Statistics Boot Camp

Dr. Kelly Avery
Institute for Defense Analyses
DATAWorks 2019

April 10, 2019
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

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

did someone say
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

<table>
<thead>
<tr>
<th>Scale of measurement</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nominal</td>
<td>Named categories</td>
<td>Wheel Type (Tracked, Wheels)</td>
</tr>
<tr>
<td>Ordinal</td>
<td>Ordered categories</td>
<td>Place in a race (First, Second, Third)</td>
</tr>
<tr>
<td>Interval</td>
<td>Ordered values, equidistant</td>
<td>Temperature (Fahrenheit, Celsius)</td>
</tr>
<tr>
<td>Ratio</td>
<td>Ordered values, equidistant, real zero</td>
<td>Distance (meters)</td>
</tr>
</tbody>
</table>

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 50\textsuperscript{th} 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:

![Bar chart showing survey 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

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

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

Population

\[ \mu \]
Parameter

Sample

Random sample

\[ \bar{X} \]
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, $\mu$, 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 ± our margin of error:

\[
\bar{X} \pm t_{\alpha} \frac{S_X}{\sqrt{N}} \quad \bar{X} \pm z_{\alpha} \frac{S_X}{\sqrt{N}} \quad \hat{p} \pm z_{\alpha} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}
\]

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 $\Rightarrow$ wider confidence intervals

Smaller sample size $\Rightarrow$ wider confidence intervals

Larger confidence level $\Rightarrow$ 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 $\mu$ and standard deviation $\sigma$, the distribution of sample means for sample size $n$ will have a mean of $\mu$ 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

<table>
<thead>
<tr>
<th>True Combat Performance</th>
<th>Observed Test Result</th>
<th>Quantity</th>
<th>Report</th>
</tr>
</thead>
<tbody>
<tr>
<td>New system is better than legacy</td>
<td>Test concludes new system is better</td>
<td>Power (95%)</td>
<td>States new system is better than legacy</td>
</tr>
<tr>
<td></td>
<td>Test concludes same or worse</td>
<td>Risk of Incorrect Conclusion (5%)</td>
<td>States system is same/worse than legacy</td>
</tr>
<tr>
<td>New system is same or worse than legacy</td>
<td>Test concludes better</td>
<td>Risk of Incorrect Conclusion (5%)</td>
<td>States system is better than legacy</td>
</tr>
<tr>
<td></td>
<td>Test concludes new system is the same or worse</td>
<td>Confidence (95%)</td>
<td>States system is same/worse than legacy</td>
</tr>
</tbody>
</table>

### Decision Table

<table>
<thead>
<tr>
<th>Truth</th>
<th>Retain Null</th>
<th>Reject Null</th>
</tr>
</thead>
<tbody>
<tr>
<td>Null is true</td>
<td>Correct Decision $1 - \alpha$</td>
<td>Type I error $\alpha$</td>
</tr>
<tr>
<td>Null is false</td>
<td>Type II error $\beta$</td>
<td>Correct Decision $1 - \beta$</td>
</tr>
</tbody>
</table>
Thinking about errors in hypothesis testing

Type I error
(false positive)

You’re pregnant

Type II error
(false negative)

You’re not pregnant
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, $\mu = 14000$. 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.

![Histogram showing weight distribution]

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

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

We obtain $t(34) = 4.43$, $p < .01$. If we set $\alpha = .05$, 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

<table>
<thead>
<tr>
<th></th>
<th>Group 1 - Old</th>
<th>Group 2 - New</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>5.60</td>
<td>8.37</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>2.03</td>
<td>2.49</td>
</tr>
<tr>
<td>Sample Size</td>
<td>20</td>
<td>20</td>
</tr>
</tbody>
</table>

![Box plot comparing old and new helmets](image)
Two-sample t-test: inference

We construct the $t$ statistic using:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$t = 3.86$$

We pool our sample variances together into one estimate

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, \( \beta_0 \), a coefficient, \( \beta_1 \), a predictor variable, \( x_i \), and random error, \( \epsilon_i \)
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

\[ y = X\beta + \epsilon \]

\( y \) is a vector of \( N \) observations on our response variable

\( X \) is an \( N \times p \) matrix of observations on \( p \) predictor variables

\( \beta \) is a \( p \times 1 \) matrix of unknown parameters

\( \epsilon \) 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.
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.

Regression equation:

\[
\text{Reaction time} = \beta_0 + \beta_1 \text{System} + \beta_2 \text{Experience} + \beta_3 (\text{System} \times \text{Experience}) + \epsilon_i
\]
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?

\[
Reaction \ time = \beta_0 + \beta_1 System + \beta_2 Experience + \beta_3 (System \ast \ Experience) + \epsilon_i
\]
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 generalized 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 from Applied Regression Analysis and Generalized Linear Models by John Fox

<table>
<thead>
<tr>
<th>Link</th>
<th>$\eta_i = g(\mu_i)$</th>
<th>$\mu_i = g^{-1}(\eta_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identity</td>
<td>$\mu_i$</td>
<td>$\eta_i$</td>
</tr>
<tr>
<td>Log</td>
<td>$\log_e \mu_i$</td>
<td>$e^{\eta_i}$</td>
</tr>
<tr>
<td>Inverse</td>
<td>$\mu_i^{-1}$</td>
<td>$\eta_i^{-1}$</td>
</tr>
<tr>
<td>Inverse-square</td>
<td>$\mu_i^{-2}$</td>
<td>$\eta_i^{-1/2}$</td>
</tr>
<tr>
<td>Square-root</td>
<td>$\sqrt{\mu_i}$</td>
<td>$\eta_i$</td>
</tr>
<tr>
<td>Logit</td>
<td>$\log_e \frac{\mu_i}{1 - \mu_i}$</td>
<td>$\frac{1}{1 + e^{-\eta_i}}$</td>
</tr>
<tr>
<td>Probit</td>
<td>$\Phi^{-1}(\mu_i)$</td>
<td>$\Phi(\eta_i)$</td>
</tr>
<tr>
<td>Log-log</td>
<td>$-\log_e[\log_e(\mu_i)]$</td>
<td>$\exp[-\exp(-\eta_i)]$</td>
</tr>
<tr>
<td>Complementary log-log</td>
<td>$\log_e[-\log_e(1-\mu_i)]$</td>
<td>$1-\exp[-\exp(\eta_i)]$</td>
</tr>
</tbody>
</table>

NOTE: $\mu_i$ is the expected value of the response; $\eta_i$ 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

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,

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

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.

<table>
<thead>
<tr>
<th>Altitude</th>
<th>Variant</th>
<th>Detect (1 Yes; 0 = No)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>High</td>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>Low</td>
<td>B</td>
<td>1</td>
</tr>
<tr>
<td>High</td>
<td>B</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
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.

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

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).
# Common statistical tests are linear models

Last updated: 02 April, 2019

See worked examples and more details at the accompanying notebook: [https://lindeloev.github.io/tests-as-linear](https://lindeloev.github.io/tests-as-linear)

<table>
<thead>
<tr>
<th>Common name</th>
<th>Built-in function in R</th>
<th>Equivalent linear model in R</th>
<th>Exact?</th>
<th>The linear model in words</th>
<th>Icon</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>y</strong> is independent of <strong>x</strong></td>
<td></td>
<td></td>
<td></td>
<td>One number (intercept, i.e., the mean) predicts <strong>y</strong>.</td>
<td><img src="https://lindeloev.github.io/tests-as-linear/icon.png" alt="Icon" /></td>
</tr>
<tr>
<td>P: One-sample t-test</td>
<td><code>t.test(y)</code></td>
<td><code>lm(y ~ 1)</code></td>
<td>✓</td>
<td>✓ for <strong>N &gt; 14</strong></td>
<td><img src="https://lindeloev.github.io/tests-as-linear/icon.png" alt="Icon" /></td>
</tr>
<tr>
<td>N: Wilcoxon signed-rank</td>
<td></td>
<td><code>wilcox.test(y)</code></td>
<td></td>
<td>- (Same, but it tests the ranked sign of <strong>y</strong>.)</td>
<td><img src="https://lindeloev.github.io/tests-as-linear/icon.png" alt="Icon" /></td>
</tr>
<tr>
<td>P: Paired-sample t-test</td>
<td><code>t.test(y, y, paired=TRUE)</code></td>
<td><code>lm(y - y ~ 1)</code></td>
<td>✓</td>
<td>✓ for <strong>N &gt; 14</strong></td>
<td><img src="https://lindeloev.github.io/tests-as-linear/icon.png" alt="Icon" /></td>
</tr>
<tr>
<td>N: Wilcoxon matched pairs</td>
<td><code>wilcox.test(y, y, paired=TRUE)</code></td>
<td><code>lm(signed_rank(y) - y ~ 1)</code></td>
<td></td>
<td>- (Same, but it tests the signed rank of <strong>y</strong>.)</td>
<td><img src="https://lindeloev.github.io/tests-as-linear/icon.png" alt="Icon" /></td>
</tr>
<tr>
<td><strong>y</strong> ~ continuous x</td>
<td></td>
<td></td>
<td></td>
<td>One intercept plus <strong>x</strong> multiplied by a number (slope) predicts <strong>y</strong>.</td>
<td><img src="https://lindeloev.github.io/tests-as-linear/icon.png" alt="Icon" /></td>
</tr>
<tr>
<td>P: Pearson correlation</td>
<td><code>cor.test()</code></td>
<td><code>lm(y ~ x)</code></td>
<td>✓</td>
<td>✓ for <strong>N &gt; 10</strong></td>
<td><img src="https://lindeloev.github.io/tests-as-linear/icon.png" alt="Icon" /></td>
</tr>
<tr>
<td>N: Spearman correlation</td>
<td><code>cor.test()</code></td>
<td><code>lm(rank(y) ~ 1 + rank(x))</code></td>
<td></td>
<td>- (Same, but with ranked <strong>x</strong> and <strong>y</strong>.)</td>
<td><img src="https://lindeloev.github.io/tests-as-linear/icon.png" alt="Icon" /></td>
</tr>
<tr>
<td><strong>y</strong> ~ discrete x</td>
<td></td>
<td></td>
<td></td>
<td>An intercept for group 1 (plus a difference if group 2) predicts <strong>y</strong>.</td>
<td><img src="https://lindeloev.github.io/tests-as-linear/icon.png" alt="Icon" /></td>
</tr>
<tr>
<td>P: Two-sample t-test</td>
<td><code>t.test(y, y, var.equal=TRUE)</code></td>
<td><code>lm(y ~ 1 + G)</code></td>
<td>✓</td>
<td>✓ for <strong>N &gt; 11</strong></td>
<td><img src="https://lindeloev.github.io/tests-as-linear/icon.png" alt="Icon" /></td>
</tr>
<tr>
<td>N: Welch's t-test</td>
<td><code>t.test(y, y, var.equal=FALSE)</code></td>
<td><code>lm(y ~ 1 + G)</code></td>
<td></td>
<td>- (Same, but with one variance per group instead of one common.)</td>
<td><img src="https://lindeloev.github.io/tests-as-linear/icon.png" alt="Icon" /></td>
</tr>
<tr>
<td>N: Mann-Whitney U</td>
<td><code>wilcox.test(y, y)</code></td>
<td><code>lm(signed_rank(y) ~ 1 + G)</code></td>
<td></td>
<td>- (Same, but it predicts the signed rank of <strong>y</strong>.)</td>
<td><img src="https://lindeloev.github.io/tests-as-linear/icon.png" alt="Icon" /></td>
</tr>
<tr>
<td><strong>y</strong> ~ discrete x</td>
<td></td>
<td></td>
<td></td>
<td>An intercept for group 1 (plus a difference if group #1) predicts <strong>y</strong>.</td>
<td><img src="https://lindeloev.github.io/tests-as-linear/icon.png" alt="Icon" /></td>
</tr>
<tr>
<td>P: One-way ANOVA</td>
<td><code>aov(y ~ group)</code></td>
<td><code>lm(y ~ 1 + G)</code></td>
<td>✓</td>
<td>✓ for <strong>N &gt; 11</strong></td>
<td><img src="https://lindeloev.github.io/tests-as-linear/icon.png" alt="Icon" /></td>
</tr>
<tr>
<td>N: Kruskal-Wallis</td>
<td><code>kruskal.test(y ~ group)</code></td>
<td><code>lm(rank(y) ~ 1 + G)</code></td>
<td></td>
<td>- (Same, but it predicts the rank of <strong>y</strong>.)</td>
<td><img src="https://lindeloev.github.io/tests-as-linear/icon.png" alt="Icon" /></td>
</tr>
<tr>
<td>P: One-way ANOVA</td>
<td><code>aov(y ~ group)</code></td>
<td><code>lm(y ~ 1 + G)</code></td>
<td>✓</td>
<td>✓ for <strong>N &gt; 11</strong></td>
<td><img src="https://lindeloev.github.io/tests-as-linear/icon.png" alt="Icon" /></td>
</tr>
<tr>
<td>N: Mann-Whitney U</td>
<td><code>wilcox.test(y, y)</code></td>
<td><code>lm(signed_rank(y) ~ 1 + G)</code></td>
<td></td>
<td>- (Same, but with a slope on <strong>x</strong>.)</td>
<td><img src="https://lindeloev.github.io/tests-as-linear/icon.png" alt="Icon" /></td>
</tr>
<tr>
<td>Note: this is discrete AND continuous. ANOVA's are ANOVA's with a continuous <strong>x</strong>.</td>
<td><img src="https://lindeloev.github.io/tests-as-linear/icon.png" alt="Icon" /></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P: Two-way ANOVA</td>
<td><code>aov(y ~ group * sex)</code></td>
<td><code>lm(y ~ 1 + G + S + G*S)</code></td>
<td>✓</td>
<td>✓ for <strong>N &gt; 11</strong></td>
<td><img src="https://lindeloev.github.io/tests-as-linear/icon.png" alt="Icon" /></td>
</tr>
<tr>
<td></td>
<td></td>
<td><code>lm(y ~ 1 + G + S + G*S)</code></td>
<td></td>
<td>Interaction term: changing sex changes the <strong>y</strong> ~ group parameters.</td>
<td><img src="https://lindeloev.github.io/tests-as-linear/icon.png" alt="Icon" /></td>
</tr>
<tr>
<td>Note: <strong>G</strong> is an indicator (0 or 1) for each non-intercept levels of the group variable. Similarly for <strong>S</strong> when sex. The first line with <strong>G</strong> is main effect of group, the second (with <strong>S</strong>) for sex and the third is the group + sex interaction. For two levels (e.g. male/female), line 2 would just be &quot;<strong>S</strong>&quot;, and line 3 would be <strong>S</strong>; multiplied with each <strong>G</strong>.</td>
<td><img src="https://lindeloev.github.io/tests-as-linear/icon.png" alt="Icon" /></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>y</strong> ~ discrete x</td>
<td></td>
<td></td>
<td></td>
<td>Interaction term: (Same as Two-way ANOVA.)</td>
<td><img src="https://lindeloev.github.io/tests-as-linear/icon.png" alt="Icon" /></td>
</tr>
<tr>
<td>P: Chi-square test</td>
<td><code>chisq.test(groupXsex_table)</code></td>
<td><code>glm(y ~ 1 + G + S + G*S)</code></td>
<td>✓</td>
<td>✓ Same two-way ANOVA.</td>
<td><img src="https://lindeloev.github.io/tests-as-linear/icon.png" alt="Icon" /></td>
</tr>
<tr>
<td>N: Goodness of fit</td>
<td><code>chisq.test(y)</code></td>
<td><code>glm(y ~ 1 + G + S + G*S)</code></td>
<td>✓</td>
<td>✓ Same as One-way ANOVA and see Chi-Square note.</td>
<td><img src="https://lindeloev.github.io/tests-as-linear/icon.png" alt="Icon" /></td>
</tr>
</tbody>
</table>

List of common parametric (P) non-parametric (N) tests and equivalent linear models. The notation **y** ~ 1 + **x** is R shorthand for **y** = 1 + **x** which most of us learned in school. Models in similar colors are highly similar, but really, notice how similar they all are across colors! For non-parametric models, the linear models are reasonable approximations for non-small sample sizes (see "Exact" column and click links to see simulations). Other less accurate approximations exist, e.g., Wilcoxon for the sign test and Goodness-of-fit for the binomial test. The signed rank function is `signed.rank = function(x) sign(x) * rank(abs(x))`. The variables **G** and **S** are "dummy coded" indicator variables (either 0 or 1) explaining the fact that when **G** = 1 between categories the difference equals the slope. Subscripts (e.g., **G2** or **y1**)) indicate different columns in data. `lm` requires long-format data for all non-continuous models. All of this is exposed in greater detail and worked examples at [https://lindeloev.github.io/tests-as-linear](https://lindeloev.github.io/tests-as-linear).

![Image](https://lindeloev.github.io/tests-as-linear/icon.png)

**Footnotes:**

[a] See the note to the two-way ANOVA for explanation of the notation.

[b] Same model, but with one variance per group: `gls(y ~ 1 + G, weights = varIdent(form = ~1|group), method="ML")`.  

Jonas Kristoffer Lindeløv

[https://lindeloev.net](https://lindeloev.net)
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 Science Process

Inferential Statistics

Data Collection
Thank you!

Contact Info:
Dr. Kelly Avery – kavery@ida.org

Resources:
• https://testscience.org/
• H. Wickham and G. Grolemund, “R for Data Science.” https://r4ds.had.co.nz/
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 \( \mu \)

Variance \( \sigma^2 \)

Common applications:
- Performance
Lognormal distribution

**PDF**

\[
\frac{1}{\sqrt{2\pi}\sigma x} e^{-\frac{\left(\log x - \mu\right)^2}{2\sigma^2}}
\]

**Mean**

\[
e^{\mu + \frac{\sigma^2}{2}}
\]

**Variance**

\[
e^{2\mu + \sigma^2} \left(e^{\sigma^2} - 1\right)
\]

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

<table>
<thead>
<tr>
<th>PMF</th>
<th>$e^{-\lambda} \frac{\lambda^x}{x!}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$\lambda$</td>
</tr>
<tr>
<td>Variance</td>
<td>$\lambda$</td>
</tr>
</tbody>
</table>

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 \( \hat{p} = \frac{X}{n} \).

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

We may construct our test statistic using \( z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \).
Test of one proportion

Scenario: Suppose we have a system for transmitting messages. We want to know if this new system is worse than $P(\text{Transmission}) = .80$, 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: $X = 38; n = 50; \hat{p} = .76; p_0 = .80$
Compute statistic and determine probability

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

\[ z = -0.71 \]

**Conclusion:** We cannot conclude that the new system is worse than the old system.
Matching the test to the research question
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
Selecting the correct test

Goal

- Description of one group
  - Comparison of one group to a hypothetical value
    - R, I
    - O
    - ND
    - NND
    - One-sample t test
    - Wilcoxon test
  - Mann-Whitney test
    - Unpaired t test
    - Paired groups
    - Chi-square test
    - Binomial test
    - Fisher's test (chi-square for large samples)
  - R, I
  - O
  - ND
  - NND
  - Mean, SD
  - Median, interquartile range
  - McNemar's test
  - Chi-square test
  - One-way ANOVA

- Comparison of two groups
  - Unpaired groups
  - Matched Groups
  - Matched
    - R, I
    - O
    - N
    - Cochran Q
    - Friedman test
    - Repeated measure ANOVA

- Prediction
  - From another measured variable
    - R, I
    - O
    - N
    - Nonparametric regression
    - Simple linear regression
    - Simple logistic regression
    - Contingency coefficients
  - From several measured or binomial variables
    - R, I
    - O
    - N
    - Multiple logistic regressions

R, I = Ratio and Interval data   O = Ordinal data   N = Nominal data
N = Normal distribution   NND = Non normal distribution
Selecting the correct test
Selecting the correct test

Type of question:

- Do data match an expected ratio?
  - **Type of data:** discrete, categorical (counts, frequencies)
  - **a priori expectation:** yes / no
  - one variable, one sample
    - two or more categories
  - two variables, both variables have two categories
    - two or more categories

- Is there an association between two variables?
  - **Type of data:** non-parametric (nominal, ordinal, interval)
  - two variables: explained var. under control or with smaller error than resp. var.
  - two variables: neither one under control and with similar error
  - # categories two, more than two
  - data: unpaired, Kruskal-Wallis

- Do samples come from the same or different populations?
  - **Type of data:** non-parametric
  - one treatment variable
    - one-way non-parametric analysis
  - two treatment variables: one variable
    - data: paired, Wilcoxon paired test
  - multi-way parametric analysis (possibly nested)

- Do samples come from the same or different populations?
  - **Type of data:** parametric
  - one treatment variable:
    - two-way non-parametric analysis
  - two treatment variables:
    - data: unpaired, paired

Always check the literature for the details!

- **:** requires homogeneous variances (F test)
- **:** check for normally distributed residuals
- **:** requires similarly shaped distributions

If you have data for which no test seems available, try to transform your data.
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.
“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!”

t = 1.13, p = .26

D = 0.33, p = .07

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