INSTITUTE FOR DEFENSE ANALYSES

Statistics Bootcamp DataWorks 2019

Heather M. Wojton, Project Leader
Rebecca M. Medlin, Project Leader

Kelly M. Avery
Stephanie T. Lane

OED Draft

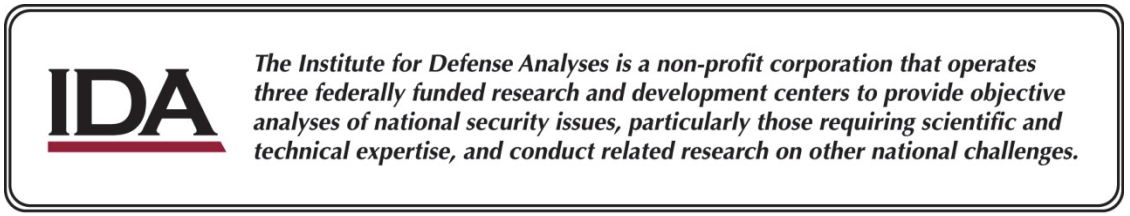
March 2019

Approved for public release.
Distribution is unlimited.

IDA Non-Standard Document
NS D-10565

Log: H 2019-000146

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882



About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-14-D-0001, Task BD-9-2299(90), “Test Science Applications,” and C9087, “T&E Knowledge Exchange," for the Office of the Director, Operational Test and Evaluation. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgments
The IDA Technical Review Committee was chaired by Mr. Robert R. Soule and consisted of Thomas H. Johnson, Heather M. Wojton, and Stephanie T. Lane from the Operational Evaluation Division.

For more information:
Heather M. Wojton, Project Leader
hwojton@idar.og • (703) 845-6811

Robert R. Soule, Director, Operational Evaluation Division
rsoule@ida.org • (703) 845-2482

Copyright Notice
© 2019 Institute for Defense Analyses
4850 Mark Center Drive, Alexandria, Virginia 22311-1882 • (703) 845-2000

This material may be reproduced by or for the U.S. Government pursuant to the copyright license under the clause at DFARS 252.227-7013 (a)(16) [Jun 2013].

INSTITUTE FOR DEFENSE ANALYSES

IDA Non-Standard Document NS D-10565

Statistics Bootcamp DataWorks 2019

Heather M. Wojton, Project Leader
Rebecca M. Medlin, Project Leader

Kelly M. Avery
Stephanie T. Lane

Executive Summary

Statistics can be a polarizing and confusing topic. However, the basic concepts of statistics are essential for extracting meaning from data in an objective and rigorous way. In the test community, we frequently use statistical methods to draw inferences with respect to topics ranging from system performance to human factors. This tutorial is geared toward new analysts in the field, or those who need a refresher on statistical ideas and basic methodologies.

The tutorial begins with some basic definitions and techniques for summarizing and simplifying data. Measures of variability and central tendency are the focus areas, and the slides include examples of when to choose one measure over another. The section concludes with a graphical depiction of the process of statistical inference.

Point and interval estimation and the foundations of statistical inference are the next topics in the tutorial. Examples and animated simulations demonstrate core concepts, such as the law of large numbers, the interpretation of confidence intervals, and the central limit theorem.

The tutorial then describes the general process of hypothesis testing and explains some common statistical tests. It begins with simple tests of one and two means, before introducing various types of regression modeling. Throughout the section, new concepts are reinforced via examples relevant to operational testing. The tutorial concludes with a few tips, including best practices for data collection and visualization.

Statistics Boot Camp

Dr. Kelly Avery
Institute for Defense Analyses
DATAWorks 2019

April 10, 2019

Google (2018)

why is statistics so

why is statistics so **hard**
why is statistics so **important**
why is statistics so **boring**
why is statistics so **hard to say**
why is statistics so **confusing**
why is statistics so **important for public**

Google Search

Google (2019)

why is statistics so

why is statistics so **hard**
why is statistics so **boring**
why is statistics so **important for public health**
why is statistics so **hard reddit**
why is statistics so **confusing**
why is statistics so **difficult**
why is statistics so **hard for me**
why is statistics so **easy**
why is statistics so **important**

Google Search I'm Feeling Lucky

Report inappropriate predictions

blossom271828 6 points · 2 years ago

The challenging part about statistics is not the calculations or the technical details, but rather the process of understanding what the thing you just calculated tells you about the world.

Outline of boot camp

- Summarizing and simplifying data
- Point and interval estimation
- Foundations of statistical inference
- The process of hypothesis testing
- Common statistical tests
- A few closing tips

Thinking about variables

Variables are characteristics that are observed or manipulated in a study

We frequently think about variables in terms of being **continuous** or **discrete**

- **Continuous** variables have fractional amounts
  - e.g., height, distance, time
- **Discrete** variables are distinct, separate, countable values
  - e.g., number of missiles, crew rank

The type of variable has implications for how we *describe*, *analyze*, *visualize*, and *characterize* our data.

Scales of measurement

Scale of measurement	Definition	Example
Nominal	Named categories	Wheel Type (Tracked, Wheels)
Ordinal	Ordered categories	Place in a race (First, Second, Third)
Interval	Ordered values, equidistant	Temperature (Fahrenheit, Celsius)
Ratio	Ordered values, equidistant, real zero	Distance (meters)

When possible, we prefer interval and ratio measures, as they contain more information

Descriptive statistics – simplify and summarize

Where do scores tend to fall (**central tendency**) and how spread out are they (**variability**)?

Central tendency aims to describe the centrality, or the typical score of a distribution

- Mean, median, and mode are measures of central tendency

Variability aims to quantify the spread of a distribution, or the typical spread away from the center

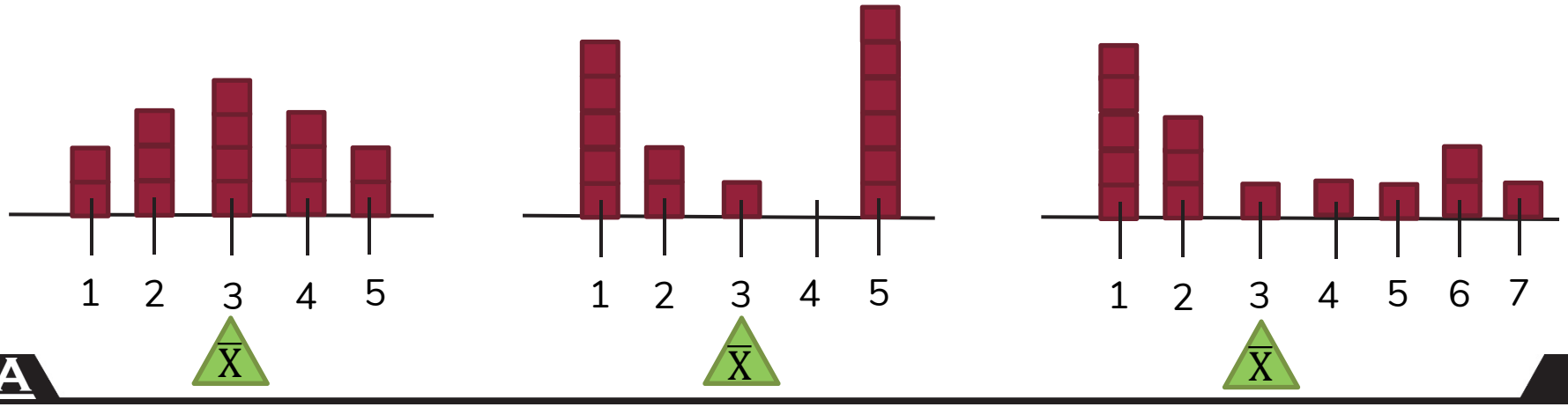
- Standard deviation, variance, range, and interquartile range are measures of variability

Central tendency

The **mean** is the average score of the distribution

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{N}$$

It is useful for describing symmetric distributions, and is frequently thought of as the “balance point” of a distribution.



Central tendency

The **median** is the 50th percentile, or the middle score in the distribution

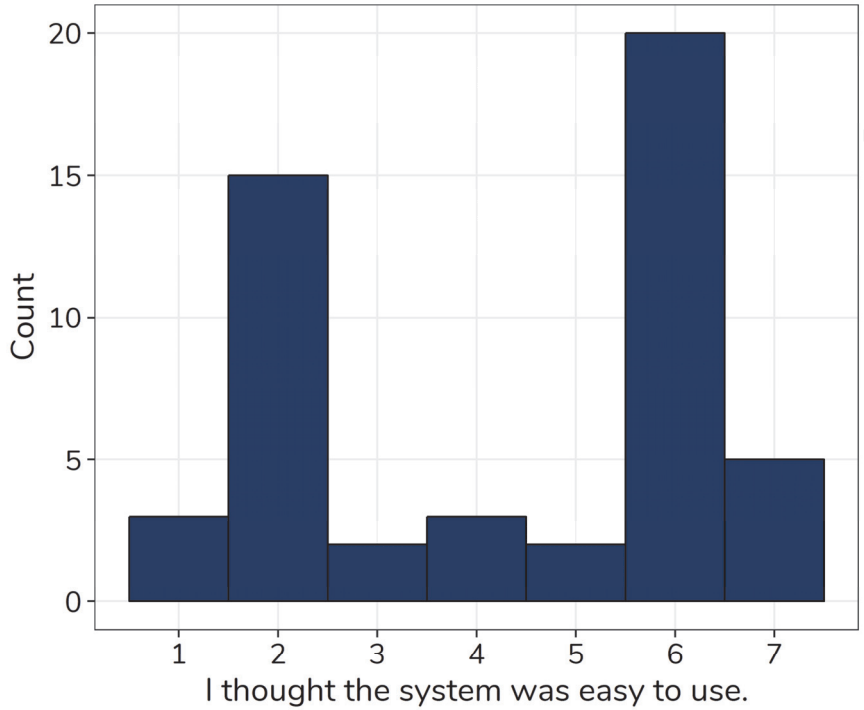
- It is useful for describing skewed distributions
- It is robust to extreme scores, meaning that it is minimally affected by outlying observations

The **mode** is the most commonly occurring score

- It is useful for describing nominal data and multi-modal distributions

Central tendency – reporting

Suppose we introduce a new system and administer a usability survey to operators. We ask operators to rate the usability of the system on a 1-7 scale. We plot our data and see the following results:



Which measure of central tendency should we report?

Variability – characterizing spread

Standard deviation is the typical spread of scores away from the mean, and is a common metric for quantifying variability

s denotes sample;
σ denotes population

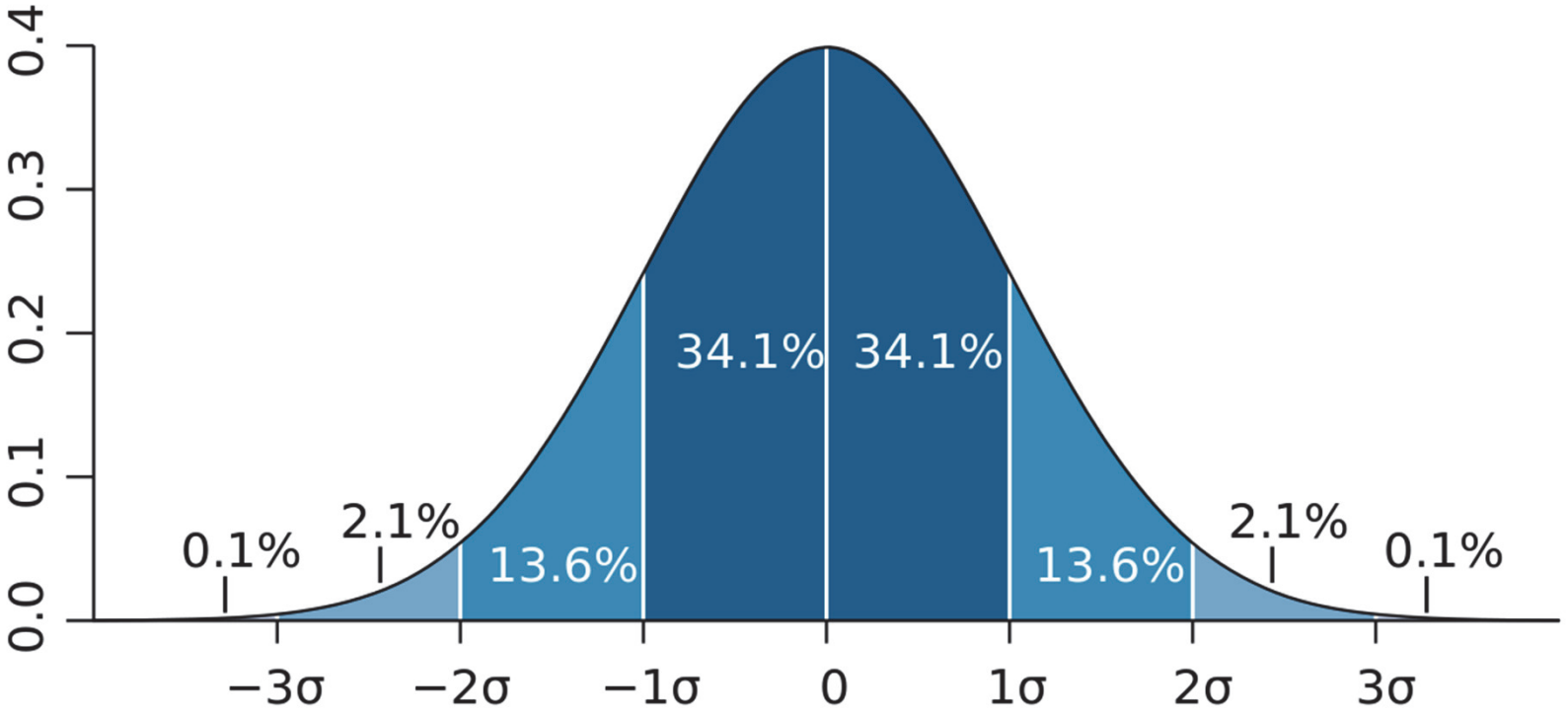
$$s_X = \sqrt{\frac{\sum_{i=1}^{N}(X_i - \bar{X})^2}{N-1}}$$

Like the mean, it takes all scores into account. Therefore, it is influenced by extreme scores or outliers.

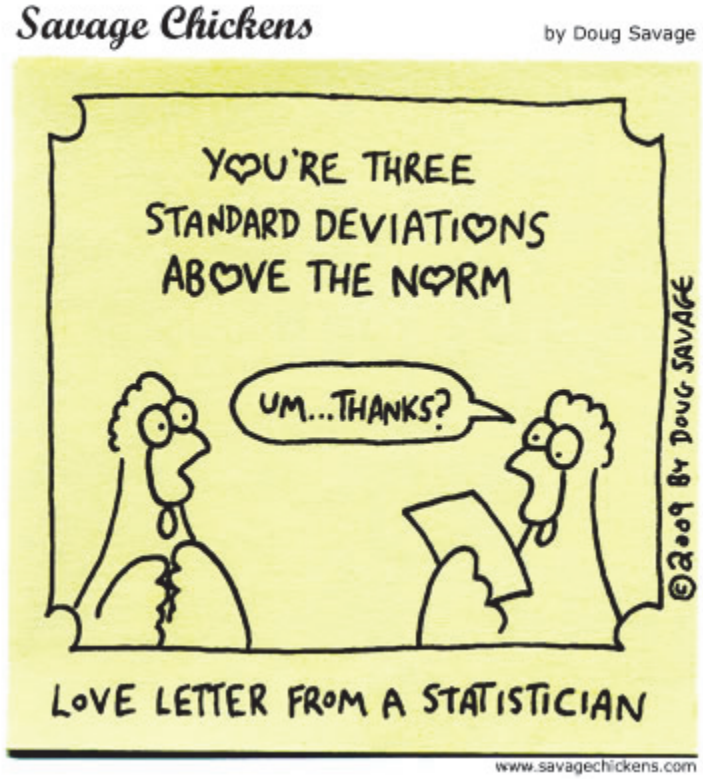
The variance is the squared standard deviation, s_X^2. We prefer reporting the standard deviation, as it is in the original units of the variable (e.g., yards, miles, points, etc.).

If we convert our scores to standard scores, we can easily quantify their distance away from the mean

$$z = \frac{X - \mu}{\sigma}$$



If we convert our scores to standard scores, we can easily quantify their distance away from the mean



Beyond descriptive statistics

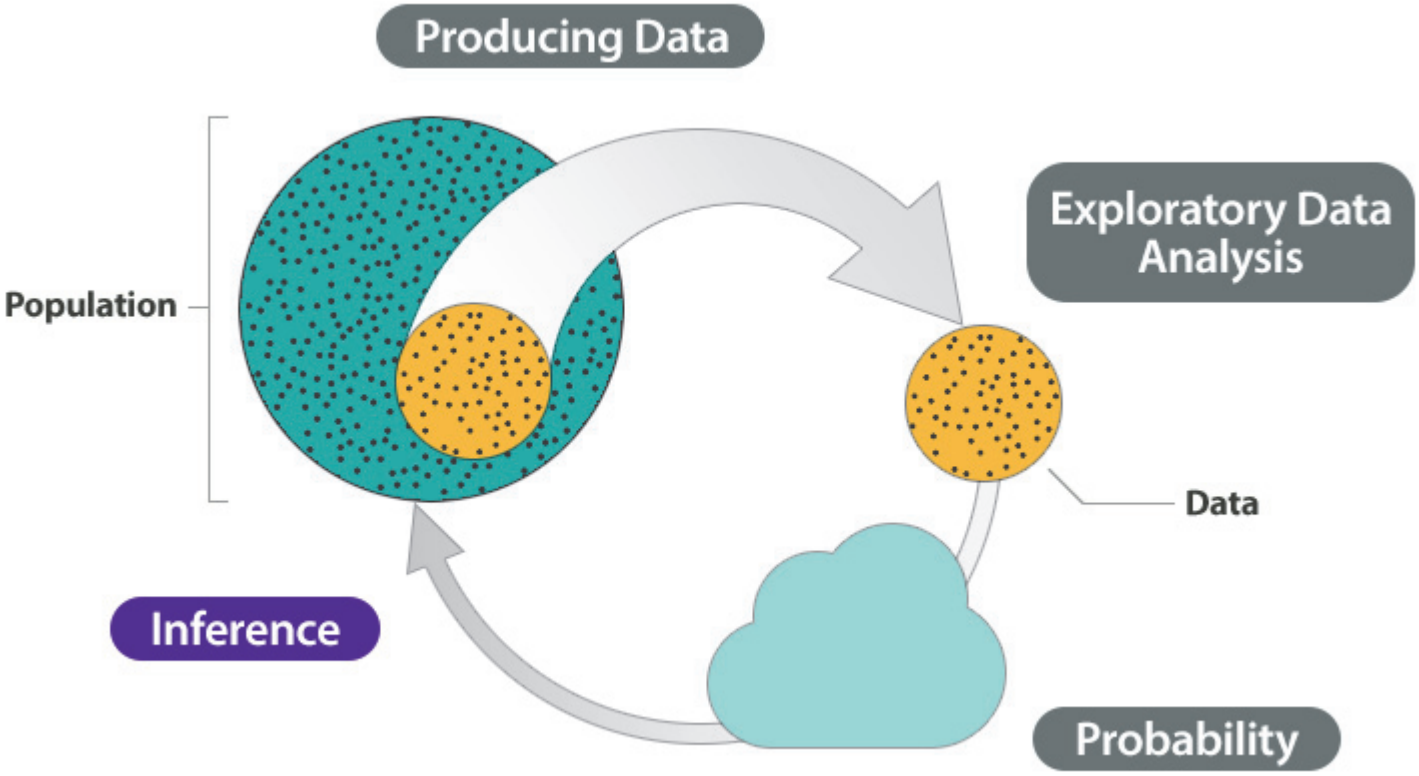
Descriptive statistics are useful, and are a great starting point for describing and understanding your data.

But as an analysis tool, they can only get us so far.

What if we are interested in not only **describing** the sample, but also in **drawing inference** to the broader population?

For this question, we need inferential statistics.

Inferential statistics



Outline of boot camp

- Summarizing and simplifying data
- **Point and interval estimation**
- Foundations of statistical inference
- The process of hypothesis testing
- Common statistical tests
- A few closing tips

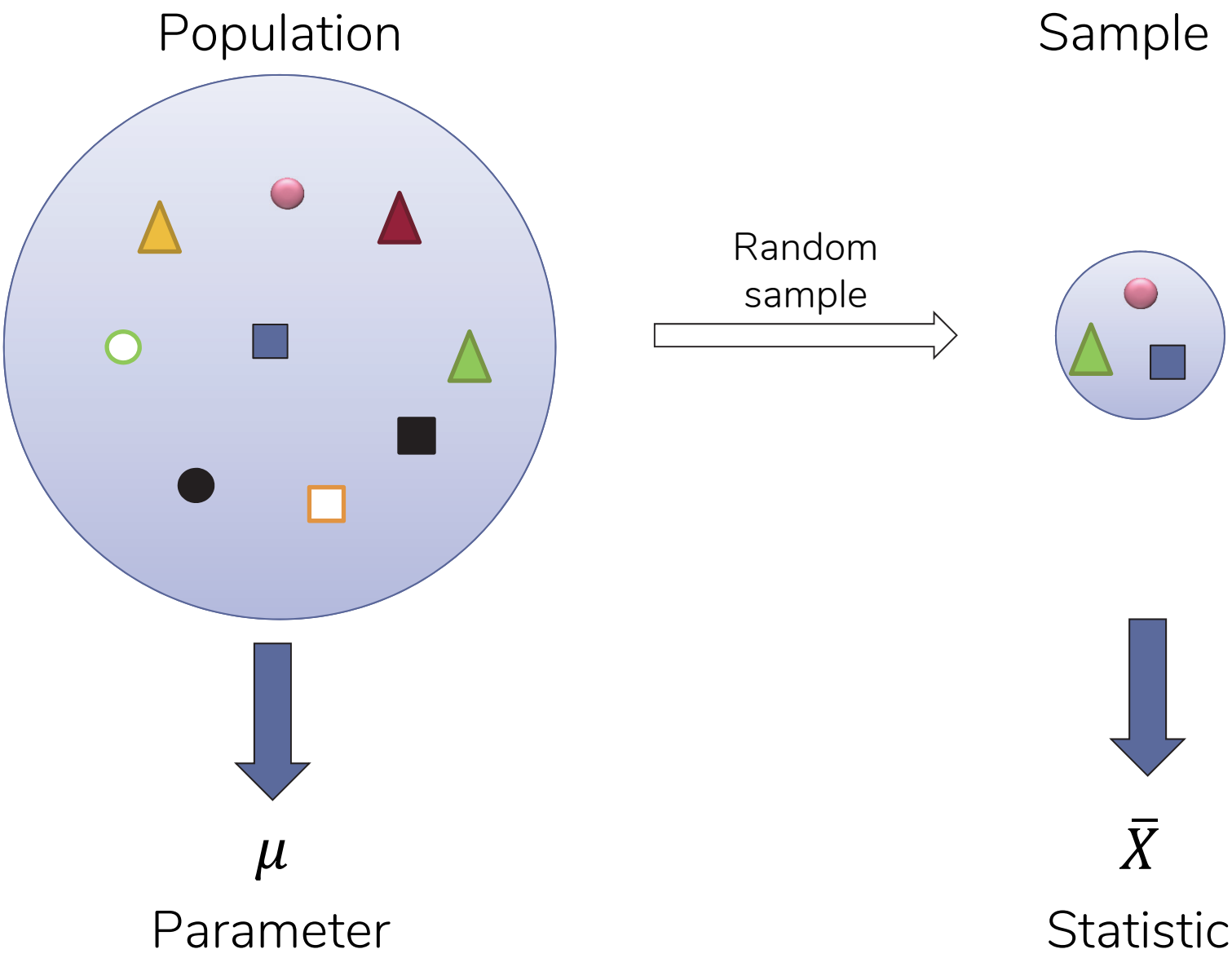
Point estimation

In point estimation, we are interested in estimating some unknown **parameter** using a statistic (e.g., a mean) that we calculate from our sample data.

The **parameter** is the value that summarizes the entire population.

We use our sample statistic to draw inferences about the value in the population – the population parameter.

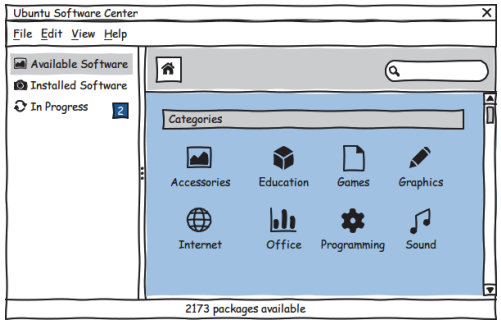
Understanding vocabulary – parameter versus statistic



Examples of point estimation

We have a new communications system. We measure the time it takes to transmit a message. From the sample of data we collect during testing, we compute a mean of $\bar{X} = 0.48$ seconds.

Our best estimate of the population mean, μ, is .48 seconds.



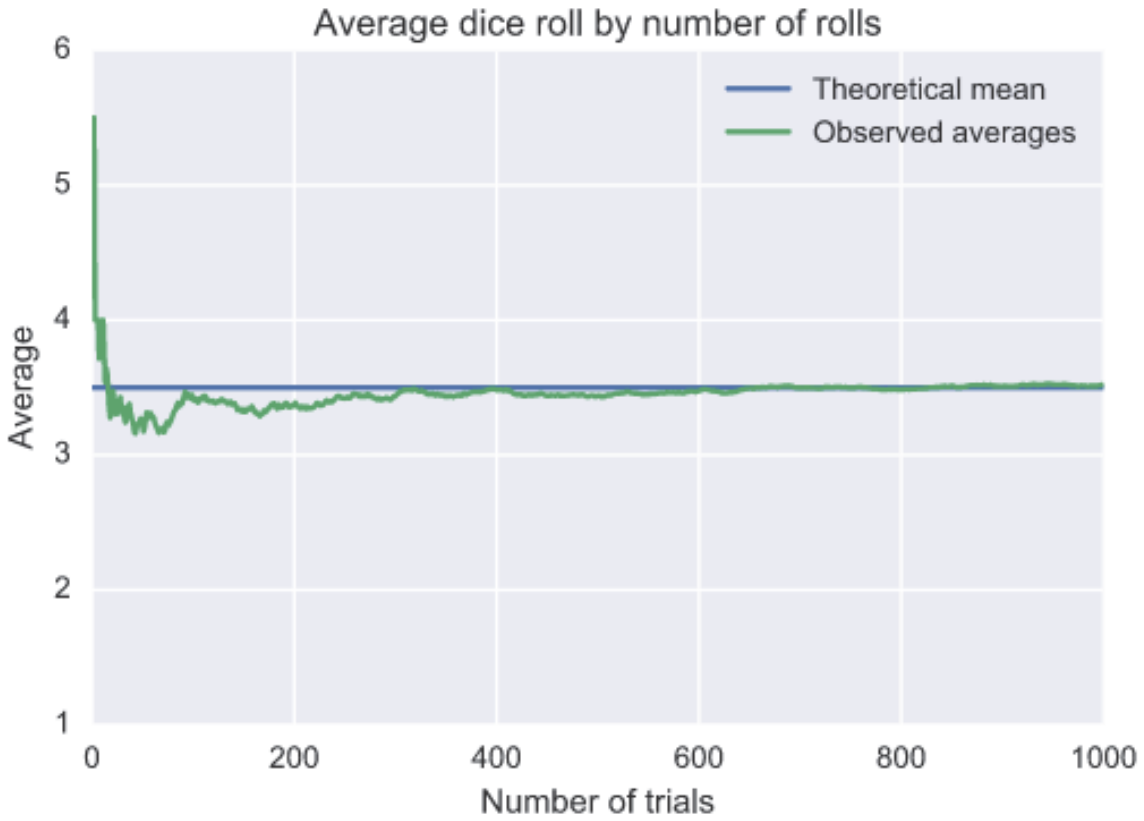
We have a new user interface for a software program. We observe whether operators could complete their mission using the new interface. We find that $27/30$, or 90% of operators, were able to successfully complete their mission.

Our best estimate of the population proportion, p_0, is .90.

How do we know our sample statistic (e.g., mean, proportion) is the best estimate of the population parameter?

Law of large numbers

As the number of our observations increases, the difference between the sample mean and population mean goes to zero



Our point estimate is a good starting point for quantifying a population parameter.

But might you also want to know…

…how much uncertainty is there?

Interval estimation

Interval estimation

Interval estimation provides us a range of values that have a specific likelihood of containing our population value

A confidence interval is constructed using our sample statistic $\pm$ our margin of error

$$\bar{X} \pm t_{\frac{\alpha}{2}} \frac{s_X}{\sqrt{N}} \qquad \bar{X} \pm z_{\frac{\alpha}{2}} \frac{s_X}{\sqrt{N}} \qquad \hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

The margin of error quantifies the degree of uncertainty in our estimate

What does a confidence interval tell us?

"If we were to repeat the study many times, 95% of the confidence intervals we constructed would contain the value of the true population parameter"

"This is why I hate statisticians."

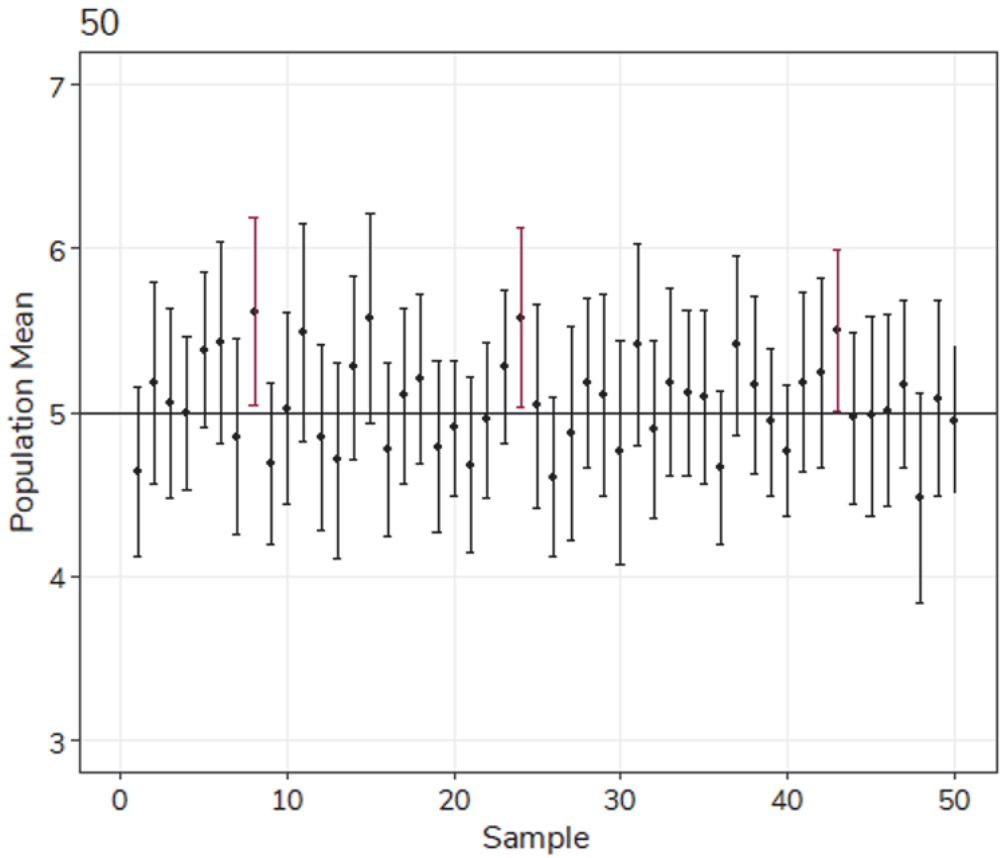
Audience poll: how does this interpretation make you feel?

What does a confidence interval tell us?

Let's unpack this with a quick demonstration.

Suppose the true population mean is $\mu = 5$.

I take a sample from the population, compute the sample statistic and margin of error, and construct a confidence interval.



Factors affecting confidence interval width

More variability ➔ wider confidence intervals

Smaller sample size ➔ wider confidence intervals

Larger confidence level ➔ wider confidence intervals

A wider confidence interval reflects more uncertainty in our estimate of the population parameter.

Outline of boot camp

- Summarizing and simplifying data
- Point and interval estimation
- **Foundations of statistical inference**
- The process of hypothesis testing
- Common statistical tests
- A few closing tips

~~The magical number 30~~ Central limit theorem

For a population with mean μ and standard deviation σ, the distribution of sample means for sample size n will have a mean of μ and a standard deviation of $\sigma/\sqrt{n}$ and will approach a normal distribution as n approaches infinity.

What does this buy us?

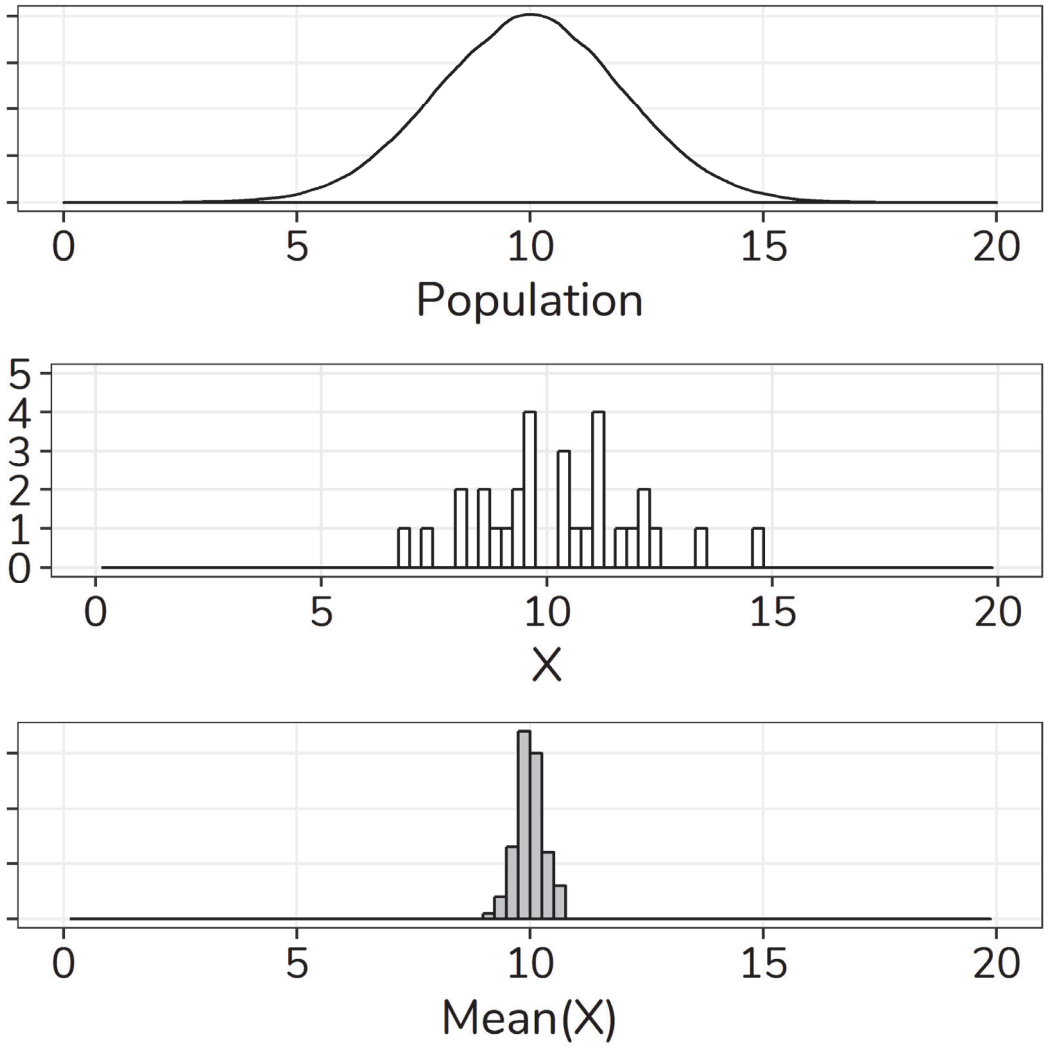
We can perform inference even
if we don't know the shape of the population distribution!

Demonstrating the CLT

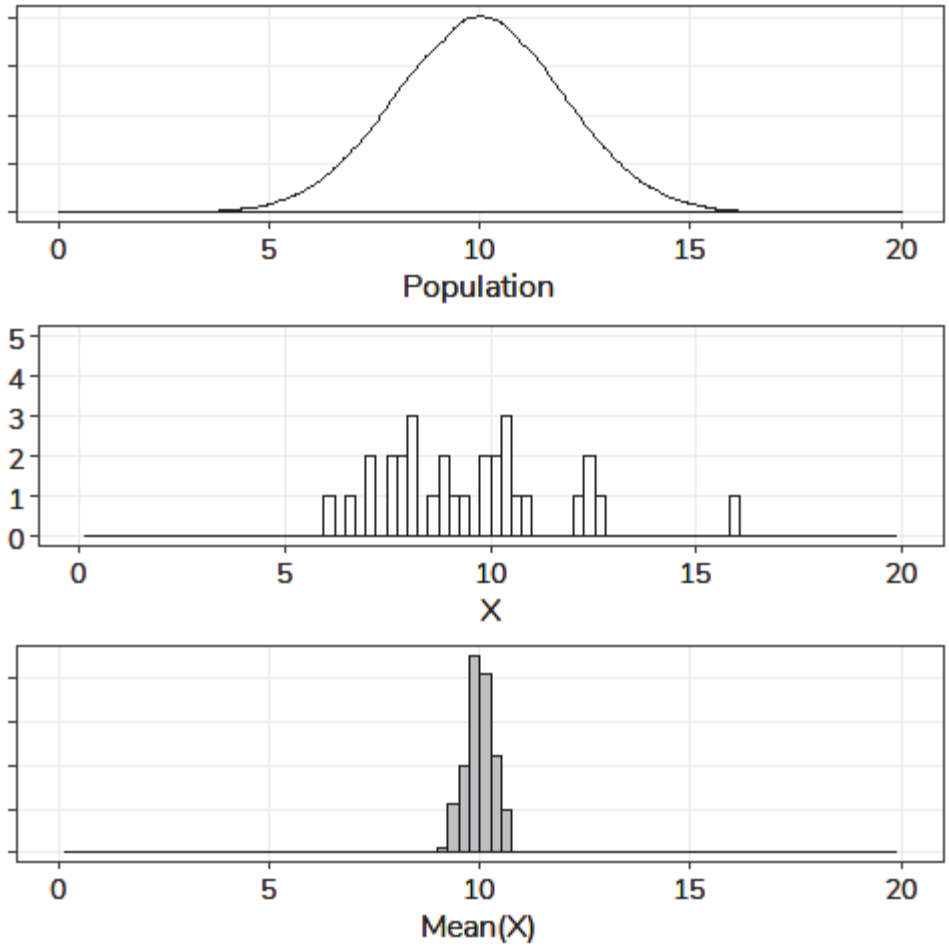
Suppose our population distribution is normally distributed, $\mu = 10; \sigma = 2$

We take samples of $N = 30$, with replacement, from the population

The sampling distribution of means is normally distributed, $\mu = 10; \sigma = \frac{2}{\sqrt{30}}$



Demonstrating the CLT via simulation

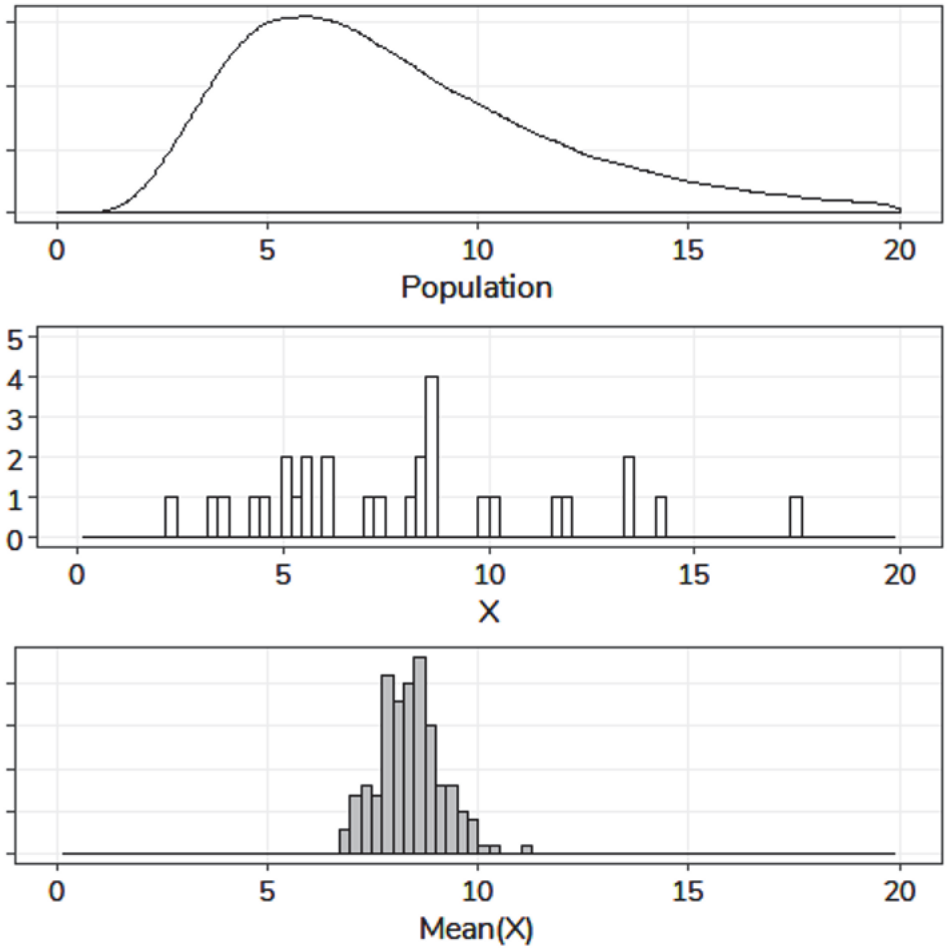


The sample means are normally distributed

But remember!

We said that the central limit theorem allowed us to know the shape of the sampling distribution of means, even if the population distribution was not normal

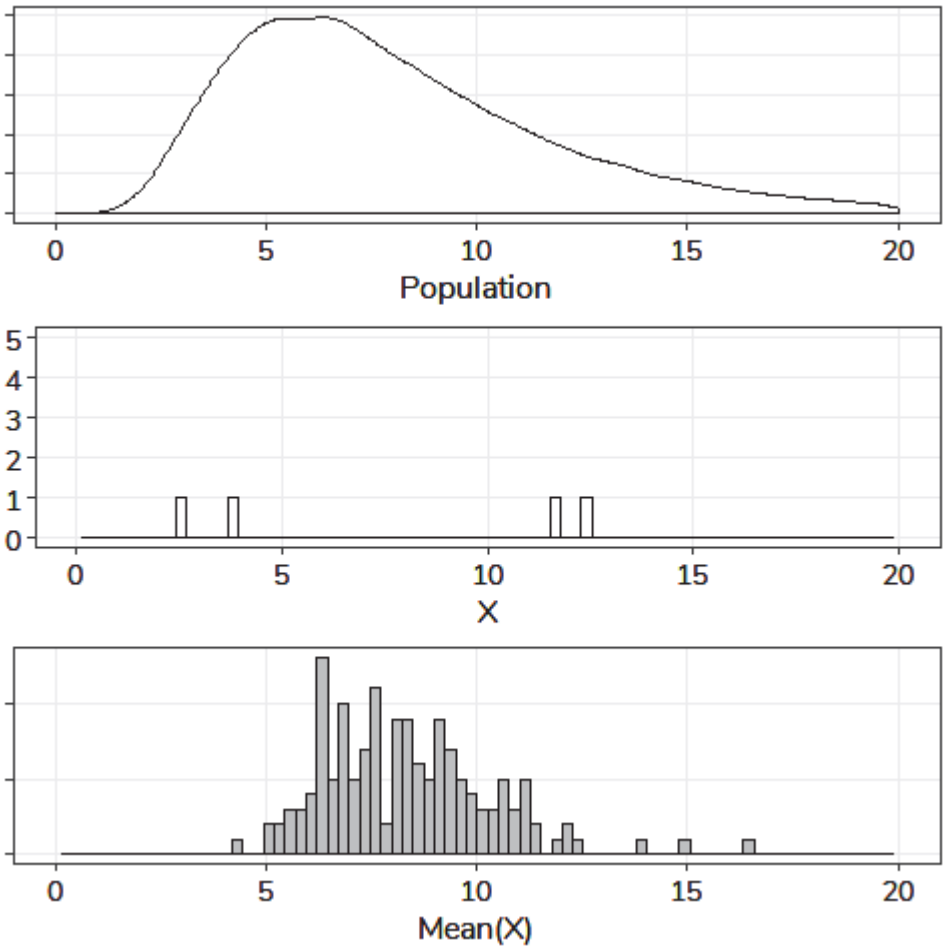
What if the population distribution is not normal?



The sample means are still normally distributed

What if our population distribution is not normally distributed and our sample size is < 30?

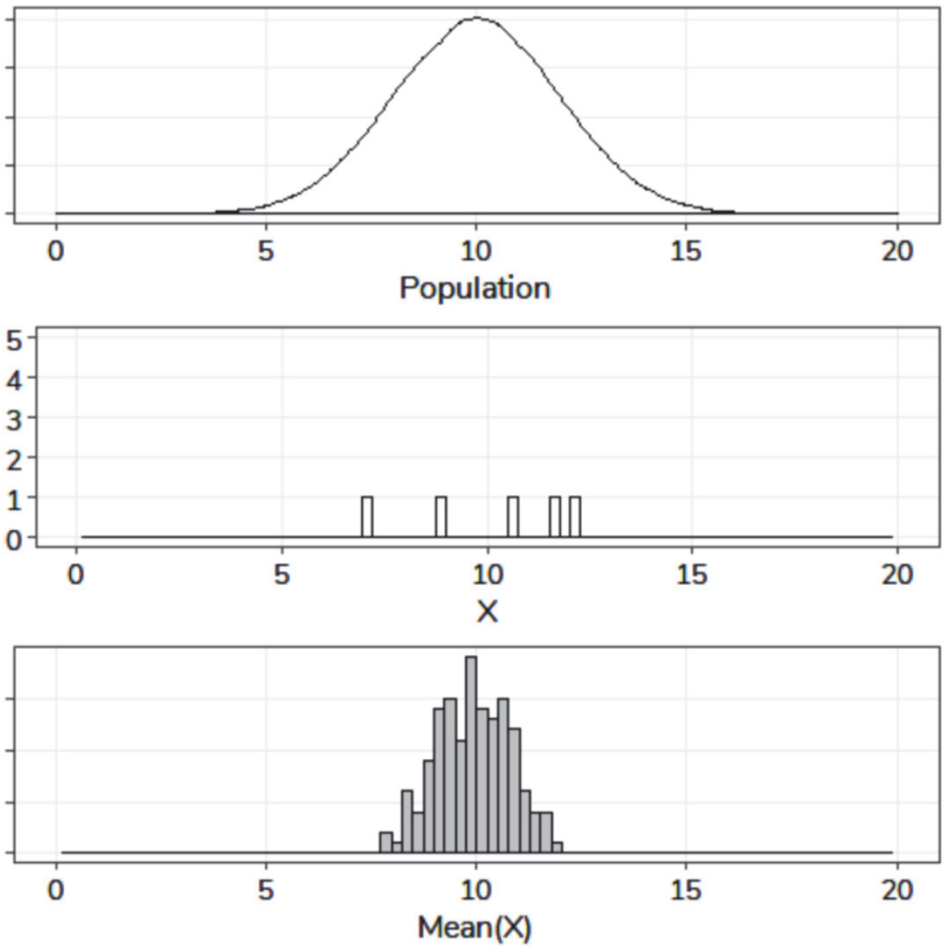
The sampling distribution will resemble the population distribution



We shouldn't use methods designed for normal theory here

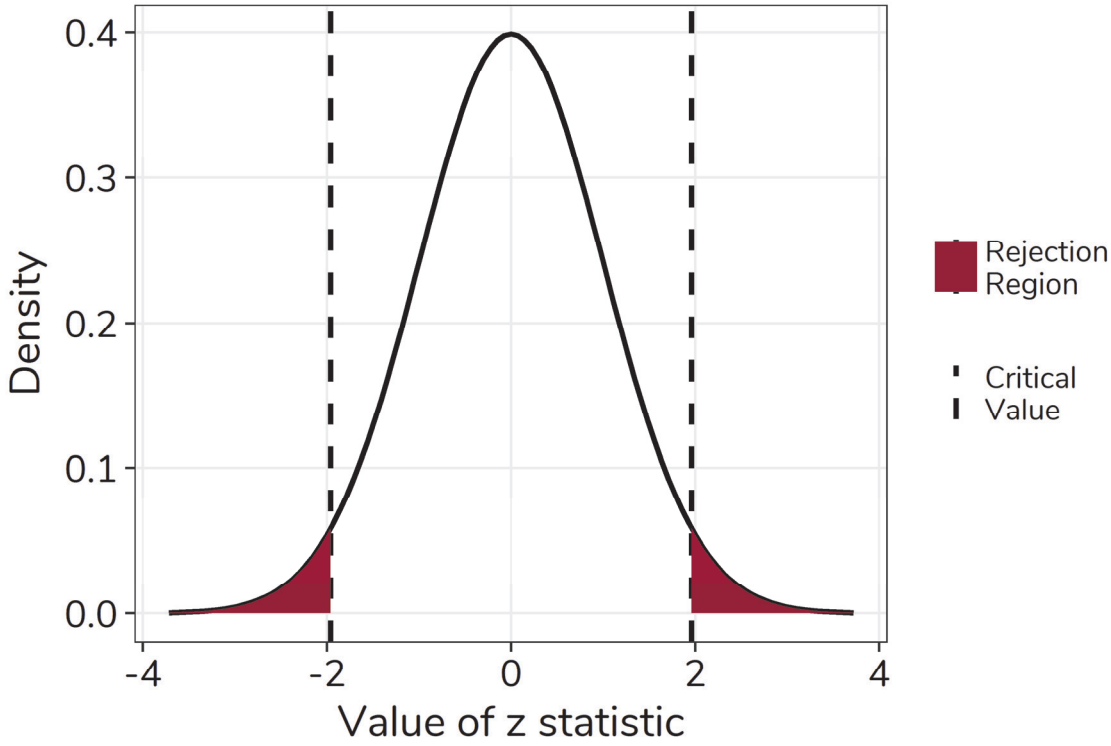
What if the population is normally distributed and our sample size is small?

The sampling distribution of means will still be symmetric and approximately normal



We can more or less use methods based on normal theory here

The CLT is fundamental to the use of sampling distributions



Bottom line: the CLT allows us to perform inference on means even when we don't know the population distribution

Outline of boot camp

- Summarizing and simplifying data
- Point and interval estimation
- Foundations of statistical inference
- **The process of hypothesis testing**
- Common statistical tests
- A few closing tips

Hypothesis testing is a process of evaluating a claim that we make about the population

Hypothesis testing

Step 1: State your hypotheses

Step 2: Set the acceptable level of risk (alpha)

Step 3: Collect data + compute test statistic

Step 4: Determine probability

Step 5: State your research conclusion

Step 1: State your hypotheses

We state two **mutually exclusive** hypotheses: the null hypothesis and the alternative hypothesis

- The null hypothesis is the hypothesis of "no change," "no difference," or "no relationship"

- The alternative hypothesis states that there is a change, a difference, or a relationship

Hypotheses can be directional or nondirectional, corresponding to one- and two-tailed tests, respectively

Step 2: Set the acceptable level of risk

There is no single correct answer to this!

In academic contexts, $\alpha = .05$ is frequently used as an acceptable level of risk

In defense research, sometimes $\alpha = .20$ is used as an acceptable level of risk

In pharmaceutical research, $\alpha = .01$ or even $\alpha = .001$ might be used!

The acceptable level of risk depends on the context. It should be set at the beginning, prior to analysis.

Step 3: Collect data and compute statistic

There are many test statistics that can be computed during this stage, depending on your hypotheses.

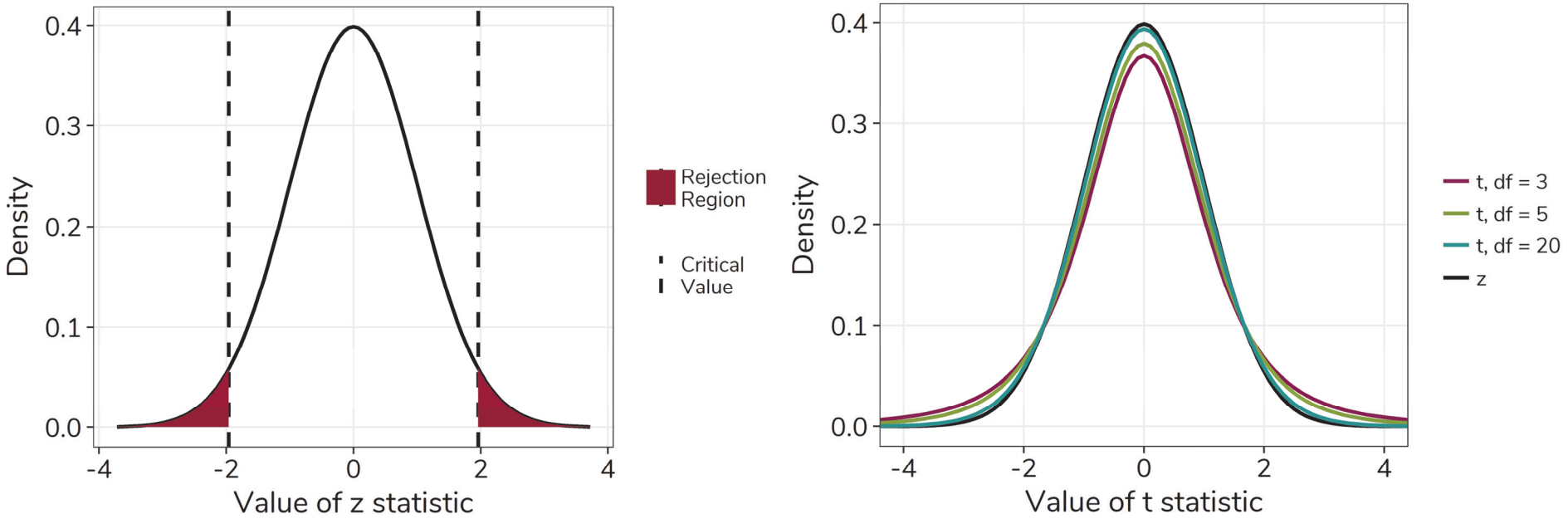
We will focus on two today: z and t

- z is used when we know both the population mean and standard deviation, and we are testing our sample mean against the population mean

- t is used when we do not know the population standard deviation, and we have to estimate it from our sample data

Step 4: Determine probability of your statistic

To determine probability, we compare our statistic to the sampling distribution of our statistic



In other words: we compare our result to the results we would have observed by chance if the null hypothesis was true

Step 5: Draw a research conclusion

Make a decision about the null hypothesis

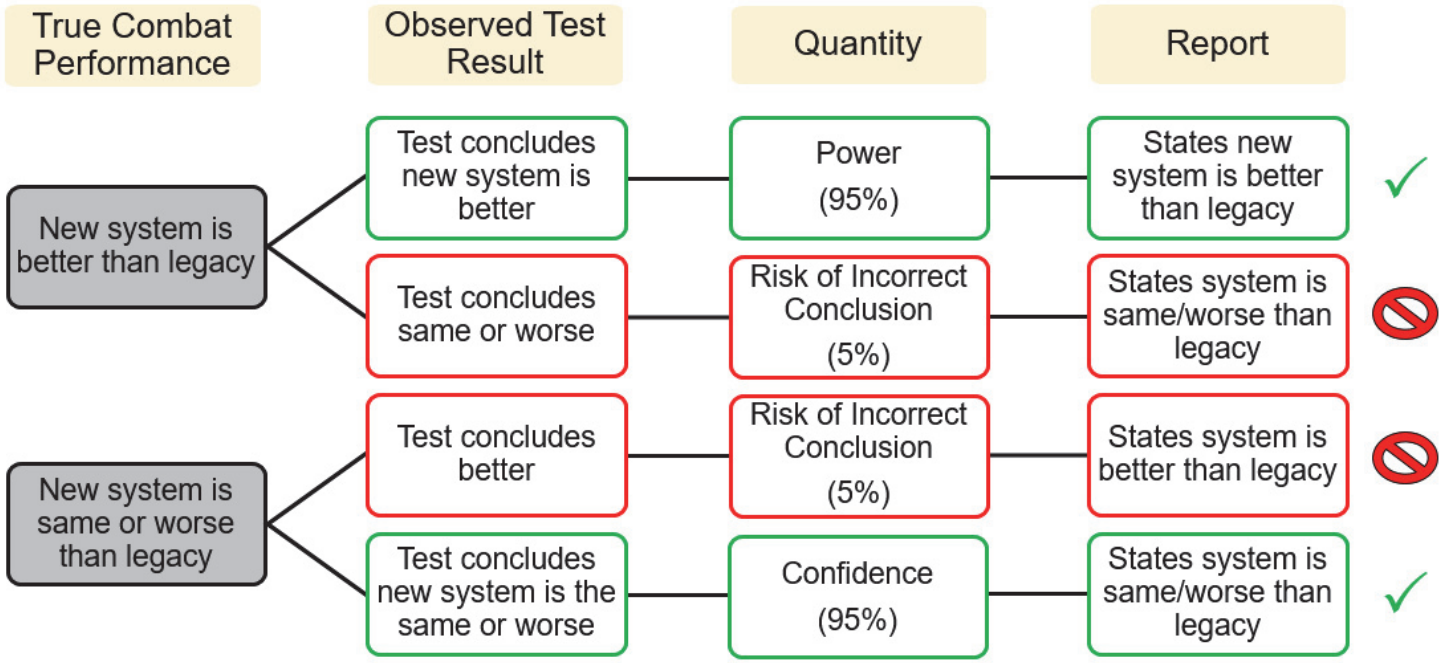
- Reject if $p < \alpha$
- Retain if $p > \alpha$

Draw a conclusion in plain English!

We rejected the null hypothesis saying that there was no effect of training on performance, p = .032.

Training significantly affected performance. Individuals who underwent training had better performance.

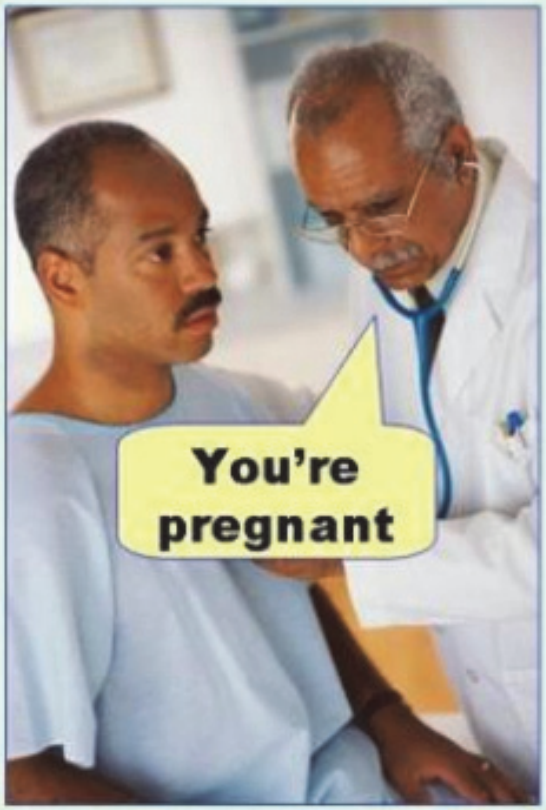
Thinking about errors in hypothesis testing



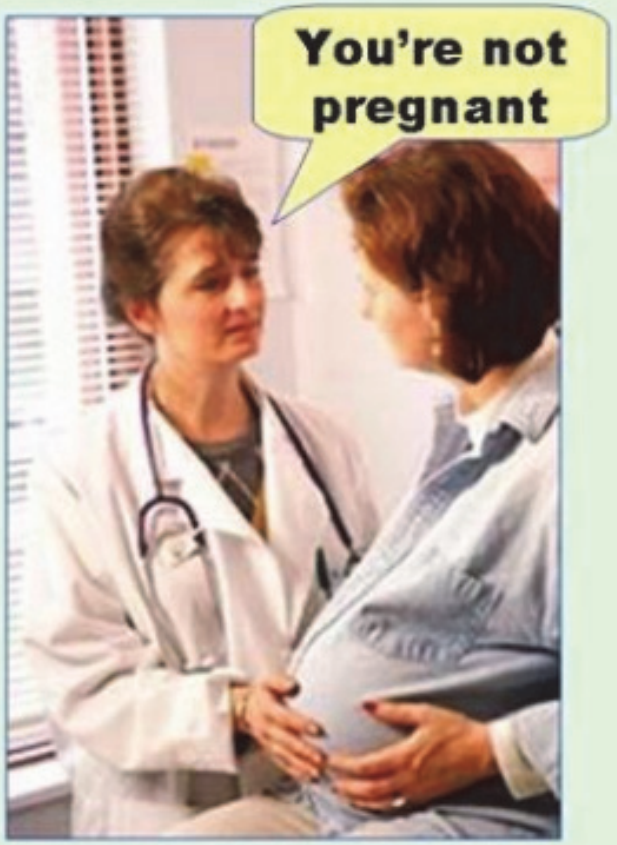
		Decision	
		Retain Null	Reject Null
Truth	Null is true	Correct Decision $1-\alpha$	Type I error α
	Null is false	Type II error β	Correct Decision $1-\beta$

Thinking about errors in hypothesis testing

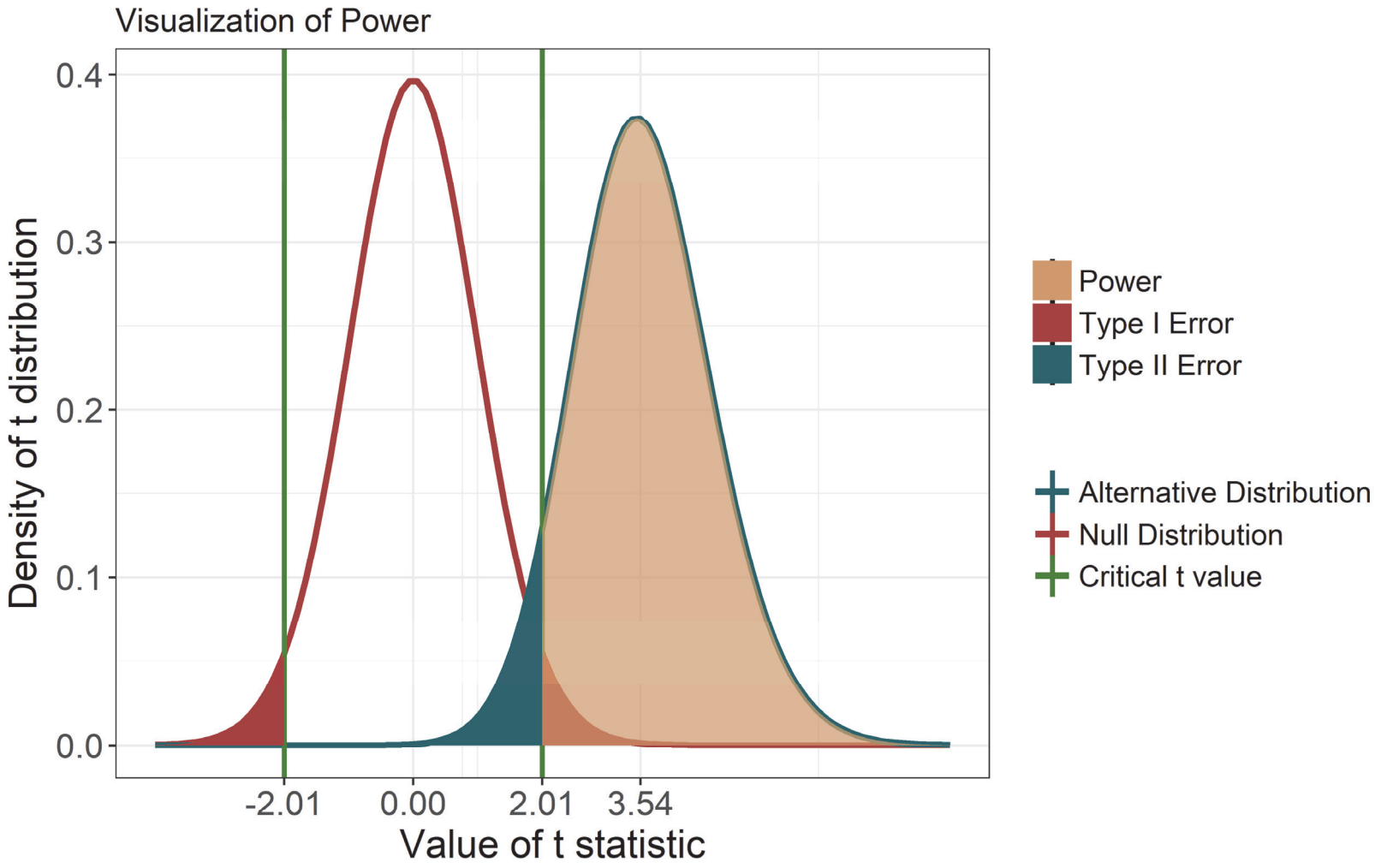
Type I error
(false positive)



Type II error
(false negative)



Visualizing power, Type I error, Type II error



These five steps describe the process of hypothesis testing.

Let’s start with one of the most basic hypothesis tests.

Outline of boot camp

- Summarizing and simplifying data
- Point and interval estimation
- Foundations of statistical inference
- The process of hypothesis testing
- **Common statistical tests**
- A few closing tips

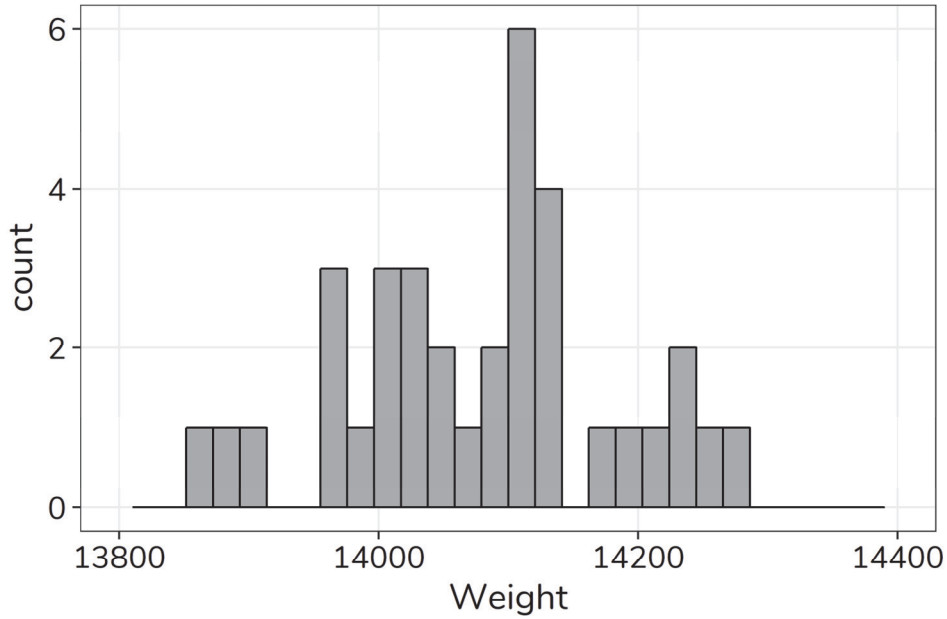
Tests of one mean

Tests of one mean

Scenario: From the manufacturer, we know that the mean weight is specified to be 14000. This is our population mean, μ = 14000. Suppose we have a set of N = 35 vehicles.

Research question: Do we have evidence that our sample of vehicles is significantly different from the population?

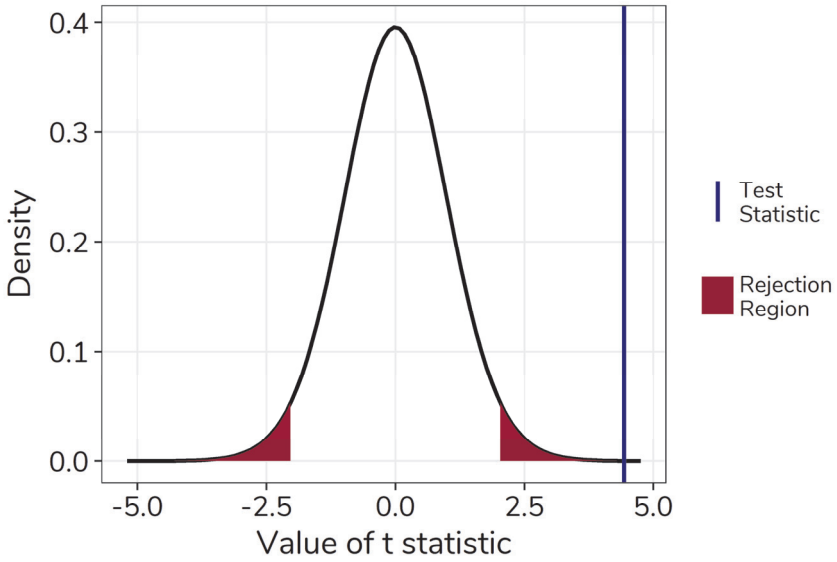
We begin by visualizing our distribution and computing basic descriptive statistics



In our sample, the mean weight was $\bar{X}$ = 14076.6 and the standard deviation was s_X = 102.40.

We proceed to a formal hypothesis test

We construct the t statistic using



$$t = \frac{\bar{X} - \mu}{\frac{s_X}{\sqrt{n}}}$$

What is the difference I observed?

What is the range of differences I expected?

$$t = 4.43$$

We obtain t(34) = 4.43, p < .01. If we set α = .05, then we reject the null hypothesis.

Conclusion: Our sample of vehicles is significantly heavier than the population.

Tests of two means

Tests of two means – motivation

Suppose we are not interested in comparing our sample to the population, but we are instead interested in comparing our samples to each other

For instance, what if we are interested in…

…comparing a new system to an old system?
…comparing variant A to variant B?
…comparing demographic groups?

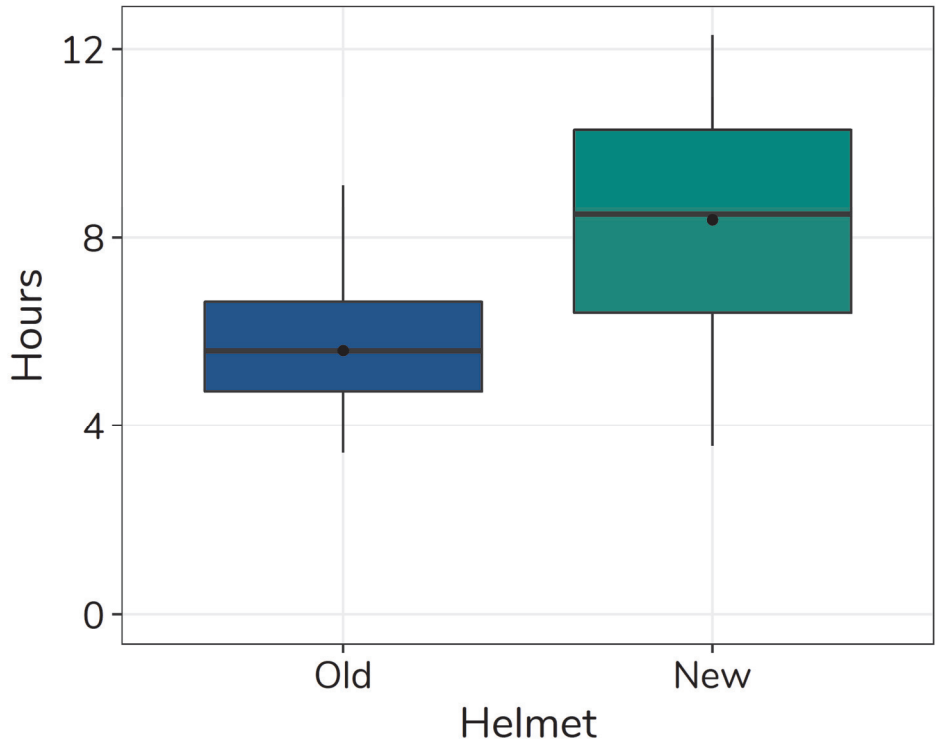
Two-sample t-test: helmet testing

Scenario: We are interested in comparing the comfort of a new helmet to an old helmet. We assign 20 operators to wear the old helmet and 20 operators to wear the new helmet. We observe the number of hours until operators remove the helmet due to discomfort for both the old and new helmet.

Research question: Is the new helmet an improvement upon the old helmet?

Two-sample t-test: describe and visualize

	Group 1 - Old	Group 2 - New
Mean	5.60	8.37
Standard Deviation	2.03	2.49
Sample Size	20	20



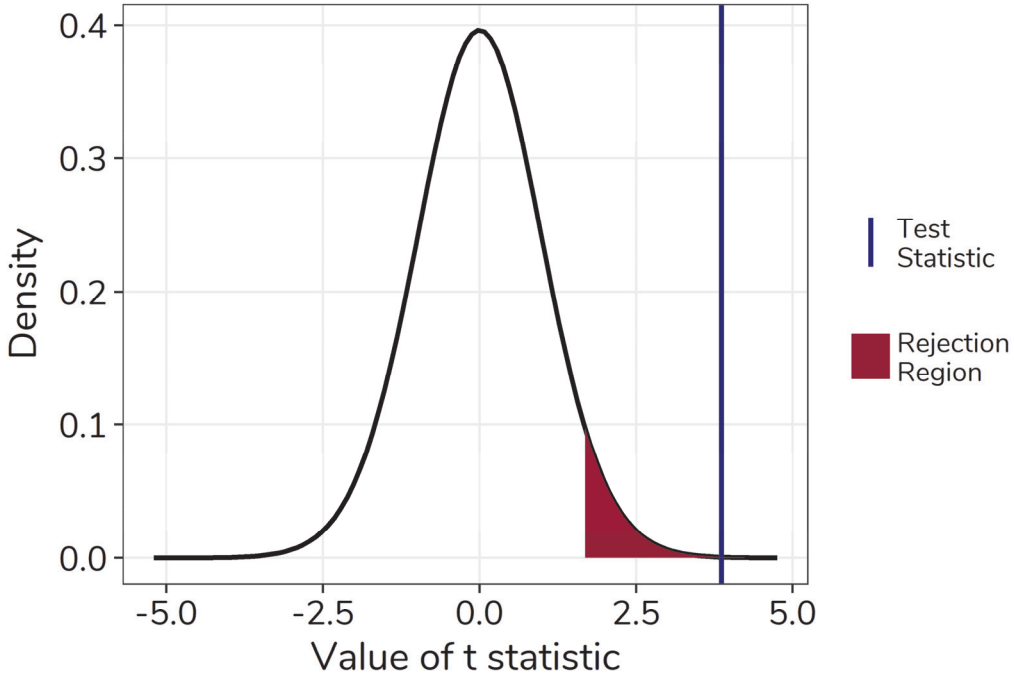
Two-sample t-test: inference

We pool our sample variances together into one estimate

We construct the t statistic using:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_p^2 \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$t = 3.86$$



We obtain t(38) = 3.86, p < .01. At $\alpha = .05$, we reject the null hypothesis.

Conclusion: The new helmet significantly reduces discomfort.

Analogous tests exist for proportions

In testing, we are frequently interested in estimating the underlying probability of an event

… what is the probability of a successful missile launch?
… what is the probability of a successful message transmission?
… what is the probability of a successful torpedo hit?

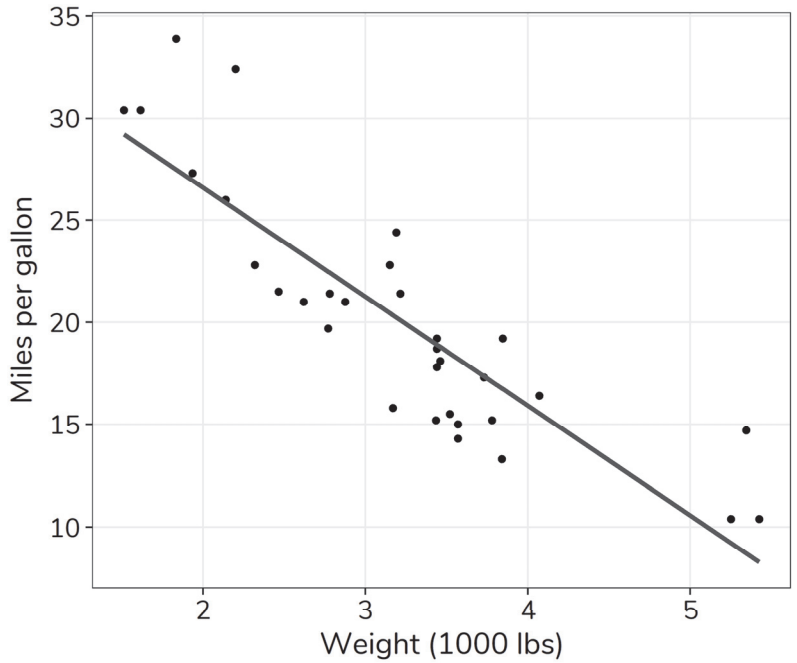
Though we generally prefer continuous metrics as response variables, sometimes we can't avoid using a 0/1 outcome variable.

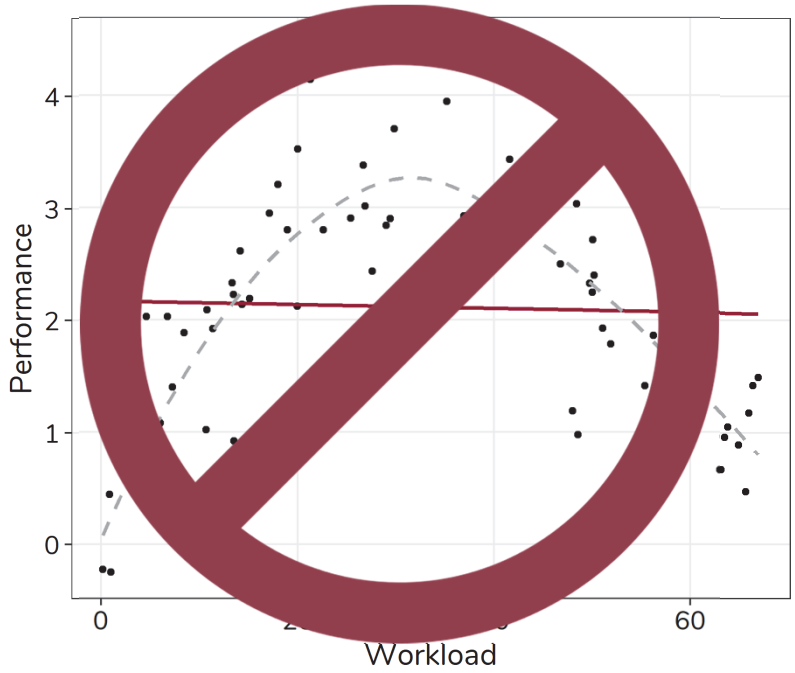
Correlation

Correlation

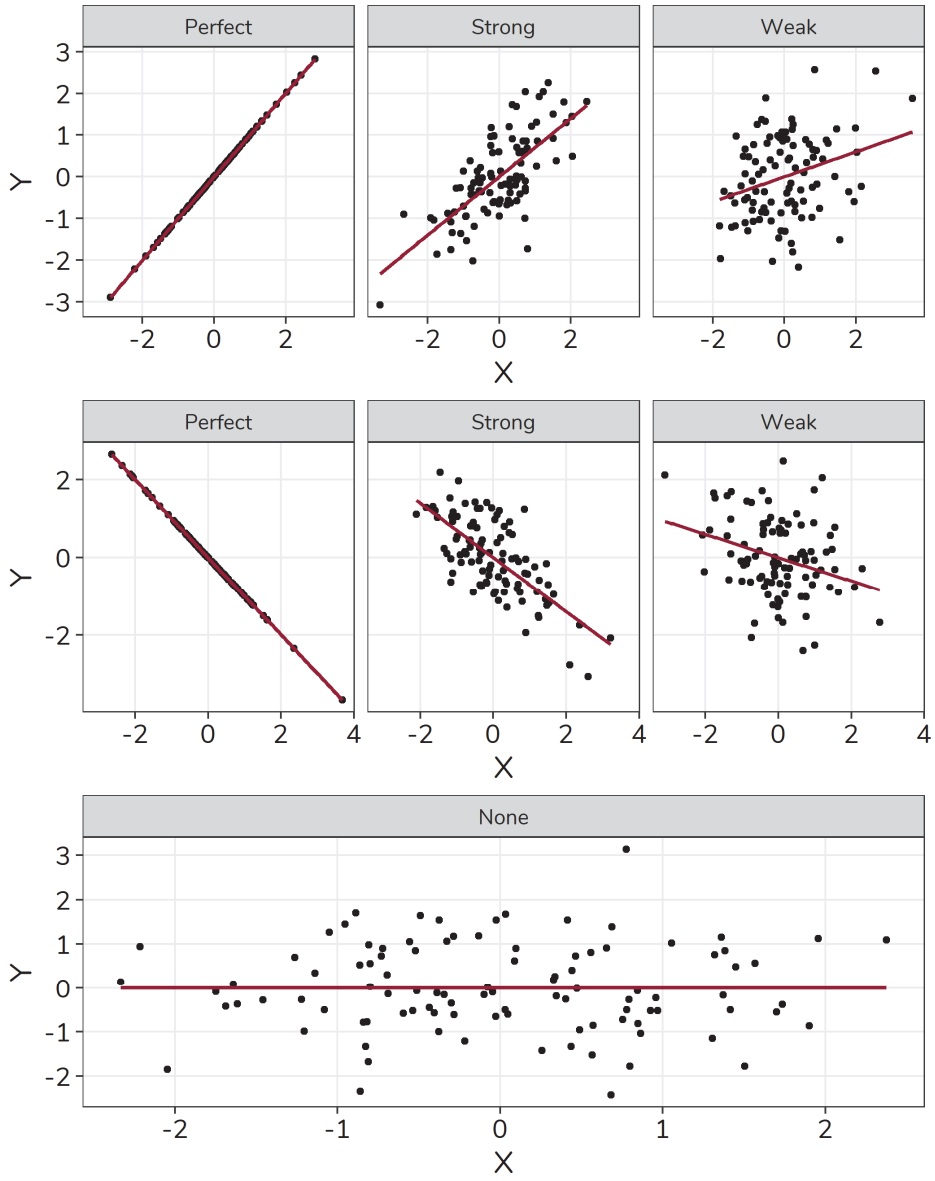
A correlation measures the strength of *linear* relationship between two variables.

Its value ranges from -1 to 1, where absolute values closer to 1 reflect a stronger linear relationship





Visualizing correlations



General linear model

General Linear Model

General: widely applicable to estimating and testing hypotheses about parameters

Linear: the function is a linear function of the parameters

Model: it provides a description of the relationship between one response and one or more predictor variables

The general linear model

In its simplest form, the GLM can be expressed as:

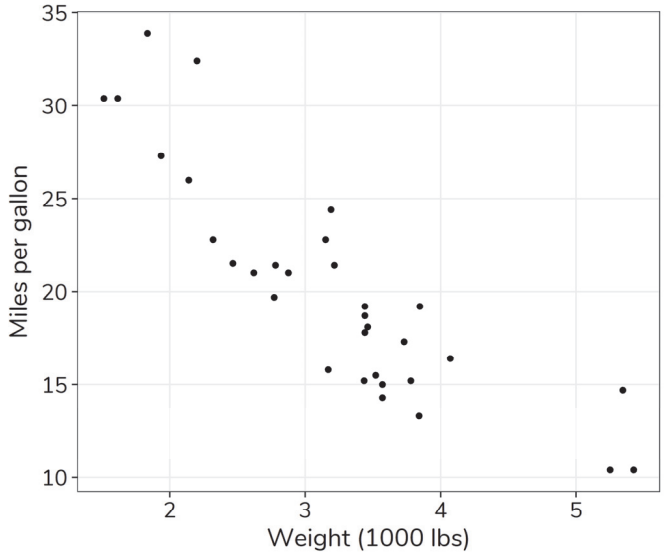
$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

The above equation expresses a model in which we can express an individual's outcome, y_i, as a function of an intercept, β_0, a coefficient, β_1, a predictor variable, x_i, and random error, ϵ_i

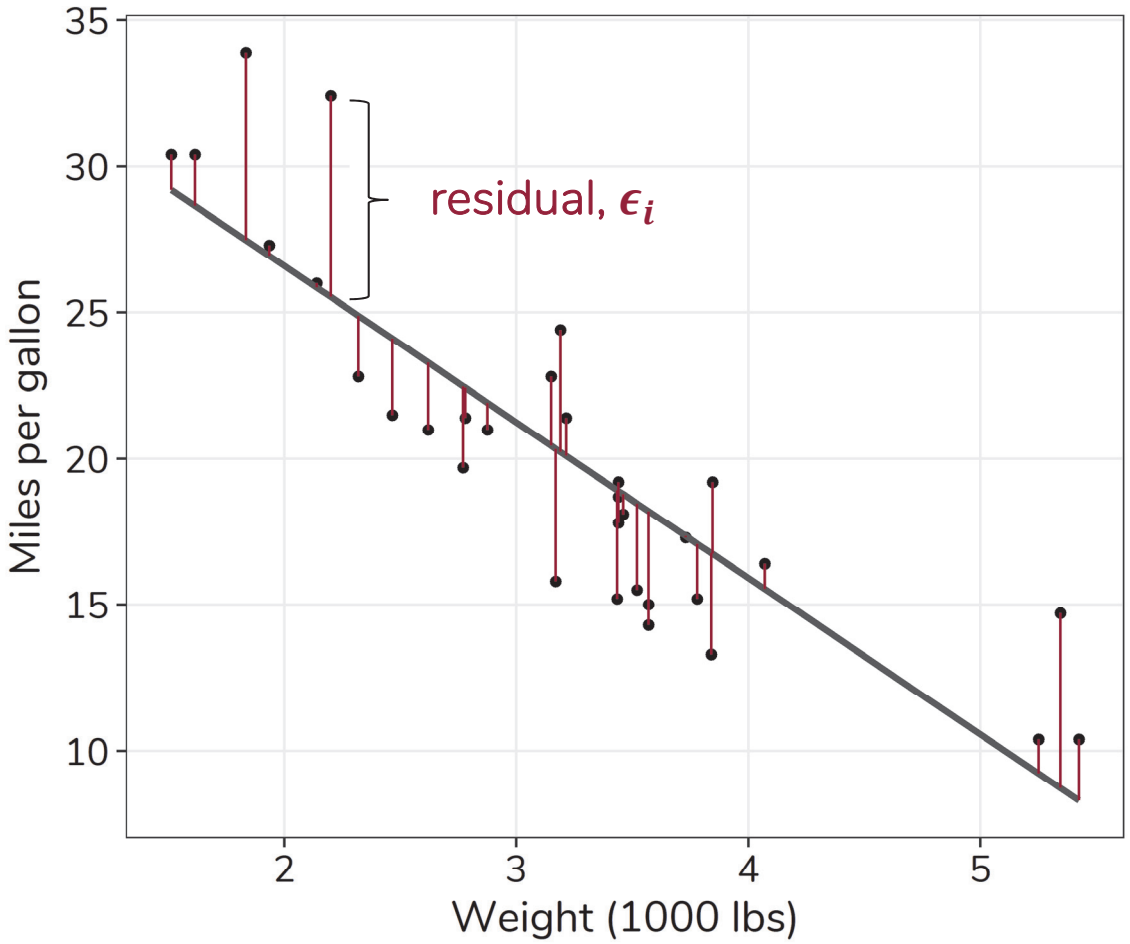
Simple linear regression

For one response variable and one predictor variable, the phrase "simple linear regression" is often used

If we have two continuous variables, we can plot a line of best fit to characterize the relationship between our predictor and our outcome variable



Ordinary least squares – line of best fit



$$y_i = \beta_0 + x_i\beta_1 + \epsilon_i$$

The general linear model – beyond simple linear regression

$$\underset{N\times 1}{\boldsymbol{y}} = \underset{N\times p}{\boldsymbol{X}}\underset{p\times 1}{\boldsymbol{\beta}} + \underset{N\times 1}{\boldsymbol{\epsilon}}$$

$\boldsymbol{y}$ is a vector of N observations on our response variable

$\boldsymbol{X}$ is an $N \times p$ matrix of observations on p predictor variables

$\boldsymbol{\beta}$ is a $p \times 1$ matrix of unknown parameters

$\boldsymbol{\epsilon}$ is a vector of N subject-specific deviations from the expected value

General linear model: Assumptions

- ✓ **Homoskedasticity** – constant variance about any value of the regression function (e.g., deviation for each unit has the same variance)

- ✓ **Independence** – errors are statistically independent
  - If violated, check sampling and/or design

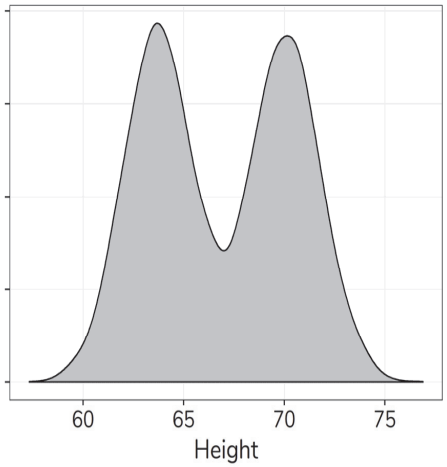
- ✓ **Linearity** – the expected values (means) are linear functions of the parameters
  - Linear: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ 
  - Linear: $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$ 
  - Not linear: $y_i = \beta_0 + x_i^{\beta_1} + \epsilon_i$ 

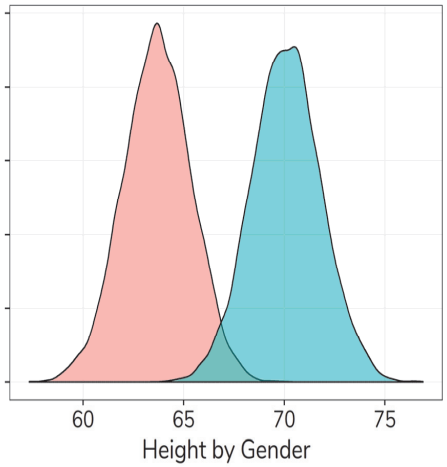
- ✓ **Existence** – finite mean and variance

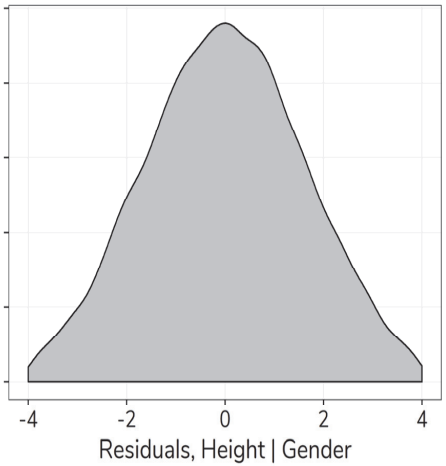
Important note about assumptions

For least squares estimation to be valid, we only need these four assumptions.

However, to be able to perform inference (e.g., hypothesis tests), we must make one final assumption – **Gaussian errors,** $\epsilon_i \sim N(0, \sigma^2)$







These five assumptions (H-I-L-E-Gauss) allow for estimation and inference in regression

General linear model – example

Scenario: We have a notional system with a new user interface designed to be simpler and clearer to operators. We want to know if the new system improves reaction time compared to the old system. We also ask operators to rate their experience on a 1-7 scale.

In multiple regression, we have a hypothesis test for each effect

Regression equation:

$$Reaction\ time = \beta_0 + \beta_1 System + \beta_2 Experience + \beta_3 (System * Experience) + \epsilon_i$$

main effect | main effect | interaction effect

General linear model– example, continued

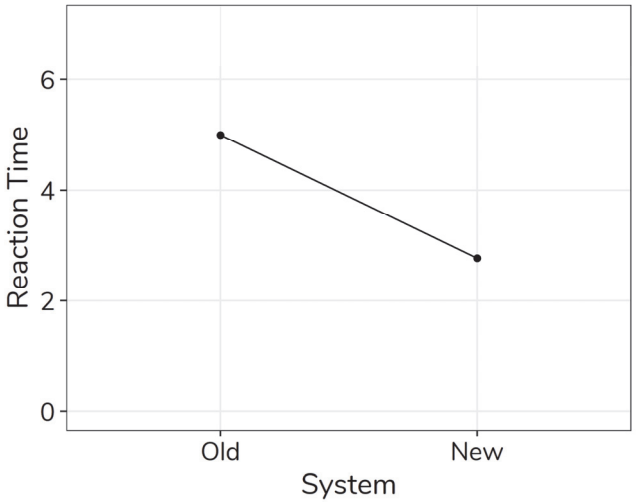
The linear model allows us to test multiple hypotheses simultaneously. With our setup, we have three questions:

1) Does the new system improve reaction time?
2) Does operator experience matter?
3) Does the effect of system (new vs. legacy) *depend* on operator experience?

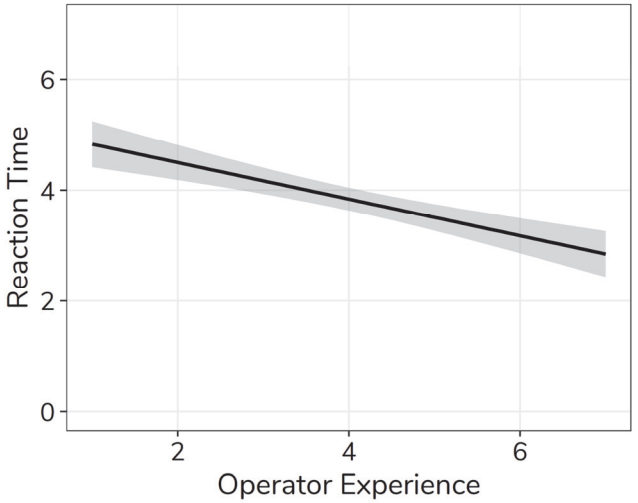
$$Reaction\ time = \beta_0 + \beta_1 System + \beta_2 Experience + \beta_3 (System * Experience) + \epsilon_i$$

Understanding effects in multiple regression – interpreting main effects

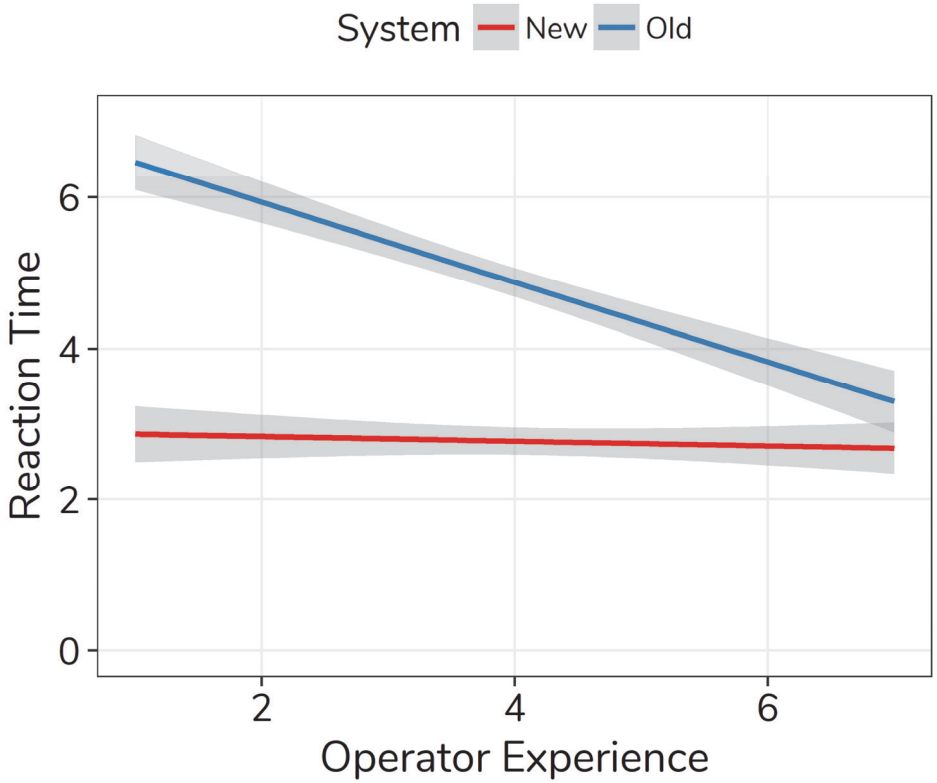
Effect of system: the reaction time is faster for the new system than the old system.



Effect of operator experience: more experience is associated with faster reaction time.



Understanding effects in multiple regression – interpreting interaction effects



Conclusion: The new system improves reaction time, particularly for individuals with low experience

Generalized linear model

Generalized linear model (GLM)

The general**ized** linear model allows for the linear model to be related to our response variable via a **link function**

Therefore, we can extend the linear modeling framework to response variables that are not normally distributed

Common examples include:

- Poisson regression (outcome is a count)
- Logistic regression (outcome is binary)

The generalized linear modeling framework opens up a large number of possibilities

Table 15.1 Some Common Link Functions and Their Inverses

Link	$\eta_i = g(\mu_i)$	$\mu_i = g^{-1}(\eta_i)$
Identity	μ_i	η_i
Log	$\log_e \mu_i$	e^{η_i}
Inverse	μ_i^{-1}	η_i^{-1}
Inverse-square	μ_i^{-2}	$\eta_i^{-1/2}$
Square-root	$\sqrt{\mu_i}$	η_i^2
Logit	$\log_e \frac{\mu_i}{1-\mu_i}$	$\frac{1}{1+e^{-\eta_i}}$
Probit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$
Log-log	$-\log_e[-\log_e(\mu_i)]$	$\exp[-\exp(-\eta_i)]$
Complementary log-log	$\log_e[-\log_e(1-\mu_i)]$	$1-\exp[-\exp(\eta_i)]$

NOTE: μ_i is the expected value of the response; η_i is the linear predictor; and $\Phi(\cdot)$ is the cumulative distribution function of the standard-normal distribution.

Generalized linear model – the link function

We can express a generalized linear model using a linear predictor,

$$\eta_i = \beta_0 + \beta_1 x_{1i}$$

with a link function $g(\cdot)$ that describes the relationship between the mean $E(y_i) = \mu_i$ and the linear predictor,

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{1i}$$

We can express the normal GLM in this way using

$$\eta_i = \beta_0 + \beta_1 x_{1i}$$

with an identity link function,

$$g(\mu_i) = \mu_i$$

Logistic regression

- Two outcomes (success or failure)
- Events are independent
- Constant underlying probability of success

If our outcome is binary,

$$y_i \sim Binomial(N_i, p_i)$$

then our linear predictor can be expressed as:

$$\eta_i = \beta_0 + \beta_1 x_{1i}$$

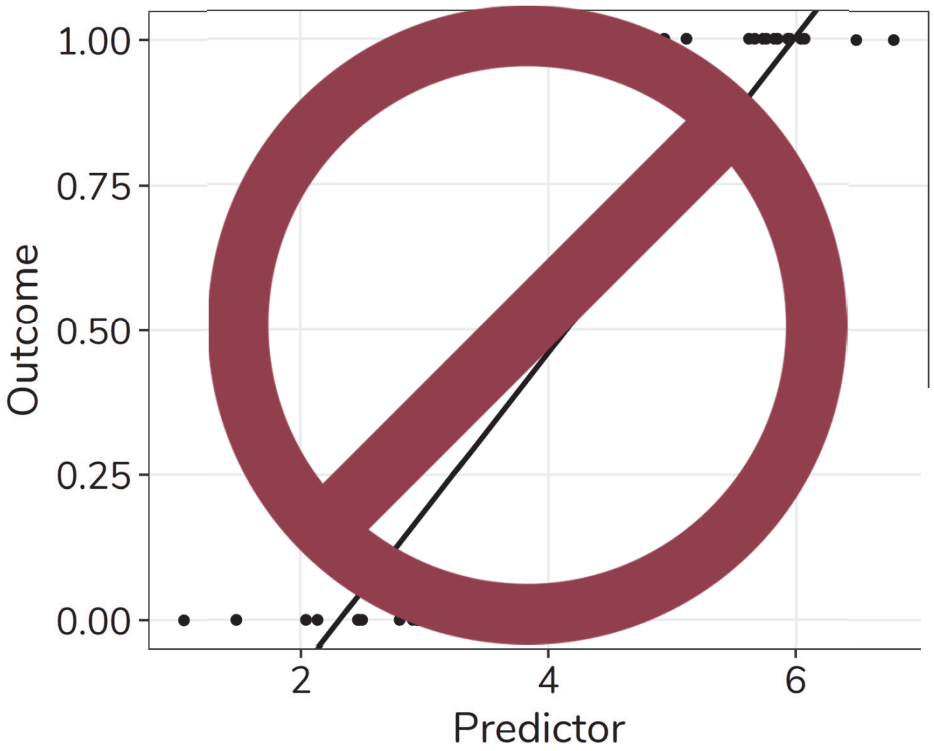
with a logit link function,

$$g(\mu_i) = logit(\mu_i) = \log\left(\frac{\mu_i}{1-\mu_i}\right),$$

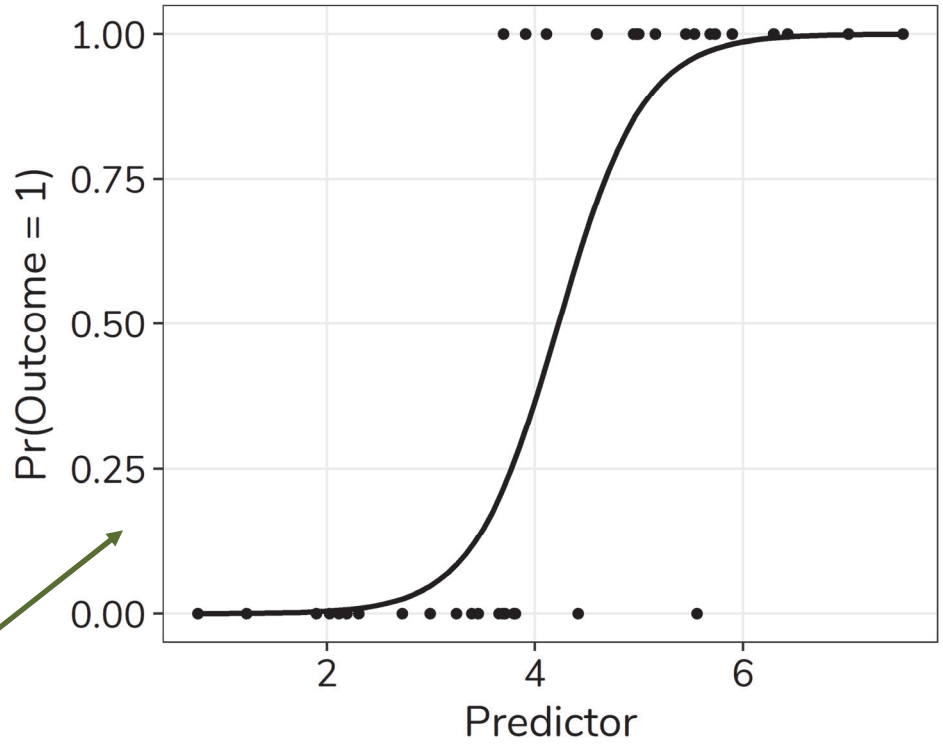
yielding the regression equation,

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \ldots \beta_k x_k$$

Visualizing the logit link function



Don't try to fit a straight line through 0/1 data!



Instead use a logit link function to map our linear predictor onto the binary response

Logistic regression – example

For instances where we have a 0/1 outcome (hit/miss; success/fail; detect/non-detect), we use logistic regression to understand variation in our response variable as a function of our test factors

Altitude	Variant	Detect (1 Yes; 0 = No)
Low	A	1
High	A	1
Low	B	1
High	B	0
…	…	…

Logistic regression – example, continued

We are interested in three effects:

1) The main effect of altitude
2) The main effect of variant
3) The interaction between altitude and variant

From this output table, we see that only altitude significantly predicts detection.

term	estimate	std.error	statistic	p.value	Odds Ratio (OR)	OR 2.5 %	OR 97.5 %
(Intercept)	-1.012	0.413	-2.450	0.014	0.364	0.152	0.785
AltitudeLow	2.621	0.641	4.091	**0.000**	13.750	4.195	53.155
VariantB	-0.178	0.597	-0.298	0.766	0.837	0.253	2.714
AltitudeLow:VariantB	0.178	0.915	0.195	0.846	1.195	0.196	7.333

Because the estimates are currently *log odds*, we can exponentiate them to compute an odds ratio. The odds ratio for 'AltitudeLow' tells us that the odds of detection for low altitude is 13.75 times higher than for high altitude.

The general linear model represents a powerful framework for evaluating our research hypotheses, and encompasses a variety of statistical tests, including t-tests, ANOVA, ANCOVA, and multiple regression.

The generalized linear model (GLM) is an extension that allows the linear model to be related to an outcome variable via a link function, and includes logistic regression, Poisson regression, and multinomial regression (among others).

Outline of boot camp

- Summarizing and simplifying data
- Point and interval estimation
- Foundations of statistical inference
- The process of hypothesis testing
- Common statistical tests
- **A few closing tips**

- ✓ Make sure the test you select reflects your research question
- ✓ Follow good practices of data visualization
- ✓ Carefully consider outliers. Don’t just delete!
- ✓ Remember that good statistical analysis depends on good data collection

Matching the test to the research question

When selecting the correct test, there are several important questions to consider

Are you looking for…

… a difference?

- a difference in means?
- a difference in medians?
- a difference in variances?

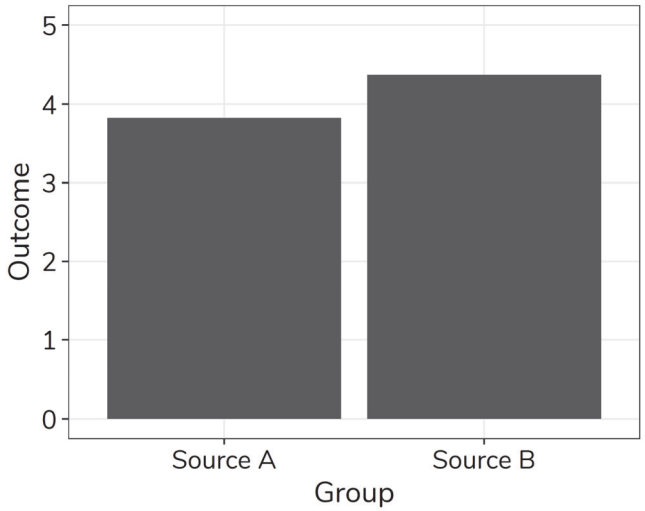
… a relationship?

- a linear relationship?
- a nonlinear relationship?
- an interaction?

Don’t be afraid to perform multiple tests!

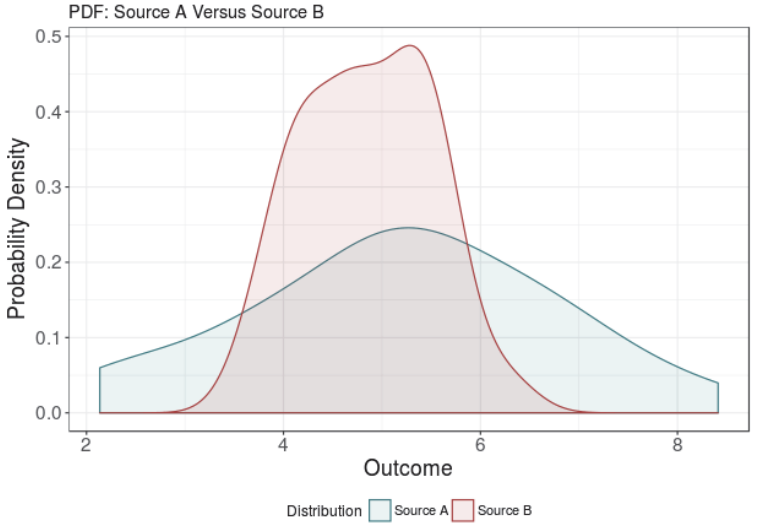
“I have a continuous outcome variable and I want to see if there is a difference between source A and source B.”

The means of the two sources are not significantly different from each other….



t = 1.13, p = .26

….But the variances are!

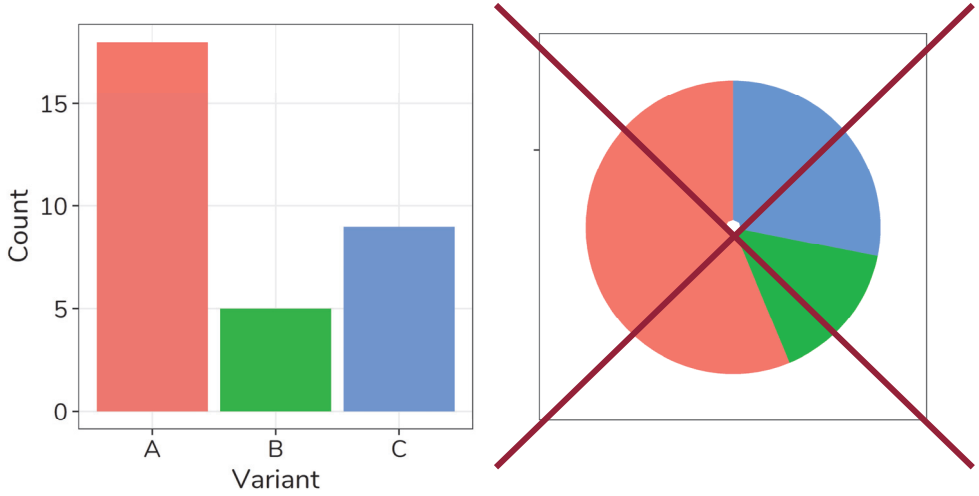


D = 0.33, p = .07

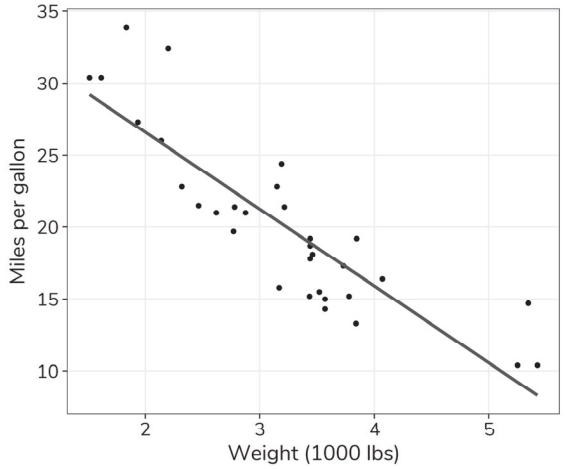
Data visualization

Meaningful data visualization

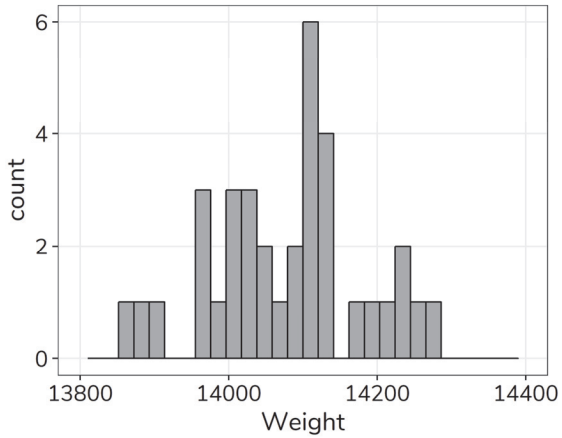
One discrete variable

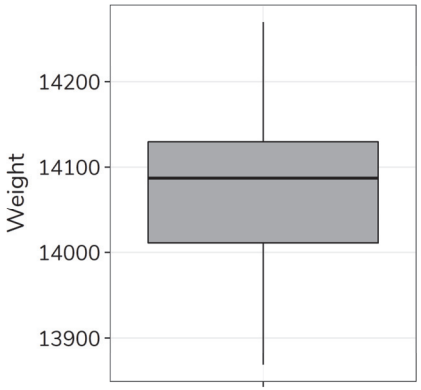


Two continuous variables – relationship

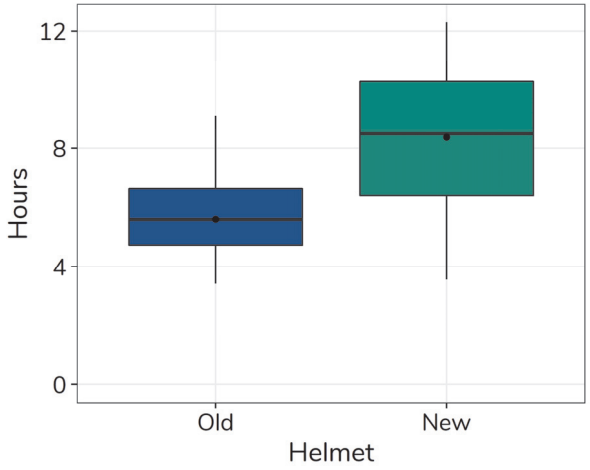


One continuous variable



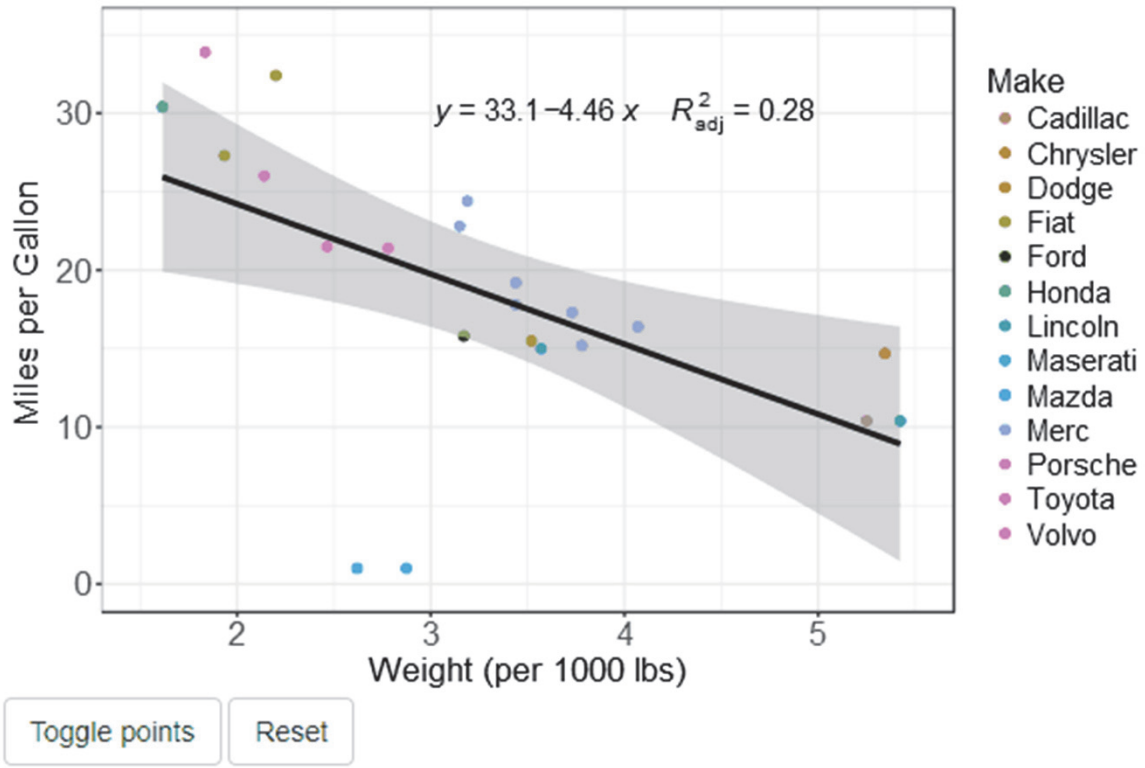


Two continuous variables – difference



Outliers

Carefully consider outliers and don't exclude valid data



With a small sample size, even a few data points can heavily influence our results!

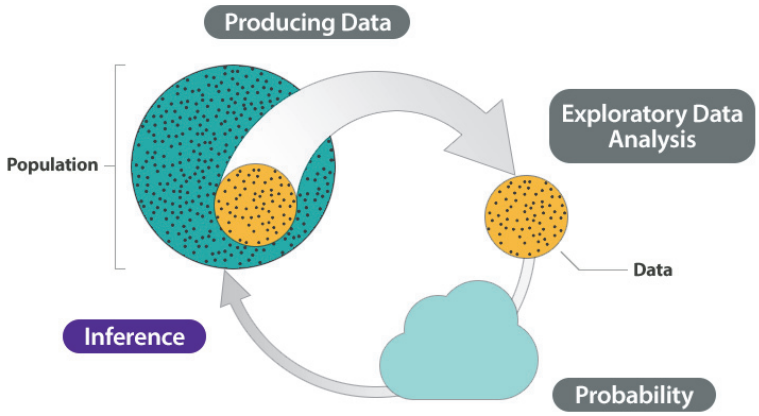
Collecting data to support your analysis

The type of analysis you want to perform drives the type and amount of data you should collect

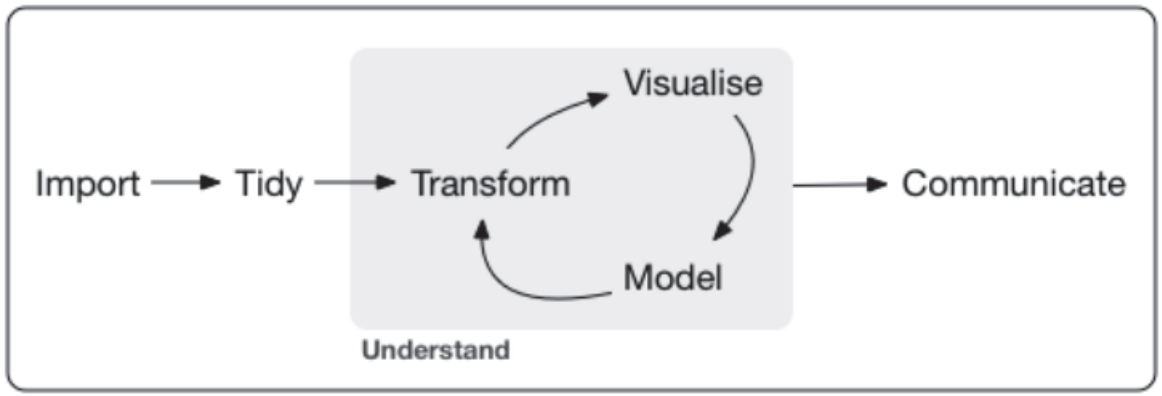
Design of Experiments (DOE) principles are fundamental to good data collection

Data Collection

Inferential Statistics



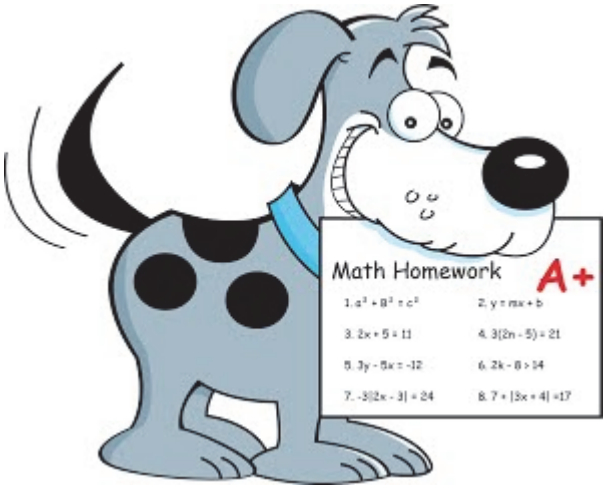
Data Science Process



Thank you!

Contact Info:

Dr. Kelly Avery – kavery@ida.org



Resources:

- https://testscience.org/
- D. S. Moore, “The Basic Practice of Statistics.” Palgrave MacMillon, 2010.
- D. C. Montgomery, “Design and Analysis of Experiments.” John Wiley & Sons, 1990.
- H. Wickham and G. Grolemund, “R for Data Science.” https://r4ds.had.co.nz/


IDA

Backups

Common distributions

Probability distributions

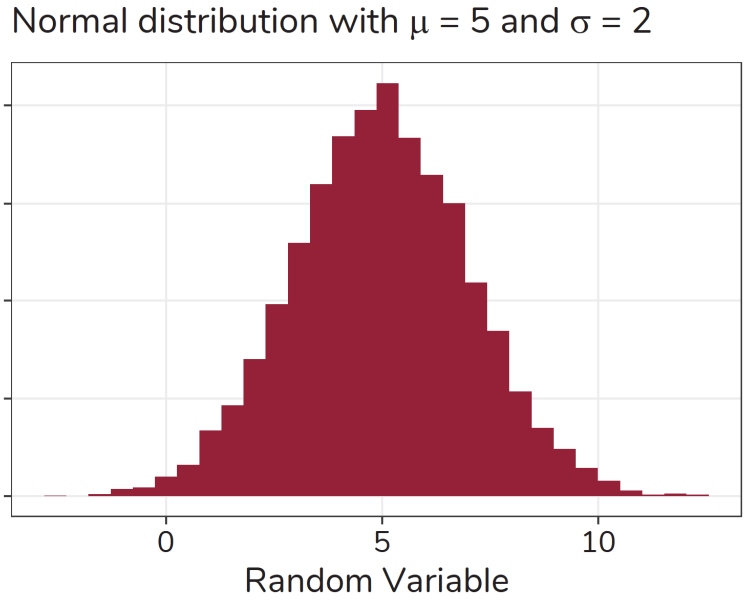
A probability distribution describes a random variable, X. We typically think of distributions being **continuous** or **discrete**.

- A random variable X has a **continuous** distribution if the range of X is infinite and uncountable
  - e.g., normal, lognormal

- A random variable X has a **discrete** distribution if the range of X is countable
  - e.g., binomial, Poisson

Normal distribution

PDF	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
Mean	μ
Variance	σ^2

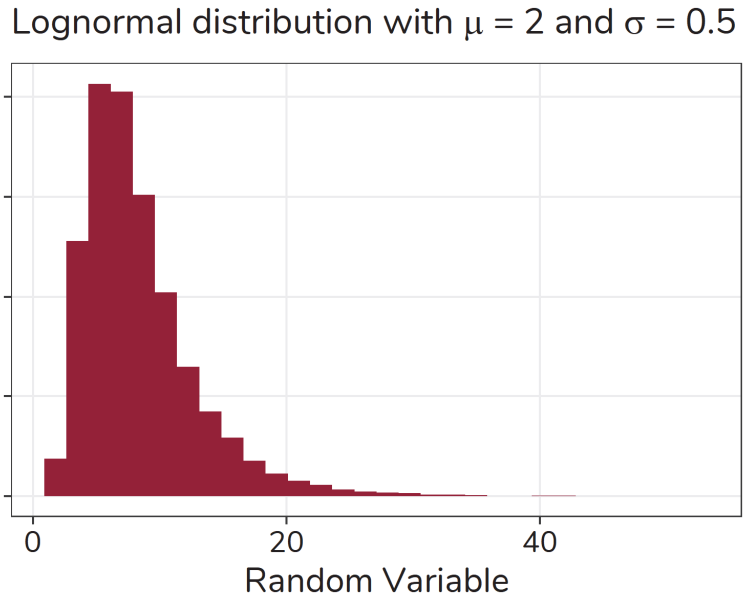


Common applications:

- Performance

Lognormal distribution

PDF	$\frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\log x-\mu)^2}{2\sigma^2}}$
Mean	$e^{\mu+\frac{\sigma^2}{2}}$
Variance	$e^{2\mu+\sigma^2}[e^{\sigma^2}-1]$

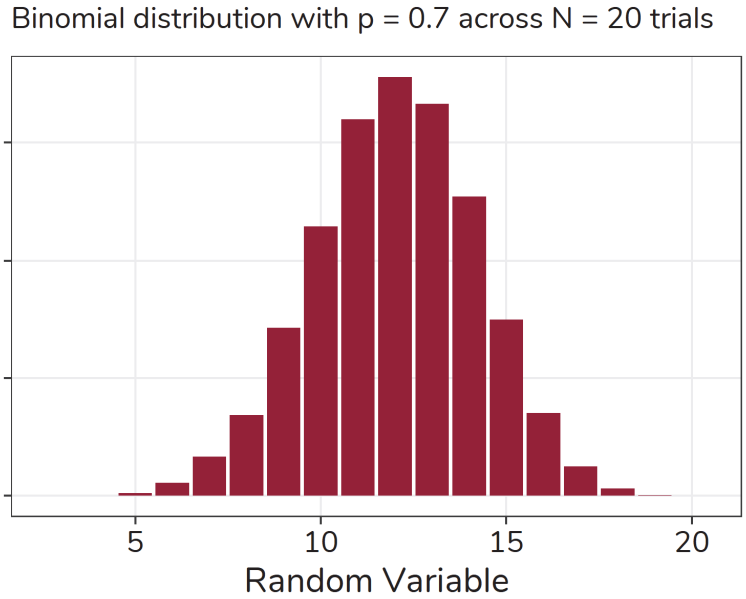


Common applications:

- Time to detect
- Miss distance

Binomial distribution

PMF	$\binom{n}{x} p^x (1-p)^{n-x}$
Mean	np
Variance	$np(1-p)$

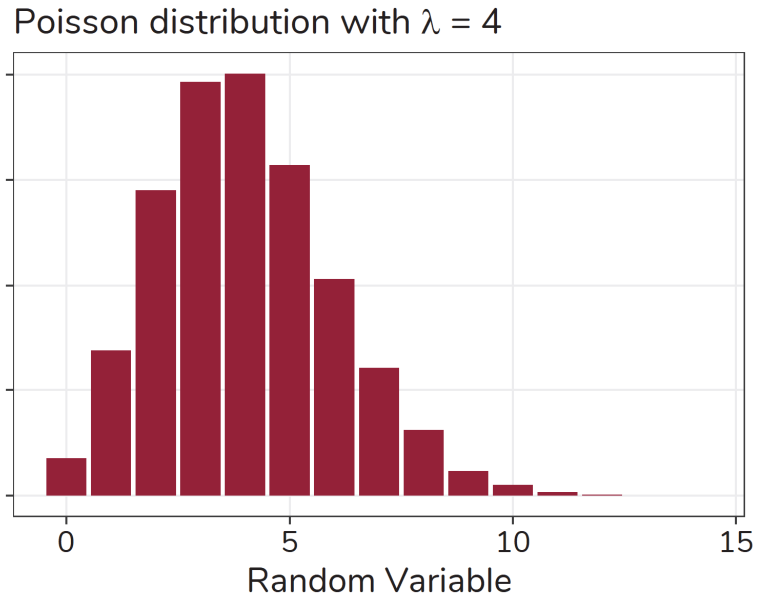


Common applications:

- Hit/miss
- Success/fail
- Detect/non-detect

Poisson distribution

PMF	$\frac{e^{-\lambda}\lambda^{x}}{x!}$
Mean	λ
Variance	λ



Common applications:

- Count data

Test of one proportion

Binomial distribution

For the binomial distribution, the following criteria must be met:

- ✓ Each event yields one of two outcomes – success or failure
- ✓ Each event is independent (memory-less, like a coin flip)
- ✓ The underlying probability of success, p_0, is constant across events

Performing inference on proportions

We can perform inference by comparing our observed sample proportion to the sampling distribution

The mean of the sampling distribution is $\hat{p} = \frac{X}{n}$

- X: Number of successes
- n: Number of trials

And the standard deviation is $\sqrt{\frac{p_0(1-p_0)}{n}}$

We may construct our test statistic using $\mathrm{z} = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$

Test of one proportion

Scenario: Suppose we have a system for transmitting messages. We want to know if this new system is worse than P(Transmission) = .80, our population estimate for the legacy system. During the OT of this new system, we observe that 38 of 50 messages transmitted successfully.

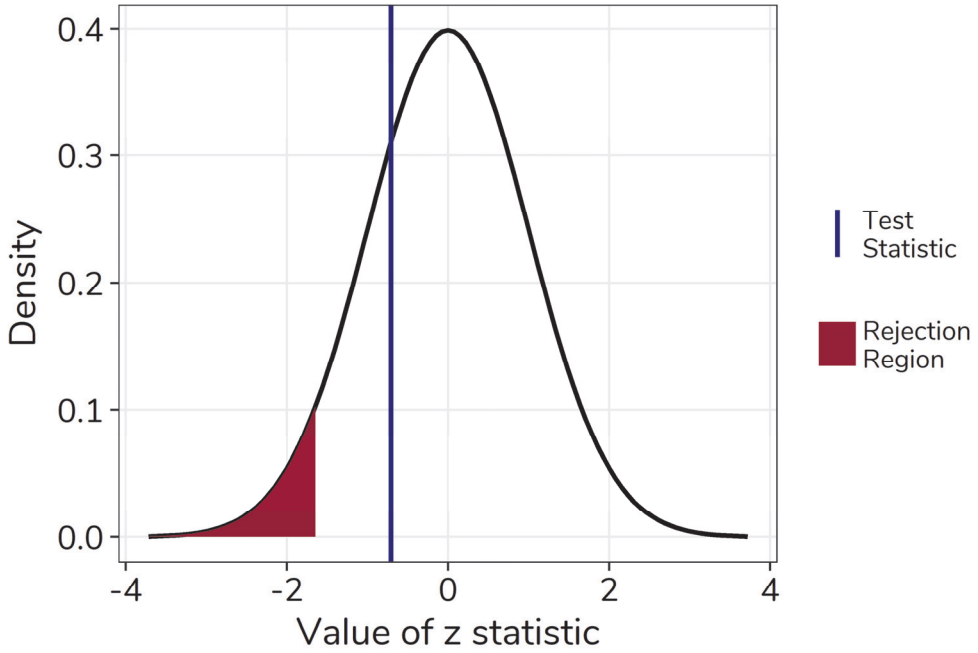
Research question: Is the new system worse than the old system?

Relevant quantities: $\mathrm{X} = 38; n = 50;\ \hat{p}\ =\ .76;\ p_0 =\ .80$

Compute statistic and determine probability

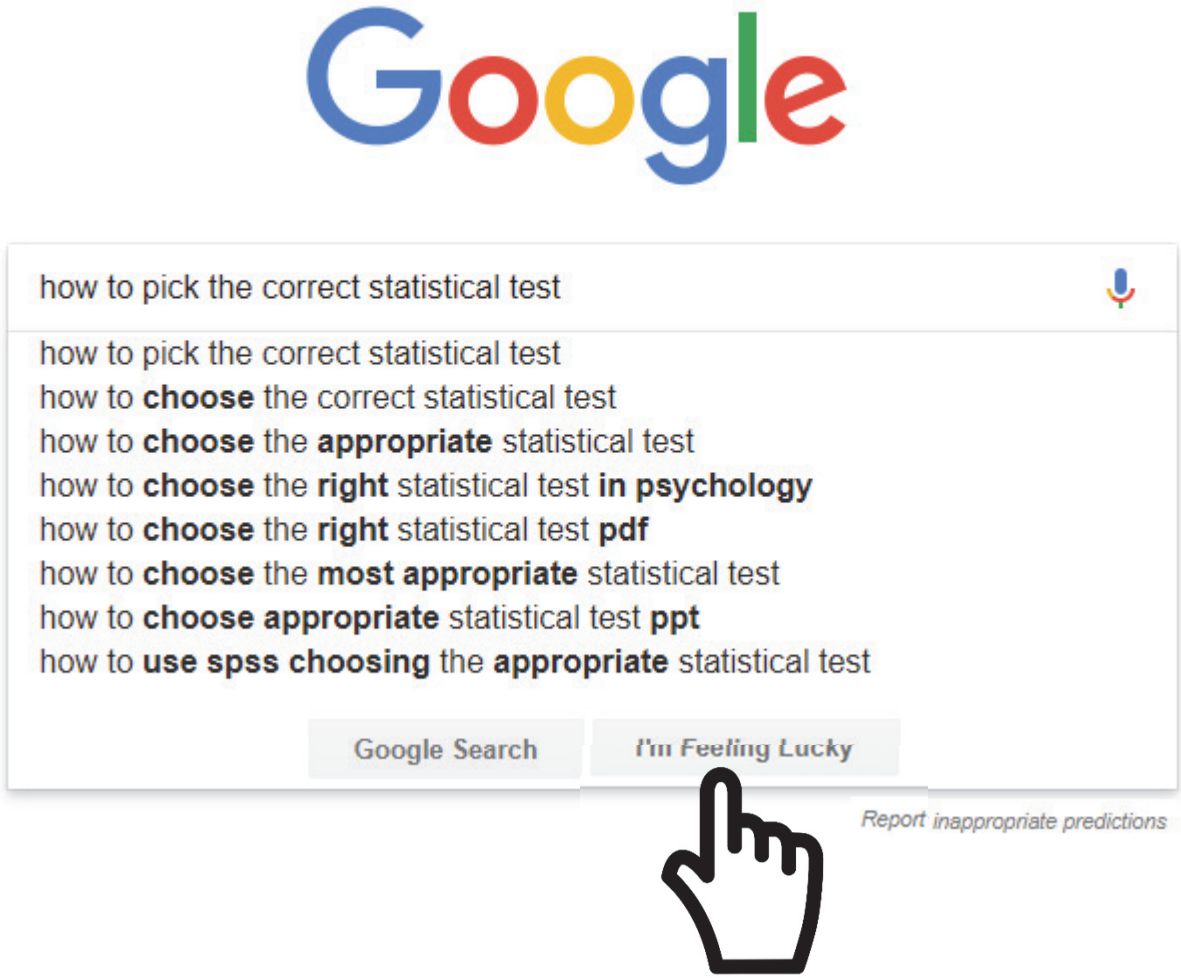
$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

$$z = -.71$$

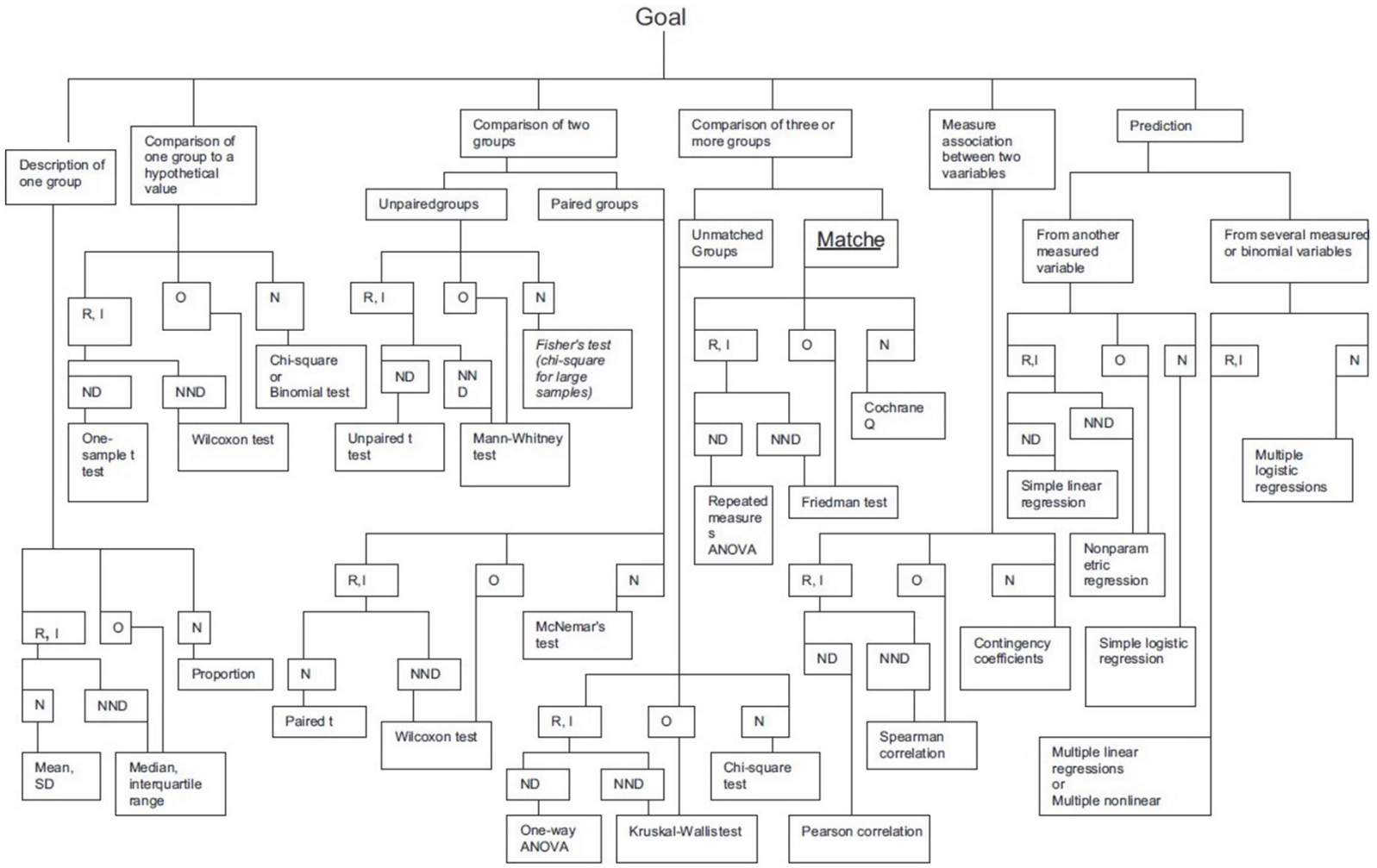


Conclusion: We cannot conclude that the new system is worse than the old system.

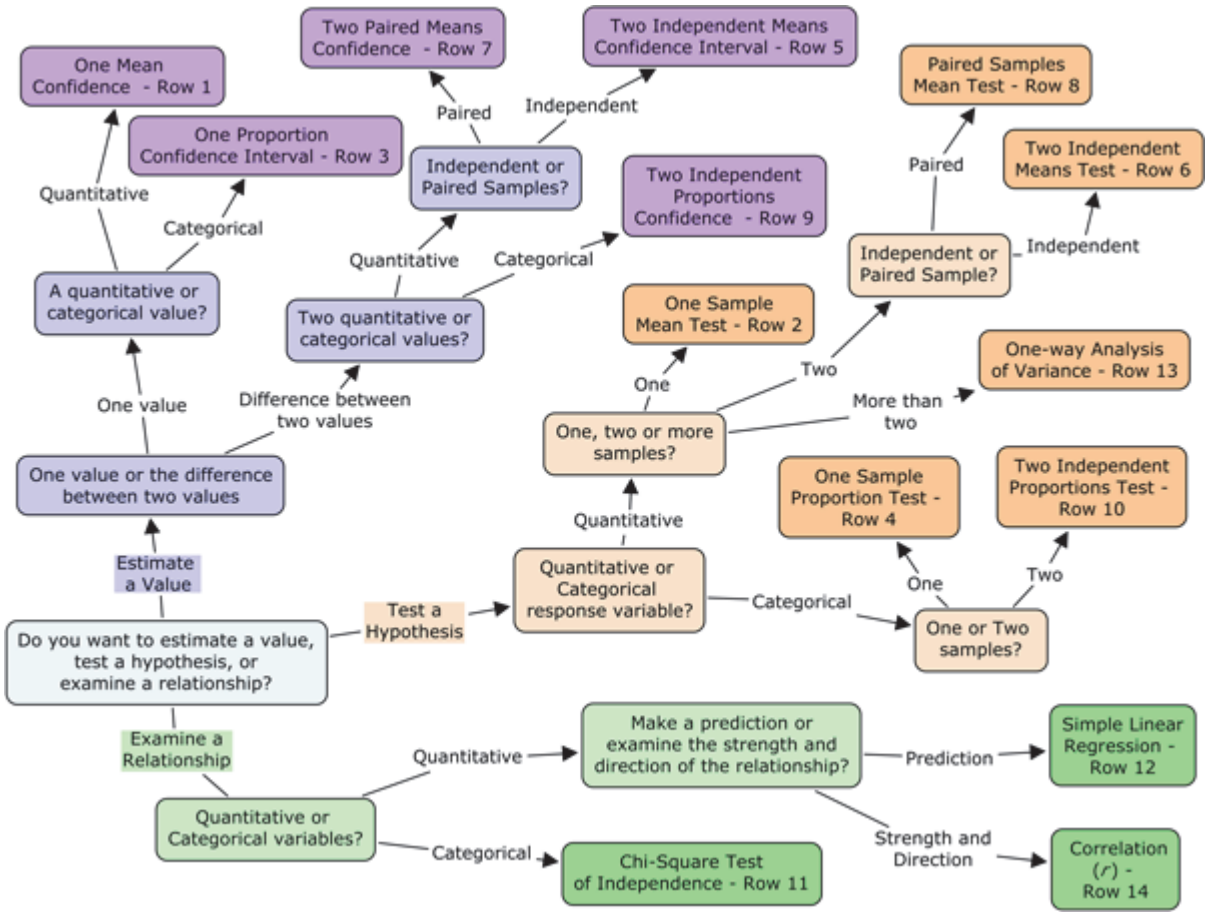
Matching the test to the research question


Google
how to pick the correct statistical test
how to pick the correct statistical test
how to choose the correct statistical test
how to choose the appropriate statistical test
how to choose the right statistical test in psychology
how to choose the right statistical test pdf
how to choose the most appropriate statistical test
how to choose appropriate statistical test ppt
how to use spss choosing the appropriate statistical test
Google Search
I'm Feeling Lucky
Report inappropriate predictions

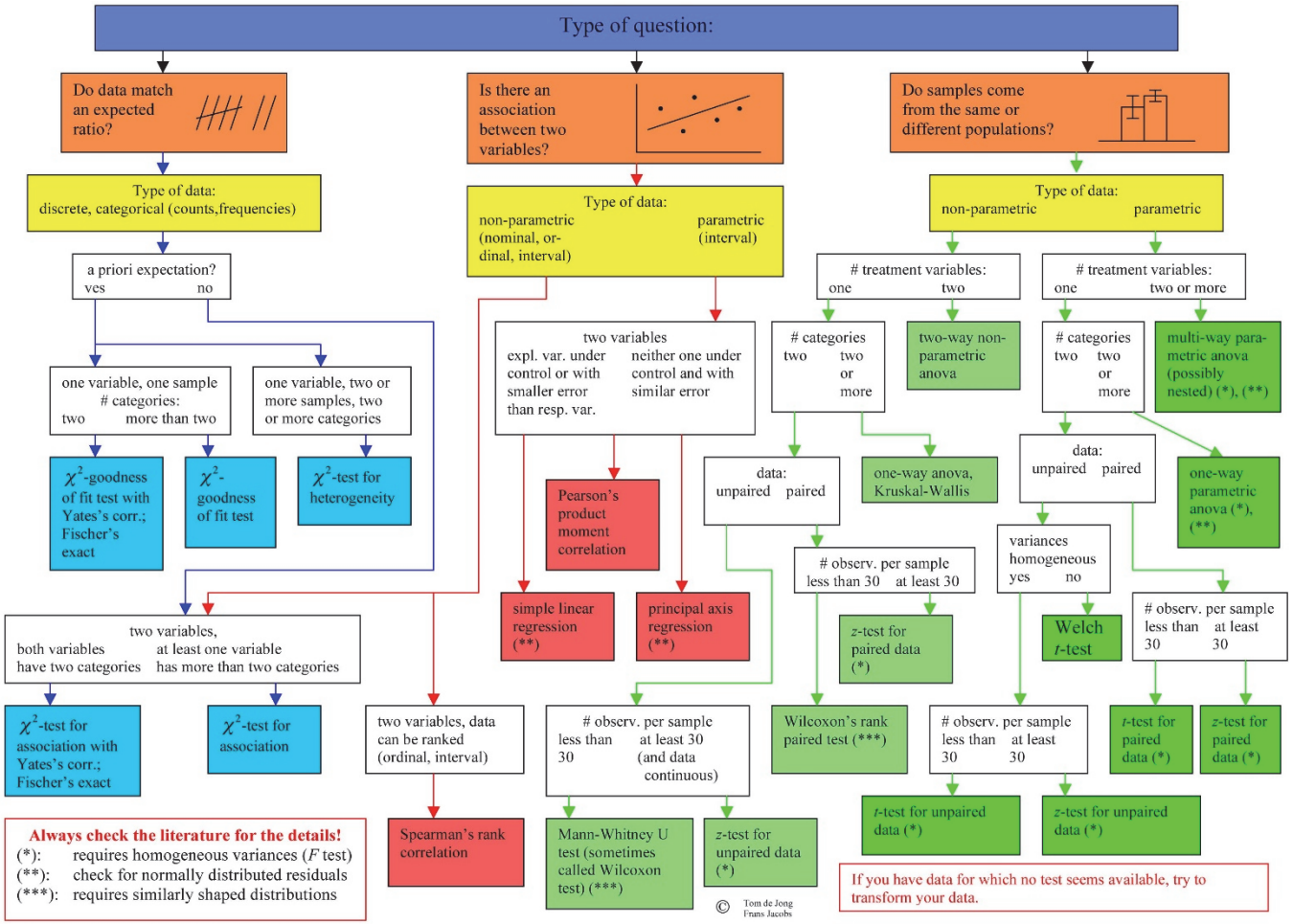
Selecting the correct test



Selecting the correct test



Selecting the correct test



These flowcharts can be useful (I'll admit).

However, adhering strictly to their use might derail you from thinking critically about the **question** you were actually trying to answer.

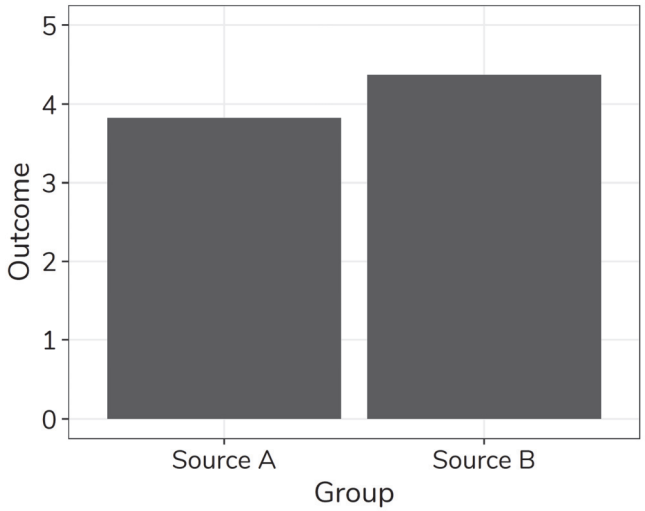
Flowchart reasoning gone wrong

"I have a continuous outcome variable and I want to see if there is a difference between source A and source B."

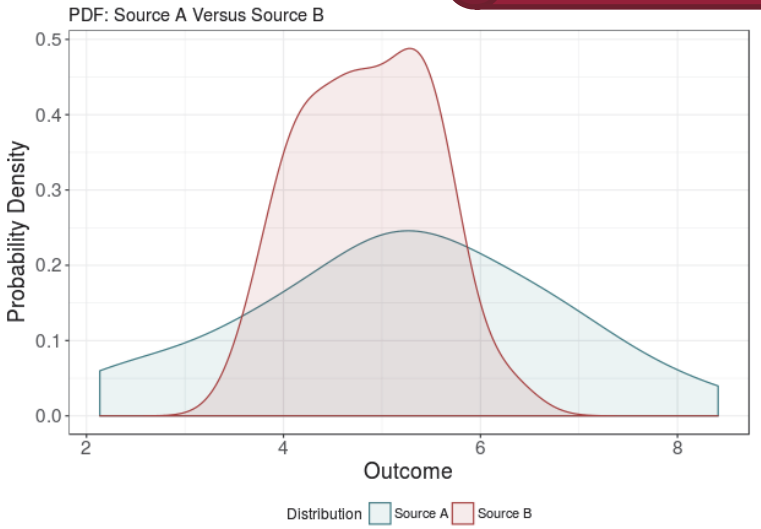
one journey through a flowchart later…

"I'll do a two-sample t-test!"

If you had also cared about a difference in **spread**, you would have missed it with the t-test!



t = 1.13, p = .26



D = 0.33, p = .07

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS**.

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
XX-03-2019	OED Draft	

4. TITLE ANDSUBTITLE

Statistics Bootcamp DATAWorks 2019

5a. CONTRACT NUMBER
HQ0034-14-D-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Kelly M. Avery (OED); Stephanie T. Lane (OED);

5d. PROJECT NUMBER
BD-9-2299(90) & C9087

5e. TASK NUMBER
22990 & C9087

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

8. PERFORMING ORGANIZATION REPORT NUMBER

D-10565-NS

H 2019-000146

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Director, Operational Test and Evaluation
1700 Defense Pentagon
Washington, DC 20301-1700

10. SPONSOR/MONITOR'S ACRONYM(S)
DOT&E

11. SPONSOR/MONITOR'S REPORT NUMBER

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release. Distribution is unlimited.

13. SUPPLEMENTARY NOTES
Project Leaders: Wojton, Heather M. and Rebecca M. Medlin

14. ABSTRACT

In the test community, we frequently use statistics to extract meaning from data. These inferences may be drawn with respect to topics ranging from system performance to human factors. In this mini-tutorial, we will begin by discussing the use of descriptive and inferential statistics, before exploring the basics of interval estimation and hypothesis testing. We will introduce common statistical techniques and when to apply them, and conclude with a brief discussion of how to present your statistical findings graphically for maximum impact.

15. SUBJECT TERMS

Operational Testing; Hypothesis Testing

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Heather Wojton (OED)
Unclassified	Unclassified	Unclassified	Unlimited	133	**19b. TELEPHONE NUMBER** *(include area code)* (703) 845-6811