Hypothesis Testing: Permutation Tests

Hypothesis Testing

Recall:

- Start by assuming a skeptical null hypothesis and state the alternative hypothesis you are testing for
- Set a significance level for how unusual your observation must be to reject the null
- Gather data
- Under the assumption of the null, see how unusual the data you observed would be (i.e. find the *p*-value)
- If highly unusual (i.e., below the significance level), reject the null hypothesis in favor of the alternative hypothesis

Hypothesis Testing

How do we determine how unusual our data is?

- Can be done analytically if we know the sampling distribution of the statistic of interest
- But, this requires assumptions - approximate normality or large enough sample size
- While the sampling distribution of the mean is easy to find, this is not true for all statistics

Permutation testing allows us to bypass both the assumptions required or the derivation necessary to use the true sampling distribution of our statistic.

Permutation Testing

Example: We suspect that there will be more crashes on average on Saturdays compared to Sundays in Davidson County.

With permutation testing, we must use a strong null hypothesis:

H_0: The distributions of crashes on Saturday and Sunday are identical

Since we are speculating that there are more crashes on Saturdays, we will have a one-tailed alternative hypothesis.

H_1: The distribution of crashes on Saturday has a larger mean than the distribution of crashes on Sunday

Permutation Testing

We also need to set the **significance level**.

That is, how unusual does our observed data under the assumption of the null hypothesis for us to reject that null hypothesis.

For this example, we'll use the 5% significance level.

That is, if what we observe would happen less than 5% of the time when the null hypothesis is true, then we'll reject the null hypothesis in favor of the alternative hypothesis.

Permutation Testing

Original Data

Saturday		
82	61	72
98	69	71
64	81	78
53	83	91

Sunday		
44	75	63
89	68	72
66	73	52
54	76	82

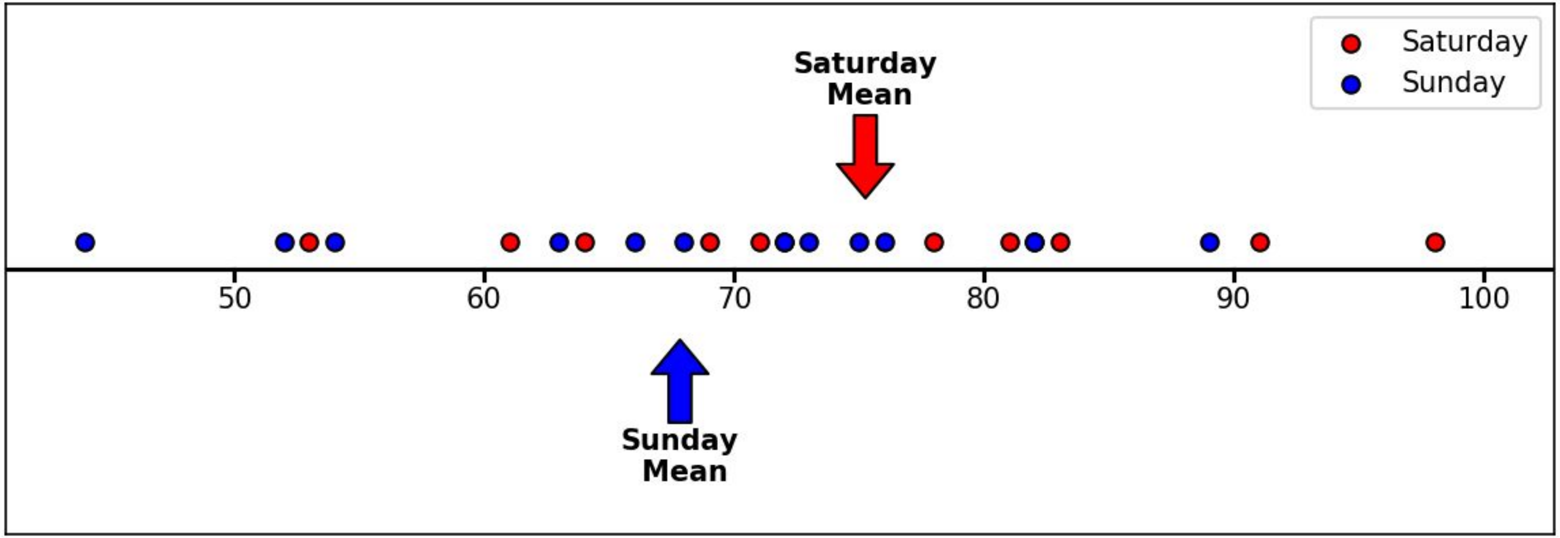
Mean = 75.25

Mean = 67.83

Observed Difference: 75.25 - 67.83 = 7.42

Permutation Testing

Original Data



Observed Difference: 75.25 - 67.83 = 7.42

Permutation Testing

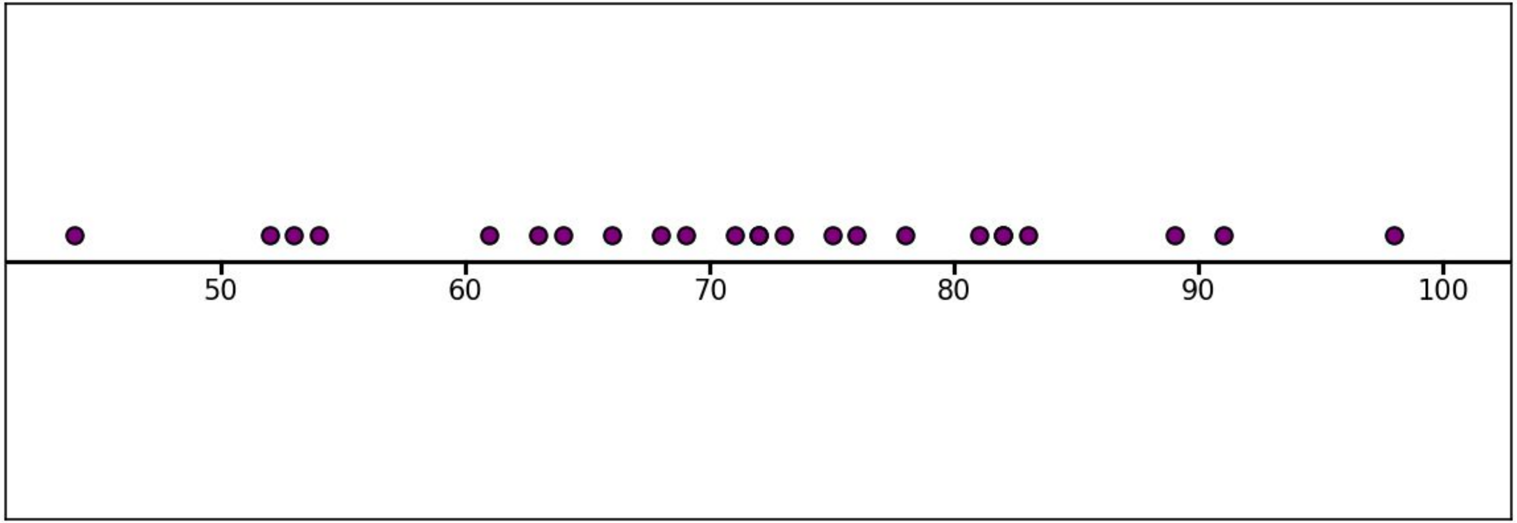
Saturday/Sunday					
82	61	72	44	75	63
98	69	71	89	68	72
64	81	78	66	73	52
53	83	91	54	76	82

Under the assumptions of the null hypothesis we are not looking at two samples but instead a single sample from the Saturday/Sunday distribution.

That is, the label does not impart any additional information.

Permutation Testing

Original Data Under Null Hypothesis



Permutation Testing

Another way to interpret the null hypothesis is that the labels are irrelevant.

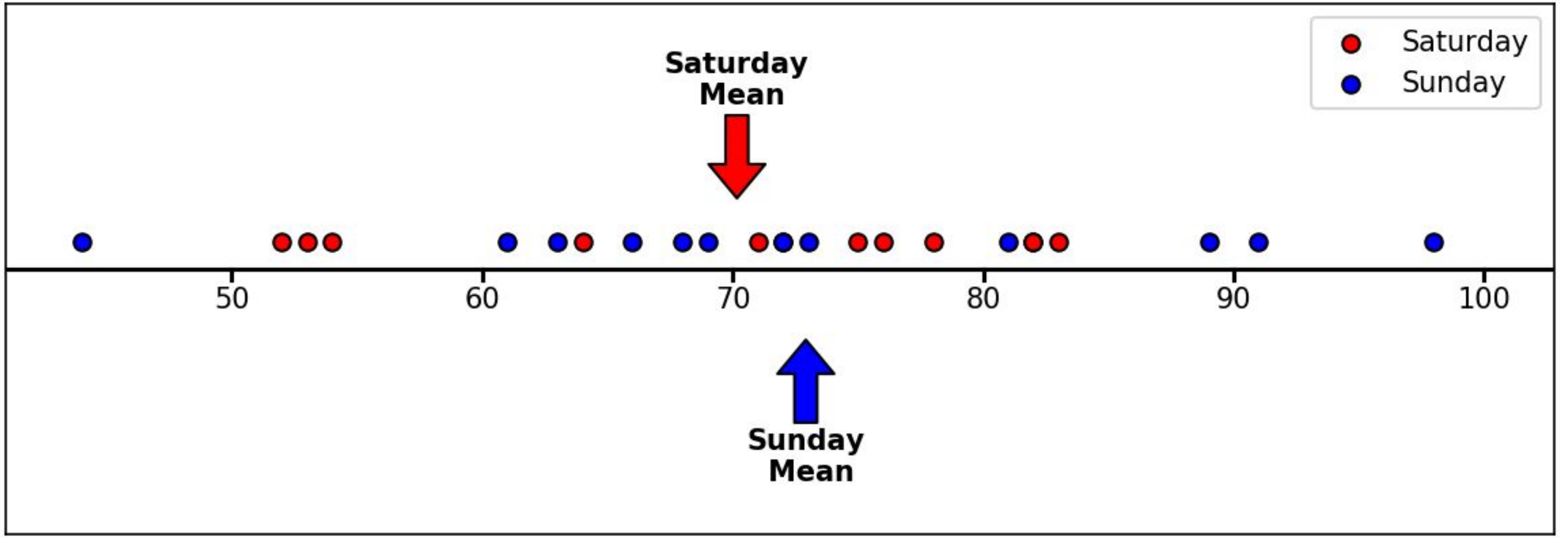
Even if we randomly assigned labels, we would still have a valid sample.

Let’s see what it would look like in a couple of cases if we did just randomly assign labels.

This technique is called permutation testing because we are permuting the observed values.

Permutation Testing

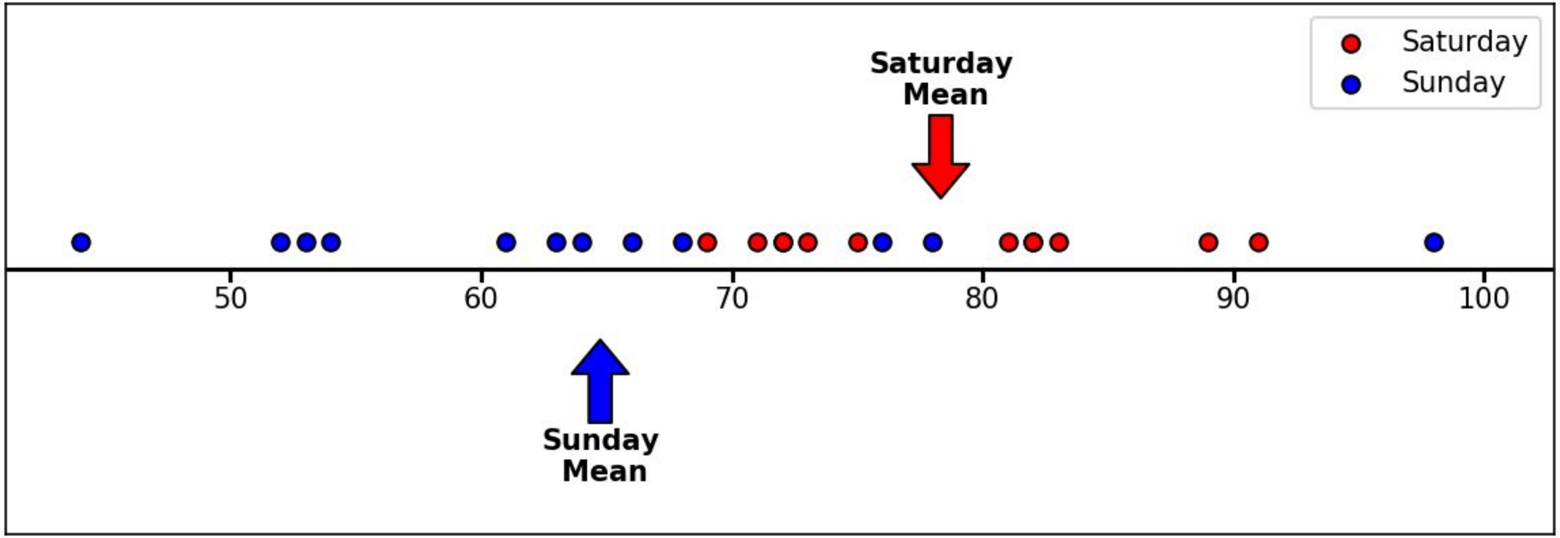
Permuted Data:



Difference: 70.17 - 72.92 = -2.75

Permutation Testing

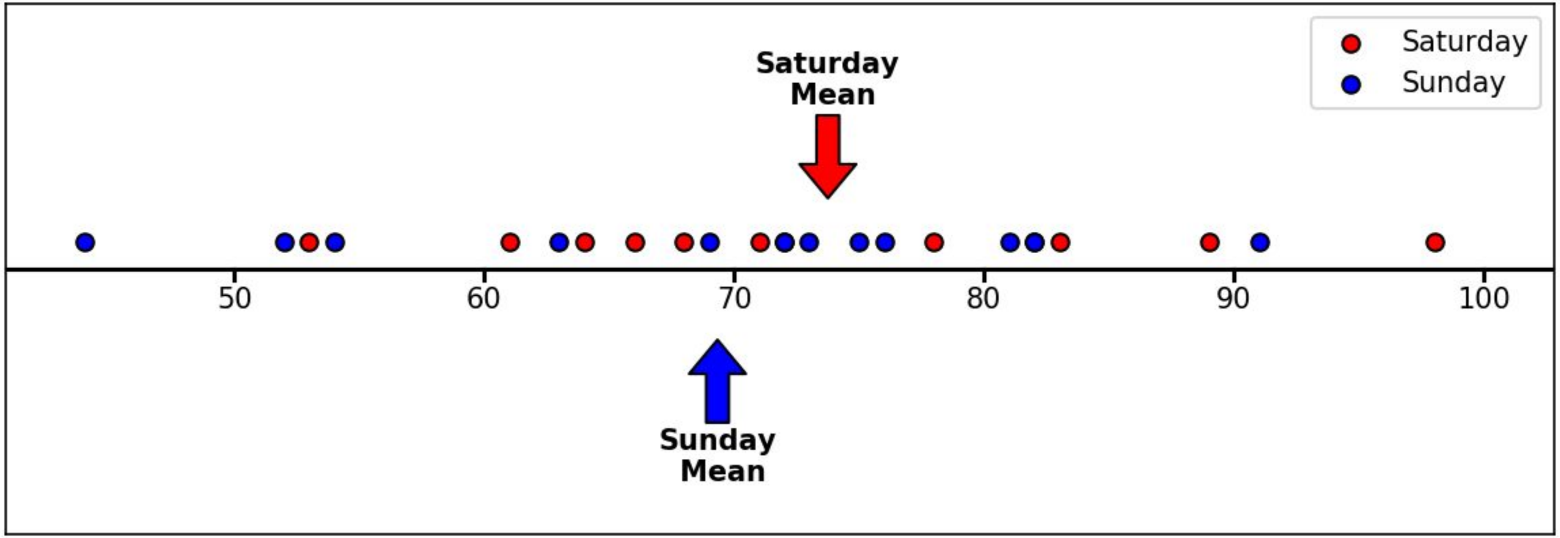
Another possible permutation:



Difference: 78.33 - 64.75 = 13.58

Permutation Testing

Another possible permutation:



Difference: 73.75 - 69.33 = 4.42

Permutation Testing

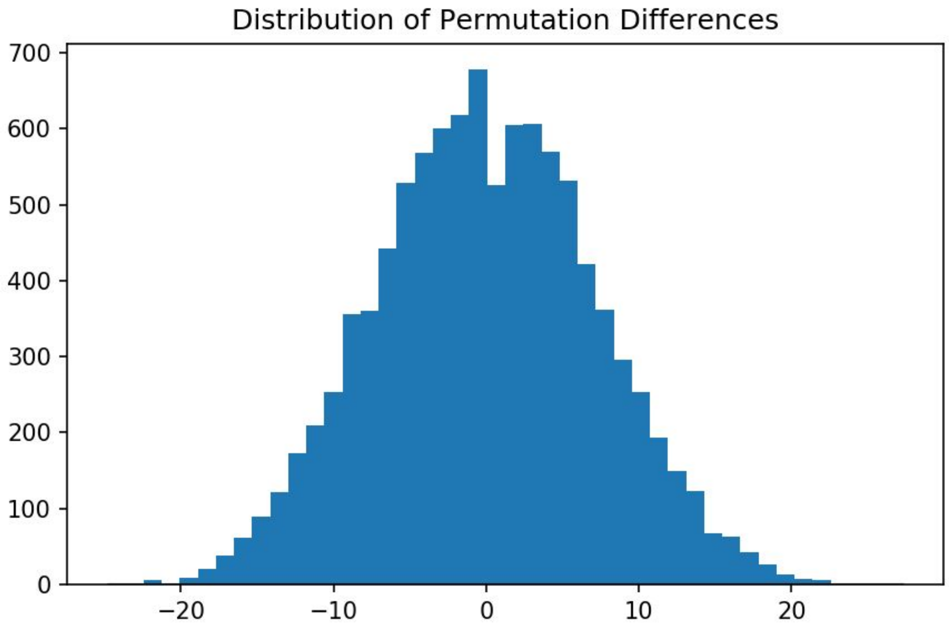
To determine whether or not to reject the null hypothesis, we need to measure how "unusual" our observed data would be if the null hypothesis were true.

In permutation testing, we do this by looking at all of the possible assignments of labels (here, Saturday/Sunday) and see for what percentage of the time we get a more extreme difference in means between groups than what we observed.

Note that there are 2,704,156 possible ways to shuffle the Saturday/Sunday labels, so we usually just take 10,000 or so random shufflings and use this to approximate the distribution of differences.

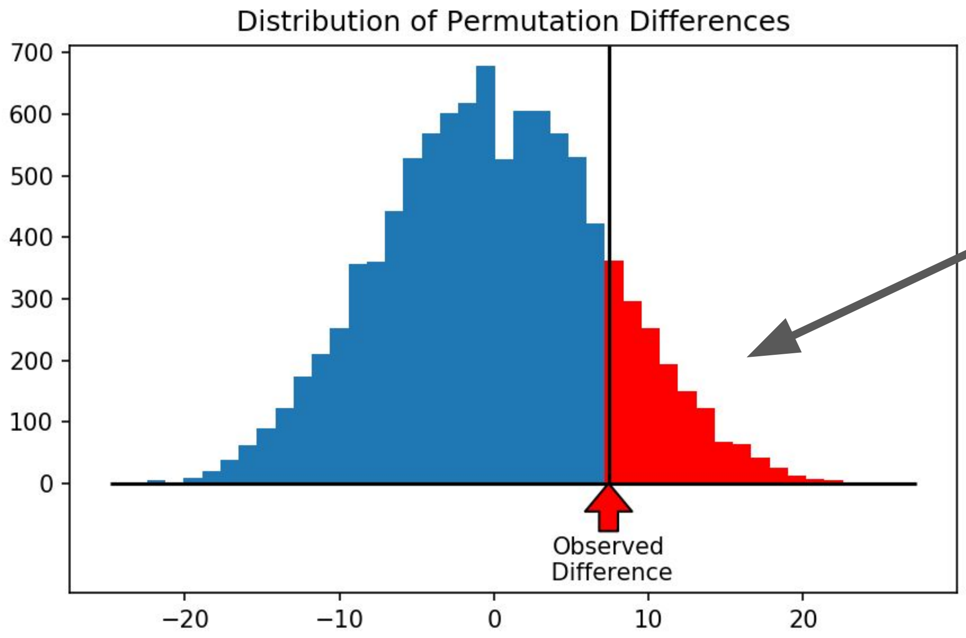
Permutation Testing

If we shuffle labels 10,000 times, we end up with a distribution of differences that looks like this:



Permutation Testing

We need to ask how unusual what we observed was. That is, how often did we end up with a difference at least as large as what we observed.



Red Proportion = 0.1509

Permutation Testing

What we see is that if the distribution on Saturday and Sunday was identical, we would obtain a difference at least as large more that 15% of the time.

This is not *that* unusual and does not meet the 5% threshold to reject the null hypothesis.

We must conclude that our data does not provide enough evidence to conclude that there is a difference in the distribution of crashes on Saturdays vs. Sundays.