Resampling Methods

The Best Way to Test for a Difference in Means (~~Independent Samples~~)

- As Kupe understands it (heard it at a conference), originally statistical testing was based on resampling methods. So, for example, to test for a difference in two means, we would do resampling as we did for "**Green Test vs. Yellow Test**".

- Back in the day, computers were not around, so resampling really wasn't feasible. The two-sample *t* test was a workaround.

- In statistics, an exact test is a test were all assumptions upon the derivation of the distribution of the test statistic are completely met. This usually isn't the case.

- In statistics, an approximate test is one in which the approximation may be made as close as desired by making the sample size large enough.

- In statistics, a parametric test is an exact test in which the parametric assumptions are fully met.

- In practice, nonparametric tests are reserved for tests that do not rest on parametric assumptions.

- In practice, most implementations of nonparametric tests use software that use asymptotical algorithms for obtaining the significance value (P-value), which makes implementation non-exact.

- A permutation test (randomization test / re-randomization test / exact test) is a test in which the distribution of the test statistic under the null hypothesis is obtained by calculating all possible values of the test statistic under rearrangements of the labels on the observed data points.

- Permutation tests lead to exact significance levels.
- Confidence intervals can be derived from the tests.
- The theory evolved in the 1930s from the works of Fisher and Pitman.
- The t test, the F test, the Z test for proportions, and the χ^2 test are all obtained from theoretical probability distributions – most likely, the assumptions and conditions are probability not met precisely, so in reality, we've been approximating P-values all along.

Testing the Difference in Two Means Using Resampling (Independent Samples) on StatCrunch

1. Load up your dataset first.
2. Hit "**StatCrunch**" →"**Applets**" → "**Resampling**" → "**Randomization Test for Two Means**"
3. Select your columns as sample 1 and sample 2 or fill in the appropriate menus.
4. Experts say that we need at least 3000 randomizations to get a good feel for the P-value. **Hit the "1000 times" button at least three times.**
5. The null hypothesis is that "**the groups have the same mean**"
6. The alternative hypothesis can be "**the groups have different means**" or "**the first group has a higher mean**" or "**the first group has a lower mean**".
7. The P-value is given the in the "**Results**" window, which is a tally of the number of times the resampling gave a more extreme difference.

Example: Does Kupe do worse during his 3rd game each week at bowling? Fire up the "***Kupe Bowling***" dataset and run the permutation test. Show all steps.

$H_0: \mu_{GAME\ 1} = \mu_{GAME\ 3}$ vs. $H_A: \mu_{GAME\ 1} > \mu_{GAME\ 3}$

$\bar{y}_{GAME\ 1} = 195.77$ and $\bar{y}_{GAME\ 3} = 178.19$

$n_{GAME\ 1} = 31$ and $n_{game\ 3} = 31$

Could this 17.58 pin difference have happened by chance?

After 10,000 randomizations P-Value = 0.0175 which means 1.75% of the time, randomizing would give a Game 1 mean that was more than 17.58 pins higher tha the game 3 mean.

1.75% is pretty unusual, so it is likely that his Game 1 mean exceeds his Game 3 mean (Reject Ho).

Explain in context the meaning of the *P*-value:

1.75% of the time, randomizing the bowling scores between Game 1 and 3 gave a difference larger than the one we actually observed, which was 17.58 pins.

Explain in context the type of error we **could** have made and what it would mean:

Since we rejected Ho, could have made a Type I error, with probability 1.75%. We concluded Kupe does better in Game 1, but really he doesn't.

Although the StatCrunch applet cannot do it currently, explain how we could derive a 95% confidence interval for the true difference in mean scores:

After 3000 + randomizations, find cutpoints in the histogram.

For a 95% 2-sided interval,

Put the smallest 2.5% of randomations as the left endpoint and the highest 2.5% as the right.

For a 95% one-sided interval, put the 5% cutpoint in the appropriate tail.

Example: Load up the "**Roller Coasters**" dataset on StatCrunch. Run the randomization test to determine if coasters with inversion have a different mean "***Drop***" compared to coasters without inversion.

Write out the hypotheses: $H_0: \mu_{WITH} = \mu_{WITHOUT}$

$H_A: \mu_{WITH} \neq \mu_{WITHOUT}$

Check side-by-side boxplots to get a sense of any difference in average drops:

Start the Applet:

Sample 1 in: "drop"
Where: Inversion = yes
Label: Inversion

Sample 2 in: "drop"
Where: Inversion = no
Label: No Inversion

Give the mean drop in roller coasters with inversion: 124.44 feet

Give the mean drop in roller coasters without inversion: 157.0686 feet

Give the difference in means: −32.628 feet

Now run the randomization 3000 times.

What is the *P*-value of this permutation test: 0.0233 Fairly low

What is your decision: Reject Ho

Write up a summary remark about roller coasters mean drop, comparing those with inversion to those without:

We are convinced that roller coasters with and without inversion differ on their mean height in the first drop.

Randomization Test for Two Proportions

- Previously, we tested for a difference in proportions by running a 2 proportion Z-test
- This method is based on $\hat{p}$ having an approximate Normal distribution, which is a very good fit if we expect to have at least 10 successes and 10 failures.
- The above method is not exact, though. For an exact test, we use the randomization test for two proportions.
- The method is similar to the randomization test for two means. Basically, we compare the observed difference in sample proportions to many simulated differences created by randomizing the data.
- If our actual difference is more extreme than most of the simulated differences, the P-Value will be low. The *P*-value is the proportion of simulated differences that are more extreme that the actual difference.

Example: Load up the "**General Social Survey 2008**" dataset on StatCrunch.

a. Is there evidence that the proportion of men earning a Bachelor's degree is different that the proportion of women earning that degree? What are the hypotheses?

$H_0: P_{men} = P_{women}$

$H_A: P_{men} \neq P_{women}$

b. Set up the applet on StatCrunch. "**Randomization Test for Two Proportions**".

Sample 1 in: HIGHEST DEGREE
Where: SEX=Male
Label: Male

Sample 2 in: HIGHEST DEGREE
Where: SEX=Female
Label: Female

Success: 3 – Bachelor

c. What proportion of men in the study have a Bachelor's? $\hat{p}_{men} = 18.62\%$

What proportion of women in the study have a Bachelor's? $\hat{p}_{women} = 16.65\%$

The difference in sample proportions is: 1.97%

This feels / doesn't feel statistically significant.

d. Randomize the data 3000 times. Interpret the *P*-value in terms of the number of more extreme randomizations and in terms of the problem. Reject or fail to reject?

*Kupe got (students will get different) 797/3000 = 26.57% or 797 times in 3000 Randomizations, the difference in proportions was more extreme than the actual 1.97% difference we observed. Since a larger difference easily can occur by chance, there is no evidence that our 1.97% difference was significant.

e. When randomizing, what were some of the biggest differences in proportions observed? Give the sample proportions (students answers will vary).

*Kupe got $\hat{P}_M = 14.21\%$, $\hat{P}_F = 20.40\%$ > −6.19% Diff

These are the *observed extremes when Randomizing.

$\hat{P}_M = 20.67\%$, $\hat{P}_F = 14.91\%$ > + 5.75% Diff.

As his biggest differences in the tails.

Example: Students' turn. Load up the "***Fall 2012 Survey 1 Student Data***" on StatCrunch to test if Math 127 parents exercise less than Math 127 non-parents. If you're raising children, is there less time for exercise? Students were asked if they rigorously exercised in the last 48 hours back in September of 2012.

a. Before running the applet, give the sample proportions of who exercises rigorously, parents and non-parents.

$\hat{P}_{parents} = \frac{4}{12} = 33.33\%$ $\hat{P}_{NON-PARENTS} = \frac{54}{105} = 51.43\%$

b. What concerns you about the amount of data collected? Why couldn't we run a two-sample z test for a difference in proportions?

For parents, only 4 exercisers, 8 non-exercisers, both under 10, so shouldn't run 2-Prop Z Test

c. Set up the applet:

Sample 1 in: Exercise
Where: Parent=Yes
Label: Parent

Sample 2 in: Exercise
Where: Parent=No
Label: Not a Parent

Success: Yes

Run it 3000 times.

d. What hypotheses are you testing? It is a one-sided test.

$H_0: P_{PARENTS} = P_{NON-parents}$
$H_A: P_{parents} < P_{NON-PARENTS}$

e. What is the P-value of the test? What can you conclude?

*Kupe got P-Value = 0.1967

Fail to Reject! No evidence the Parents exercise less than the non-parents.

Randomization Test for Correlation

- Typically, we use the correlation coefficient ___r___ to measure the strength and direction of the linear relationship between two quantitative variables.

- Though we never studied it in Math 127 or Math 128 at Cecil College, you can run a hypothesis test on a correlation to test if it is ___statistically significant___. This test is actually a t test, so there are requirements and assumptions that must be met.

- Since almost every correlation is statistically significant, your instructor thinks this boils down to common sense and context of the problem. Many times, a meaningless correlation (something like 0.3121, e.g.) ***will be*** statistically significant.

Quick Example 1 (Statistically Significant Correlation)

Load up "**2010 Hurricanes**" and start the Randomization Test for Correlation applet.

We'd like to predict "***Pressure***" of a hurricane based on its "***Max Wind***". What is the actual correlation?

$r = -0.937$

You'd be testing these hypotheses:

$H_0: \rho = 0$
$H_A: \rho \neq 0$

ρ is greek letter "rho", equivalent of r.

Run the applet and confirm that the actual correlation is statistically significant.

P-Value = 0. Never did we randomize and get a more extreme correlation.

(Had a few in the ± 0.6 range, no where close to −0.937)

Quick Example 2 (A Statistically Significant but Not a Meaningful Correlation)

Load up the "**Kupresanin Quiz 1 Data**" dataset on StatCrunch. Cecil students responded to our online survey, and among other things, were asked their "***Age***" and "***Number of Work Hours***".

What is the correlation between "***Age***" and "***Work***"? $r = +0.237$

Looking at the scatterplot and using the value of r, is there much of a relationship? Not Really

Run the applet to test if the relationship is significant.

P-Value = 0.011 suggesting statistical significant.
Relationship is weak and not meaningful

Quick Example 3 (A Not Statistically Significant and Not Meaningful Correlation)

Same dataset. What is the correlation between the number of Facebook friends and the number of credit hours students are taking?

$r = -0.0093$

Run the applet to see if the relationship is significant.

P-Value = 0.944

Not significant, not meaningful.