Lecture 5.3: Permutation Tests

STAT 1361/2360: Statistical Learning and Data Science

University of Pittsburgh
Prof. Lucas Mentch

Overview

- Last time we talked about the bootstrap and used confidence intervals to motivate the idea
- Today we'll introduce **permutation tests**
  - Form of hypothesis testing that requires far fewer assumptions than the parametric versions you're used to
  - Accordingly, we now focus on hypothesis testing as the motivation for permutation tests
- First, a quick review of the general inference framework we discussed last time:

Inference Overview

The fundamental steps in statistical inference:

1. We have some (population distribution) F with some associated parameter θ
2. We take a sample from that population and calculate an estimate $\hat{\theta}$
3. We want to use $\hat{\theta}$ to *infer* something about θ

   But ... In order to make any kind of inference, we need to know something about the distribution of $\hat{\theta}$

Inference Overview

That final point – needing to know something about the distribution of $\hat{\theta}$ – should be readily apparent in the context of hypothesis testing

Hypothesis Testing Outline

Now let's review the fundamental steps in hypothesis testing:

1. Define the hypotheses of interest (null and alternative)

Hypothesis Testing Outline

Now let’s review the fundamental steps in hypothesis testing:

1. Define the hypotheses of interest (null and alternative)
2. Define a test statistic t

Hypothesis Testing Outline

Now let's review the fundamental steps in hypothesis testing:

1. Define the hypotheses of interest (null and alternative)
2. Define a test statistic t
3. Determine what the distribution of t would be if the null hypothesis were true (i.e. under H_0)

Hypothesis Testing Outline

Now let's review the fundamental steps in hypothesis testing:

1. Define the hypotheses of interest (null and alternative)
2. Define a test statistic t
3. Determine what the distribution of t would be if the null hypothesis were true (i.e. under H_0)
4. Calculate the test statistic and determine whether it's “reasonable" to think it could have come from that distribution.
   - If not, we *reject* H_0. If so, we *fail to reject* H_0.

Hypothesis Testing Logic

- Remember, we think of H_0 as being true “by default" – we only reject H_0 if we find sufficient enough evidence to do so.
- The logic behind arguing that we have sufficient evidence to reject H_0 is:
  1. If H_0 were true, then t would have to come from this distribution

Hypothesis Testing Logic

- Remember, we think of H_0 as being true “by default" – we only reject H_0 if we find sufficient enough evidence to do so.
- The logic behind arguing that we have sufficient evidence to reject H_0 is:
  1. If H_0 were true, then t would have to come from this distribution
  2. The particular value of t we calculated seems very unlikely to have come from this distribution (depends on our chosen level of α)

Hypothesis Testing Logic

- Remember, we think of H_0 as being true “by default" – we only reject H_0 if we find sufficient enough evidence to do so.
- The logic behind arguing that we have sufficient evidence to reject H_0 is:

  1. If H_0 were true, then t would have to come from this distribution
  2. The particular value of t we calculated seems very unlikely to have come from this distribution (depends on our chosen level of α)
  3. Therefore, it’s likely that t actually came from a different distribution

Hypothesis Testing Logic

- Remember, we think of H_0 as being true “by default" – we only reject H_0 if we find sufficient enough evidence to do so.
- The logic behind arguing that we have sufficient evidence to reject H_0 is:

1. If H_0 were true, then t would have to come from this distribution
2. The particular value of t we calculated seems very unlikely to have come from this distribution (depends on our chosen level of α)
3. Therefore, it’s likely that t actually came from a different distribution
4. Therefore, H_0 must not be (or, is likely not) true

Hypothesis Testing Overview

- Thus, in order to do any kind of hypothesis test, we need to know what the distribution of our test statistic t would look like if the null hypothesis were true

** *Sidenote:* The terms "statistic" and "estimator" are referring to the same kind of fundamental object – both are just functions of the data. Thus, whether we talk about needing the distribution of an estimator $\hat{\theta}$ or a test statistic t, we're talking about the same kind of thing.

Hypothesis Testing Example

Let’s go through an example:

Suppose we have two groups of observations and we want to test whether there is a significant difference in the group means:

$$H_0 : \mu_1 = \mu_2 \iff \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 \neq \mu_2 \iff \mu_1 - \mu_2 \neq 0$$

Data (unpaired) of the form:

Group 1: $X_1, \ldots, X_{n_1}$
Group 2: $Y_1, \ldots, Y_{n_2}$

Hypothesis Testing Example

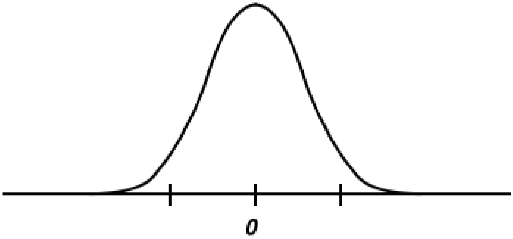
The natural test statistic to use here would be $t = \hat{\mu}_1 - \hat{\mu}_2$

How would we know the distribution of t?

1. If we assume the observations from each group come from a normal distribution:
    - Then we know $\hat{\mu}_1$ is approximately normal and $\hat{\mu}_2$ is approximately normal, so $\hat{\mu}_1 - \hat{\mu}_2$ is approximately normal
2. If n_1 and n_2 are both reasonably large
    - Then we can appeal to the central limit theorem to know that both group means and hence the difference in group means are approximately normal

Hypothesis Testing Example

Now, *if the null hypothesis were true* and either of those two situations apply, then we know $t = \hat{\mu}_1 - \hat{\mu}_2$ would come from a normal distribution with mean 0, so we can calculate our value of t and compare



Hypothesis Testing Example

BUT … what if neither of those situations apply?

This is actually a very common situation in many applied fields:

- We have two groups of observations (and often times there looks to be a clear difference in group means)
- But we only have a few observations in each group and no reason to think that observations in each group came from a normal distribution

So what do we do in this situation?

Hypothesis Testing Example

Let's say we were willing to accept that everything about the distributions from the two groups was the same except (possibly) the means

Then under the null hypothesis, the means (and thus the entire distributions) would be identical

Group 1	**Group 2**
x_1	y_1
x_2	y_2
.	.
.	.
x_{n_1}	y_{n_2}

Hypothesis Testing Example

Key Point #1: Under the null hypothesis then, why would any observation – say x_1 – appear in Group 1 instead of Group 2?

Hypothesis Testing Example

Key Point #1: Under the null hypothesis then, why would any observation – say x_1 – appear in Group 1 instead of Group 2?

There would be no reason! The distributions would be identical! Thus, under the null hypothesis, the fact that we observe any particular value in Group 1 or Group 2 is simply due to random chance.

Hypothesis Testing Example

Key Point #1: Under the null hypothesis then, why would any observation – say x_1 – appear in Group 1 instead of Group 2?

There would be no reason! The distributions would be identical! Thus, under the null hypothesis, the fact that we observe any particular value in Group 1 or Group 2 is simply due to random chance.

How can we exploit this fact?

Data Format

Let's think about the data in a different format:

Observation	Group Label
x_1	1
x_2	1
.	.
.	.
x_{n_1}	1
y_1	2
y_2	2
.	.
.	.
y_{n_2}	2

Group Label Randomness

Under the null hypothesis, the "Group Label" is random

Key Point #2: If this were true and we randomly shuffled the groups (i.e. permuted the Group Label) and recalculated $t_1^* = \hat{\mu}_1^* - \hat{\mu}_2^*$ according to these new groups, I should not expect t_1^* to be all that different from the test statistic I calculated on the original (correct) Group Labels, t

Group Label Randomness

Under the null hypothesis, the "Group Label" is random

Key Point #2: If this were true and we randomly shuffled the groups (i.e. permuted the Group Label) and recalculated $t_1^* = \hat{\mu}_1^* - \hat{\mu}_2^*$ according to these new groups, I should not expect t_1^* to be all that different from the test statistic I calculated on the original (correct) Group Labels, t

But how different is too different to think H_0 isn't true? What if we did this twice to compute t_1^* and t_2^*?

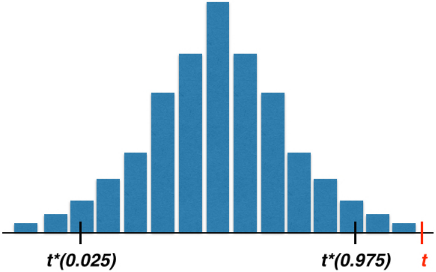
Permutation Test Idea

Idea: Repeat this a bunch of times, each time permuting the Group Labels and recomputing t_i^*. Then look at the distribution of these “permuted" test statistics and see where your original statistic t falls in that distribution.

Permutation Test Idea

Idea: Repeat this a bunch of times, each time permuting the Group Labels and recomputing t_i^*. Then look at the distribution of these "permuted" test statistics and see where your original statistic t falls in that distribution.

$\Longrightarrow$ If t lies at the extremes of this distribution, then we say it's likely it came from a different distribution, which means we reject H_0



Permutation Test Outline

Permutation Test for testing equality of group means:

1. Compute original test statistic $t = \hat{\mu}_1 - \hat{\mu}_2$

Permutation Test Outline

Permutation Test for testing equality of group means:

1. Compute original test statistic $t = \hat{\mu}_1 - \hat{\mu}_2$
2. Shuffle groups (permute group labels) and (re)compute $t_1^* = \hat{\mu}_1^* - \hat{\mu}_2^*$

Permutation Test Outline

Permutation Test for testing equality of group means:

1. Compute original test statistic $t = \hat{\mu}_1 - \hat{\mu}_2$
2. Shuffle groups (permute group labels) and (re)compute $t_1^* = \hat{\mu}_1^* - \hat{\mu}_2^*$
3. Repeat step (2) a large number of times (we'll say B times) to get $t_1^*, \ldots, t_B^*$

Permutation Test Outline

Permutation Test for testing equality of group means:

1. Compute original test statistic $t = \hat{\mu}_1 - \hat{\mu}_2$
2. Shuffle groups (permute group labels) and (re)compute $t_1^* = \hat{\mu}_1^* - \hat{\mu}_2^*$
3. Repeat step (2) a large number of times (we'll say B times) to get $t_1^*, ..., t_B^*$
4. Look at distribution of the permuted test statistics (t_i^*'s) – if the original statistic t lies within the $\alpha/2$ and $1 - \alpha/2$ quantiles, we fail to reject H_0. Otherwise we reject.

Permutation Test Outline

Permutation Test for testing equality of group means:

1. Compute original test statistic $t = \hat{\mu}_1 - \hat{\mu}_2$
2. Shuffle groups (permute group labels) and (re)compute $t_1^* = \hat{\mu}_1^* - \hat{\mu}_2^*$
3. Repeat step (2) a large number of times (we'll say B times) to get $t_1^*, ..., t_B^*$
4. Look at distribution of the permuted test statistics (t_i^*'s) – if the original statistic t lies within the $\alpha/2$ and $1 - \alpha/2$ quantiles, we fail to reject H_0. Otherwise we reject.

Note: The q quantile is simply the value T such that $100 \times q\%$ of the observed values of the t_i^* are less than or equal to T.

Other kinds of permutation tests

Permutation tests refer to the style of hypothesis test $\implies$ can be used in many situations far beyond the problem of testing for equality of group means. Let's look at an example.

Other kinds of permutation tests

Permutation tests refer to the style of hypothesis test $\implies$ can be used in many situations far beyond the problem of testing for equality of group means. Let's look at an example.

Suppose I have two variables X_1 and X_2 and I want to test if they're correlated:

$$H_0 = \text{corr}(X_1, X_2) = r = 0$$
$$H_1 = \text{corr}(X_1, X_2) = r > 0$$

How could I do this – what would I be permuting?

Permutation Test for Correlation

Permutation Test for testing correlation:

1. Compute original test statistic $r = \widehat{\text{corr}}(X_1, X_2)$
2. Fix the order of X_1, shuffle the order of X_2, and (re)compute $r_1^* = \widehat{\text{corr}}(X_1, X_2^*)$
3. Repeat step (2) a large number of times (say B times) to get $r_1^*, \ldots, r_B^*$
4. Look at distribution of the permuted test statistics (r_i^*'s) – if the original statistic r lies below the $1 - \alpha$ quantile, we fail to reject H_0. Otherwise, we reject.

Permutation Test for Variable Importance

More generally, suppose I build a model with p predictors and I want to know if some predictor – say X_1 – is important. How could we (informally) use this permutation idea to test this?

Permutation Test for Variable Importance

More generally, suppose I build a model with p predictors and I want to know if some predictor – say X_1 – is important. How could we (informally) use this permutation idea to test this?

1. Build model with original data and compute the error Err_0

Permutation Test for Variable Importance

More generally, suppose I build a model with p predictors and I want to know if some predictor – say X_1 – is important. How could we (informally) use this permutation idea to test this?

1. Build model with original data and compute the error Err_0
2. Fix $X_2, \ldots, X_p$, permute the values of X_1, rebuild the model and (re)compute the error Err_1^*

Permutation Test for Variable Importance

More generally, suppose I build a model with p predictors and I want to know if some predictor – say X_1 – is important. How could we (informally) use this permutation idea to test this?

1. Build model with original data and compute the error Err_0
2. Fix $X_2, \ldots, X_p$, permute the values of X_1, rebuild the model and (re)compute the error Err_1^*
3. Repeat step (2) a large number of times (say B times) to get $\text{Err}_1^*, \ldots, \text{Err}_B^*$

Permutation Test for Variable Importance

More generally, suppose I build a model with p predictors and I want to know if some predictor – say X_1 – is important. How could we (informally) use this permutation idea to test this?

1. Build model with original data and compute the error Err_0
2. Fix $X_2, \ldots, X_p$, permute the values of X_1, rebuild the model and (re)compute the error Err_1^*
3. Repeat step (2) a large number of times (say B times) to get $\text{Err}_1^*, \ldots, \text{Err}_B^*$
4. Look at distribution of model errors when X_1 is permuted. If Err_0 lies above the α quantile of the permuted errors, then the original model isn't doing significantly better with the true values of X_1, so we could conclude X_1 isn't that important.

Notes on Permutation Tests

- Determining which quantiles of the permutation distribution to compare the original test statistic with is going to depend on what kind of test you're doing – see last three examples (group means - middle quantiles; correlation - upper quantile; variable importance - lower quantile)
- We can actually get p-values for these tests. Recall a p-value is just the probability of seeing a test statistic *more extreme* than what you observed if H_0 is true
  - We can use the permutation distribution to estimate this $\implies$ percentage of permuted test statistics more extreme than that calculated on original data

P-value Example

Example: Suppose we're doing a permutation test for correlation. Our original test statistic is $r = 0.31$. We do 1000 permutations, each time calculating the (permuted) test statistics to get $r_1^*, \ldots, r_{1000}^*$.

Say that of those 1000 permuted test statistics, 6 are greater than 0.31. Then our p-value would be

$$p = \frac{\text{\# More Extreme Test Statistics}}{\text{Total \# Permuted Test Statistics}} = \frac{6}{1000} = 0.006$$

Notes on Permutation Tests

- The theory behind permutation tests is actually quite involved, so we aren't getting into that here

- The formal development of permutation tests usually requires that the hypotheses be something like

$$H_0 = \mathscr{F}_1 = \mathscr{F}_2$$
$$H_1 = \mathscr{F}_1 \neq \mathscr{F}_2$$

  in the case where we're testing equality between two groups, so $\mathscr{F}_1$ would be the distribution of the data from Group 1 and $\mathscr{F}_2$ the distribution of the data from Group 2

  - The *power* of the test would then depend on the choice of test statistic

Notes on Permutation Tests

- Recall that the *power* of a hypothesis test is defined as the probability of (correctly) rejecting the null hypothesis H_0 when it is false.

- The idea behind permutation tests is actually quite simple – you might ask why we even bother with the parametric-style hypothesis tests you learned in intro stats
  - Permutation tests suffer from lower power than their parametric counterparts
  - Keep in mind though that many times there is no nice parametric counterpart – back to the idea of needing to know the distribution of a statistic/estimator

Notes on Permutation Tests

Permutation tests are very often used to test group differences in small sample size settings. What can go wrong here?

1. Even if H_0 is true, might be an outlier in one of the groups
   - Could cause you to falsely reject H_0 – parametric counterparts (usually t-tests) would also suffer
2. Even if H_0 is true, all samples in one group might be smaller than all samples in another
   - Likely would cause you to falsely reject H_0; t-tests would also suffer, but may be able to overcome it if differences were small enough

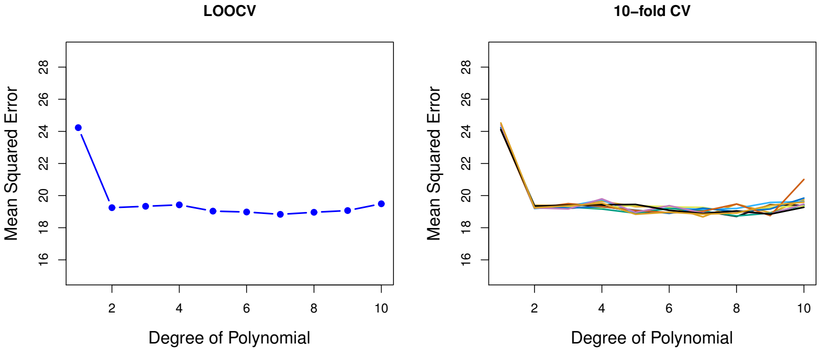
Review of Recent Material

Cross-validation

- *k*-fold Cross-validation gives us a “better” (more robust) way of estimating generalizability error of a particular model (problems with training error and test error based on one split)

- Very useful in terms of deciding which models are outperforming others
  - Can be used to compare across different kinds of models (e.g. Logistic Regression vs. LDA v.s. QDA) or choosing the “best" from a class of models (e.g. choosing the value of a tuning parameter like the *k* in *k*NN or the number of terms in linear models)

Cross-validation



ISLR Fig. 5.4: Cross-validation error (MSE) from a linear model when different degrees of polynomial for one covariate (feature) are included in the model. Left: LOOCV. Right: Several different runs of 10-fold CV.

Bootstrapping and Permutation Tests

- In order to do any kind of inference (CIs or hypothesis tests), we need to know something about the distribution of a statistic/estimator
- In many cases, there are no mathematical results that tell us what the distribution is, or the data do not satisfy the assumptions necessary in order to assume a particular distribution
- In such instances, the bootstrap and permutation tests can be used in order to perform familiar kinds of statistical inference (CIs = bootstrap; Hypothesis Tests = bootstrap or permutation tests)