**Tip** Learn how to write papers in a very clear and simple style. Whenever you can write in plain english and make the approach you are using understandable and clear. This can (sometimes) make it harder to get your papers in to journals. But simple, clear language leads to much higher use/citation of your work. Examples of great writers are: Rob Tibshirani, Brad Efron, Robert May, Martin Nowak, etc.

**Paper of the Day:** “Direct photosynthetic recycling of carbon dioxide to isobutyraldehyde” [http://www.nature.com/nbt/journal/v27/n12/abs/nbt.1586.html](http://www.nature.com/nbt/journal/v27/n12/abs/nbt.1586.html)
Today’s Outline

- Testing
- Bootstrap testing
In many scientific applications, producing/interpreting estimates and intervals will be your default. But hypothesis testing (for better or worse) plays a major role in the scientific enterprise.

- Doing tests is simple; based on the available data, we make a binary decision. Frequentist calibration of the testing “rule” considers replications of the experiment.

- Interpretation of testing results is trickier. You may need to be flexible about how testing is viewed; it may change as you work in different areas of science. Different statisticians lend more or less credibility to p-values/testing (although they are used extensively by e.g. lab scientists, the FDA, etc.)

- We will discuss permutation/bootstrap hypothesis tests today. However, similar ideas apply to parametric testing.
A Digression on Testing †

Before the formalities, some more “target practice”

BACK TO THE ARCHER SETUP. AS BEFORE, SHE AIMS AT THE ‘BULLSEYE’ (A SINGLE UNKNOWN LOCATION) AND IN 95% OF SHOTS HITS WITHIN 10CM OF IT
Before the formalities, some more “target practice”
Before the formalities, some more “target practice”
Implementing a Test†

To implement testing, we choose a test statistic \( T(X, Y) \). We use it to make a binary decision \( D \), where convention is that:

\[
D = \begin{cases} 
0, & T(X, Y) \leq c \quad \text{“not significant”} \\
1, & T(X, Y) > c \quad \text{“significant”}
\end{cases}
\]

for some pre-chosen “critical value” \( c \). Calibration of the test (ie. choosing \( c \)) is done under the “null hypothesis” \( H_0 \) a statement about the true state of Nature.

Typical null hypotheses are \( H_0 : \theta = 0 \) or perhaps \( H_0 : \theta < 0 \) - in fact any well-defined statement about the superpopulation could be used. We will focus on \( H_0 \) with some regression-based meaning.

I will write \( F|H_0 \) to denote sampling from a superpopulation where the null holds, in expressions such as \( \mathbb{E}_{F|H_0}[D] \). This is known as sampling “under the null”
Frequentist Calibration†

The “size” or “Type I error rate” of a test is defined as:

$$E_{F|H_0}[D] = \Pr_{F|H_0}[D = 1]$$

i.e. the probability of getting $D = 1$ under the null.

In practice, we choose $c$ to give a desired size, then do the test. The “target” Type I error rate is denoted $\alpha$; $\alpha = 0.05$ is “traditional”. Results with $D = 1$ are “significant at the $\alpha$ level”.

We say tests are “valid at the nominal $\alpha$” if $E_{F|H_0}[D] \leq \alpha$ and “exact” if $E_{F|H_0}[D] = \alpha$..., but some authors reverse these terms. Tests with $E_{F|H_0}[D] \leq \alpha$ “control the Type I error rate below $\alpha$”
Frequentist Calibration

Frequentist tests and intervals are closely connected.

Suppose you had a valid confidence interval “recipe” $CI(Y, X)$ for parameter $\theta$, to be used in replicate experiments. A simple binary decision for testing $H_0 : \theta = \theta_0$ is the indicator function $D = 1_{\{\theta_0 \notin CI\}}$. As $Pr_F[\theta \in CI] = 1 - \alpha$

$$Pr_F[\theta \notin CI] = \alpha$$

$$\Rightarrow Pr_F|_{\theta=\theta_0}[1_{\{\theta_0 \notin CI\}} = 1] = \alpha$$

$$\Rightarrow E_F|_{\theta=\theta_0}[D] = \alpha$$

.. and we have a test of size $\alpha$.

▶ On its own, this result does not imply sanity! What’s the size, for a CI which covers $\mathbb{R}$ with probability 0.95, and returns the value of 42 with probability 0.05?

▶ Of course, sane estimates/intervals give “sane” tests

▶ You can do the reverse as well converting tests into CI’s, but this is usually less useful
P-values, although not essential for doing tests, are nearly ubiquitous in applied work. They are a very useful “shorthand” and you should understand them.

Under the convention that larger test statistics occur “farther” from $H_0$, for observed data $Y$ we define:

$$ p = p(Y) = \Pr_{Y' \sim F|H_0}[T(Y') > T(Y)] $$

i.e. the long-run proportion of datasets, under the null, which are “more extreme” than the observed $Y$. 

![Graph showing p-values](image-url)
Notes on p-values

▶ If $p < \alpha$, the result is “significant”, otherwise it is “not significant” $p$ is the “largest $\alpha$ at which the result would be significant” - i.e. it summarizes the tests you could do.

▶ If $p$ is small, you could say the data are “inconsistent” with $H_0$ (not the same as an inconsistent estimator). Avoid writing about “evidence” or “support”, unless these terms express what you really mean. Writing that the data “suggest” conclusions is safe(r).

▶ Small $p$ occur (i) when $H_0$ is true and something unusual happened, (ii) when $H_0$ is not true. On its own, a small $p$ does not distinguish these two things

▶ For discrete $Y$, $T(Y) = T(Y')$ can happen, discreteness of $p$ is only interesting in small samples, and is often ignored

▶ Replicates may include random $X$ or fixed.
What’s the distribution of a $p$-value? Under the null:

\[
\Pr_{Y \sim F|H_0}[p(Y) < \alpha] = \Pr_{Y \sim F|H_0}[\Pr_{Y' \sim F|H_0}[T(Y') > T(Y)] \leq \alpha]
\]

\[
= \Pr_{Y \sim F|H_0}[1 - F(T(Y)) \leq \alpha]
\]

\[
= \Pr_{Y \sim F|H_0}[T(Y) > F^{-1}(1 - \alpha)]
\]

\[
= 1 - F(F^{-1}(1 - \alpha)) = \alpha
\]

where $F(\cdot)$ denotes the cumulative distribution function of $T(Y)$ under the null.

- $\Leftrightarrow$ $p$-values are uniform on $[0, 1]$ under the null
- $p$-values are given by the “tail area” of the distribution of (replicate) $T(Y')$ beyond observed $T(Y)$ under the null
- $p$-values are (complicated) functions of the observed data
- Just as with intervals, if your assumptions about $F$ are wrong, your putative $p$-values will not have “nominal” behavior. As with intervals, approximations are widely used.
People often forget $p$-values are just functions of the data:

Formally, the $p$ value is a 1-1 function of $T(Y)$

We usually focus on how the lower quantiles of $p(Y)$ depends on $\theta$ - but the spread can also be useful to know.
The simplicity of calculating $p(Y)$, seeing if its $< 0.05$ and reporting on the underlying “truth” means that many non-experts think that inference is entirely “$p$’s and $t$’s”

You know this isn’t true. P-values can be a useful summary of the data, with nice properties (particularly useful in applications where confidence intervals may be unwieldy).

But...at the end of the day as sane applied statisticians you are looking for sensible/scientifically meaningful patterns, which are often more usefully summarized by confidence intervals and point estimates.
Testing: different decisions†

Frequentist tests report $D = 1$ whenever $p \leq \alpha$. You should report $p$, so your reader can pick their own $\alpha$.

But what $\alpha$? And what does “significant” mean practically? Testing output may be used for different goals. We may want:

- A summary of whether the data is consistent with a particular hypothesis (a pure significance test - just calculate $p$)
- A decision (yes/no) whether the data is consistent with a particular hypothesis (a test of significance - if $p < \alpha$ we reject $H_0$)
- A decision (yes/no) on which of two hypotheses is best supported by the data (a hypothesis test - if $p < \alpha$ we reject $H_0$ in favor of $H_1$, else “accept” $H_0$. $H_1$ is the alternative hypothesis. )

These goals are all different!
Testing: different decisions

For either type of “significance” test

- There is no alternative hypothesis. Informally, you can view these as a “screening tool” indicating how noteworthy the data is, or whether (yes/no) we declare it “interesting”
- You must choose $T(\cdot)$ to detect deviations from $H_0$ which are of interest!
- Power = $\Pr_F[D = 1]$, the probability of rejecting when $H_0$ does not hold. In regression settings, this is a function of the true $\theta$ and $n$.
- Power = $1 - \Pr_F[D = 0]$ i.e. 1-the probability of a Type II error.
- Think about power before doing tests! In a lower power setting, if $D = 0$ what two things may have happened? When $n$ is huge and $D = 1$ what may have happened?
- Unthinking adherence to $p$ is not good statistics/science. Don’t expect readers to “reject the null” without convincing science (the more money on the line, the more this is true).
Testing: different decisions

For hypothesis tests:

- There has to be an alternative hypothesis, $H_1$. You should choose $T(\cdot)$ based on this alternative.
- The test is calibrated under $H_0$, not $H_1$. It is not an “even-handed” comparison. $H_0$ represents a “default” outcome.
- Power matters, and real-world interpretations should use your prior knowledge.
- In e.g. drug licensing, getting $p < \alpha$ (twice!) is necessary but not sufficient for a “yes” decision, even with well-powered studies. Some scientific justification is again required; $p$ alone will not convince competent reviewers.
Testing: different decisions

Which type of testing would you use?

▶ You work with a mouse lab. They did an experiment to assess whether a specific “candidate” gene is essential to life.

▶ A consulting client comes to you with a totally novel chemical process. They have some (unanalyzed) provisional data, and want to know whether to proceed with more experiments.

▶ Prior to analysis of a new risk factor, your co-authors want to check if age, sex, etc. are unusual in your sample (this is known as a “Table 1” analysis in some circles)

▶ A consulting client provides you with a complex dataset and wants to know “what’s interesting”?

▶ In a large trial, you obtain data consistent with a new drug being moderately effective (not strongly, not weakly). You are asked, “is it significant”? 
Testing: what p-values are not (the fair warning slide)†

There is a tremendous confusion over *p*-values. The following hold very generally

- They do not represent $\Pr(H_0 \text{ is true})$ or $\Pr(\theta > 0)$. Statements like this don't even make sense in a frequentist setting - so avoid writing them or something that could be interpreted this way.

- They are not “measures of evidence”. They are a summary of the testing decision you could make, a and a potentially useful function of the data. “Evidence” is a loaded word to some statisticians!

- $p \leq 0.05$ is not “proof” of anything.

- $p \neq 0$ (Why?) - don’t round *p* values down. You can write, e.g. $p < 10^{-4}$
Calculating a P-value (Parametric)

Suppose we calculate a statistic $T(Y)$ to test the null hypothesis $H_0 : \theta = 0$ and we know that under the null $T(Y') \sim t_{19}$. How do we calculate the p-value?

This calculation preserves we know the distribution $t_{19}$, what if we want to make fewer assumptions about the null distribution?
Calculating a P-value (Parametric)

Suppose we calculate a statistic $T(Y)$ to test the null hypothesis $H_0 : \theta = 0$ and we know that under the null $T(Y') \sim t_{19}$. How do we calculate the p-value?

This calculation presumes we know the distribution $t_{19}$, what if we want to make fewer assumptions about the null distribution?
Calculating a P-value (Permutation)

Suppose we observe survival times on 16 mice in a treatment and a control group:

\[ X = (94, 197, 16, 38, 99, 141, 23) \]

\[ Y = (52, 104, 146, 10, 51, 30, 40, 27, 46) \]

\( X \sim F \) and the values \( Y \sim G \) and we want to test: \( H_0 : F = G \).

The difference of means is \( \hat{\theta} = \bar{x} - \bar{y} = 30.63 \). One way to calculate a p-value would be to make a parametric assumption.

Another clever way, devised by Fisher is permutation. Define the vector \( g \) to be the group labels \( g_i = 1 \) if treatment \( (X) \) and \( g_i = 0 \) if control \( (Y) \). Then pool all of the observations together into a vector

\( v = (94, 197, 16, 38, 99, 141, 23, 52, 104, 146, 10, 51, 30, 40, 27, 46) \)
Calculating a P-value (Permutation)

If there are \( n \) samples from \( F \) and \( m \) samples from \( y \) for a total of \( N = m + n \) samples, then there are \( \binom{N}{n} \) possible ways to assign the group labels. Under \( F = G \) it is possible to show that - conditional on the observed values - each of these is equally likely\(^1\).

We can write our statistic

\[
\hat{\theta} = \bar{x} - \bar{y} = \frac{1}{n} \sum_{g_i=1} v_i - \frac{1}{m} \sum_{g_i=0} v_i
\]

The permutation null distribution of \( \hat{\theta} \) is calculated by forming all permutations of \( g \) and recalculating the statistic:

\[
\hat{\theta}^* = \frac{1}{n} \sum_{g_i^*=1} v_i - \frac{1}{m} \sum_{g_i^*=0} v_i
\]

\(^1\)See for example Lehmann Hypothesis Testing
Calculating a P-value (Permutation)

Then a permutation p-value is the permutation probability that $|\hat{\theta}^*|$ exceeds $|\hat{\theta}|$. $\Pr(\hat{\theta}^* \geq \hat{\theta}) = \frac{\#\{ |\hat{\theta}^*| \geq |\hat{\theta}| \}}{\binom{N}{n}}$.

This is an exact calculation, just like we did before with the exact plug-in estimator. But usually $\binom{N}{n}$ is humongous, so we use Monte Carlo.

1. Sample the labels $g_i$ with replacement to get $B$ permuted sets of group labels $g_i^*b$

2. Evaluate the permutation statistics

   $\hat{\theta}^*b = \frac{1}{n} \sum_{g_i^*b=1} v_i - \frac{1}{m} \sum_{g_i^*b=0} v_i$

3. Calculate the permutation p-value:

   $\hat{p}_{perm} = \frac{\#\{ |\hat{\theta}^*b| \geq |\hat{\theta}| \}}{B}$
Example 1: Loess Smoothing

This boils down to a similar procedure.
Calculating a P-value (Bootstrap)

- Draw $B$ samples with replacement of size $n + m$ from the values $v = [x; y]$. Call the first $n$ observations $x^*b$ and the last $m$ observations $y^*b$.
- Evaluate $\hat{\theta}^b = \bar{x}^b - \bar{y}^b$ on each sample.
- Approximate the $p$-value by:

$$\hat{p}_{boot} = \#\{|\hat{\theta}^b| \geq |\hat{\theta}|\} / B$$

Note this is very similar to the permutation based approach. We just sample with replacement.

In the case of the permutation test, if we had studentized the statistic (e.g. set $\hat{\theta} = \frac{\bar{x} - \bar{y}}{\bar{\sigma} \sqrt{1/n + 1/m}}$) where $\bar{\sigma}$ is an estimate of the pooled s.d., we would get the same permutation $p$-value.

For the bootstrap this doesn’t hold - you get more accurate testing with the studentized statistic.
Calculating a P-value (Bootstrap)

This algorithm still tests the null hypothesis that $F = G$. What if we wanted to be more specific? Say if we wanted to test that $\mu_F = \mu_G$?

We could use this algorithm:

- Let $\hat{F}$ put equal probability on the points $\tilde{x}_i = x_i - \bar{x} + \bar{v}$ for $i = 1, 2, \ldots, n$ and $\hat{G}$ put equal probability on the points $\tilde{y}_i = y_i - \bar{y} + \bar{v}$ for $i = 1, 2, \ldots, m$, where $\bar{x}$ and $\bar{y}$ are the group means and $\bar{v}$ is the mean of the combined sample.

- Form $B$ bootstrap data sets $(x^{*b}, y^{*b})$ where $x^{*b}$ is sampled with replacement from $\{\tilde{x}_i\}$ and $y^{*b}$ is sampled with replacement from $\{\tilde{y}_i\}$.

- Evaluate $T = \frac{\bar{x}^{*b} - \bar{y}^{*b}}{\sqrt{\sigma_1^{2*b} + \sigma_2^{2*b}}}$

- Approximate the $p$-value by:

$$\hat{p}_{boot} = \#\{|\hat{\theta}^{*b}| \geq |\hat{\theta}|\}/B$$
Bootstrap vs. Permutation

The preceding examples illustrate differences between the two approaches:

▶ A permutation test exploits a special symmetry that exists under the null hypothesis to create a permutation distribution.

▶ The permutation test is exact, it is the exact probability of obtaining a test statistic as extreme, having fixed the values of the combined sample.

▶ In contrast, the bootstrap estimates the probability mechanism under the null hypothesis, and then samples from it.

▶ The estimate is not an exact probability, but does not require the special symmetry, so can be applied much more generally.

▶ For instance, in the two-sample problem a permutation test can only test the null hypothesis $F = G$, while the bootstrap can test equal means with possibly unequal variances.
Bootstrap Hypothesis Test For a Linear Model

Suppose we assume a mean model of the form:

$$E[Y_i|X_i = x_i, Z_i = z_i] = \beta_0 + \sum_{j=1}^{J} \beta_j x_{ij} + \sum_{k=1}^{K} \gamma_k z_{ik}$$

covariates of interest \hspace{1cm} adjustment covariates

using least squares. Suppose we wish to test the null hypothesis that $\beta_1 = \beta_2 = \cdots = \beta_J = 0$. One approach would be to form the F-like-statistic:

$$F = \frac{RSS^0 - RSS}{RSS}$$
Bootstrap Hypothesis Test For a Linear Model

Where:

\[
RSS_0 = \sum_{i=1}^{n} \left( y_i - \hat{\beta}_0^0 - \sum_{k=1}^{K} \hat{\gamma}_k^0 z_{ik} \right)^2
\]

\[
RSS = \sum_{i=1}^{n} \left( y_i - \hat{\beta}_0 - \sum_{j=1}^{J} \hat{\beta}_j x_{ij} - \sum_{k=1}^{K} \hat{\gamma}_k z_{ik} \right)^2
\]

where \( \beta_0^0 \) and \( \gamma_k^0 \) are obtained when fitting the null model:

\[
\mathbb{E}[Y_i | X_i = x_i, Z_i = z_i] = \beta_0 + \sum_{k=1}^{K} \gamma_k z_{ik}
\]

adjustment covariates
Bootstrap Hypothesis Test For a Linear Model

Then an algorithm for testing the significance of $F$ could be the following:

1. Calculate the residuals from the full model
   \[ \epsilon_i = y_i - \hat{\beta}_0 - \sum_{j=1}^{J} \hat{\beta}_j x_{ij} - \sum_{k=1}^{K} \hat{\gamma}_k z_{ik} \]

2. Calculate the fitted values under the null
   \[ \hat{y}_i^0 = \hat{\beta}_0^0 + \sum_{k=1}^{K} \hat{\gamma}_k^0 z_{ik} \]

3. Sample $B$ times with replacement from $\epsilon_i$ and add the values back to the null fit to generate null data: $y_i^{*b} = \hat{y}_i^0 + \epsilon_i^{*b}$

4. Recalculate the statistic $F^{*b}$ on all bootstrap samples

5. Approximate the $p$-value by:

\[ \hat{p}_{boot} = \#\{ |\hat{\theta}^{*b}| \geq |\hat{\theta}| \} / B \]
Bootstrap Hypothesis Test For a Linear Model

These tests are common in some areas (like genomics) where the null distribution is frequently not clearly known.

They still make assumptions (e.g., the data are i.i.d. $F$)

The big problem here is creating $Y^*$ that are sampled under the null hypothesis. The standard bootstrap doesn't do this, it samples under the truth approximately.

The $\alpha$-level accuracy of these procedures is not exact (upsets e.g. the FDA)

Also “fiddling” with $\tilde{F}$ requires choices, try not to be influenced by “peeking” at the data. The FDA will not let you do this; they demand code in advance.

There are now permutation based approaches to the same problem: “COVARIATE-ADJUSTED NONPARAMETRIC ANALYSIS OF MAGNETIC RESONANCE IMAGES USING MARKOV CHAIN MONTE CARLO” [http://www.bepress.com/jhubiostat/paper187/]
Stamps! A slightly more complicated example

Here is a case where simple normal theory/permutation tests can’t be used. Data on the thickness in millimeters of 485 stamps, printed in 1872. Stamps from that year are thought to have been printed on a mixture of papers. How many papers were used?
Stamps! A slightly more complicated example

We could fit a Gaussian kernel density estimate. Denoting the data by $x_1, \ldots, x_n$ the estimate is defined by:

$$
\hat{f}(t; h) = \frac{1}{nh} \sum_{i=1}^{n} \phi \left( \frac{t - x_i}{h} \right)
$$

where $\phi(t) = \frac{1}{\sqrt{2\pi}} \exp(-t^2/2)$. The parameter $h = \sigma$ is the window size and determines the amount of smoothing. Larger $h$ means more smooth estimates. The idea is adding up lots of little Gaussian distributions at each $x_i$. 

![Graphs showing density estimates with different bandwidths.](image)
Stamps! A slightly more complicated example

We will use the fact that as the bandwidth $h$ increase, the number of modes is non-increasing. We will consider testing $H_0$: number of modes = 1, versus number of modes $> 1$.

Since the number of modes decreases as $h$ increases, there is a smallest value of $h$ such that $\hat{f}(t; h)$ has one mode. Call this $\hat{h}_1 \approx 0.00673$. 

![Graph showing the number of modes decreasing as bandwidth increases.](image-url)
Stamps! A slightly more complicated example

We will use $\hat{f}(t; \hat{h}_1)$ as the estimated null distribution. It is the density estimate “closest” to our data that is consistent with $H_0$. We will rescale the estimate to have the same variance as the sample variance and call it $\hat{g}(\cdot; \hat{h}_1)$.
Stamps! A slightly more complicated example

We will use as a test statistic $\hat{h}_1$ since a large value of $\hat{h}_1$ means that a lot of smoothing is necessary to get one mode. Now we just need to generate bootstrap samples. To do this, remember we need to sample from the null distribution $\hat{g}(\cdot; \hat{h}_1)$ (the smooth bootstrap). This seems complicated, but can be made simpler because of the form of the Gaussian kernel estimate.

We sample $y_1^*, \ldots, y_n^*$ from the stamp values and set:

$$x_i^* = \bar{y}^* + (1 + \hat{h}_1^2/\hat{\sigma}^2)^{-1/2}(y_i^* - \bar{y}^* + \hat{h}_1 \epsilon_i)$$

Where $\bar{y}^*$ is the mean of the bootstrap samples, $\hat{\sigma}^2$ is the plug in estimate of the variance of the data, and $\epsilon_i$ are standard normal random variables.
Our bootstrap distribution looks like this:
And our approximate $p$-value is:

$$\Pr_{\hat{g}(\cdot; \hat{h}_1)}\{\hat{h}_1^* > \hat{h}_1\} \approx \frac{3}{500} = 0.006$$
Bootstrap + Permutation: Summary

- Bootstrap and permutation methods are more common in some disciplines than in others.
- They are very commonly used for calculating standard errors/confidence intervals.
- They can be used to evaluate more complicated statistics than just simple regression parameters.
- Care should still be taken to correctly model dependence, there is no substitute for careful thought and sane analyses.
- Permutation gives exact p-values, but may be less flexible. The bootstrap is more flexible, but cannot be interpreted as exact (some people don’t like this).