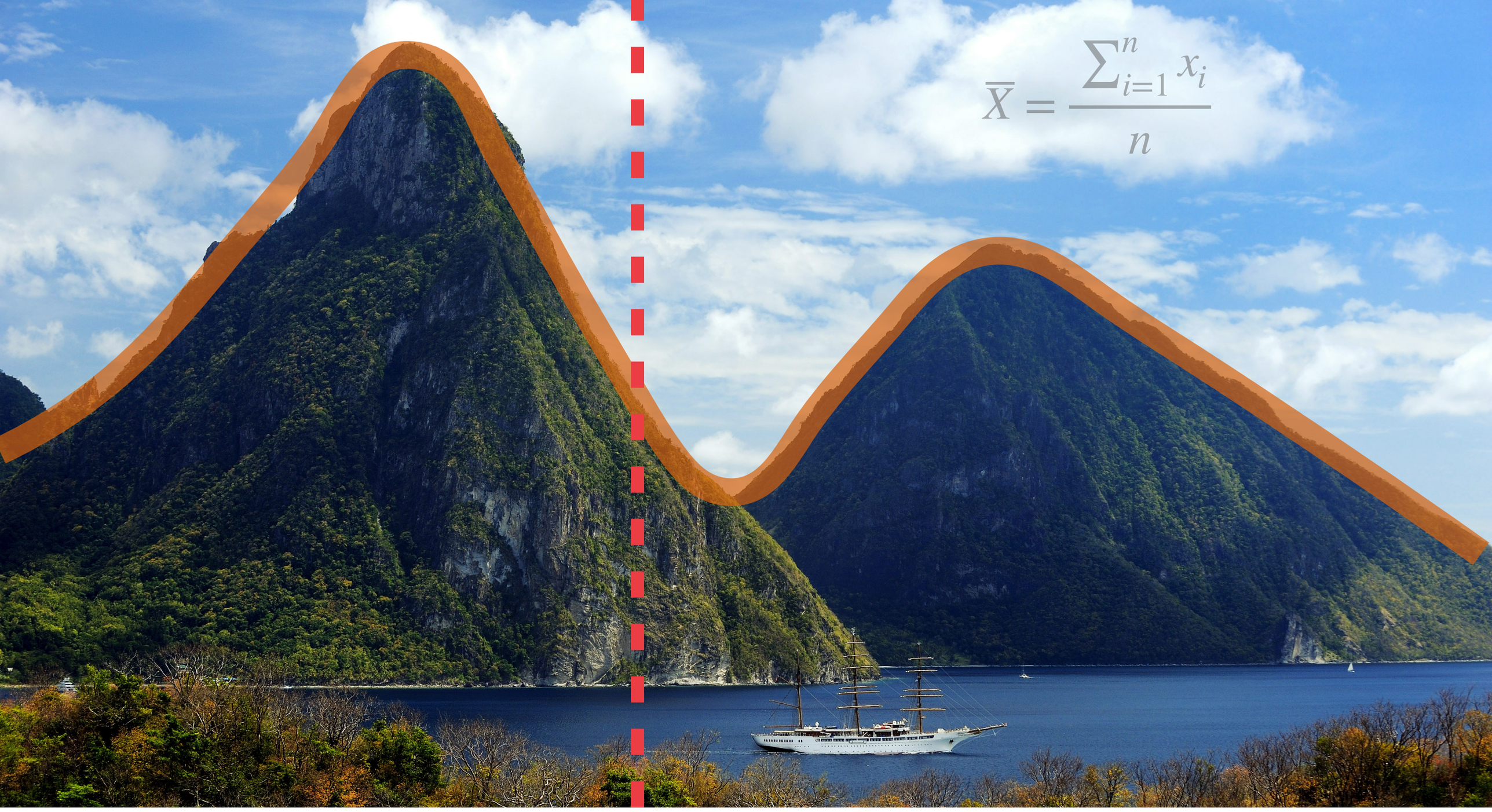


Building statistical models

Methods 1, E2021 - Lecture 3
Tuesday 14/9/2021
Fabio Trecca

Attendance registration

Check in using the PIN-code on the blackboard

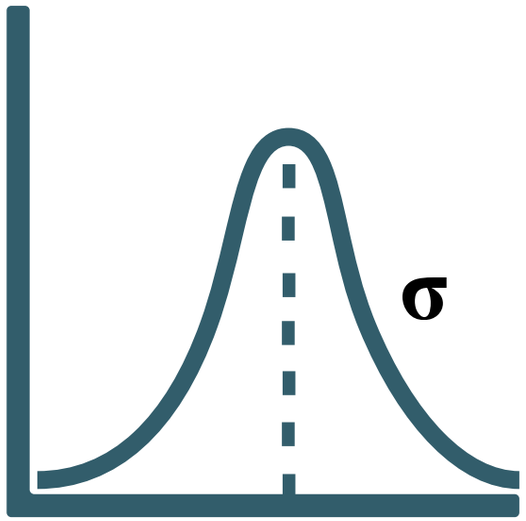
Recap: The normal distribution (1)

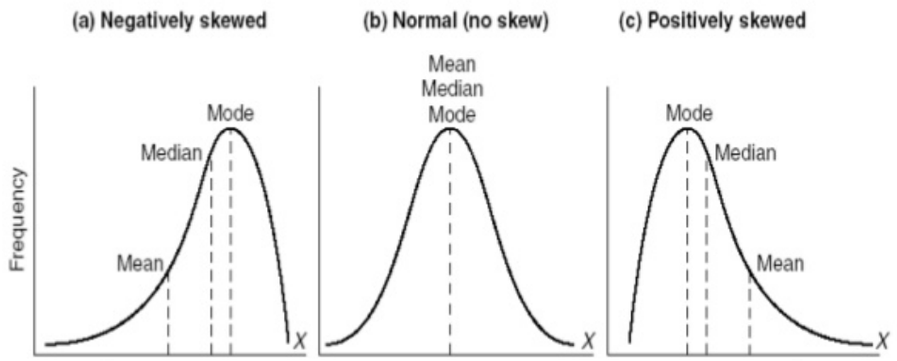
- Symmetrical gravitation toward the mean with decreasing N of data points as we approach the tails

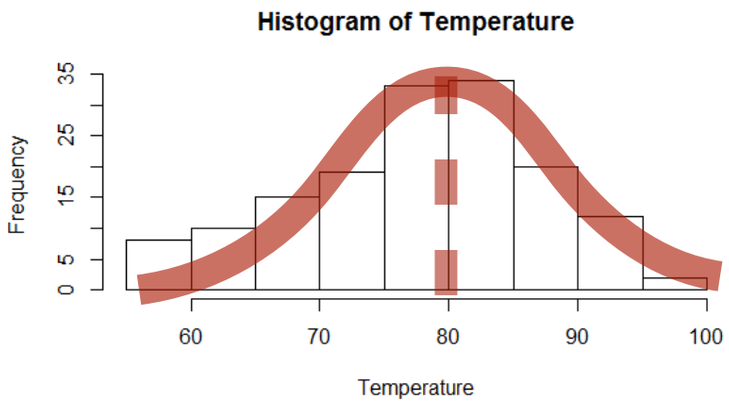
- Many cognitive and behavioural processes are normally distributed

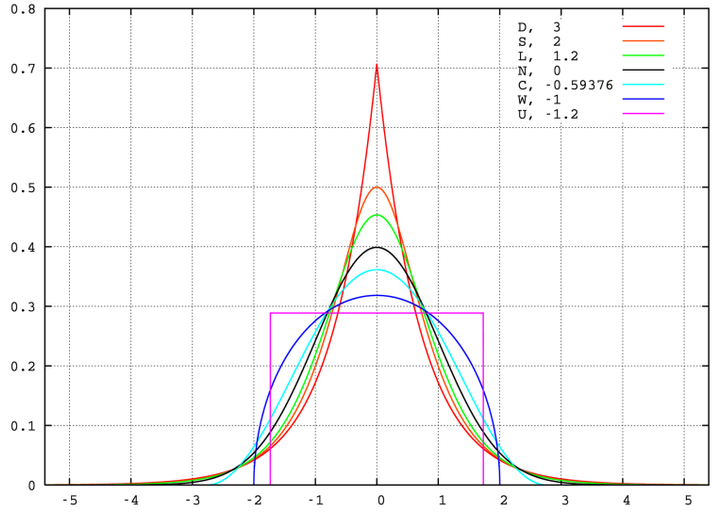
- Defined by two parameters: mean (μ) and standard deviation (σ)

- Results from sum of independent events/ factors







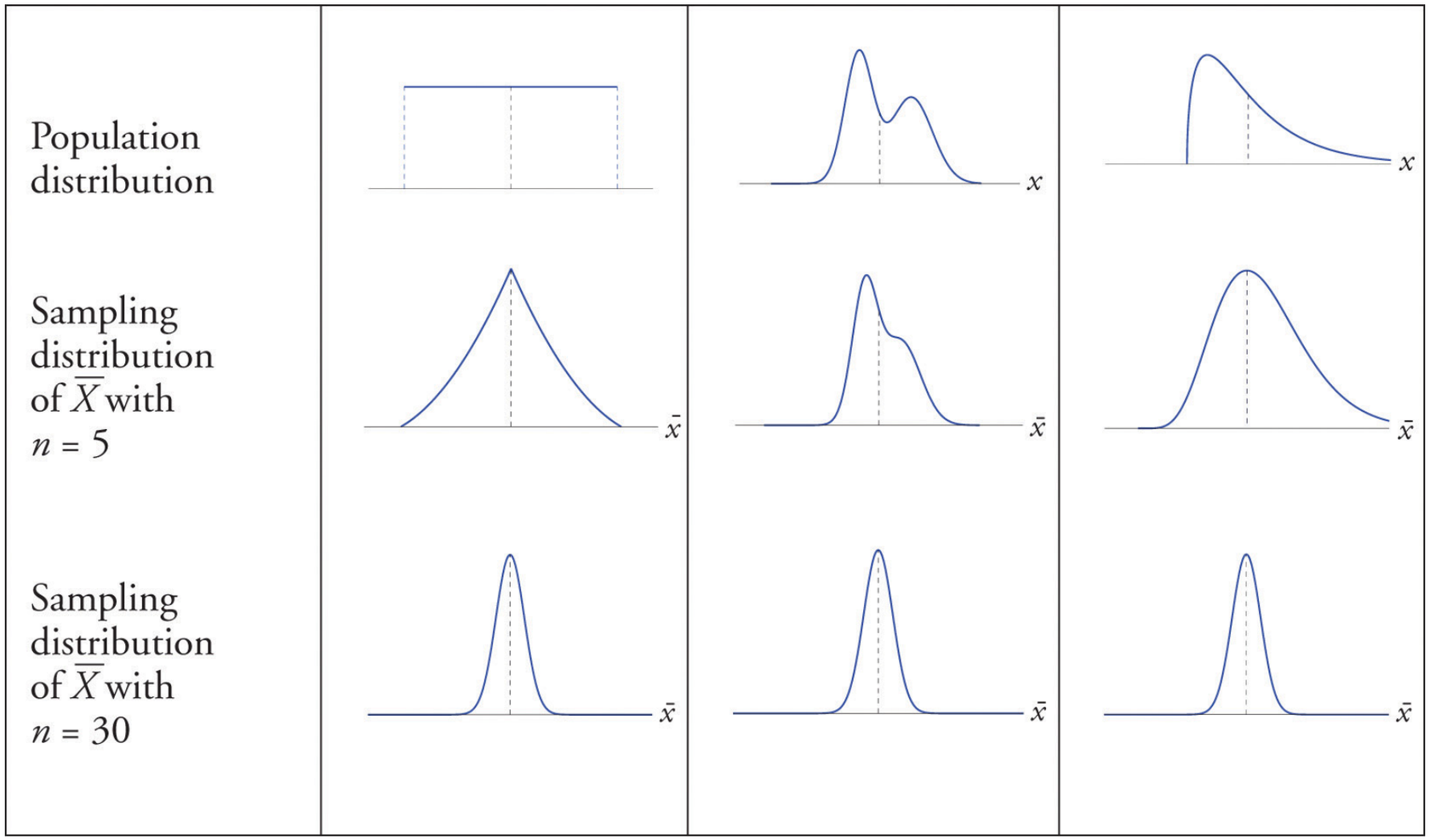


Demo

Recap: The normal distribution (2)

- "The primary significance of the normal distribution is that many chance phenomena are at least approximately described by a member of the family of normal probability density functions. If you were to collect a thousand snowflakes and weigh each one, you would find that the distribution of their weights was accurately described by a normal curve. If you measured the strength of bones in wildebeests, again you are likely to find that they are normally distributed. Why should this be so? [. . .] **It turns out that if we add together many random variables, all having the same probability distribution, the sum (a new random variable) has a distribution that is approximately normal**. [...] This result is formally called the **Central Limit Theorem**, and it provides the theoretical basis for why so many variables that we see in nature appear to have a probability density function that approximates a bell-shaped curve. If we think about **random biological or physical processes, they can often be viewed as being affected by a large number of random processes with individually small effects**. The sum of all these random components creates a random variable that converges on a normal distribution regardless of the underlying distribution of processes causing the small effects." (Denny & Gaines, 2000, pp. 82–83)

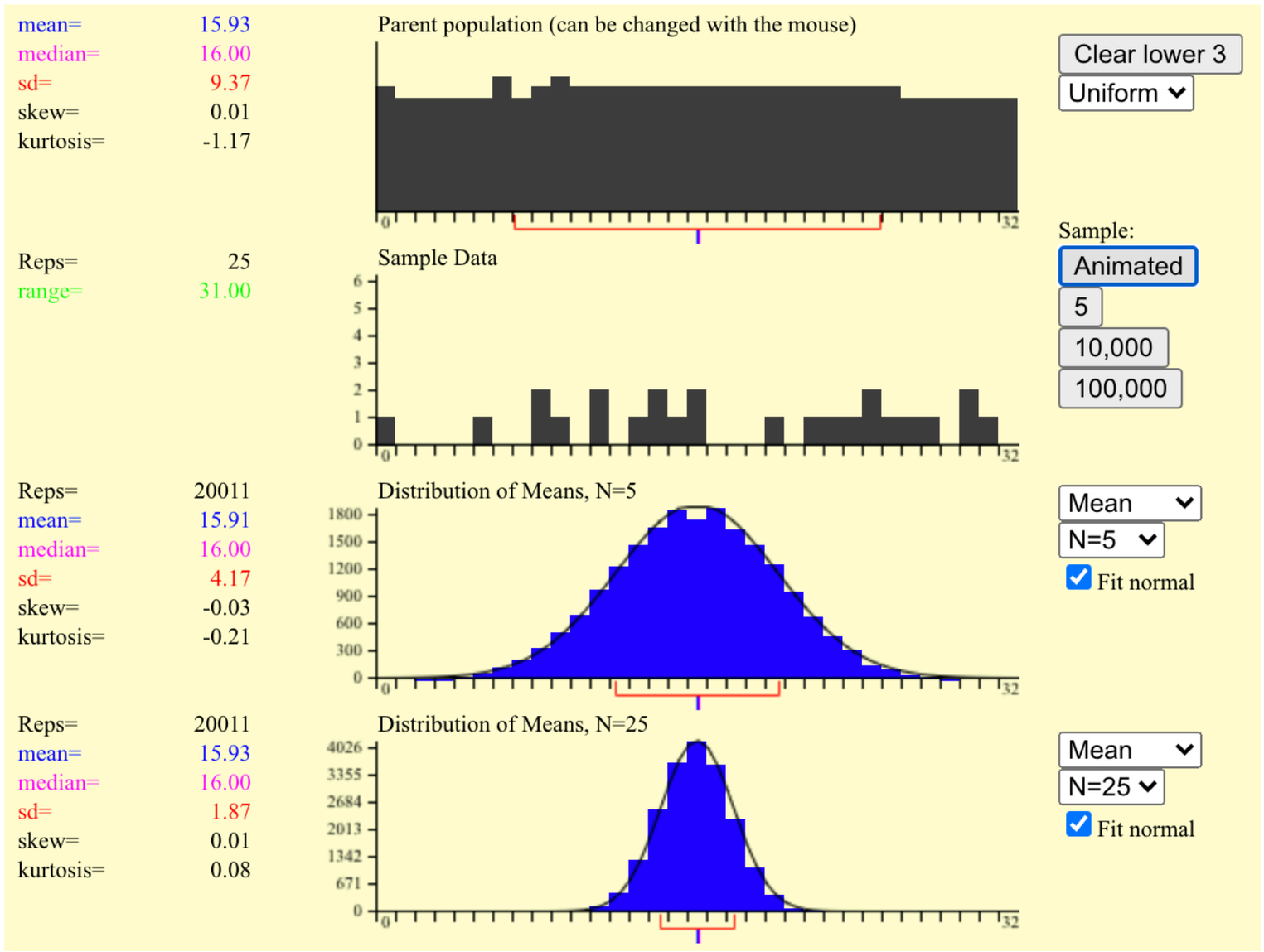
Central Limit Theorem:
Sampling distributions of the sample mean



https://saylordotorg.github.io/text_introductory-statistics/s10-02-the-sampling-distribution-of-t.html

Demo

- http://onlinestatbook.com/stat_sim/sampling_dist/index.html

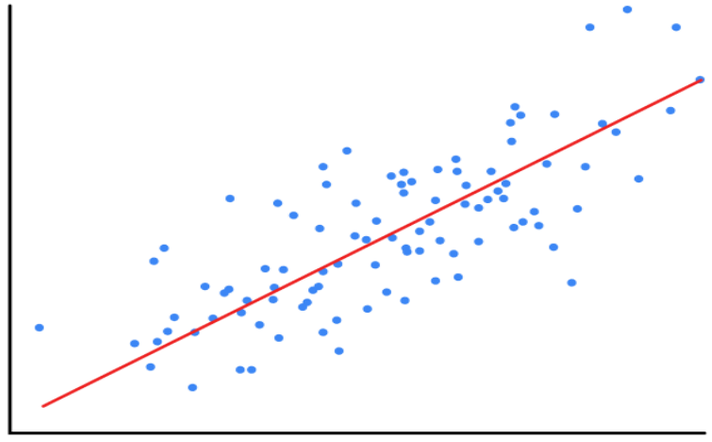


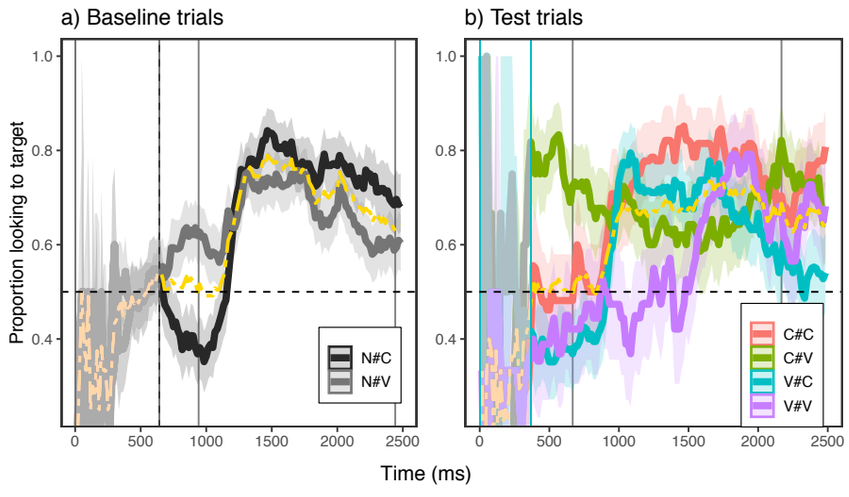
Building statistical models

- We fit models to the data in order to summarize the data

- Models represent real-world phenomena

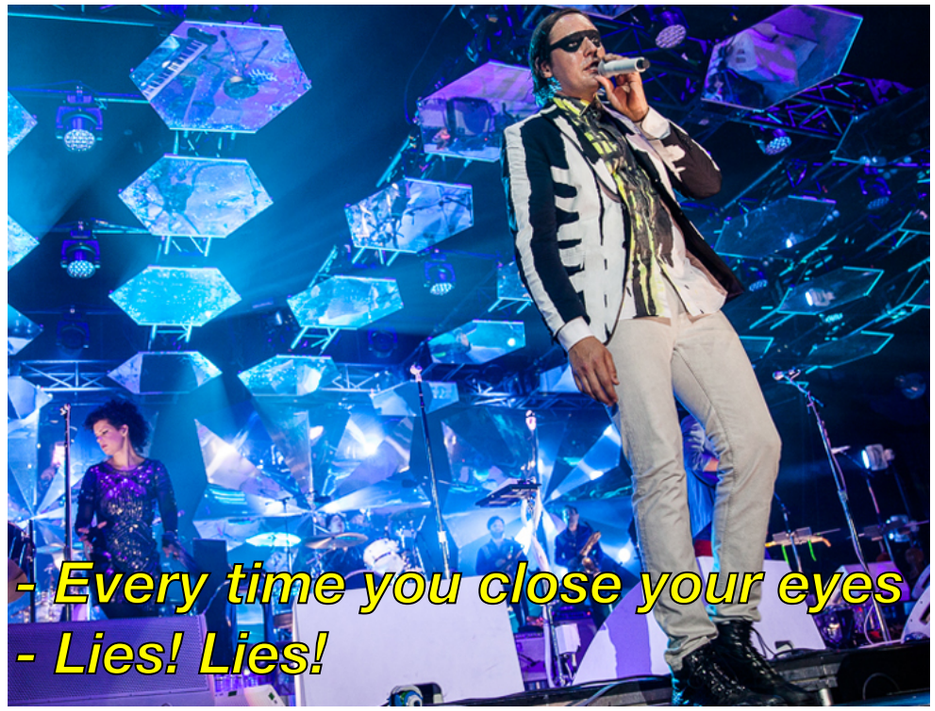
- Models are “wrong” by definition
  - hypothetical/summary
  - don’t necessarily include true values from the data (→ prediction)





Models are lies

- *"The average Danish couple has 1.67 children"*
- *"The average Dane drinks 9.1 liter of pure alcohol/year"*
- *"The average CogSci20 student listens to 13.74194 hours of music/week"*
- *"The average CogSci20 holds their breath for 52.61113 seconds"*
- Hypothetical values: not values that I observe in my dataset



A very useful model: The linear model

…participant *i* from a sample of participants

our statistical model

$$\textbf{Outcome}_i = \textbf{model} + \textbf{error}_i$$

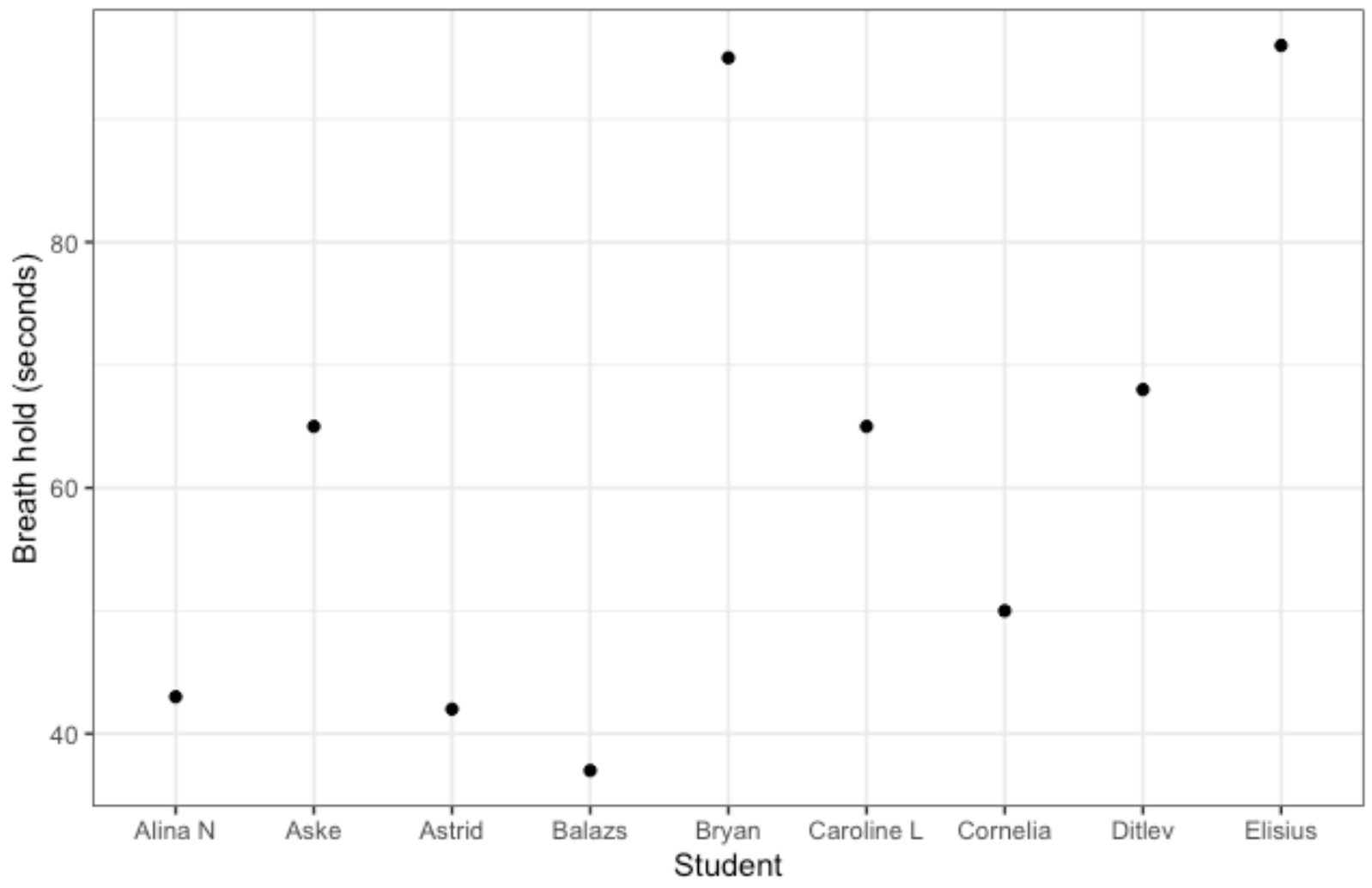
a certain observation/data point collected from…

is predicted by

and some error (= deviation from the model) for each participant *i*

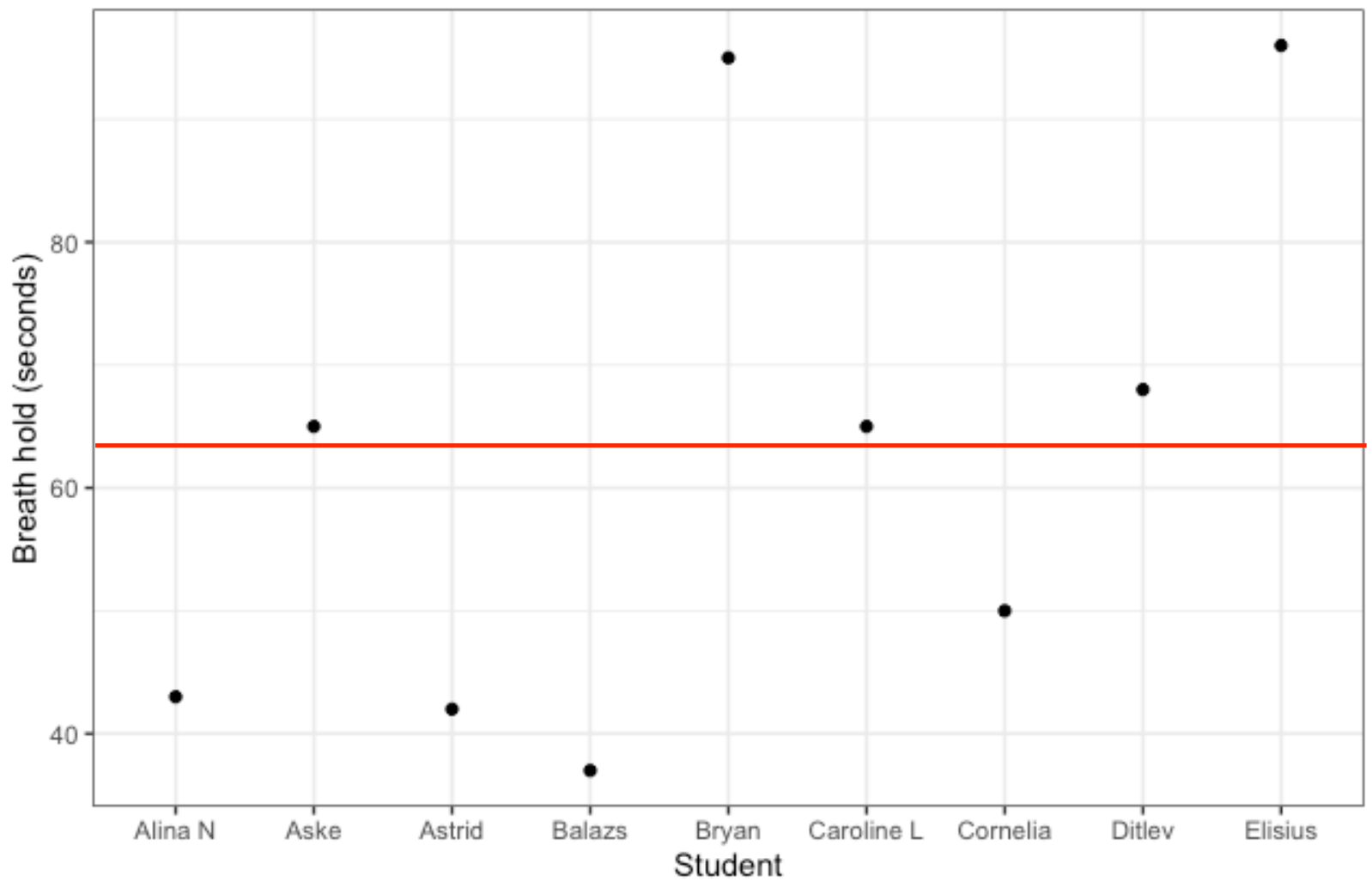
Building statistical models (3)

- E.g.: How good are CogSci21 students at holding their breath?



Building statistical models (3)

- E.g.: How good are CogSci21 students at holding their breath?



$\bar{x} = 62.33$

Mean: the simplest statistical model (1)

- $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

- $\bar{x}$ = the mean of the sample (not the population)

- $\sum_{i=1}^{n} x_i$ = the sum of each person's *x* for all *i*'s from *1* to *n*

- n = the total number of observations

Mean: the simplest statistical model (2)

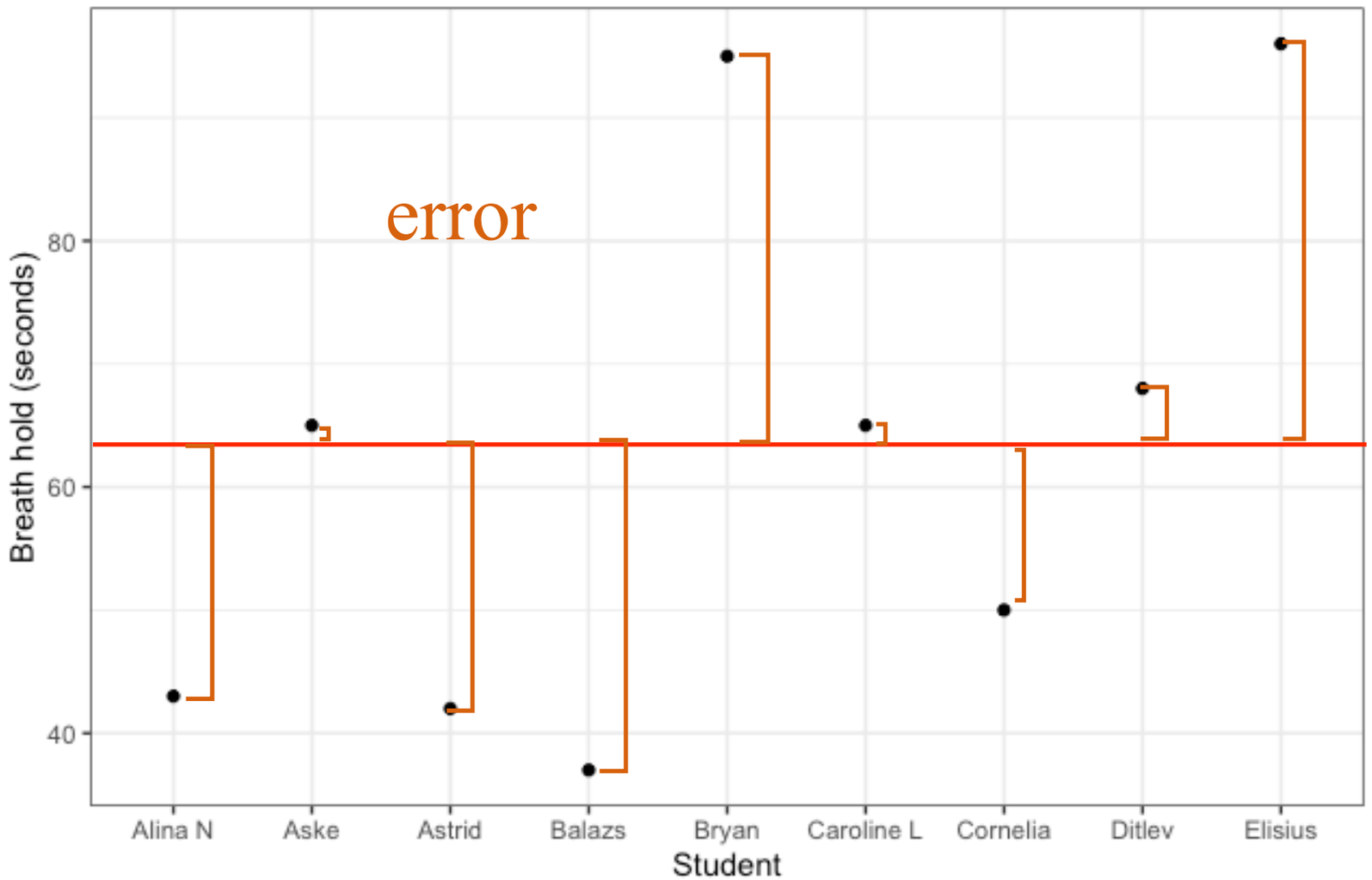
- In everyday terms:
  - Take all the values
    - `values <- c(43, 65, 42, 37, 95, 65, 50, 68, 96)`
  - Sum them up
    - `sum(values) = 561`
  - Divide it by N
    - `561/9 = 62.33`
  - …or just do:
    - `mean(c(43, 65, 42, 37, 95, 65, 50, 68, 96))`
    - `mean(df$breathhold)`

Mean as model

- $Outcome_i$ = model + $error_i$
- Breathhold = 62.33 + $error_i$
- Can be used **descriptively** or **inferentially** (= predictively)

The error part (1)

- E.g.: How good are CogSci21 students at holding their breath?



The error part (2)

$$\mathbf{Outcome_i = model + error_i}$$

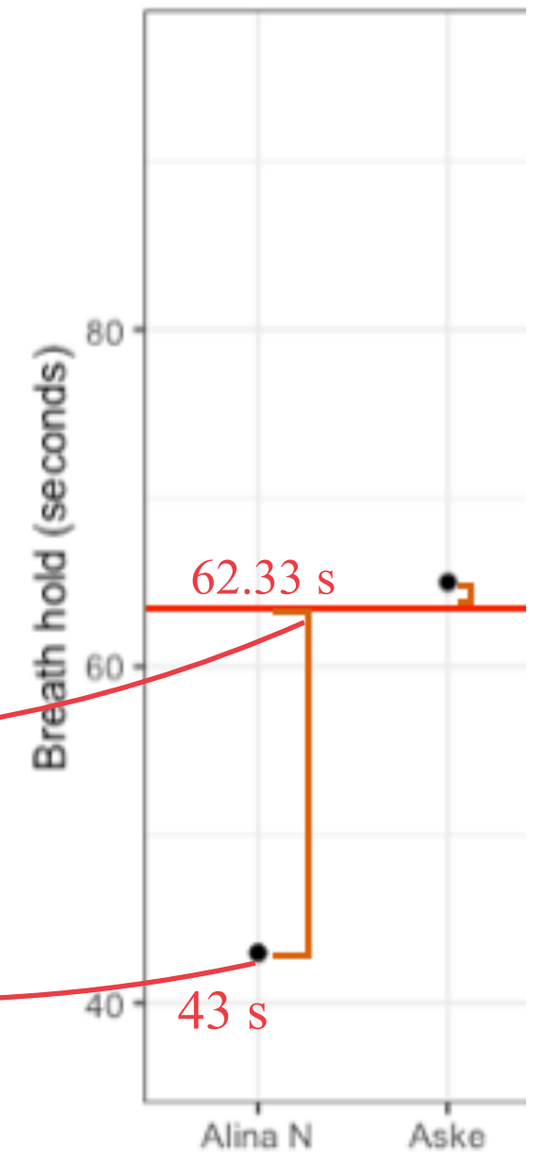
$$\mathbf{x_i = 62.33 + error_i}$$

How do we measure error?

- We call this error **deviance**
- Difference between mean and actual value for person *i*
- deviance = $x_i - \bar{x}$
- deviance = outcome_i - mean

How do we measure error?

- We call this error **deviance**
- Difference between mean and actual value for person *i*
- deviance = $x_i - \bar{x}$
- deviance = outcome$_i$ - mean



Sum of squared errors

SS

- How do I quantify the total error in a data set?
- Sum of deviance? ± values cancel each other out 😕
- → Sum of squared deviances (*SS*)

- $SS = \sum (x_i - \bar{x})(x_i - \bar{x}) = \sum (x_i - \bar{x})^2$

- *Note: large numbers gain more weight*

Variance

s^2

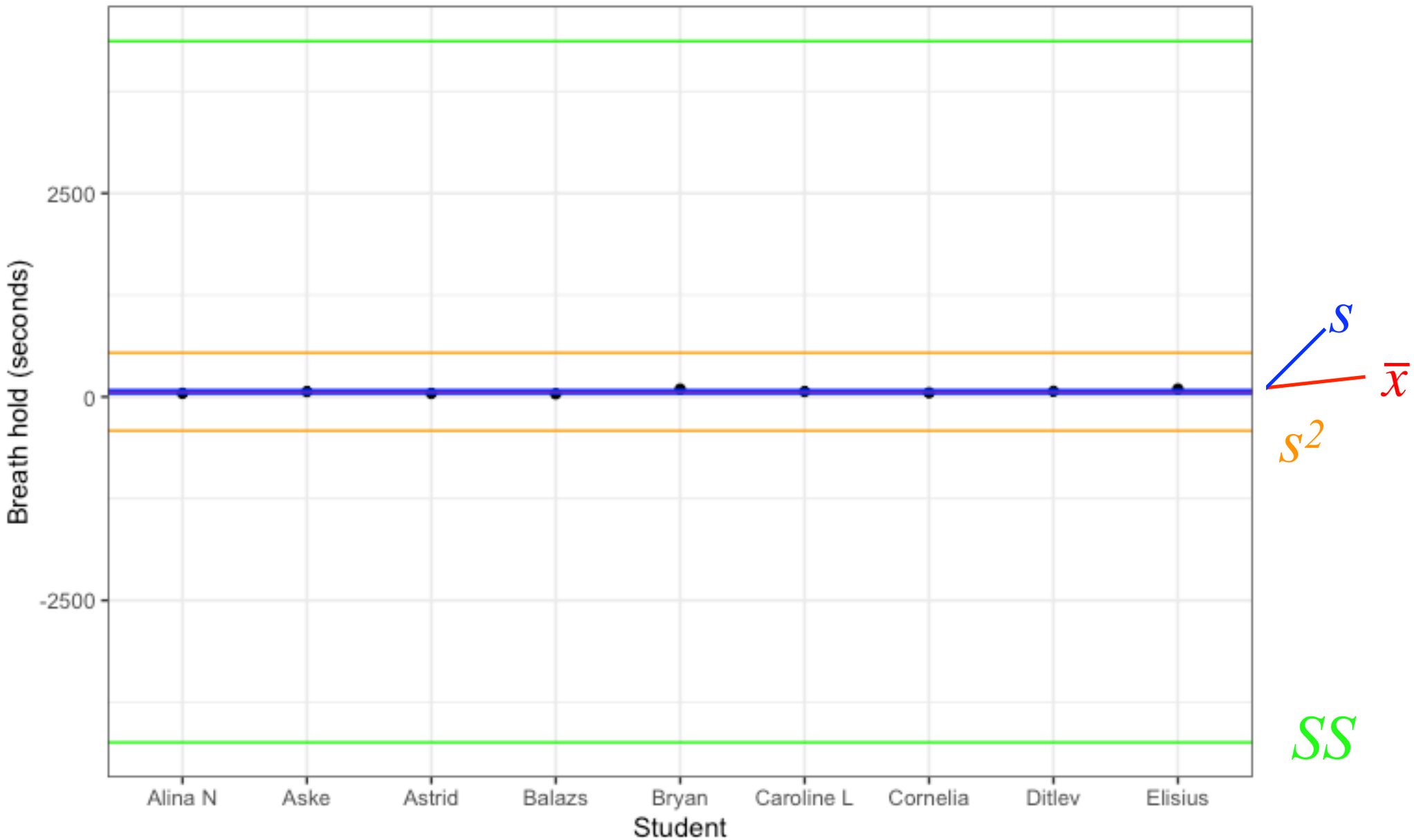
- SS is good, but too dependent on N
- → Variance (s^2) = mean deviance
- SS divided by N-1
- (the -1 part has to do with degrees of freedom — more on that later on)
- $$s^2 = \frac{SS}{N-1} = \frac{\sum (x_i - \bar{x})^2}{N-1}$$

Standard deviation

S

- s^2 is in units squared — not very meaningful
- What if we want to go back to the original scale?
- Take the square root of s^2
- $$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N - 1}}$$
- This is the **standard deviation**
- **A measure of whether the sample mean is a good model of the sample**

Recap: Measures of “error”



Excercise

- Help me find the following values for **breathhold**:
  - mean
  - sum of squared deviances (SS)
  - variance (s^2)
  - standard deviation

Descriptive vs. inferential statistics (1)

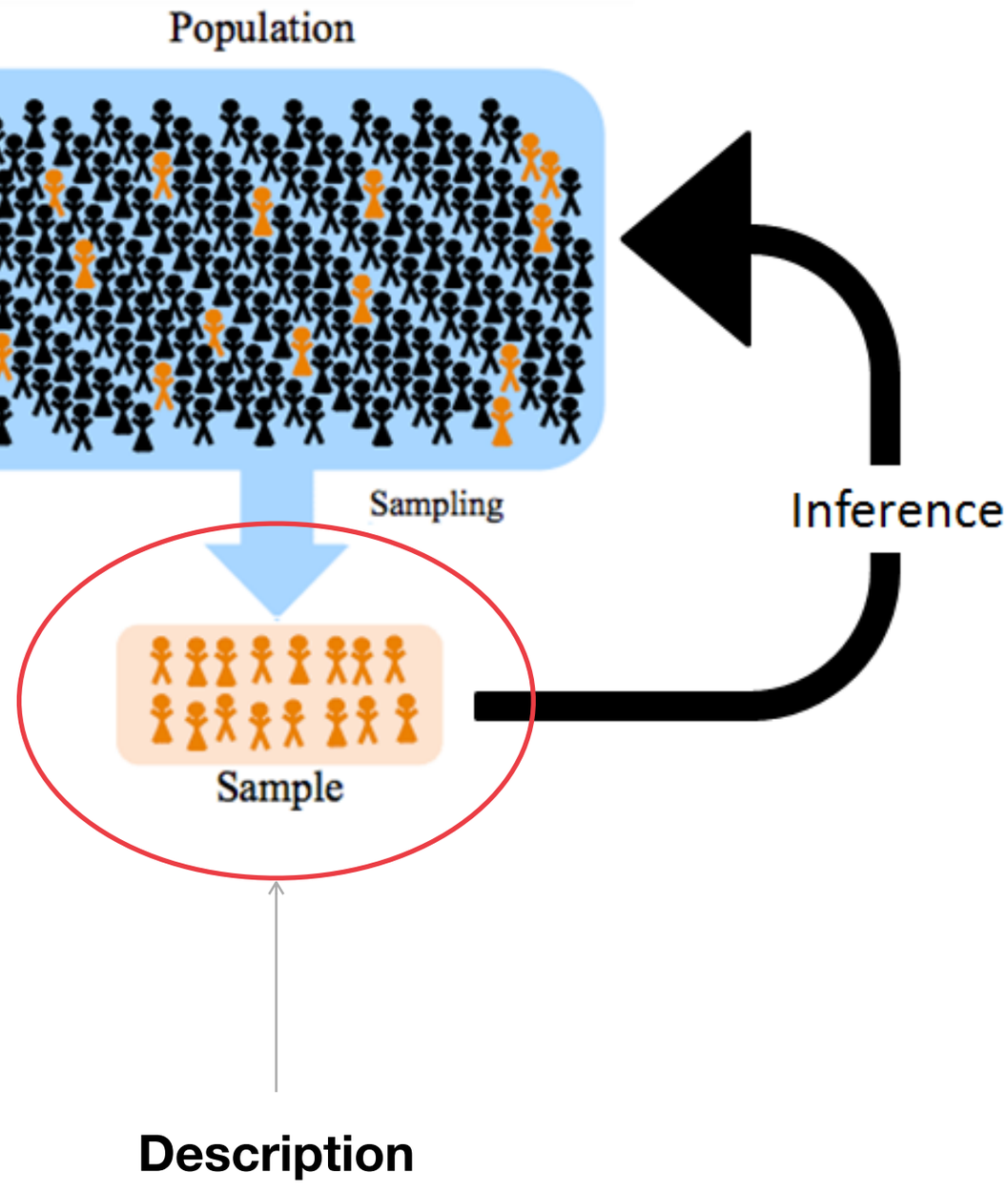
- $\bar{x}$ and s only describe the sample

- However, we want to know stuff about the population (μ and σ)

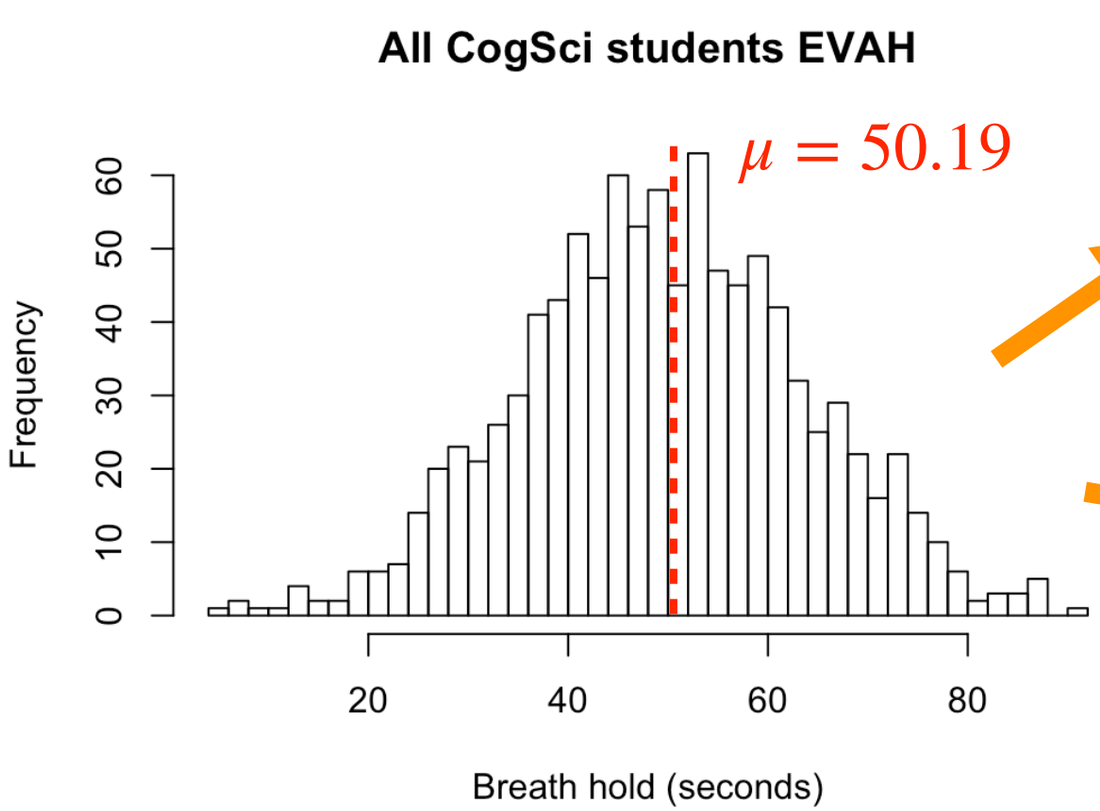
- Inferential statistics:

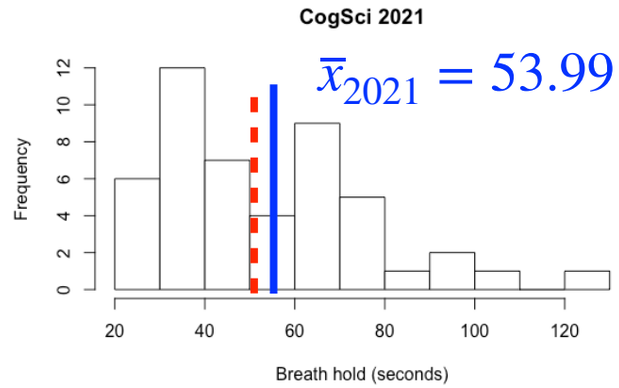
  - **Can we use $\bar{x}$ and s to learn something about μ and σ?**

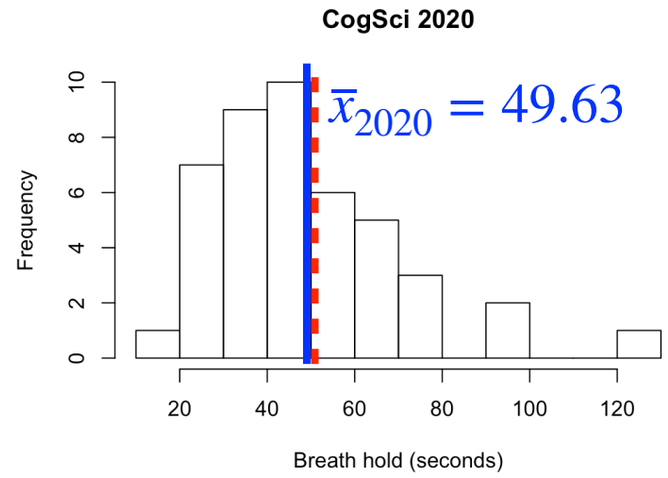
  - **How likely is it that $\bar{x}$ = μ?**

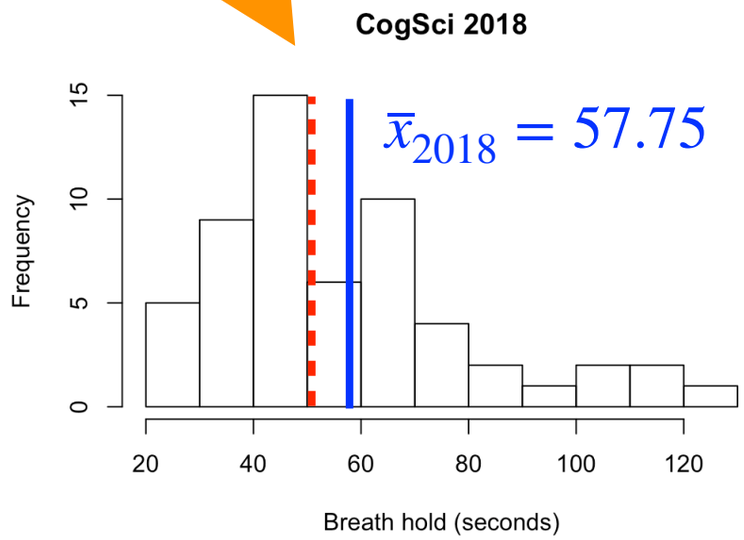


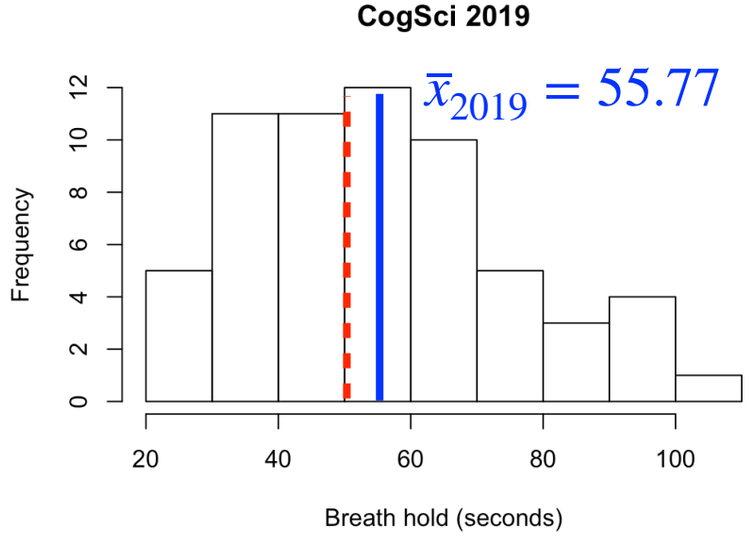
Descriptive vs. inferential statistics (2)





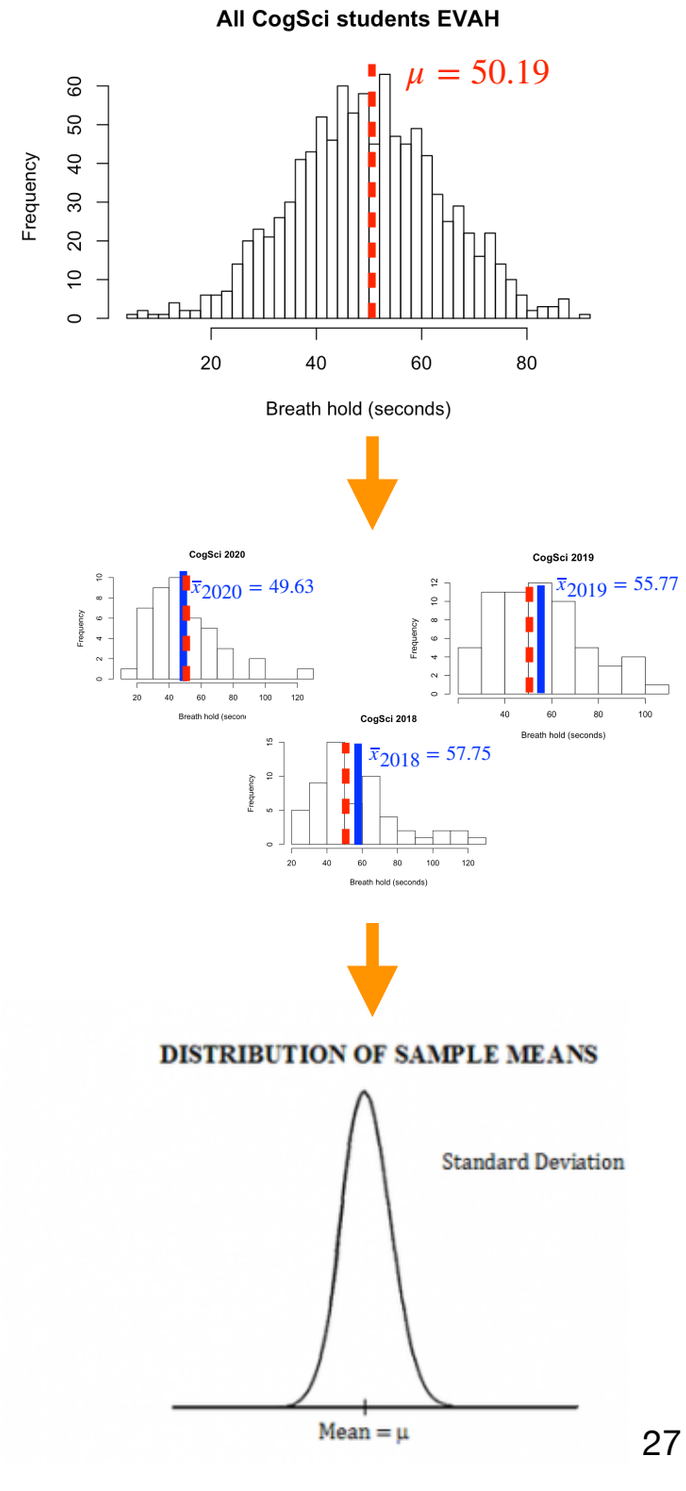






Descriptive vs. inferential statistics (3)

- The true mean μ is unobserved
- The mean of each sample (2021, 2020, 2019, etc.) will be different
- The sampling distribution of the sample means ($\bar{x}$) will approach normality with its own μ and σ^2
- Variance of the sampling distribution of the sample means:
  - $$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{N}$$
- Standard deviation of the sampling distribution of the sample means (standard error of the mean):
  - $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}}$$



Standard Error of the Mean (1)

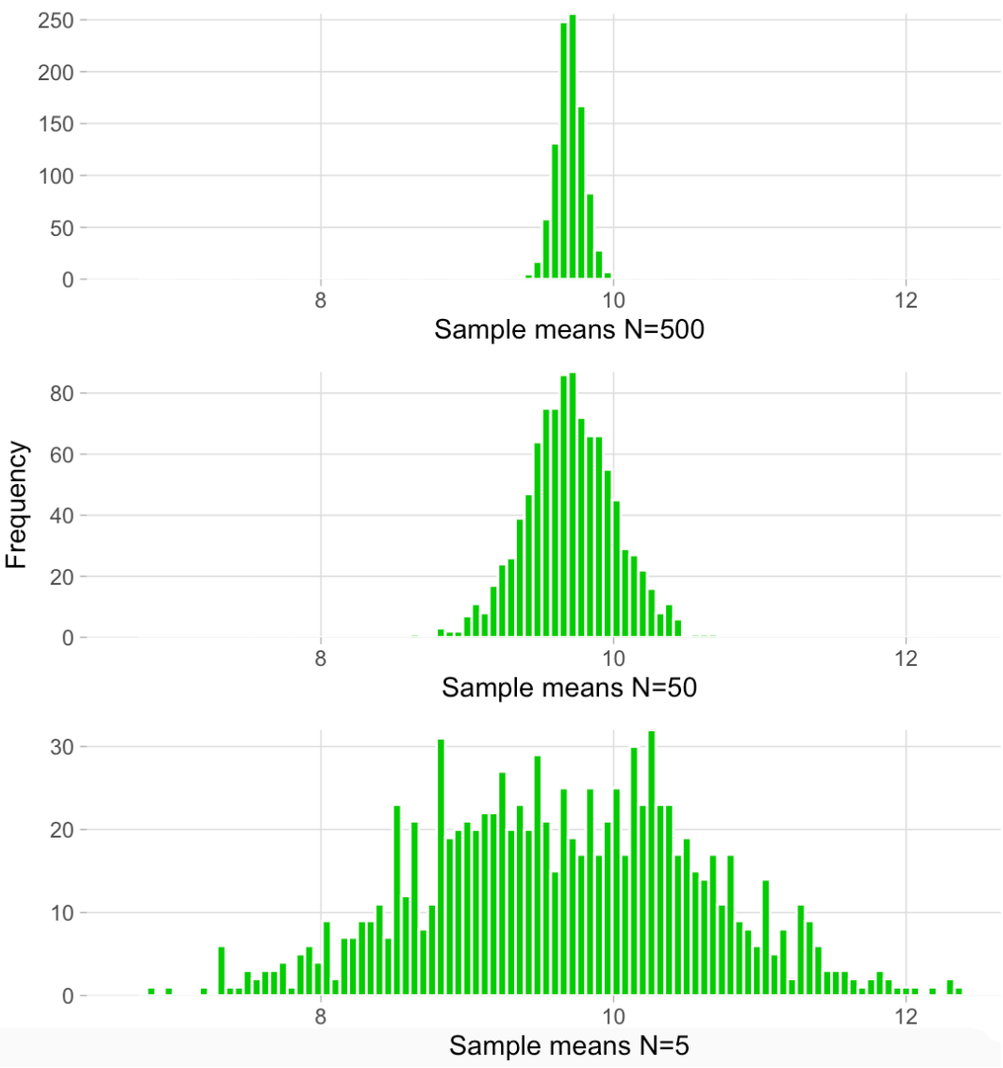
$\sigma_{\bar{x}}$

- How precisely does the sample statistic $\bar{x}$ approximate the population statistic μ?

- $\sigma_{\bar{x}}$ is a measure of the **uncertainty of a single sample value as an estimate of the population value**

- Becomes smaller as the sample size increases

- Lower $\sigma_{\bar{x}}$ means better estimation of μ



Standard Error of the Mean (2)

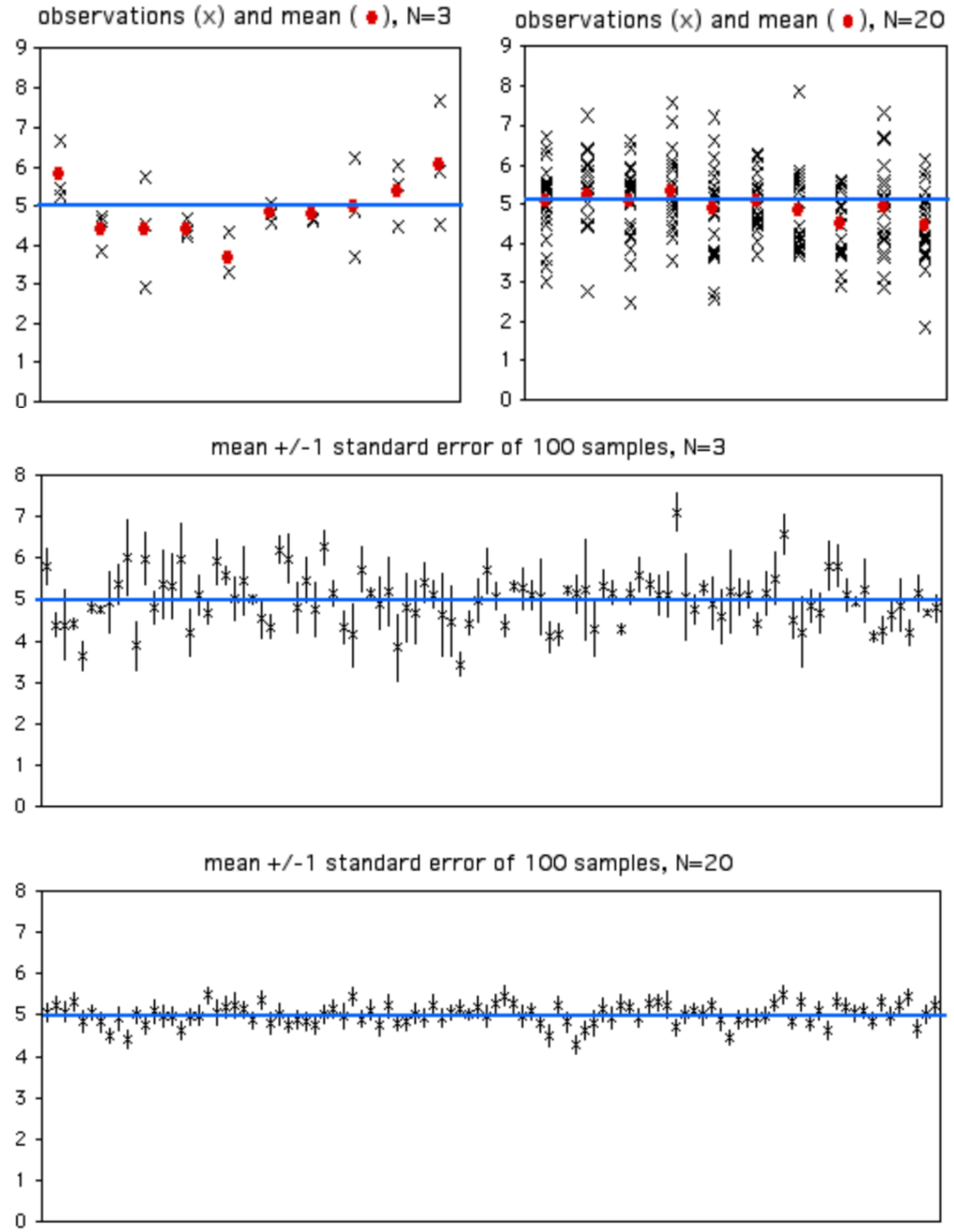
- Why should we care?

- Normally, we only have one estimate of the population mean from our single sample

- $\sigma_{\bar{x}} \approx \frac{s}{\sqrt{n}}$ if:

  - The sample is roughly normally distributed

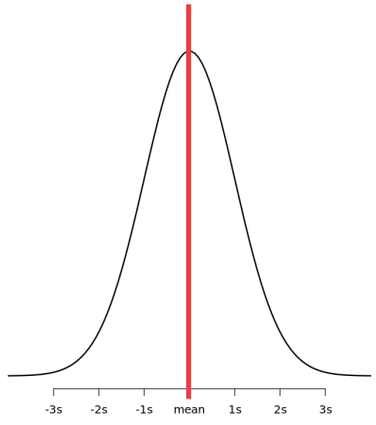
  - n is large (at least ~20/30 measurements)



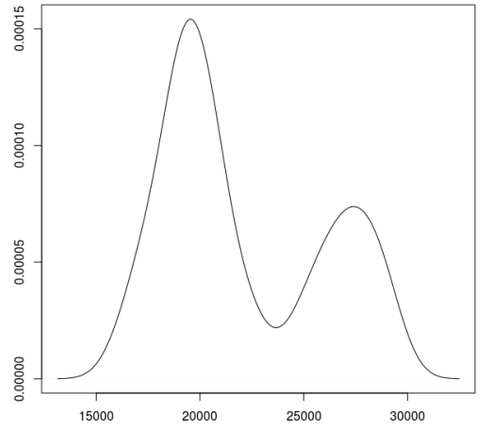
http://www.biostathandbook.com/standarderror.html

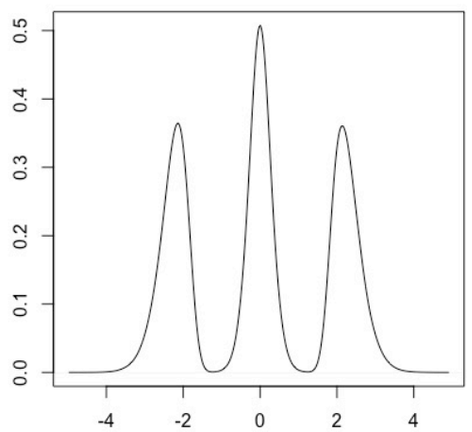
Is the model a good fit for my data?

- Is the model descriptive of the data?
- **Standard Deviation (*s*) as a reflection of whether the mean is a good model of the observed data**



- **Standard Error of the Mean ($\sigma_{\bar{x}}$) as a reflection of whether my sample mean is a good model of the true mean of the population**
- You will learn more ways to quantify "goodness of fit" throughout the program





Thursday

- We will continue our data mining exercise…
- … and look at a few new tools