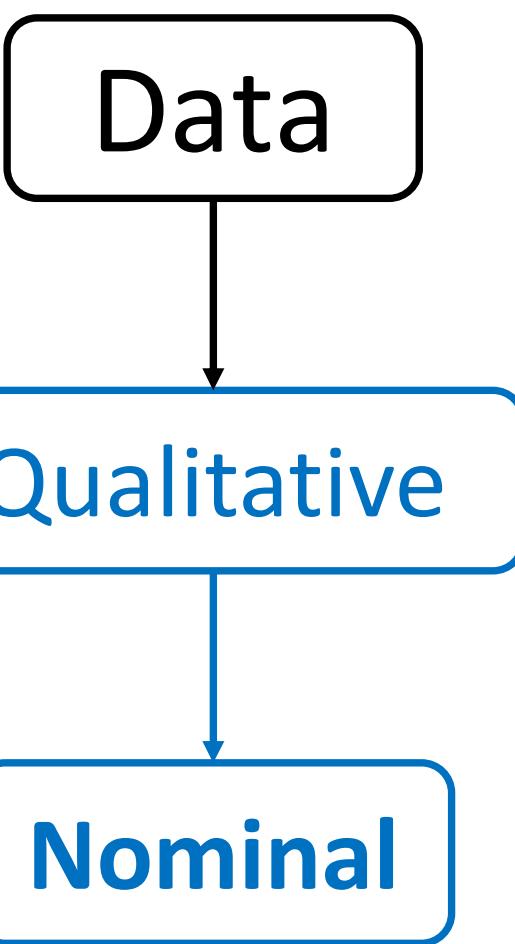
Ex.:

- Number of rain days in a month
- Number of viable chick per clutch
- Number of dens occupied by red foxes

# Continuous vs discrete

The distinction can be blurred:

- # of cells per ml blood: integer therefore discrete, but thousands of cells so behave as continuous
- Distance is continuous, but if you measure flushing distance of geese to the nearest 5m it will behave as discrete



- Partition data into classes
  - NOT ordered
  - Also called factors with levels
  - Special case: binary data
  - Also called categorical data
- Ex.:
- Fur color (black, grey, spotted)
  - Species (lynx, wolf, fox)
- Ex. binary:
- Occurrence (present, absent)
  - Sex (Male, Female)

Data

Qualitative

Ordinal

- Partition data into classes
- Intrinsically ordered
- Also called “Ranked” data

Ex.:

- Body condition index (High, medium, low)
- Index of growth rate (slow, medium, fast)
- Timing of molt (before, during after migration)

*One more type you will likely  
need...*

# Derived or computed data

- Calculated from 2 (or more) raw data:
  - Ratio
  - Proportion
  - Percentage
  - rate



Derived or computed data

- Loss of accuracy (rounding)
- Loss of information

$$5/10 = 0.5$$
$$500/1000 = 0.5$$

But are they the same?



2% of my 50 animals are sick...

But, really, that is only 1 animal



## Derived or computed data

- Awkward to model (often bounded – ex. Proportion: [0-1], percentage: 0-100%)
- Yet we often use derived data in biology: we'll see tricks to deal with them later

# Describe your data with numbers

- Central tendency and spread
- What is a typical value of my dataset?
- How scattered is my data?
- How do my data typically differ from the mean?
- How skewed is my data?

# Quantitative data

# Quantitative data

## 1. Central tendency

# Describing and summarizing the data – central tendency

- The arithmetic mean or average:

- by far the most commonly used

- sample  $\bar{X} = \frac{\sum X}{N}$

- $\mu$  for population

# central tendency

- The median:
  - The next most common measure
  - Middle value of ranked data
  - I will not scare you off with an equation ;)

# central tendency

- The geometric mean:
  - Good for right-skewed datasets (example lognormal data)
  - **Geometric Mean** =  $\sqrt[n]{x_1 x_2 \dots x_n}$
  - Always smaller than the arithmetic mean

# central tendency

- The harmonic mean:

- **Harmonic Mean Formula** = 
$$\frac{n}{\left( \frac{1}{x_1} + \frac{1}{x_2} + \frac{1}{x_3} + \dots + \frac{1}{x_n} \right)}$$
- Hence, impossible to use if some  $x = 0$
- Rarely used
- But gives higher weight to smaller values than other means

# central tendency

- The weighted mean:
  - Higher weight to certain values
  - $\text{weighted mean} = x_1 * \text{weight}_1 + x_2 * \text{weight}_2 + \dots + x_n * \text{weight}_n$
  - Used when some values are more likely or more important than others

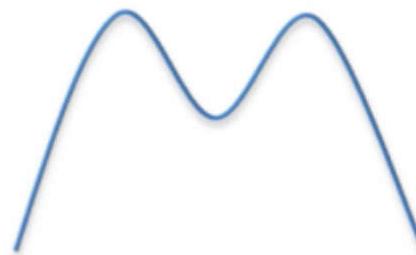
# central tendency

- The Mode:
  - The most frequent value of the dataset
  - Can be used with categorical data
  - Useful with very large datasets or data with low accuracy
  - Some datasets don't have a mode, while others have  $>1$  (multimodal)

**Unimodal**



**Bimodal**



**Multimodal**





The logo consists of a large, stylized letter 'R' in blue, centered within a white circle with a thick gray border. The background of the slide features a dark, textured pattern resembling fire or lava.

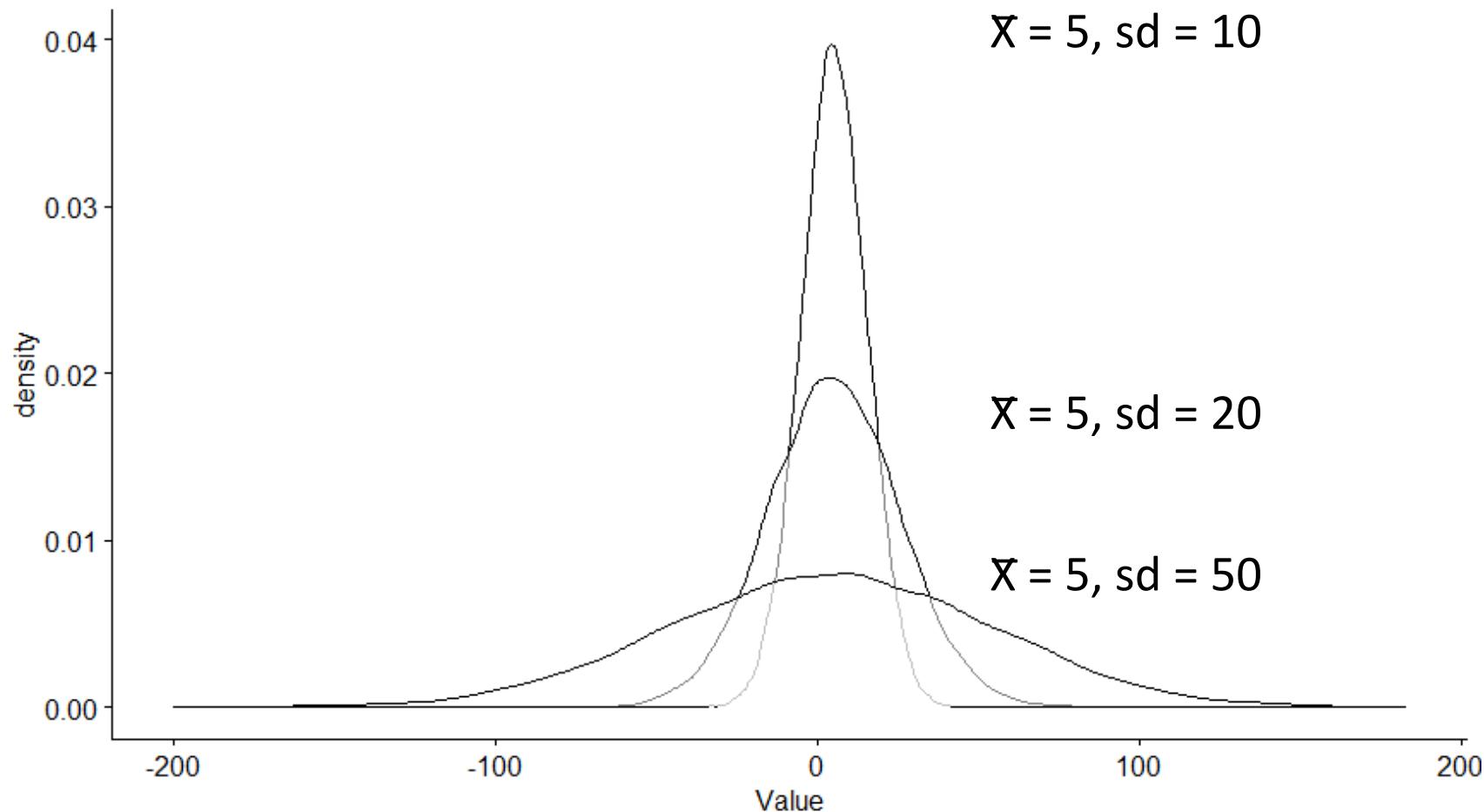
Intermission

# Quantitative data

## 1. Variability

# Describing and summarizing the data – variability

- Spread, dispersion



# Describing and summarizing the data – variability

- Many ways to characterize dispersion, but choose according to:
  - Data
  - Central tendency measure (ex. Mean)
  - Your question

# Variability

- Range:
  - The simplest one
  - $[\text{min. value} - \text{max. value}]$
  - Sensitive to outliers!

# Variability

- Variance
  - Estimate  $s^2$  of the true variance of the population ( $\sigma^2$ )
    - $$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$$
    - Rarely used as a descriptive stat
- Standard deviation (SD)
  - Estimate ( $s$ ) of the true standard deviation ( $\sigma$ )
  - Square root of Variance
  - Often used, but not ideal to compare different datasets

# Variability

- Standard Error (SE):

- Commonly used

- $$SE = \frac{\sigma}{\sqrt{n}}$$

- If you compare samples, confidence intervals are better

- Confidence intervals (CI):

- The most useful measure of dispersion
  - You can customize (90% CI, 95% CI) depending on the needs
  - 95% CI of the mean: “you are 95% sure that the interval contains the true mean of the population”
  - Measure of confidence

# Variability

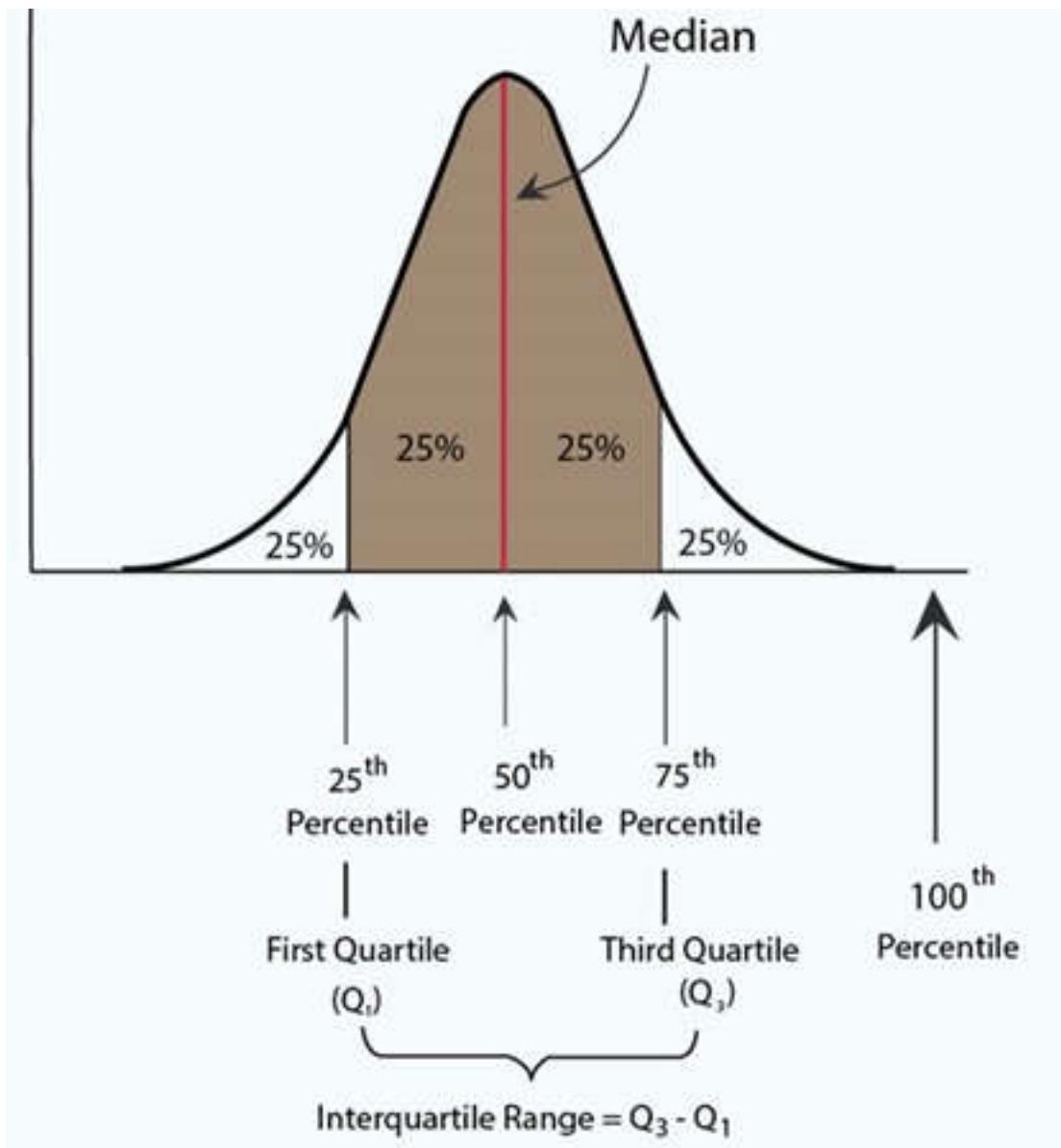
- Coefficient of variation (CV):
  - Widely used in labs to assess the precision and repeatability of an assay
  - $$CV = \frac{\sigma}{|\mu|} * 100$$
  - Relative measure of variation

# Variability

- Interquartile Range (IQR):
  - Non-parametric measure of dispersion
  - Works with ranked data
  - Use it if you describe the central tendency with a median
  - Unaffected by outliers or sample size: so more useful than the range on that aspect!
  - Useful to identify the skewness of a dataset
  - Difference between 75<sup>th</sup> and 25<sup>th</sup> percentile of the data

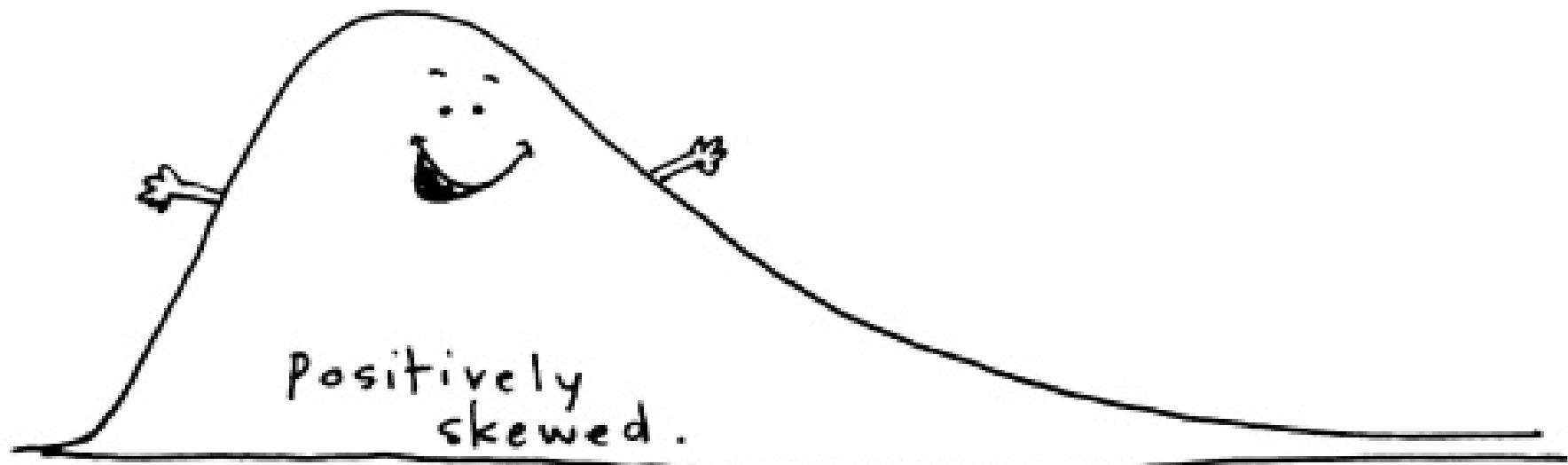
# Variability

- Interquartile Range:



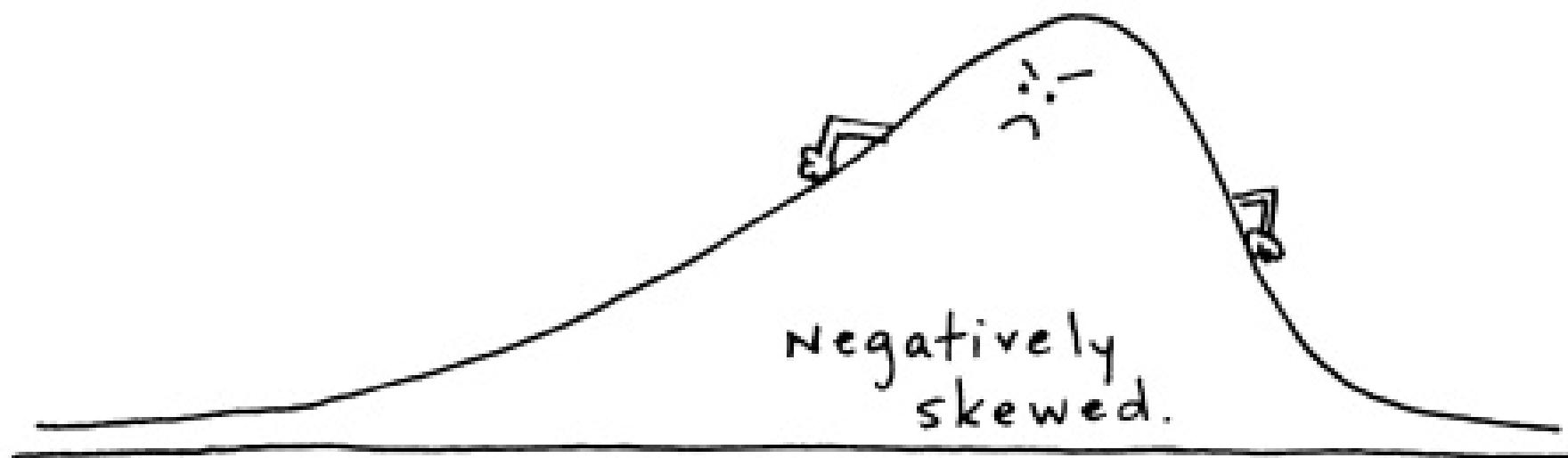
# Describing and summarizing the data – other summary statistics

- Skewness: symmetry of the dataset
  - Perfect symmetry: skewness = 0
  - Tail to the right (right-skewed):  $\text{skewness} > 0$



# Describing and summarizing the data – other summary statistics

- Skewness: symmetry of the dataset
  - Tail to the left (left-skewed):  $\text{skewness} < 0$



# Describing and summarizing the data – other summary statistics

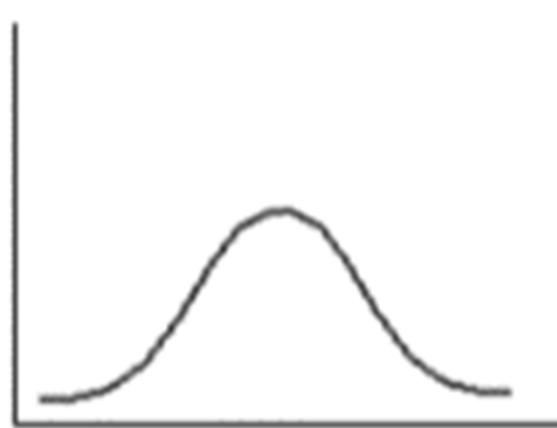
- Skewness: symmetry of the dataset
  - When the dataset is skewed: mean closer to the tail than the median (remember the lognormal R example)
  - **Not really useful to calculate skewness for small data set (i.e.,  $n < 30$ )**

# Describing and summarizing the data – other summary statistics

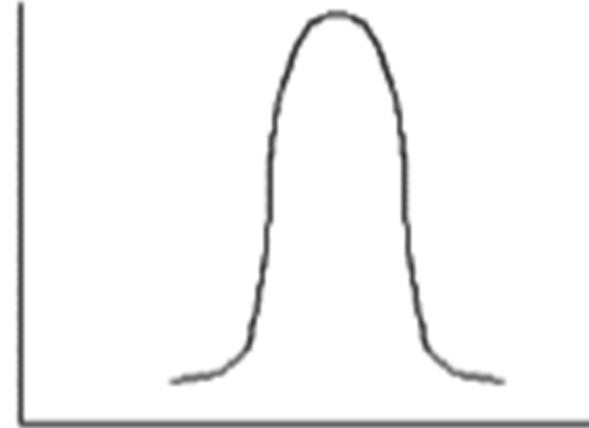
- Kurtosis: flatness of the data
  - Always in comparison to normal distribution
  - Only meaningful for large dataset
  - 2 big words: *leptokurtic* and *platykurtic*

# Describing and summarizing the data – other summary statistics

- Kurtosis: flatness of the data
  - *Leptokurtic*: data more concentrated around the mean



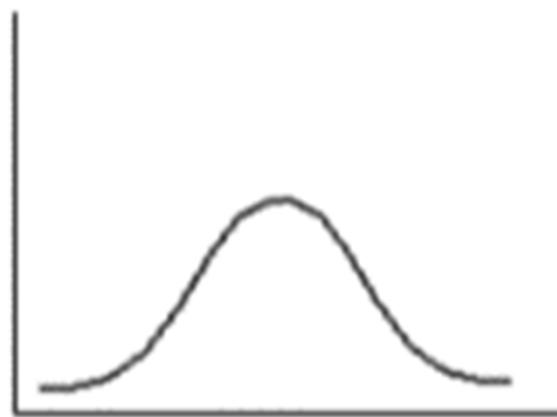
Normal curve



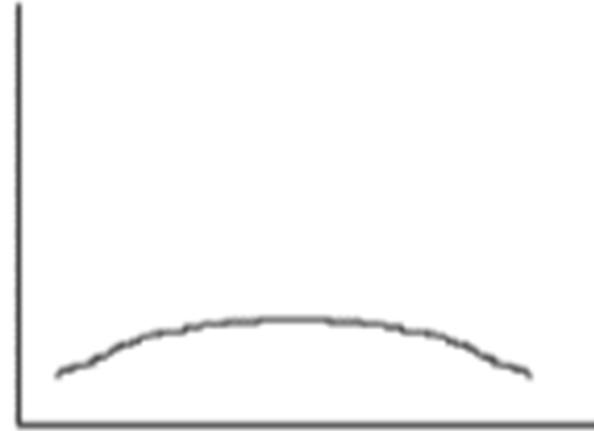
Leptokurtic curve

# Describing and summarizing the data – other summary statistics

- Kurtosis: flatness of the data
  - *Platykurtic*: thinner tails than Normal distribution, so less extreme values



Normal curve



Platykurtic curve



Intermission

# Qualitative data

# Describing categorical data

- Frequency
- Proportions and percentages
- Marginals (total counts across columns or rows in contingency tables)
- Coefficient of unalikeability (how often observations differ from each other)
- Visualization (which we'll see in a future workshop)

# Ordinal categorical data

- Central tendency:

- Median
- Mode

The mean assumes equal spacing between categories, so cannot be used.

- Variability:

- Range
- Variation ratio (the simplest)

Standard deviation, Standard error and variance all depend on the mean, and thus cannot be used.

# Nominal categorical data

- Central tendency:
  - Mode is the most common
- Spread:
  - Rarely reported
  - > 50 indices of qualitative variation
  - Variation ratio (the simplest)



Intermission

# What is a variable?

- Any property that you can **measure** or **control**
- Likely changes between objects
- **Data** results from observations or measurements you make on one or more variables

Variable

Quantitative

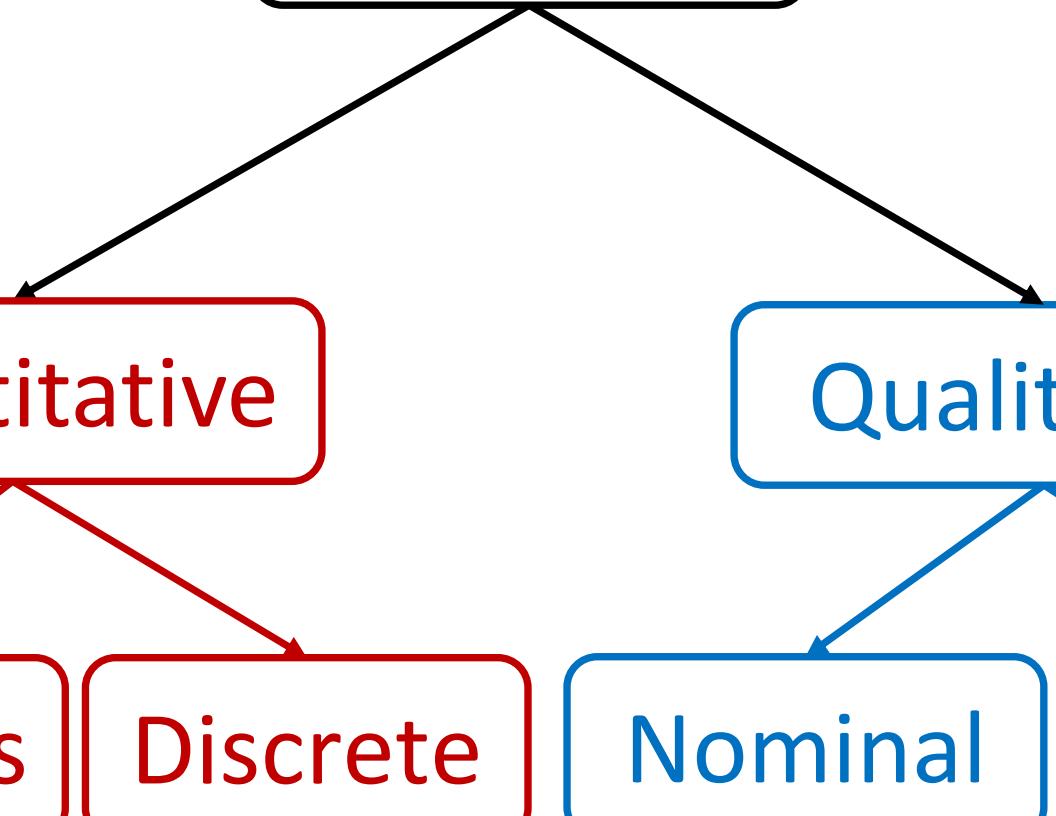
Continuous

Discrete

Qualitative

Nominal

Ordinal



# Choosing the right variable

- In hypothesis testing the choice of variable is key

Examples:

- You're interested in occurrence of a bird per site
  - Presence/absence of bird in each site
  - number of bird seen per site
- You want to know bird abundance by habitats, but some habitats are more present than others
  - number of birds per unit habitat area

# A reminder on semantics

Accuracy vs Precision

# A reminder on semantics

Precision

High

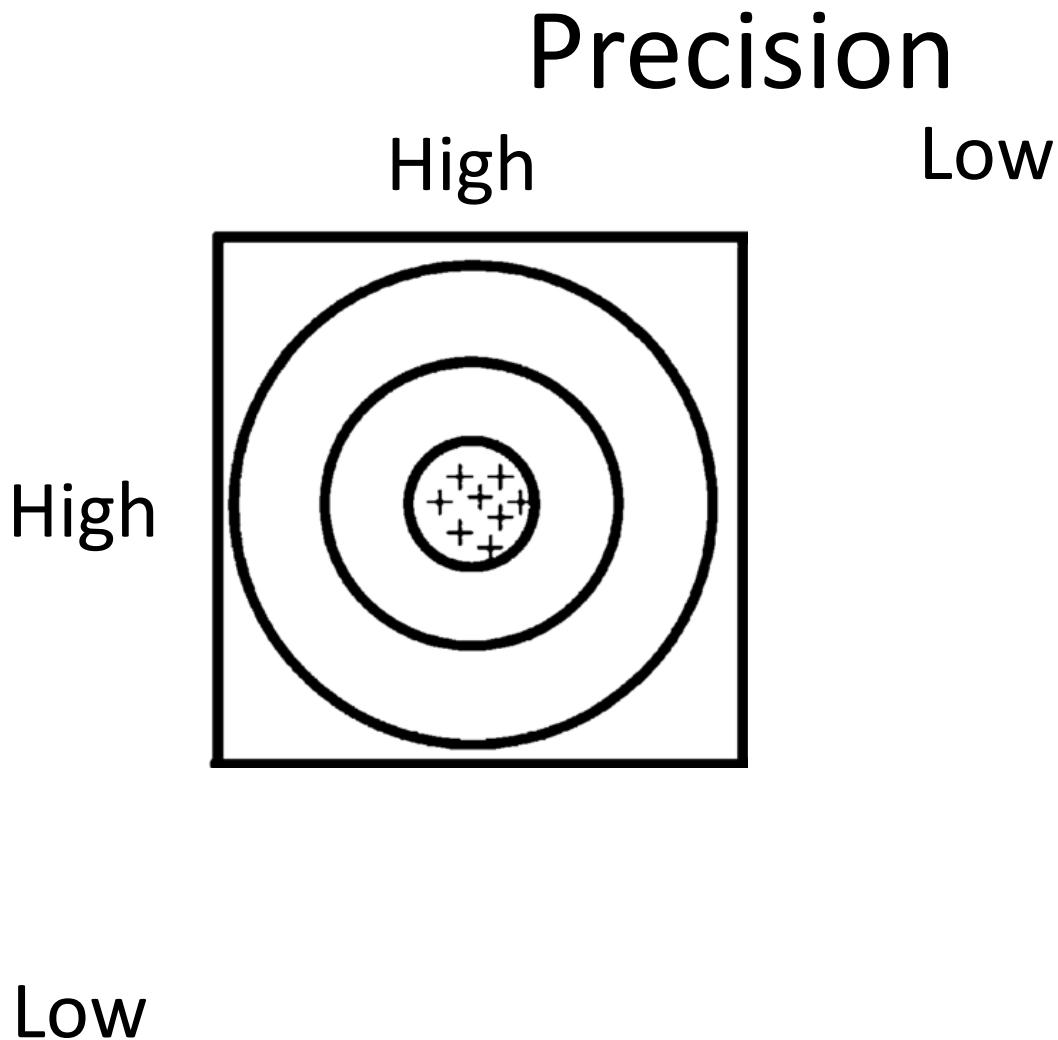
Low

High

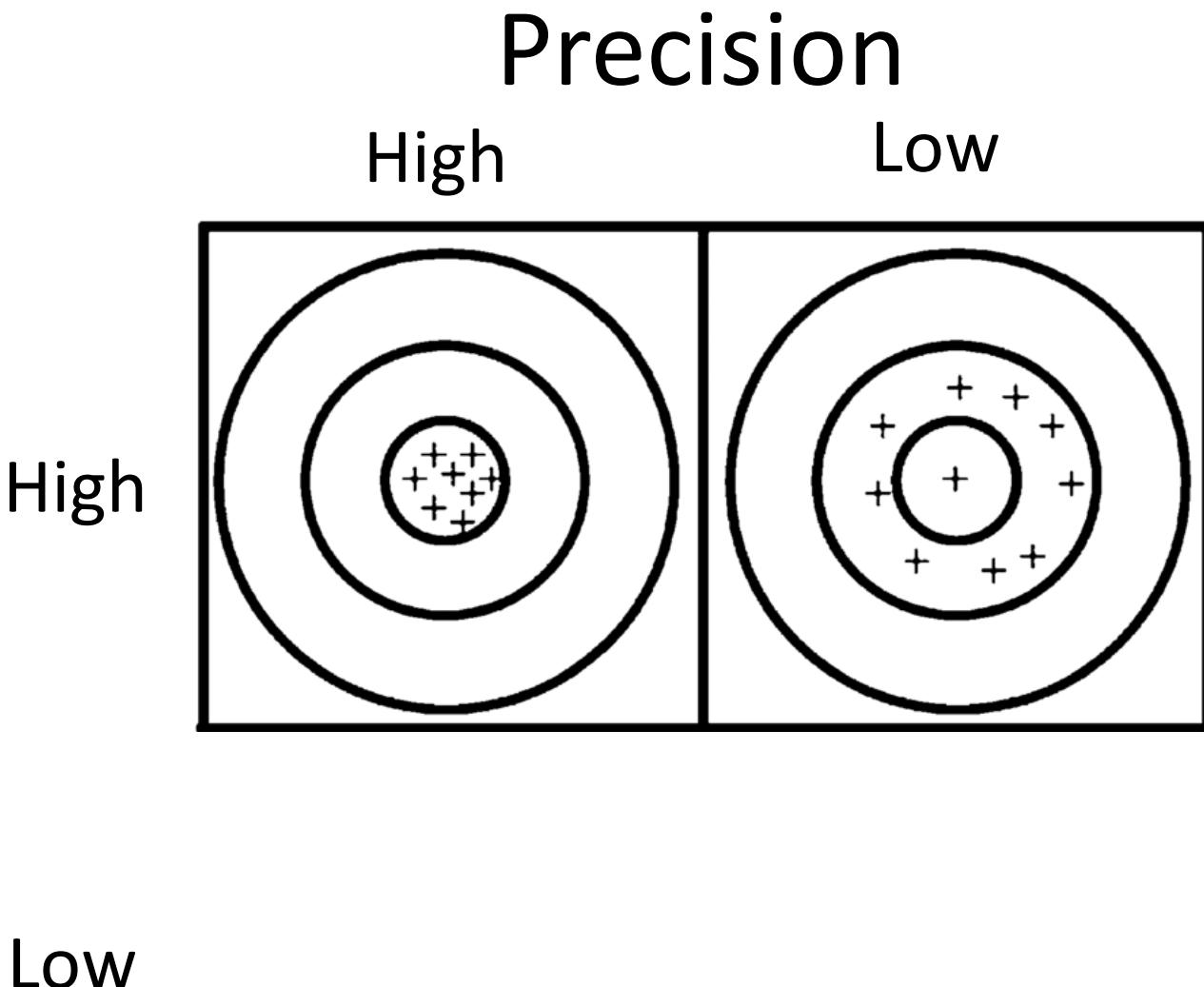
# Accuracy

Low

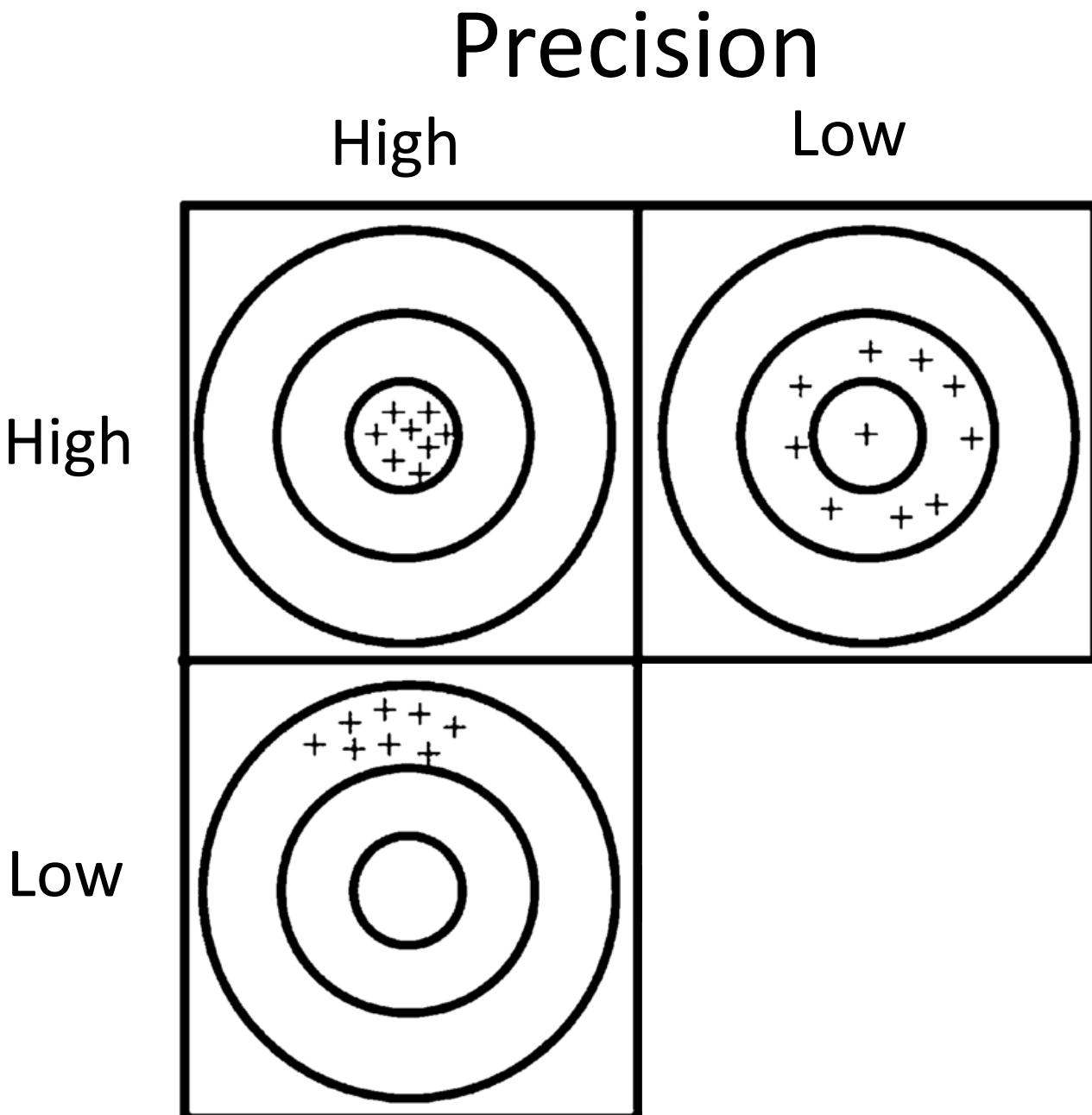
# A reminder on semantics



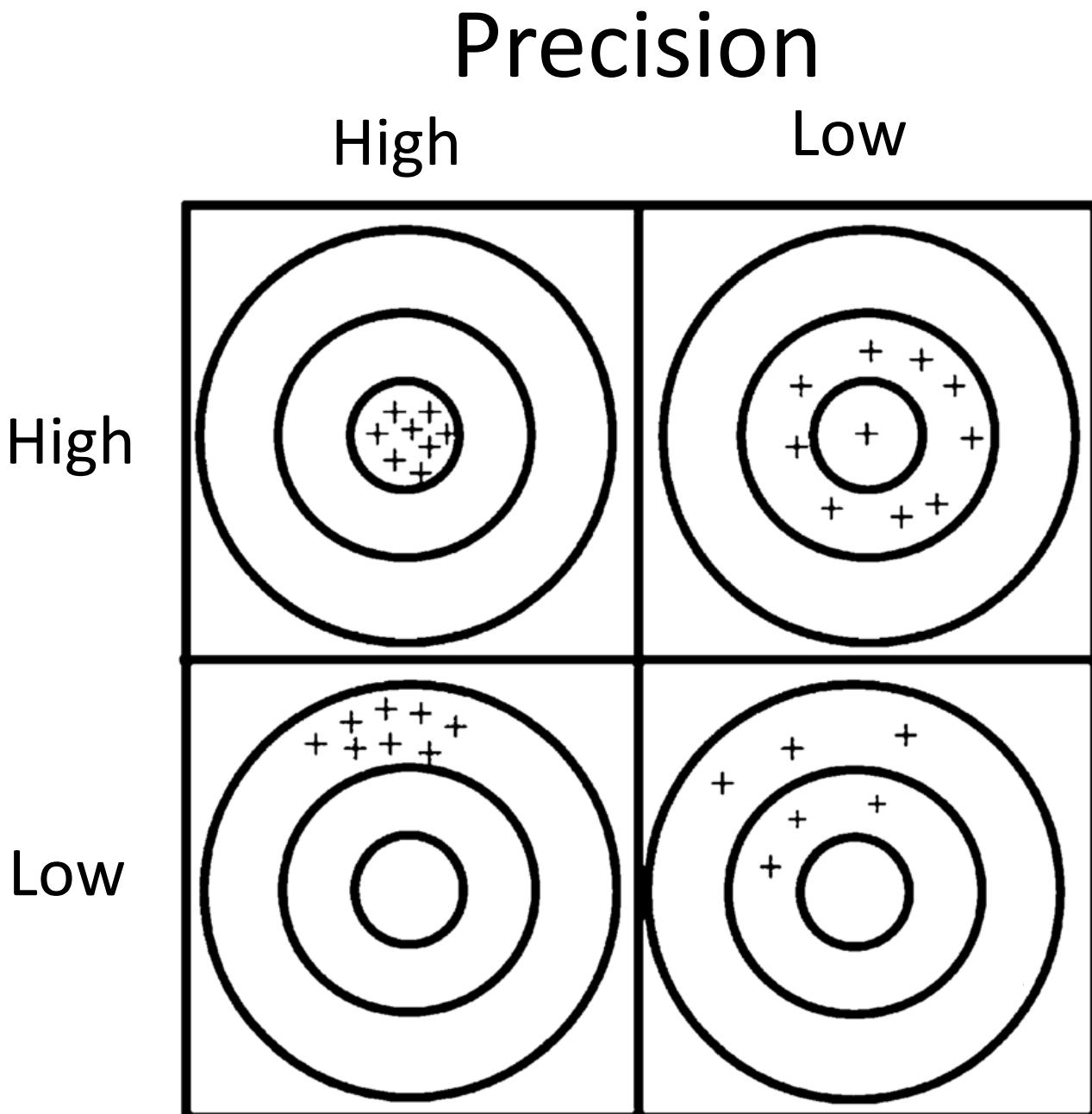
# A reminder on semantics



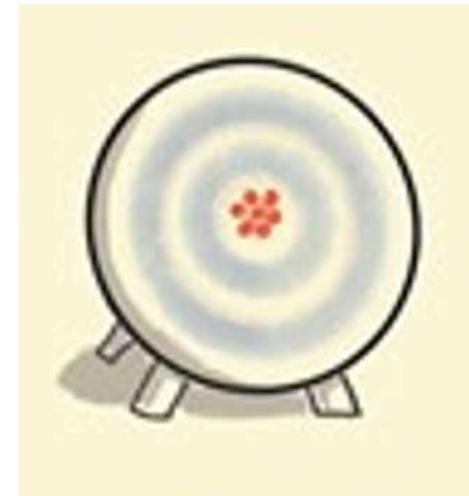
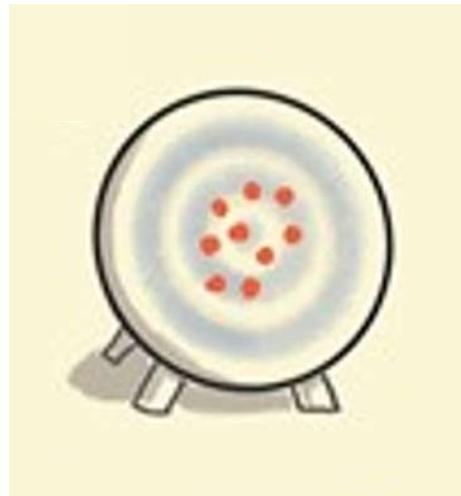
# A reminder on semantics



# A reminder on semantics

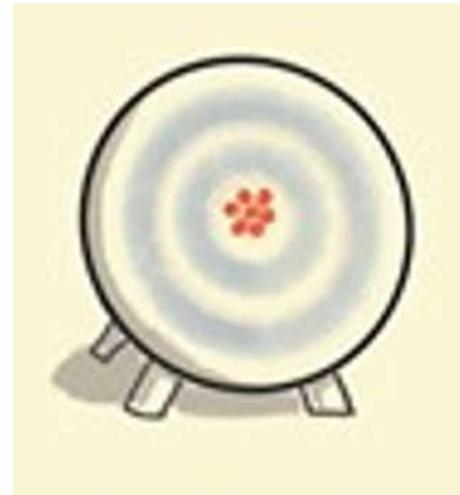


# A reminder on semantics



**Accuracy** relates to the quality of the results

# A reminder on semantics



**Precision** relates to the quality of the process by which the results are obtained

# Hypothesis testing: two types of variables

- Explanatory variable (or “independent”, or “predictors”): expected cause
- Response variable (or “dependent variables”): expected effect

In general, researchers observe changes in explanatory variables to explain effects on the response variable they measure.

# Hypothesis testing: two types of variables

The R way to write a model:

$$Y \sim X_1 + X_2$$

# Hypothesis testing: two types of variables

The R way to write a model:

$$Y \sim X_1 + X_2$$

Explanatory  
variable 1

# Hypothesis testing: two types of variables

The R way to write a model:

$$Y \sim X_1 + X_2$$

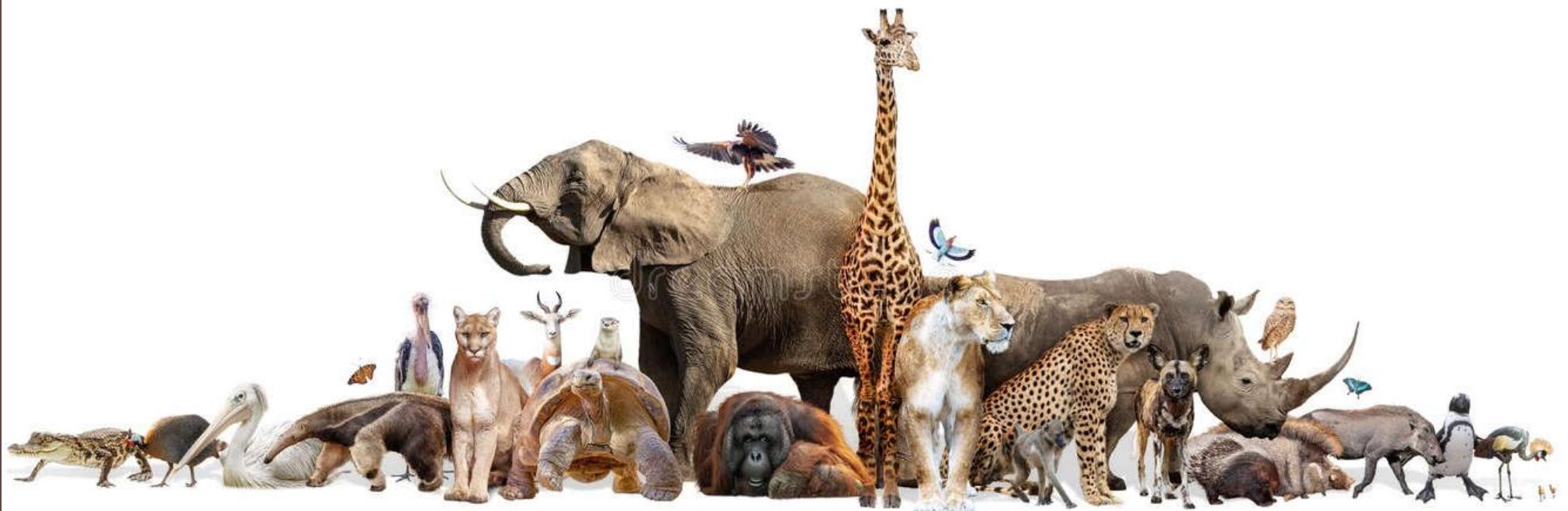
Explanatory  
variable 2

# Hypothesis testing: two types of variables

The R way to write a model:

$$\textcircled{Y} \sim X_1 + X_2$$

Response  
variable



Your turn to work:  
Find the explanatory and response variables

---

# Your turn to work

- You want to know the effect of protein content in pellets on chick growth
- Explanatory: protein content
- Response: chick growth rate



A close-up photograph of a leopard cat's face. The cat has large, expressive brown eyes and a light-colored coat with dark spots. It is wearing a silver tracking collar. The background is blurred green foliage.

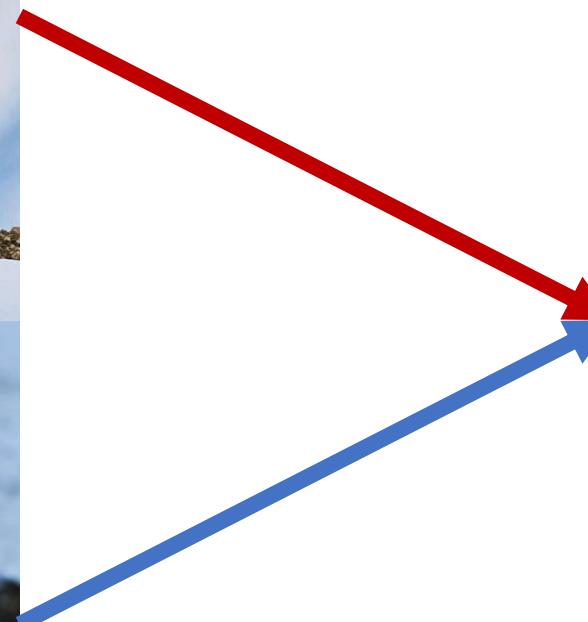
# Your turn to work

- You want to know the circadian rhythm of leopard cats
- Explanatory: hour of the day
- Response: activity rate of leopard cats

# Your turn to work

You are comparing Argali sheep's response to predators with different hunting modes

Explanatory: hunting mode



Response: prey spatial behavior

A close-up photograph of a fluffy, long-haired cat, likely a Persian or similar breed. The cat has thick, dark and light brown fur, intense yellow eyes, and a wide-open mouth showing its teeth and tongue. It appears to be hissing or meowing loudly. The background is blurred, showing a natural outdoor setting with gravel.

Questions?