
REVIEW

Overview

Review of ggplot

Review of material covered in the class so far

If there is time/interest

- Bonus features of ggplot: special geoms, animation, interactive graphics

Announcements

Midterm exam is on Thursday

- Bring a pen and a pencil
- One page (2 sides) with code and equations only!
  - You will turn in this page of notes with your exam (put your name on it)
  - You can write down conditions for hypothesis tests. Have SE formulas, etc.

Office hours this week

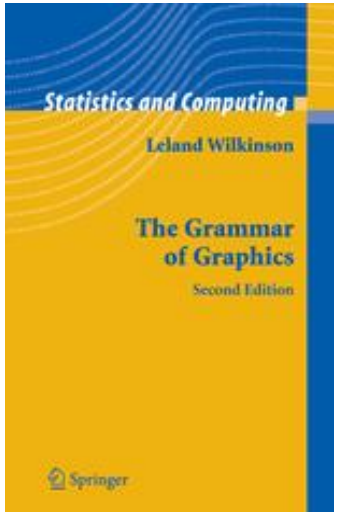
- Stephan and Nathan are doing another review session tonight
- No TA office hours since no homework

Review of the grammar of graphics and ggplot

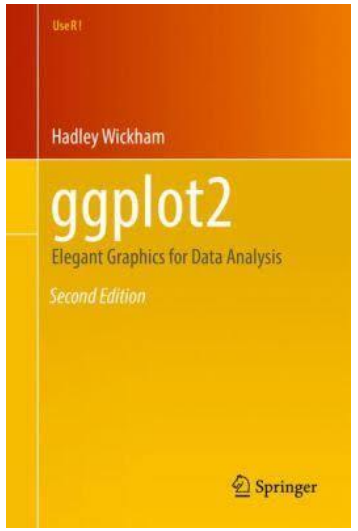
The grammar of graphics

Leland Wilkinson noticed similarities between many graphs and tried to generate a ‘grammar’ that could be used to express a graph

- i.e., a list elements that can be combined together to create a graph



Hadley Wickham implemented these ideas in R in the ggplot2 package



Graphs are composed of…

A Frame: Coordinate system on which data is placed

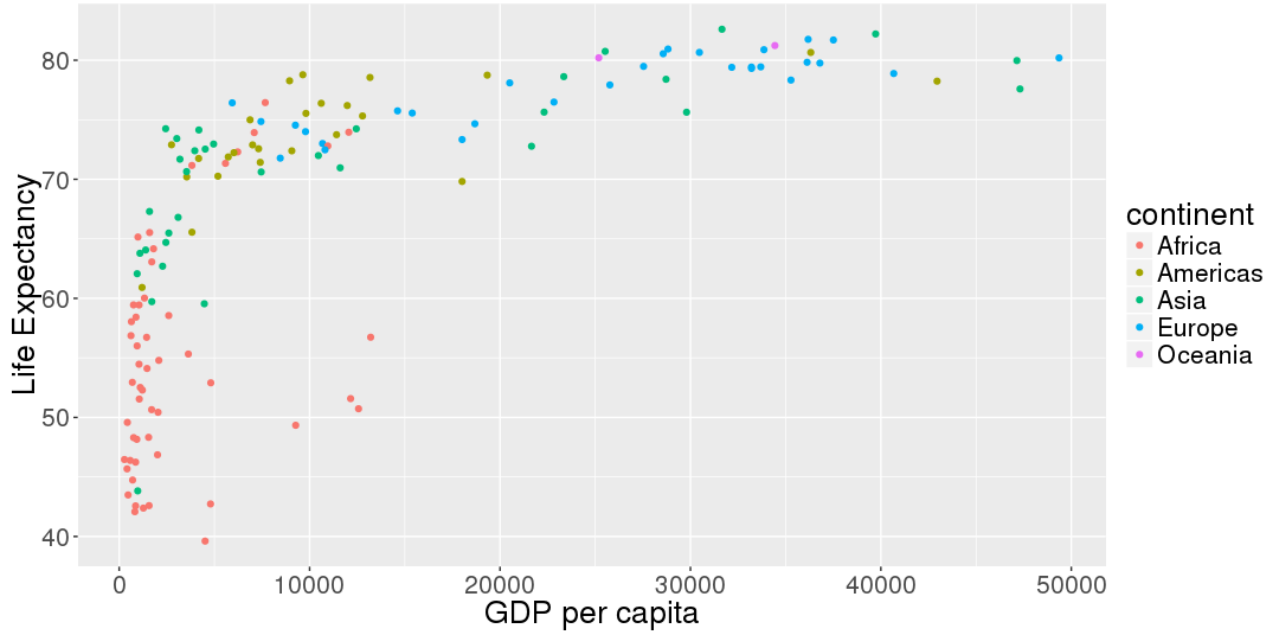
- ggplot() +

Glyphs: basic graphic unit representing cases or statistics

- Data is **mapped** onto these aesthetics such as: shape, color, size, etc. and/or aesthetics can be set to a fixed value
  - geom_point(aes(x = gdpPercap, y = lifeExp, color = continent)) geom_point(aes(x = gdpPercap, y = lifeExp), color = "red")

Scales and guides: shows how to interpret axes and other properties of the glyphs

- scale_x_continuous(trans = "log10") scale_color_brewer(type = "qua", palette = 2)



Plots can also contain…

Facets: allows for multiple side-by-side graphs based on a categorical variable

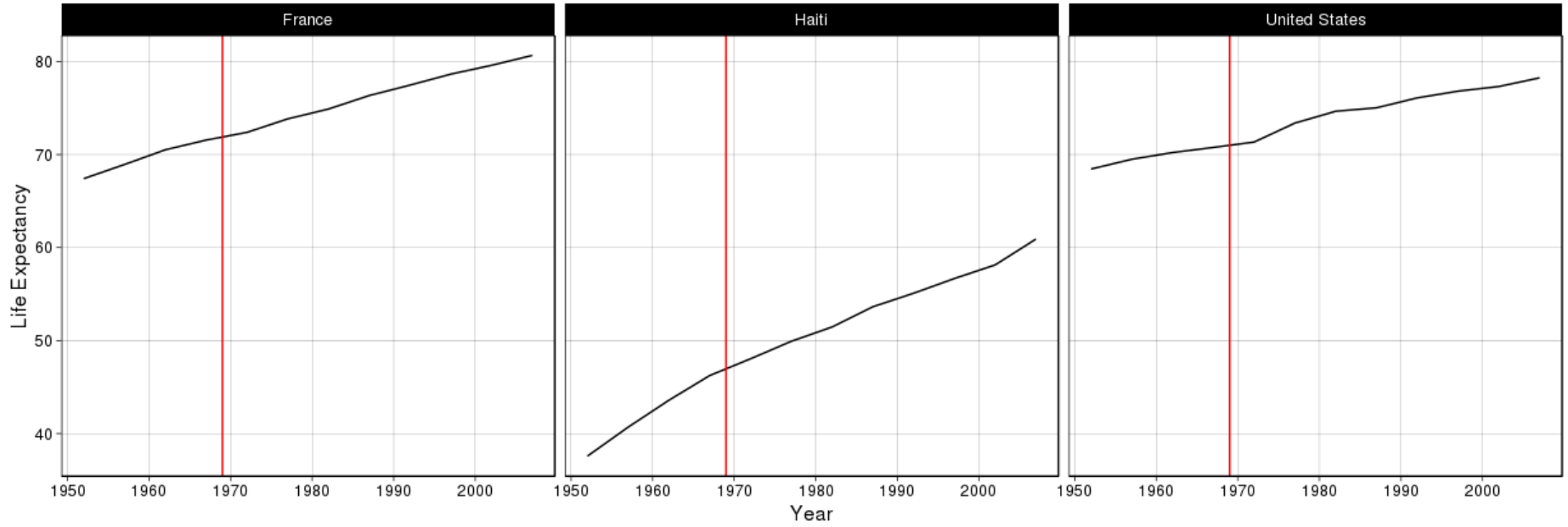
- facet_wrap(~country)

Layers: allows for more than one types of data to be mapped onto the same figure

- geom_vline(xintercept = 1969, col = "red")

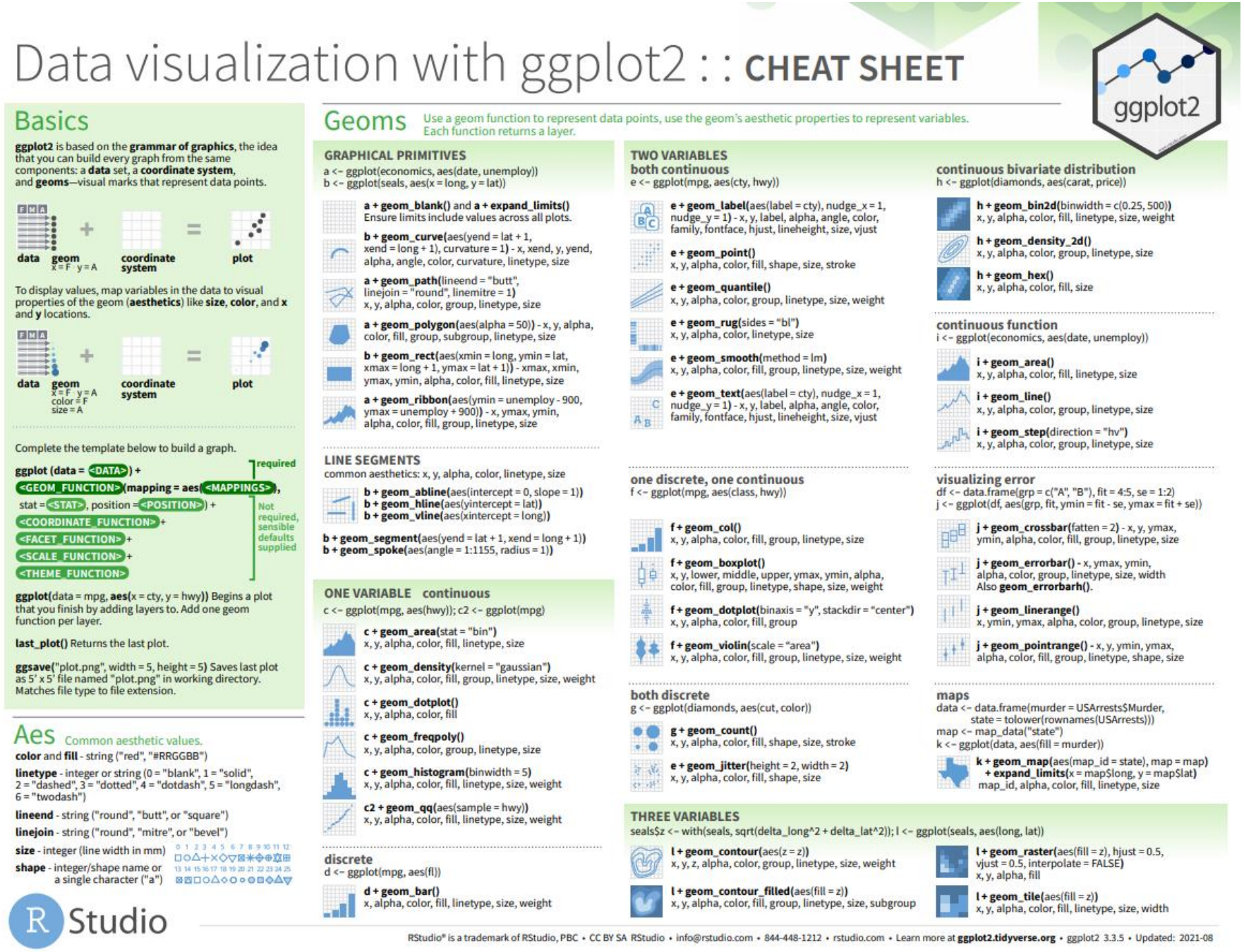
Theme: contains finer points of display (e.g., font size, background color, etc.)

- theme_wsj()



Questions?

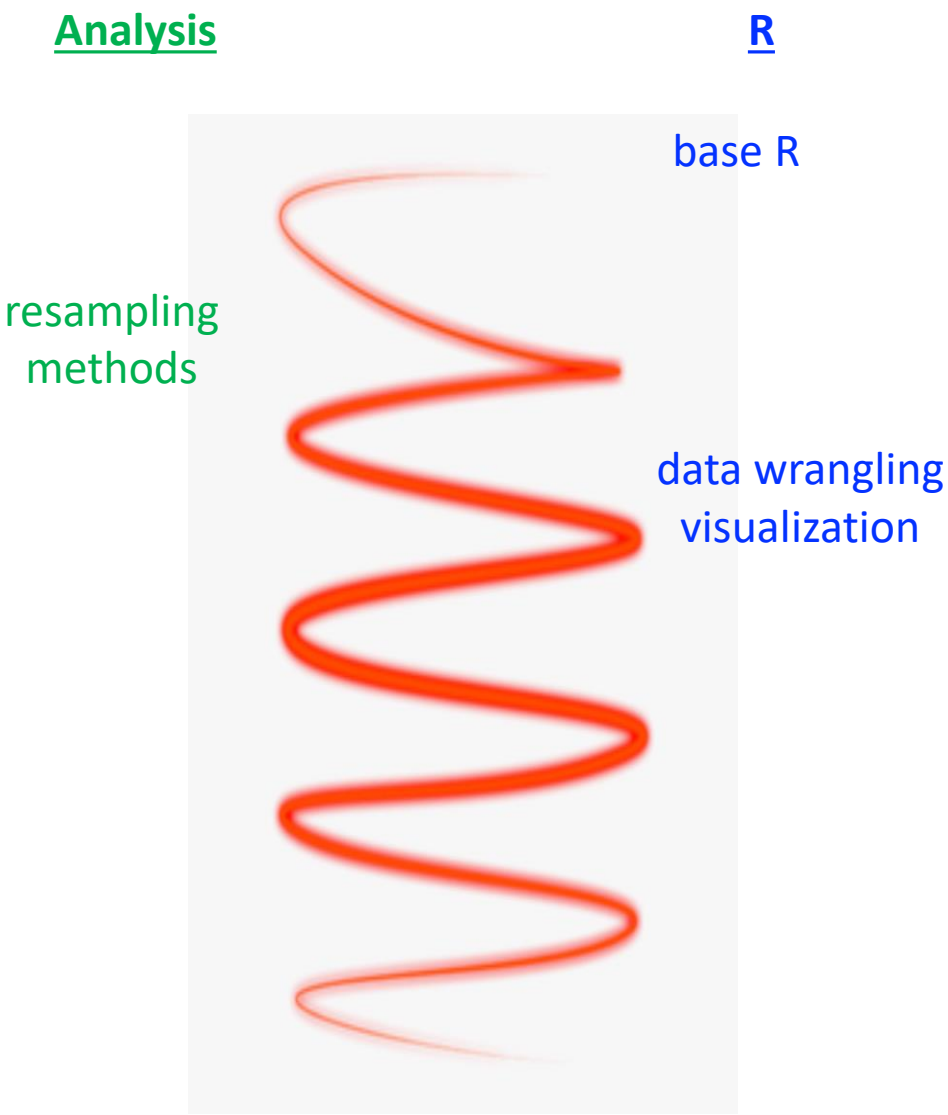
[ggplot2 cheat sheet]




TIME FOR REVIEW

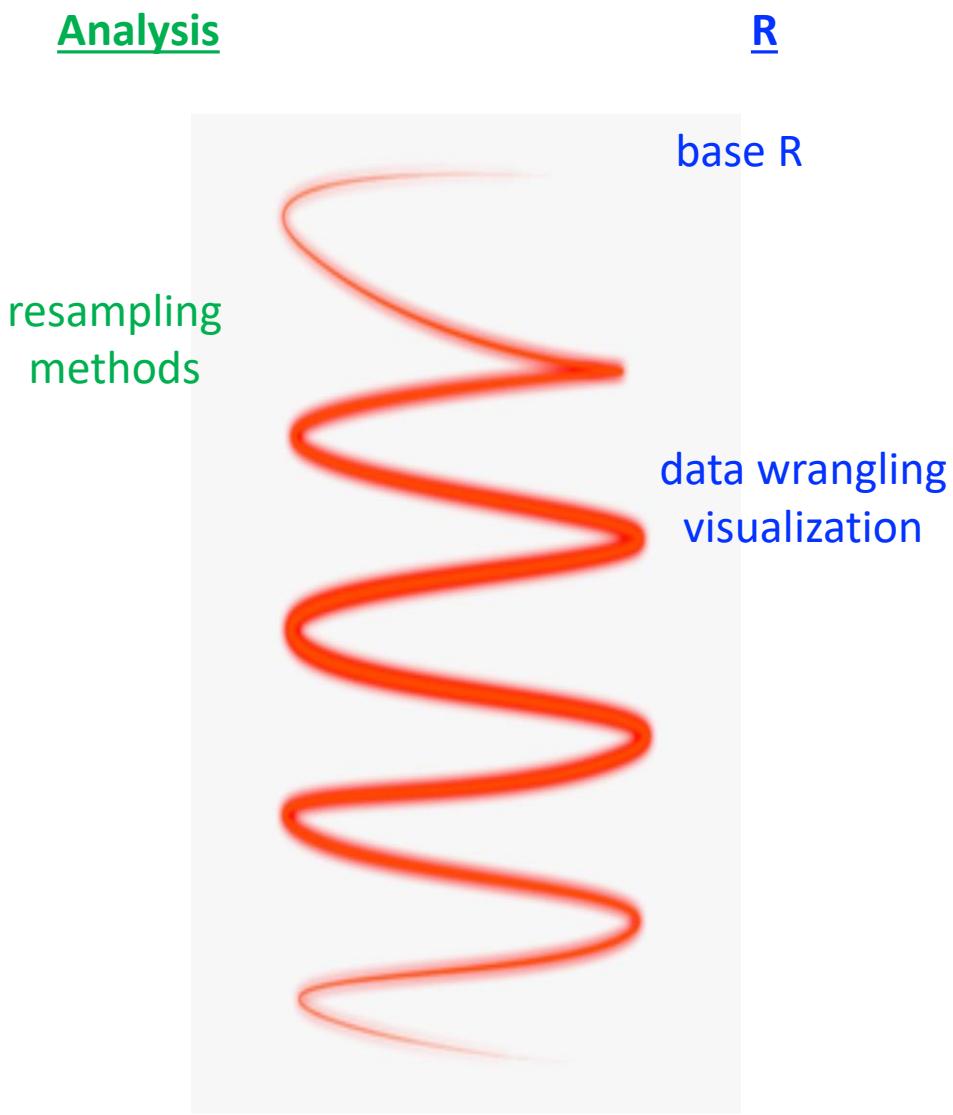
What we have covered so far…

1	Sep 1	Course overview, introduction to R, descriptive statistics
2	Sep 6-8	Review of central statistical concepts and exploratory analysis using R
3	Sep 13-15	Confidence Intervals and the bootstrap
4	Sep 20-22	Review of hypothesis tests and permutation tests in R
5	Sep 27-29	Parametric, non-parametric and theories of hypothesis testing
6	Oct 4-6	Data manipulation and visualization
7	Oct 11-13	Review and midterm exam
8	Oct 18-22	Odds and ends, October break

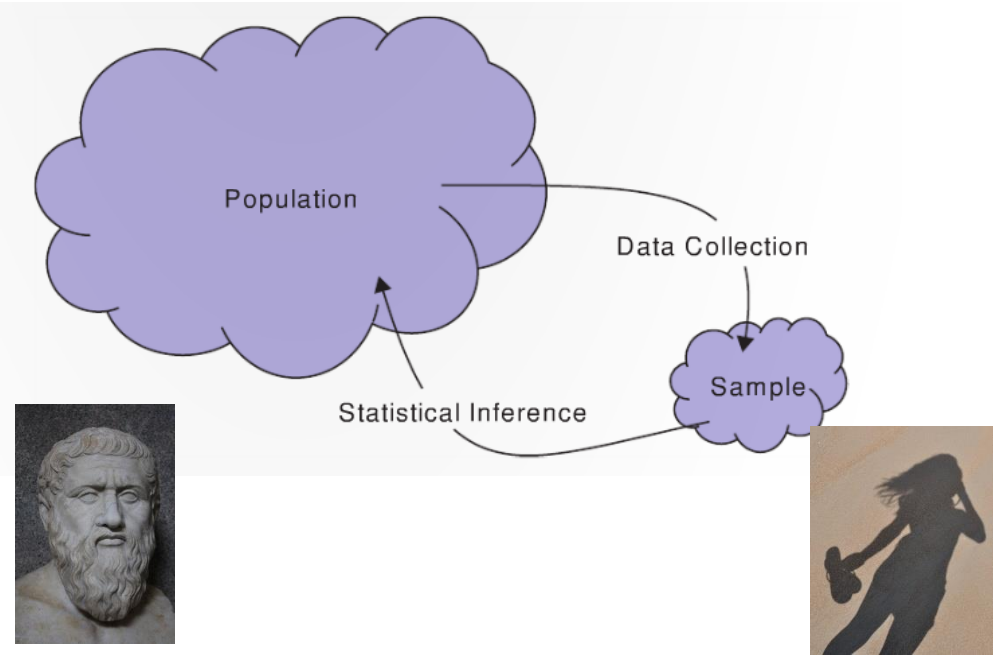


What we have covered so far…

1	Sep 1	Course overview, introduction to R, descriptive statistics
2	Sep 6-8	Review of central statistical concepts and exploratory analysis using R



Parameters and statistics commonly used symbols



$$\bar{x} = \frac{\sum_i^n x_i}{n}$$

	Population parameter (Plato)	**Sample statistic (shadow)**
Mean	μ	x̄
Standard deviation	σ	s
Proportion	π	p̂
Correlation	ρ	r
Regression slope	β	b

Base R

Basics of R

```
> my_vec <- c(5, 28, 19)
> inds_less_than_10 <- 10
```

How to plot data in base R

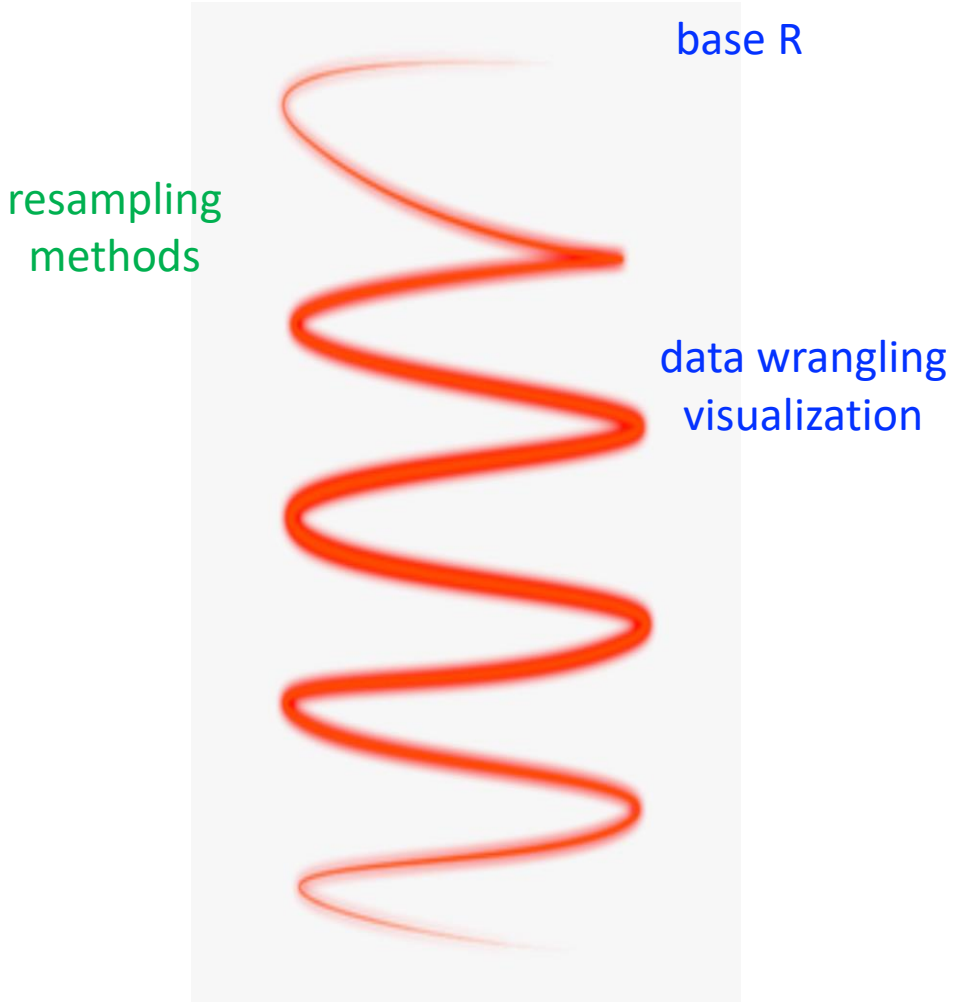
```
> drinks_table <- table(profiles$drinks)
> barplot(drinks_table)
> pie(drinks_table)
> hist(profiles$height)
```

For loops

```
my_results <- NULL
for (i in 1:100) {
        my_results[i] <- i^2
}
```

What we have covered so far…

3	Sep 13-15	Confidence Intervals and the bootstrap
4	Sep 20-22	Review of hypothesis tests and permutation tests in R
5	Sep 27-29	Parametric, non-parametric and theories of hypothesis testing

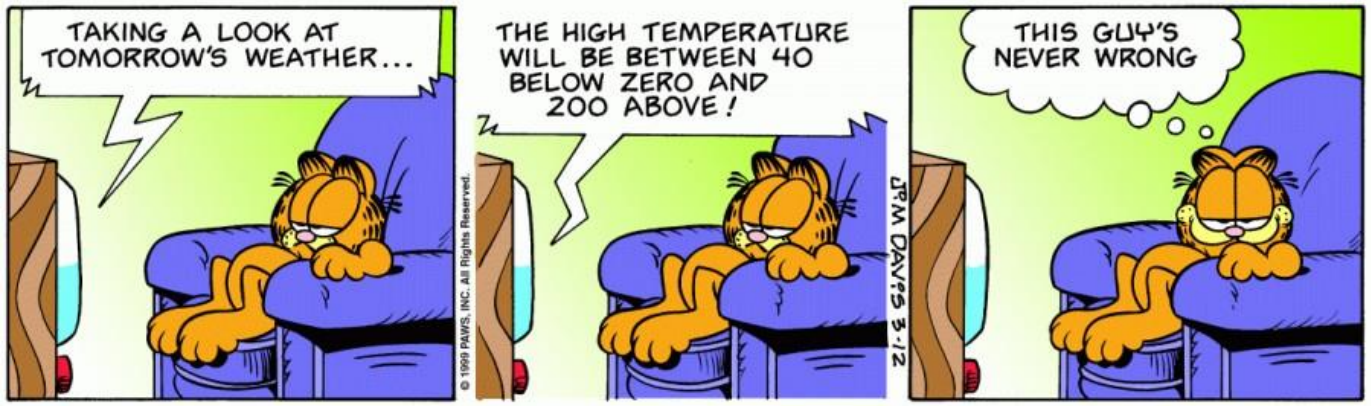


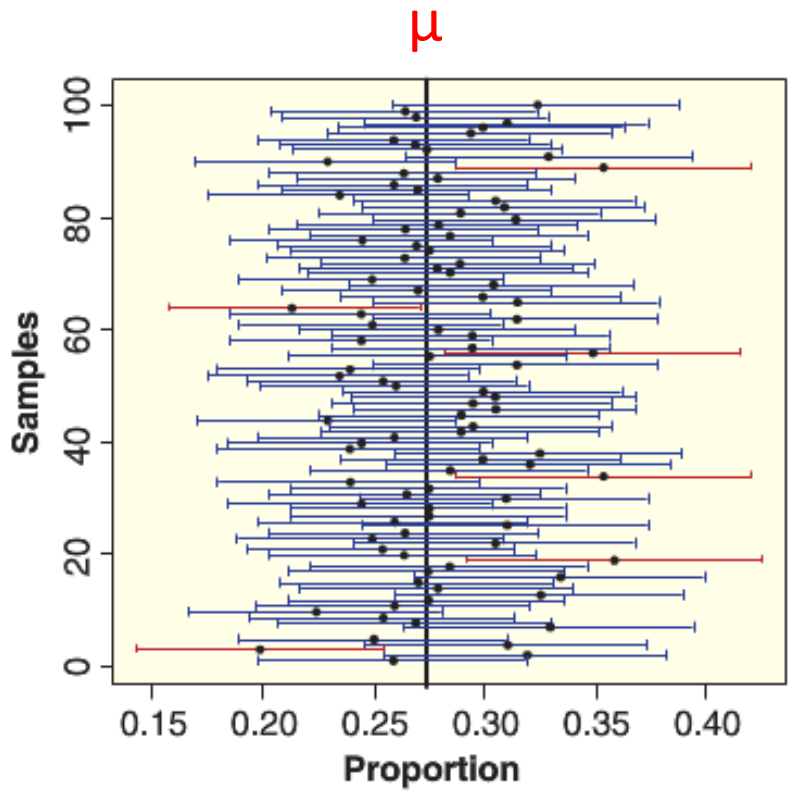
Probability and confidence intervals

Probability functions; e.g., rnorm, pnorm, dnorm, qnorm

Confidence intervals:

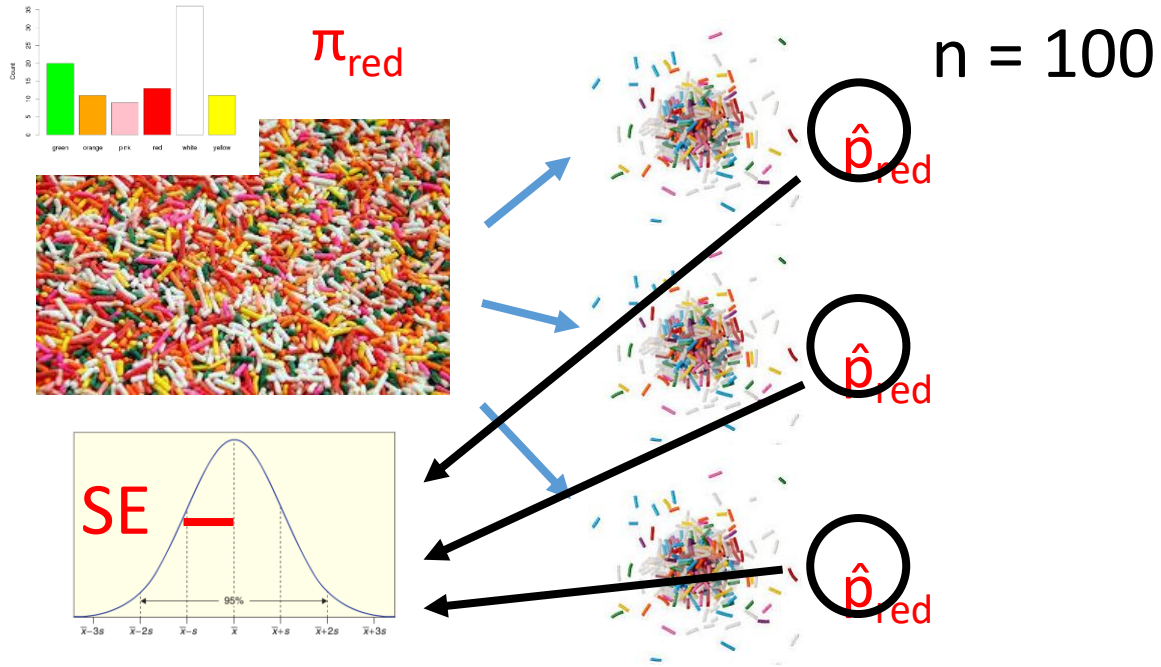
$$CI_{95} = stat \pm 2 \cdot SE$$





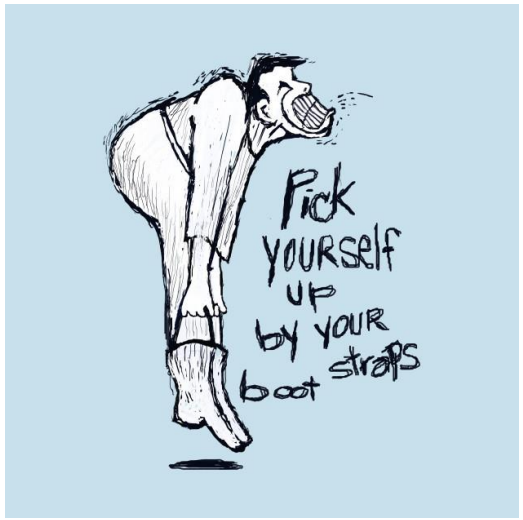
Sampling and bootstrap distributions

Sampling distribution



Sampling distribution!

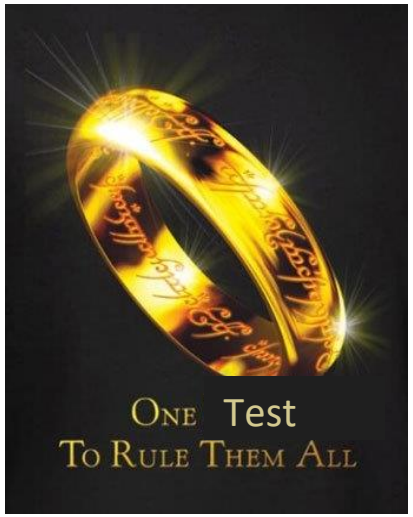
Bootstrap distribution

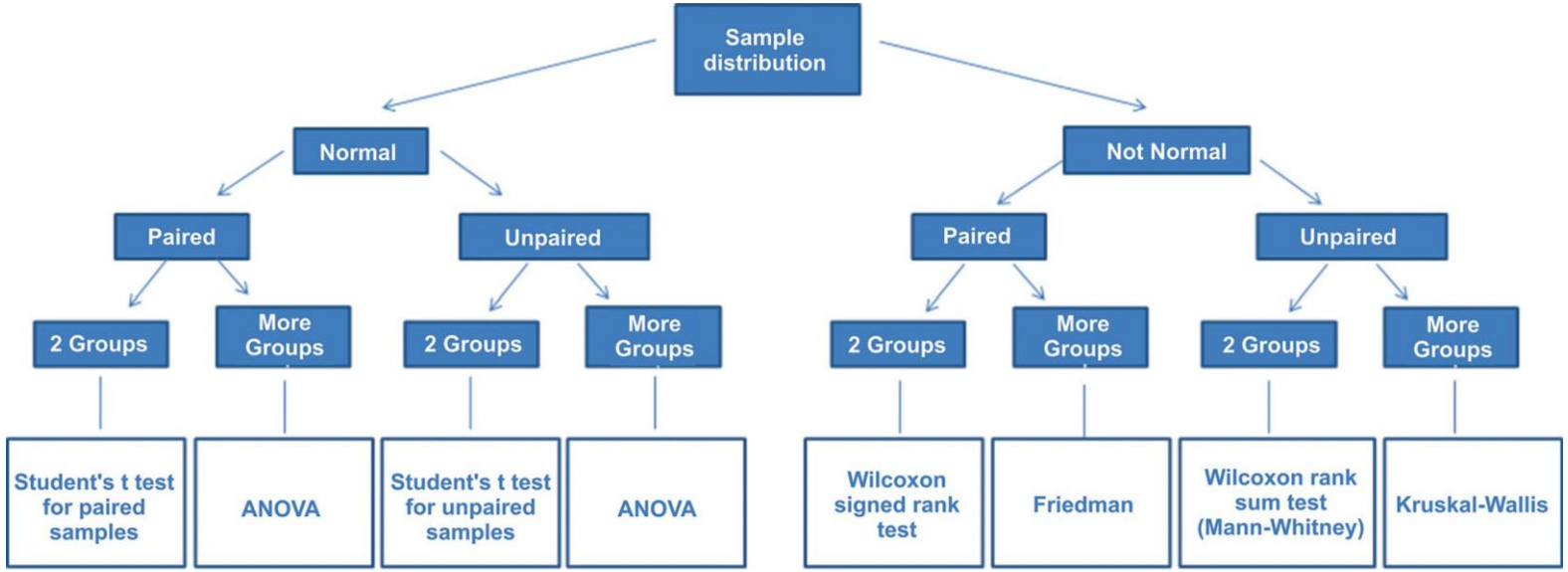


Sample with replacement from our original sample to mimic a sampling distribution

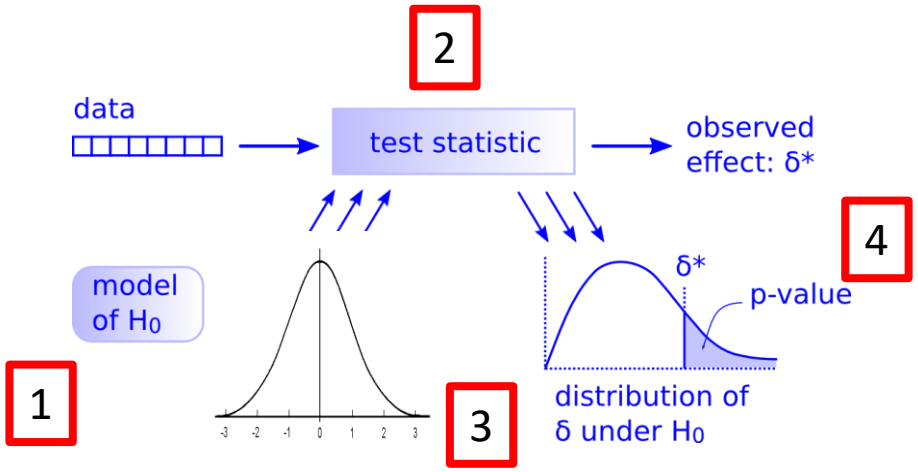
$$CI_{95} = \text{stat} \pm 2 \cdot SE^*$$

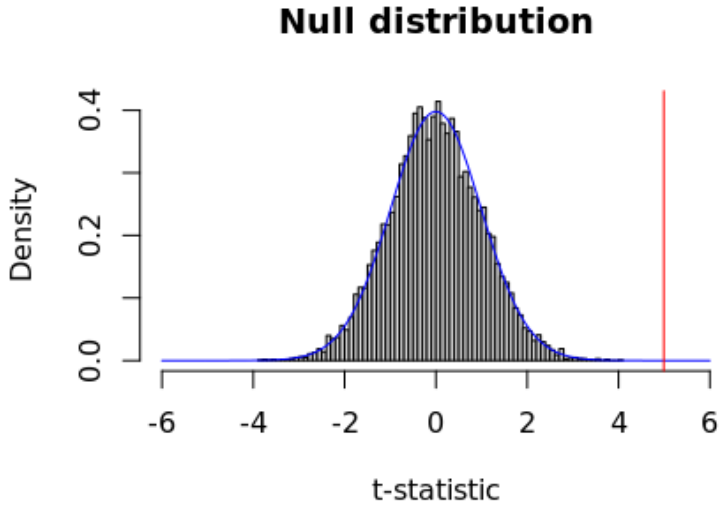
Hypothesis tests





Just need to follow 5 steps!





Randomization/permutation tests

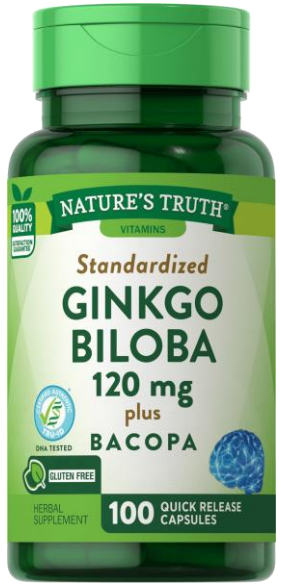
Create a null distribution through computational simulations/shuffling

- rbinom(), sample(), etc.

H_0: $\pi = 0.5$
H_A: $\pi > 0.5$

H_0: $\mu_T - \mu_C = 0$
H_A: $\mu_T - \mu_C > 0$

H_0: $\mu_i = \mu_j \ldots = \ldots \mu_k$
H_A: $\mu_i \neq \mu_j$ for some i, j



Data	1 Sample	2 Samples	> 2 Samples
Categorical data	H_0: $\pi = p_0$ H_A: $\pi \neq p_0$ Flip "coins" rflip_count()	H_0: $\pi_1 = \pi_2$ H_A: $\pi_1 \neq \pi_2$ Flip "coins" rflip_count()	H_0: $\pi_1 = p_1$, $\pi_2 = p_2$, … , $\pi_k = p_k$ H_A: At least one p_i is different than specified Flip coins rflip_count()
Quantitative data	H_0: $\mu = v_0$ H_A: $\mu \neq v_0$ resample sample(… , replace = TRUE)	H_0: $\mu_1 = \mu_2$ H_A: $\mu_1 \neq \mu_2$ Shuffle data shuffle()	H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ H_A: At least one μ_i is different Shuffle data shuffle()

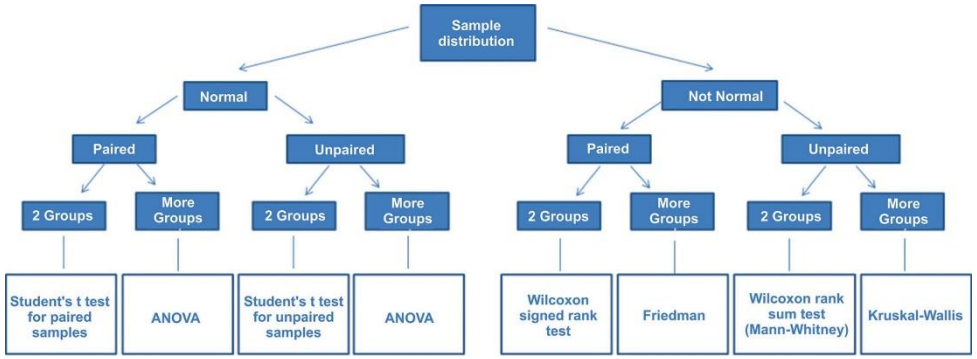
Parametric tests

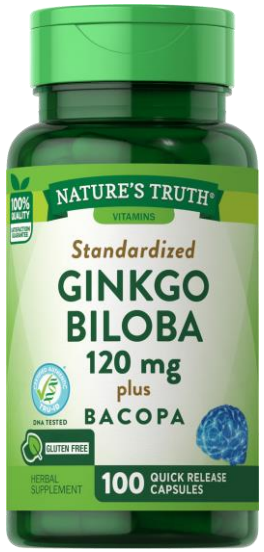
Use mathematical density functions for the null distribution

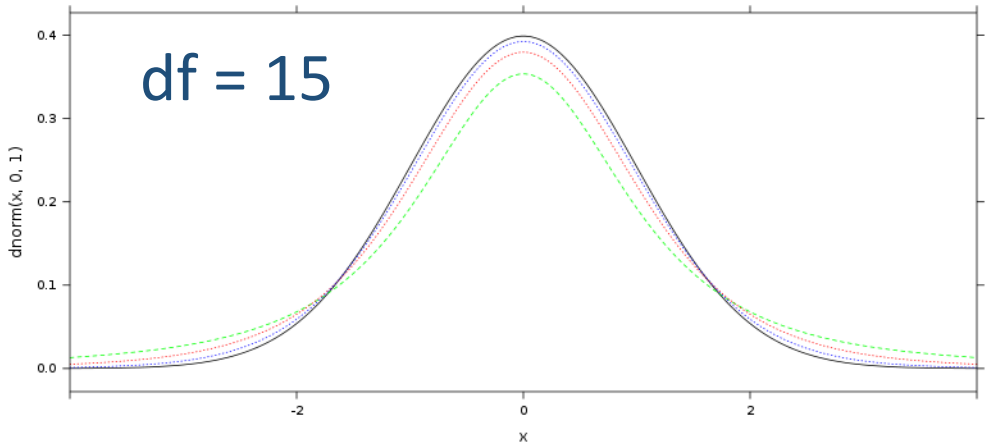
H_0: $\mu_T - \mu_C = 0$

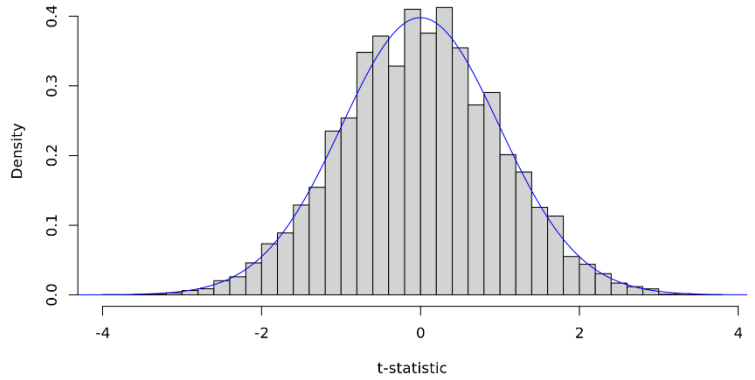
H_A: $\mu_T - \mu_C > 0$

$$t = \frac{\bar{x}_t - \bar{x}_c}{\sqrt{\frac{s_t^2}{n_t} + \frac{s_c^2}{n_c}}}$$









Data	1 Sample	2 Samples	> 2 Samples
Categorical data	H_0: $\pi = p_0$ H_A: $\pi \neq p_0$ z-test $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$	H_0: $\pi_1 = \pi_2$ H_A: $\pi_1 \neq \pi_2$ z-test or a chi-square $z = \frac{\hat{p_1} - \hat{p_2}}{\sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}}$	H_0: $\pi_1 = p_1,\ \pi_2 = p_2,\ \ldots\ , \pi_k = p_k$ H_A: At least one p_i is different than specified chi-square test $\chi^2 = \sum_{i=1}^{k} \frac{(Observed_i - Expected_i)^2}{Expected_i}$
Quantitative data	H_0: $\mu = v_0$ H_A: $\mu \neq v_0$ One sample t-test $t = \frac{\bar{x} - v_0}{s/\sqrt{n}}$	H_0: $\mu_1 = \mu_2$ H_A: $\mu_1 \neq \mu_2$ Two sample t-test $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ H_A: At least one μ_i is different Analysis of Variance $F = \frac{\frac{1}{K-1}\sum_{i=1}^{K} n_i(\bar{x}_i - \bar{x}_{tot})^2}{\frac{1}{N-K}\sum_{i=1}^{K}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2}$ $df_1 = k$, $df_2 = n - k$

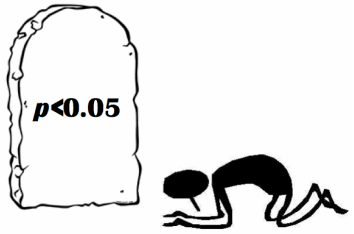
Data	1 Sample	2 Samples
Categorical Data	$SE = \sqrt{\frac{\pi(1-\pi)}{n}}$ $\hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$SE = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$ $\hat{p_1} - \hat{p_2} \pm z^* \sqrt{\frac{\hat{p_1}(1-\hat{p_1})}{n_1} + \frac{\hat{p_2}(1-\hat{p_2})}{n_2}}$
Quantitative Data	$SE = \frac{s}{\sqrt{n}}$ $\overline{x} \pm t^* \frac{s}{\sqrt{n}}$	$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ $(\overline{x_1} - \overline{x_2}) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

Theories of hypothesis testing

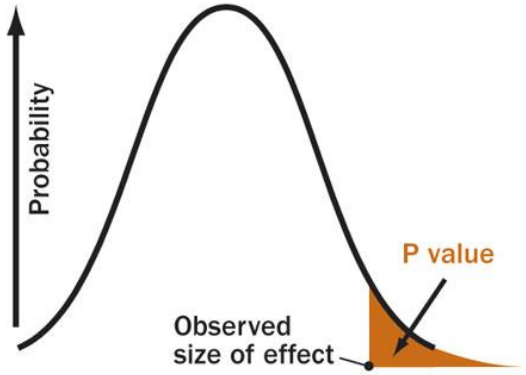
Fisher (1890-1962)

Neyman (1894-1981)

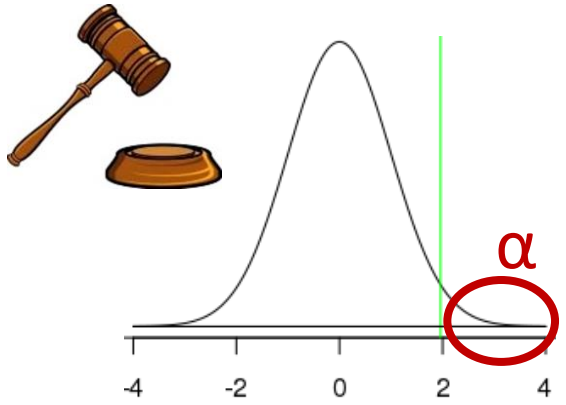
Pearson (1895-1980)

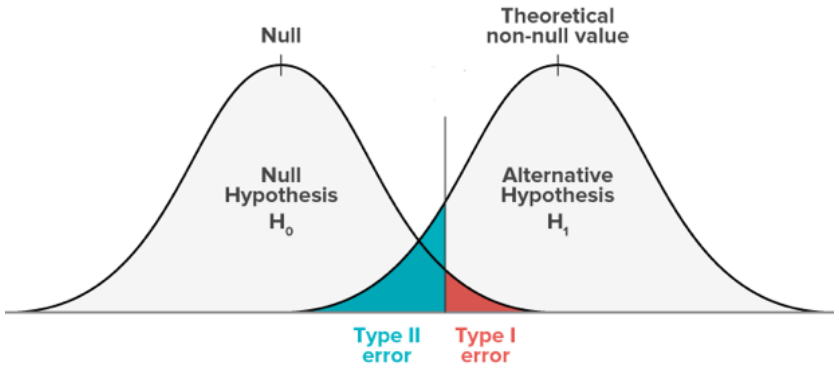


p-value a strength of evidence

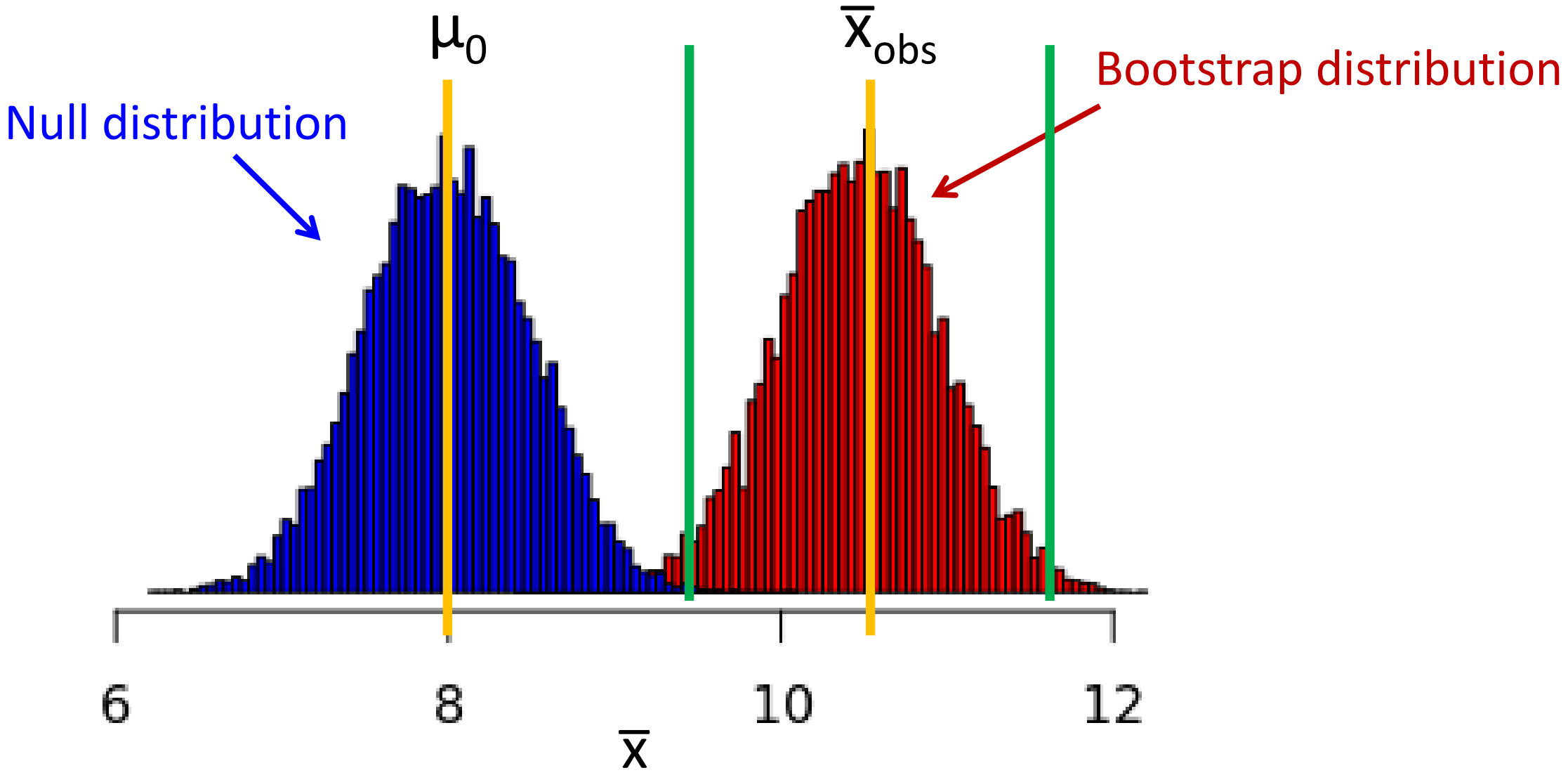


Use p-value to make a decision





Relationship between null and bootstrap distributions



Data manipulation with dplyr

dplyr is a package that has a set of verbs for transformations data

- All these function **take a data frame** and other arguments and **return a data frame**

1. filter()
2. select()
3. mutate()
4. arrange()
5. summarize()
6. group_by()

age	body_type	diet
22	a little extra	strictly anything
35	average	mostly other
38	thin	anything
23	thin	vegetarian
29	athletic	*NA*
29	average	mostly anything

```
film_results <- movies |>
    filter(title_type == "Feature Film") |>
    select(critics_score, audience_score, genre) |>
     mutate(audience_prefers =
              audience_score - critics_score) |>
     group_by(genre) |>
      summarize(mean_audience_prefers =
                  mean(audience_prefers)) |>
      arrange(desc(mean_audience_prefers))


head(film_results )
```

Grammar of graphics with ggplot

A Frame: Coordinate system on which data is placed

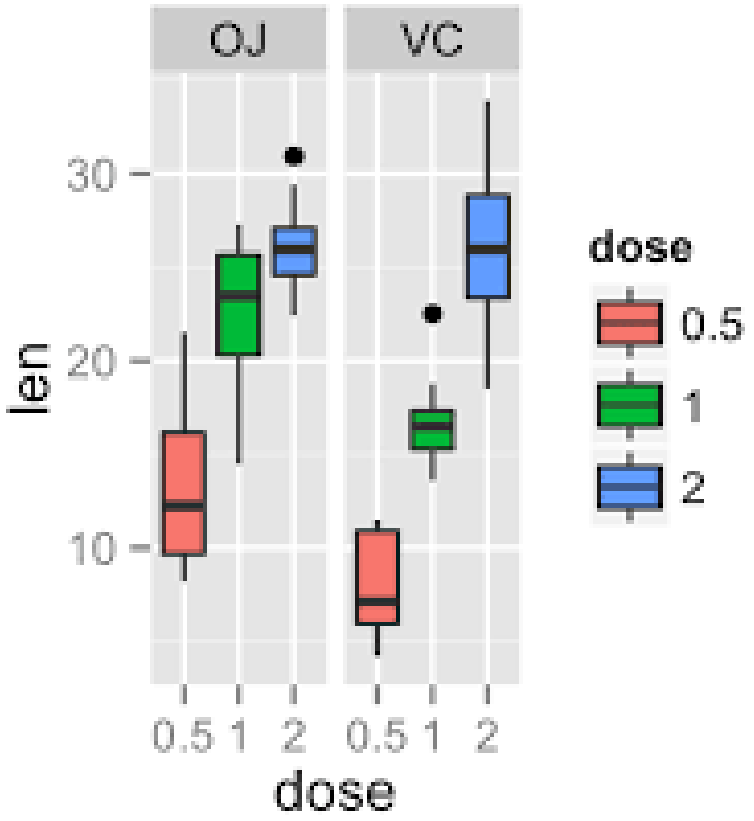
Glyphs: basic graphic unit representing cases or statistics

Scales and guides: shows how to interpret axes and other properties of the glyphs

Facets: allows for multiple side-by-side graphs based on a categorical variable

Layers: allows for more than one types of data to be mapped onto the same figure

Theme: contains finer points of display (e.g., font size, background color, etc.)



Questions