Review of Introductory Statistics

Zack Treisman

Spring 2022

WESTERN
COLORADO UNIVERSITY
MATHEMATICS & COMPUTER
SCIENCE DEPARTMENT

Philosophy

This class builds on the knowledge and skills you acquired in your previous study of Statistics. We are going to use mathematical language to develop a framework for bringing what may have seemed like a disparate collection of topics and tools into a unified way of thinking. With that framework in place we'll see what we can say about some of the problems that were too difficult for Introductory Statistics. For example, after this class you will be able to tackle

- multiple predictor variables,
- non-normal data,
- classification problems,
- and more.

Data frame → Statistical analysis → Model

Data frames are arrangements of **observations** of **variables**. An **observation** is a single unit. A **variable** is a measurement made on that unit

- Record observations as **rows** and variables in **columns**.
- Variables can be **categorical** or **numerical**.
  - Categorical variables can be **binary** or not, **ordered** or not.
  - Numerical variables can be **discrete** or **continuous**.
- Dates, times and locations merit special consideration.
- Vocabulary is not universal: Factor, case, treatment ...

```
head(Sitka) # from package MASS
```

```
##   size Time tree treat
## 1 4.51  152    1 ozone
## 2 4.98  174    1 ozone
## 3 5.41  201    1 ozone
## 4 5.90  227    1 ozone
## 5 6.15  258    1 ozone
## 6 4.24  152    2 ozone
```

Distributions

The **distribution** of a variable is a description of how often it takes each possible value.

- An **observed distribution** is what we actually see.
  - "the sample"
  - column of a data frame
  - summarized by computing means, standard deviations, medians, sample proportions, et cetera.
- A **theoretical distribution** is a mathematical guess.
  - "assume X is normally distributed"
  - "assume the probability of success is 50%"
  - might come from a formula
  - might come from randomization/ simulation

Much of statistics is concerned with comparing observed distributions to theoretical distributions.

We often discuss distributions of variables other than those explicitly in our data, such as the mean of a variable in our data, a test statistic like χ^2, or the residuals of a linear model.

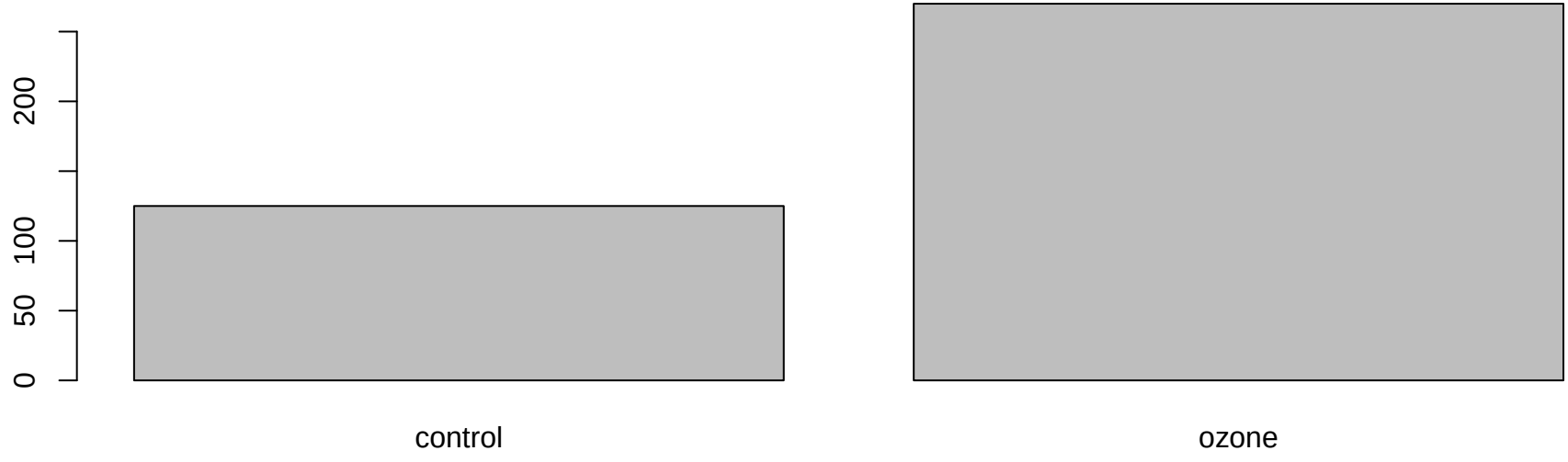
Distributions of Categorical Variables

- The distribution of a categorical variable is a list of the percentage of observations in each category.

```
table(Sitka$treat)/length(Sitka$treat)
```

```
##
##    control      ozone
## 0.3164557 0.6835443
```

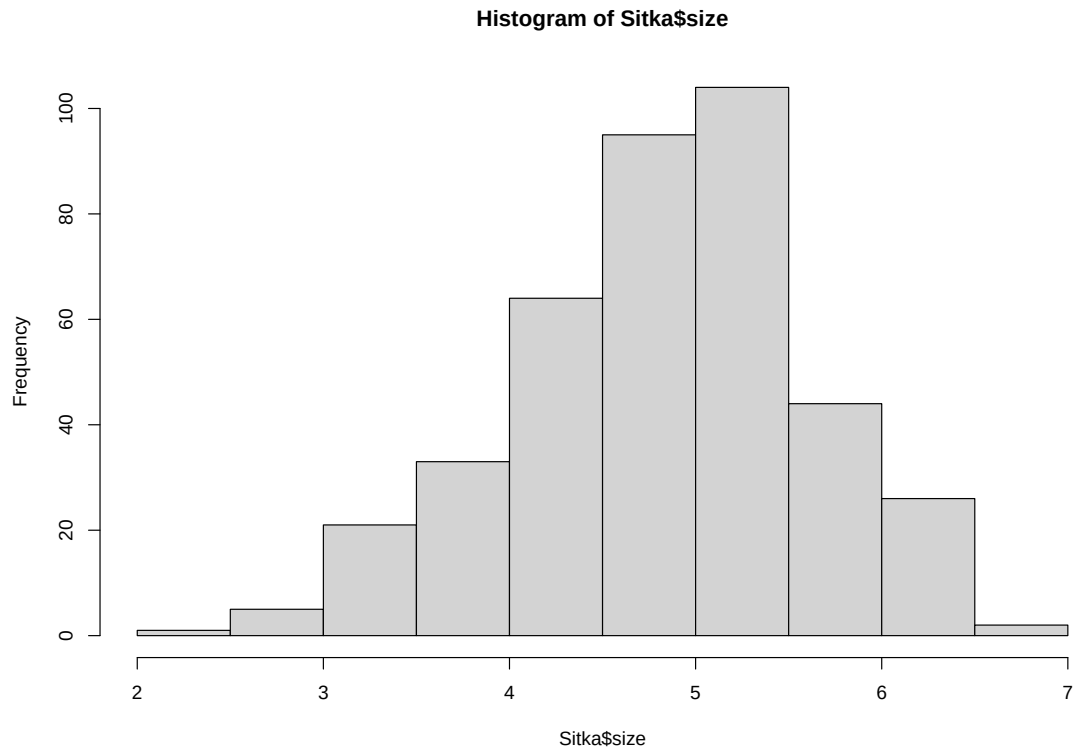
```
barplot(table(Sitka$treat))
```



Distributions of Numeric Variables

- Picture the distribution of a numeric variable with a histogram, boxplot or density estimate.

```
hist(Sitka$size)
```



- Shape: center, spread, skew, kurtosis

Summaries of Numeric Variables

```
summary(Sitka$size)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   2.230   4.345   4.900   4.841   5.400   6.630
```

```
quantile(Sitka$size, c(0.025,0.975))
```

```
##   2.5%  97.5%
## 3.2370 6.2815
```

```
sd(Sitka$size)
```

```
## [1] 0.7982084
```

```
var(Sitka$size)
```

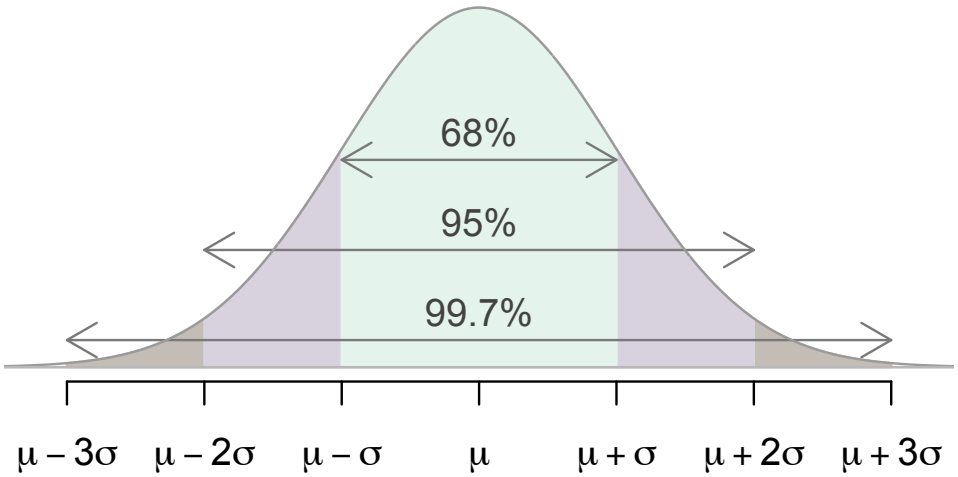
```
## [1] 0.6371367
```

```
IQR(Sitka$size)
```

```
## [1] 1.055
```

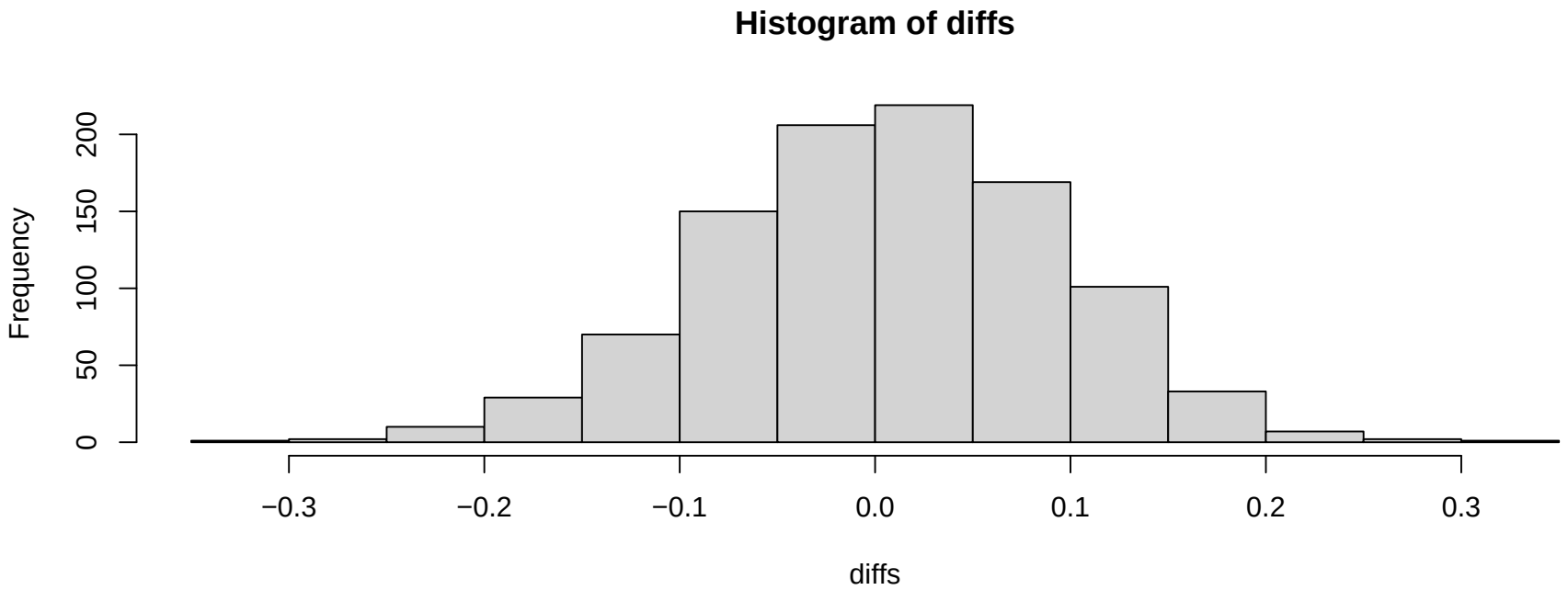
A familiar theoretical distribution: the Normal distribution

- Sums of many independent effects are normally distributed.
- Means are normally distributed.
- Proportions of successes are eventually normally distributed.
- Formula that you never use: $N(\mu, \sigma^2)(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$
- The **standard normal distribution** has $\mu = 0$, $\sigma = 1$.
- Observations on disparate scales can be standardized with z scores: $z = \frac{x - \mu}{\sigma}$



Randomization for a Theoretical Distribution

```
n <- nrow(Sitka); num_ozone <- sum(Sitka$treat=="ozone"); num_sim <- 1000
set.seed(17)                    # initialize the random number generator
diffs <- numeric(num_sim)       # initialize a vector to hold the differences
for(i in 1:num_sim){              # loop: repeat the following num_sim times
  ozone <- sample(1:n,num_ozone)          # 1. randomly select observations
  ozone_mean <- mean(Sitka$size[ozone])# 2. average selected observations
  control_mean <- mean(Sitka$size[-ozone])# 3. average the others
  diffs[i] <- ozone_mean - control_mean}  # 4. store difference in means
hist(diffs)     # plot the resulting distribution
```



Other Theoretical Distributions from Intro Stats

Your first statistics class introduced you to several useful distributions:

- t - Like the normal distribution, but adjusted for describing means of small samples.
- χ^2 - Sum of several squared standard normal distributions. Useful when discussing several proportions at once, such as when considering categorical variables with more than two possible values.
- F - Similar to χ^2, used in ANOVA.

And maybe...

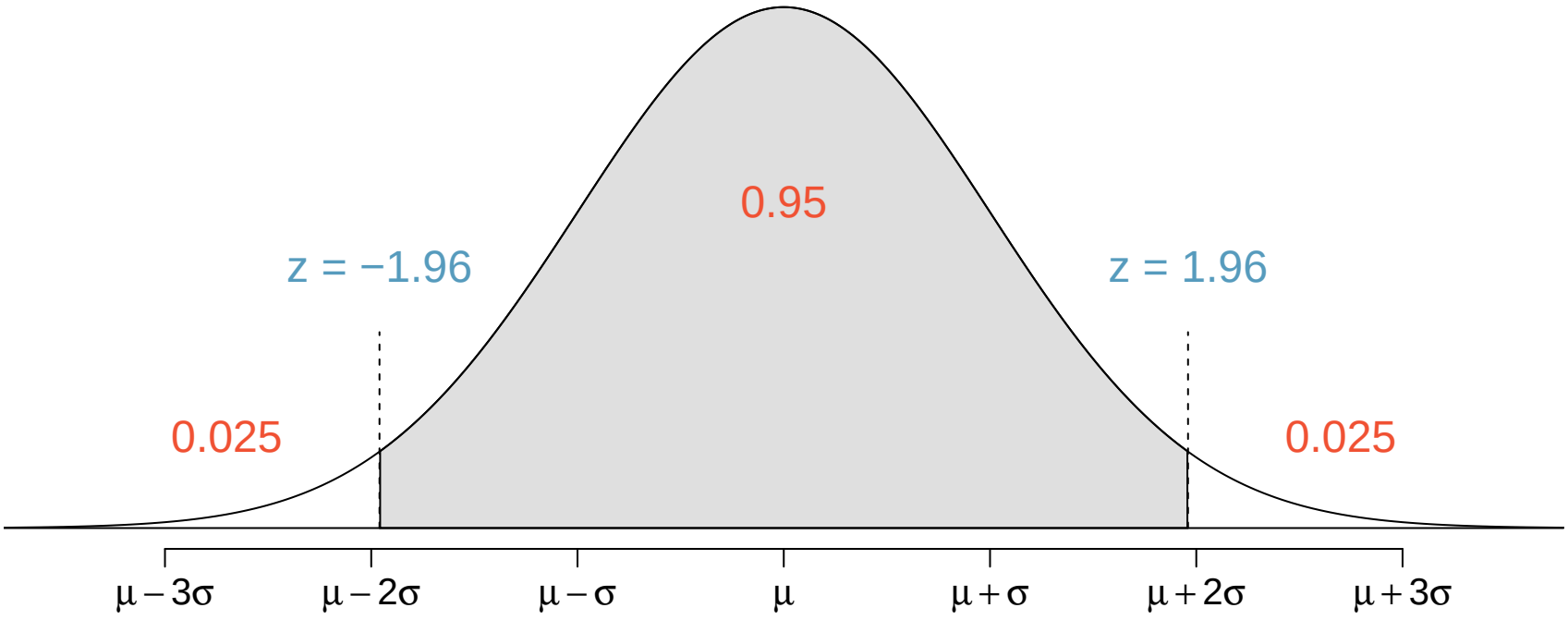
- Binomial - How many successes in n trials?
- Poisson - Count of discrete events in fixed time or space.
- Perhaps others? We'll see lots more...

Sampling Distributions

Given a data set (the sample) and a quantity that we can calculate from the data (a sample or test statistic) we propose an expected distribution for that calculated quantity (the sampling distribution).

- Sampling distributions are never observed, always theoretical.

Having a sampling distribution allows us to do inference.



Inference: Confidence Intervals

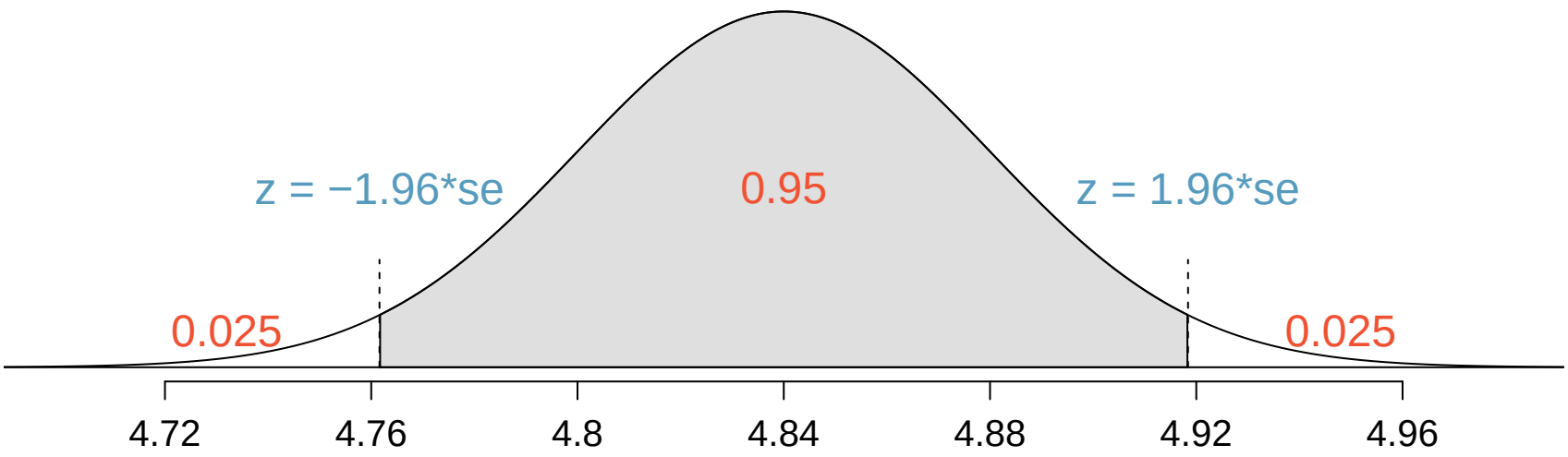
Often, we assume the shape of a sampling distribution, but not its specific parameters. By using the data to estimate those parameters, we get a guess at the sampling distribution that we can use to compute a confidence interval.

```
(x_bar = mean(Sitka$size))
```

```
## [1] 4.840785
```

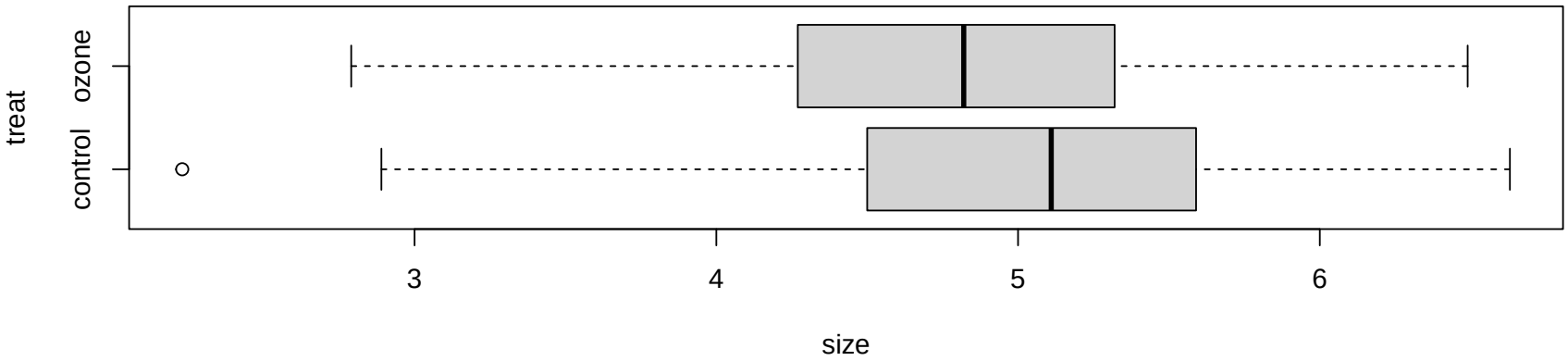
```
s = sd(Sitka$size); n = nrow(Sitka); (se = s/sqrt(n))
```

```
## [1] 0.04016222
```



Inference: Hypothesis Tests

A null hypothesis determines a sampling distribution. Compare data to that proposed distribution. If the data are sufficiently unlikely, we reject the null hypothesis.




```
##
##  Welch Two Sample t-test
##
## data:  size by treat
## t = 2.3163, df = 209.44, p-value = 0.02151
## alternative hypothesis: true difference in means between group control and g
## 95 percent confidence interval:
##  0.03144833 0.39086574
## sample estimates:
## mean in group control   mean in group ozone
##              4.985120              4.773963
```

The Key Question of Intro Stats: Which test to use?!

One variable:

- Numeric: t-test for the mean (`t.test`)
- Binary categorical: z-test or binomial test for the proportion of success (`prop.test` or `binom.test`)
- Nonbinary categorical: χ^2 test for goodness of fit (`chisq.test`)

Two variables:

- Both numeric: linear regression (`lm`)
- One numeric, one binary categorical: t-test for a difference of means (`t.test`)
- One numeric, one nonbinary categorical: analysis of variance (`lm` and `anova` or `aov`)
- Both binary categorical: z-test or binomial test for equality of proportions (`prop.test` or `binom.test`)
- Two categorical, at least one nonbinary: χ^2 test for independence (`chisq.test`)