Permutation Tests

Noa Haas
Statistics M.Sc. Seminar, Spring 2017
Bootstrap and Resampling Methods

The Two-Sample Problem

We observe two independent random samples:

$$\begin{aligned} F \rightarrow \quad & \boldsymbol{z} = (z_1, z_2, \ldots, z_n) \ \ independently\ of \\ G \rightarrow \quad & \boldsymbol{y} = (y_1, y_2, \ldots, y_m) \end{aligned}$$

And we wish to test the *null hypothesis* of no difference between F and G,

$$H_0 : F = G$$

The Two-Sample Problem

- An *hypothesis test*, of which a permutation test is an example, is a formal way of deciding whether or not the data decisively reject H_0
- An hypothesis test begins with a *test statistic* $\hat{\theta}$. For example, the mean difference - $\hat{\theta} = \bar{z} - \bar{y}$
- We will assume here that if H_0 is not true, we expect to observe large values of $\hat{\theta}$ than if H_0 is true – the larger the value of $\hat{\theta}$ we observe, the stronger the evidence against H_0
- The *achieved significance level* is defined as:

$$ASL = Prob_{H_0}\{\hat{\theta}^* \geq \hat{\theta}\}$$

  - Where the random variable $\hat{\theta}^*$ has the distribution of $\hat{\theta}$ if H_0 is true
- The hypothesis test of H_0 consist of computing ASL, and seeing if it is smaller than some predetermined threshold (.05, for example)

The Two-Sample Problem

A traditional hypothesis test for the problem:

$$F = N(\mu_T, \sigma^2), \qquad G = N(\mu_C, \sigma^2)$$

Then the null hypothesis is

$$H_0: \mu_T = \mu_C$$

And we know that under H_0

$$\hat{\theta} = \bar{z} - \bar{y} \sim N\left(0, \sigma^2\left(\frac{1}{n} + \frac{1}{m}\right)\right)$$

So we can compute ASL:

$$ASL = 1 - \Phi\left(\frac{\hat{\theta}}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}\right)$$

The Two-Sample Problem

How do we calculate the ASL when the null hypothesis H_0 doesn't specifies a single null distribution?

In most problems the null hypothesis $F = G$ leaves us with a family of possible null hypothesis distributions, rather than just one.

For instance, when $\hat{\theta} \sim N\left(0, \sigma^2(1\backslash n + 1\backslash m)\right)$, the null hypothesis family includes all normal distributions with expectation 0 (since we don't know the true value of σ).

In the normal situation, we could approximate the ASL using Student's t distribution. But this only works in the normal case.

Permutation Tests

- A permutation test (also called a randomization test, or an exact test) is a type of statistical significance 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.
- The idea was introduced by R.A. Fisher in the 1930's, more as a theoretical argument supporting Student's t-test than as a useful statistical method in its own right.
- The basic idea is attractively simple and free of mathematical assumptions.

Fisher's Permutation Test

We define $N = n + m$, and let $\boldsymbol{v} = (v_1, v_2, \ldots, v_N)$ be the combined and ordered vector of the observed values $\boldsymbol{z}, \boldsymbol{y}$. Also let $\boldsymbol{g} = (g_1, g_2, \ldots, g_N)$ be the vector that indicates the observations' original labels.

There are $\binom{N}{n} = \frac{N!}{n!m!}$ possible $\boldsymbol{g}$ vectors.

Permutation Lemma: *Under $H_0: F = G$, the vector $\boldsymbol{g}$ has probability $1/\binom{N}{n}$ of equaling any one of its possible values*

Let $\boldsymbol{g}^*$ indicate any one of the $\binom{N}{n}$ possible vectors of $\boldsymbol{g}$, and $\hat{\theta}^* = \hat{\theta}(\boldsymbol{g}^*)$

The permutation ASL is defined as:

$$ASL_{perm} = Prob_{perm}\{\hat{\theta}^* \geq \hat{\theta}\} = \#\{\hat{\theta}^* \geq \hat{\theta}\}/\binom{N}{n}$$

Approximating ASL_{perm}

1. Choose B independent vectors $\boldsymbol{g}^*(1), \boldsymbol{g}^*(2), \ldots, \boldsymbol{g}^*(B)$, each being randomly selected **without replacement** from the set of all $\binom{N}{n}$ possible such vectors
2. Evaluate the permutation replications of $\hat{\theta}$ corresponding to each permutation vector, $\hat{\theta}^*(b), b = 1,2, \ldots, B$
3. Approximate ASL_{perm} by

$$\widehat{ASL}_{perm} = \#\left\{\hat{\theta}^*(b) \geq \hat{\theta}\right\}/B$$

How should we choose B?

The Mouse Data

Sixteen mice were randomly assigned to a treatment group or a control group. Shown are their survival times, in days, following a test surgery.

Group	Data			Mean	Estimated Standard Error
Treatment:	94	197	16	86.86	**25.24**
	38	99	141		
	23				
Control:	52	104	146	56.22	**14.14**
	10	51	30		
	40	27	46		

Here, $\hat{\theta} = \bar{z} - \bar{y} = 30.63$, and the "pooled variance" estimator is $\bar{\sigma}$

$$= \left\{ \frac{\left[\sum_{i=1}^{n} (z_i - \bar{z})^2 + \sum_{j=1}^{m} (y_j - \bar{y})^2 \right]}{[n+m-2]} \right\}^{1/2} = 54.21$$

So the ASL according to Student's t-test is:

$$ASL = Prob\left\{ t_{14} > \frac{30.63}{54.21\sqrt{1/9 + 1/7}} \right\} = .141$$

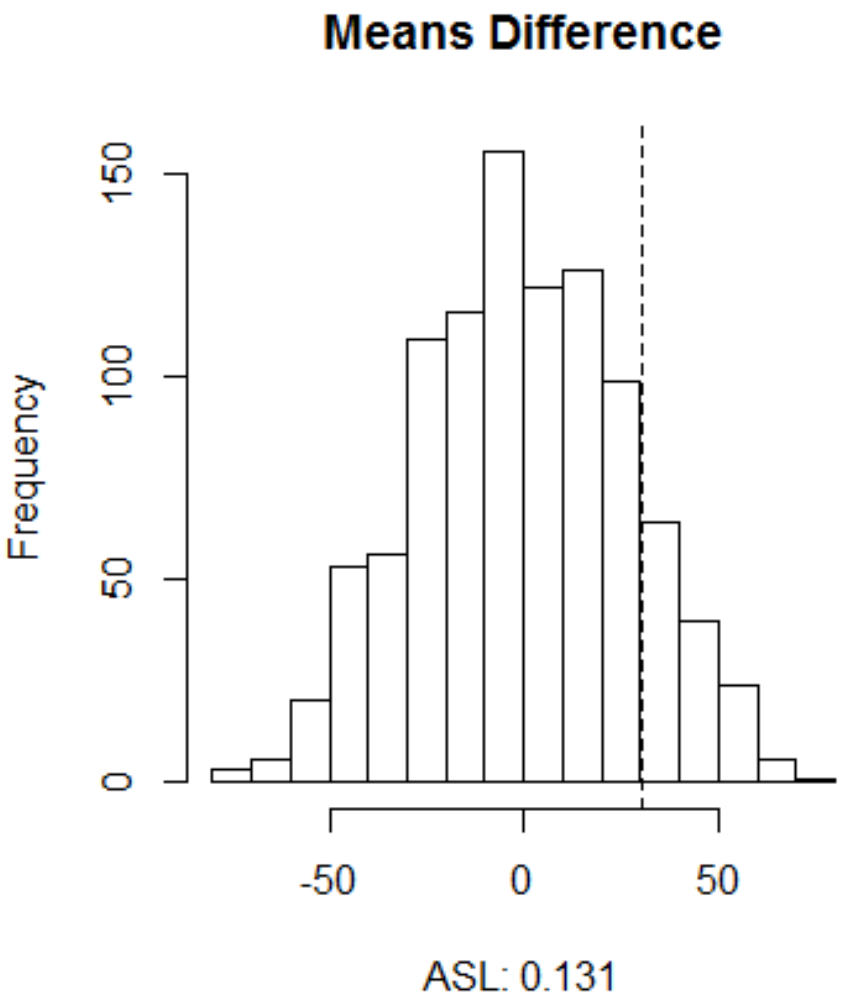
The Mouse Data

Based on $B = 1000$ permutations, the estimated ASL_{perm} is

$$\widehat{ASL}_{perm} = \frac{131}{1000} = .131$$

Although there are no normality assumptions underlining ASL_{perm}, the result is close to the t-test's.

Fisher's main point in introducing permutation tests was to support the use of Student's test in non-normal applications.



Number of Replications

Let $A = ASL_{perm}$ and $\hat{A} = \widehat{ASL}_{perm}$, then $B * \hat{A}$ is the number of time that the $\hat{\theta}^*(b)'s$ values exceeded the observed $\hat{\theta}$, and so has a binomial distribution:

$$B * \hat{A} \sim Bin(B, A)$$

$$E(\hat{A}) = A; Var(\hat{A}) = \frac{A(1-A)}{B}$$

The coefficient of variation of $\hat{A}$ is

$$CV_B(\hat{A}) = \left[\frac{(1-A)/A}{B}\right]^{1/2}$$

Controlling $CV_B(\hat{A})$ means controlling the affect of the Monte Carlo error on our estimation of ASL_{perm}

Number of Replications

$[(1-A)/A]^{1/2}$ as a function of A:

A:	.5	.25	.1	.05	.025
$[(1-A)/A]^{1/2}$:	1	1.73	3.00	4.36	6.24

Number of permutations required to make $CV\left(\widehat{ASL}\right) = k$, as a function of the true ASL_{perm}:

ASL_{perm}:	.5	.25	.1	.05	.025
B for $k = .1$:	100	299	900	1901	3894
B for $k = .2$:	25	75	225	476	974

Permutation Test's Accuracy

In general, if $H_0: F = G$ is true, then

$$Prob_{H_0}\{ASL_{perm} < \alpha\} = \alpha$$

For any value of α between 0 and 1, except for small discrepancies caused by discreteness of the permutation distribution.

This applies to any statistic $\hat{\theta}$ (median, trimmed means…)

"Accuracy" means that ASL_{perm} won't tend to be misleadingly small when H_0 is true (maintaining type I error).

But what about the power?

Choosing a poor test statistic $\hat{\theta}$ will result in small probability of rejecting H_0 when it is false – low power

Other Test Statistics

We can look at the ratio of estimated variances:
$\hat{\theta} = log\left(\hat{\sigma}_z^2/\hat{\sigma}_y^2\right)$, where, in such case we have no a priori reason to for believing $\hat{\theta}$ will be greater than one rather than less than one.

In this situation, it is common to compute *two-sided* ASL:

$$\widehat{ASL}_{perm}(two-sided) = \#\left\{\left|\hat{\theta}^*(b)\right| > |\hat{\theta}|\right\}/B$$

For example, in the mouse data, after 1000 permutation replications of $\hat{\theta} = log\left(\hat{\sigma}_z^2/\hat{\sigma}_y^2\right)$ (log transformation has no effect on the permutation results), we get

$$\widehat{ASL}_{perm}(two-sided) = .305$$

Other Test Statistics

We run 4 different permutation tests, according to 4 different test statistics – $\hat{\theta}_1 = \bar{z} - \bar{y}; \hat{\theta}_2 = \bar{z}_{.15} - \bar{y}_{.15}; \hat{\theta}_3 = \bar{z}_{.25} - \bar{y}_{.25}; \hat{\theta}_4 = \bar{z}_{.5} - \bar{y}_{.5}$, each using same 1000 replications. Then we rank the evidence against H_0 according to

$$\hat{\phi} = \min_k \left\{ \widehat{ASL}_k \right\}, k = 1,2,3,4$$

Small values of $\hat{\phi}$ more strongly contradict H_0, but it isn't true that the ASL_{perm} based on $\hat{\phi}$ equals to $\min_k \left\{ \widehat{ASL}_k \right\}$

Instead, for each k and $b = 1, \dots, B$ we calculate

$$A_k^*(b) = \frac{1}{B} \sum_{i=1}^{B} I_{\left\{ \hat{\theta}_k^*(i) \geq \hat{\theta}_k^*(b) \right\}}$$

Then let

$$\hat{\phi}^*(b) = \min_k \{ A_k^*(b) \}$$

Which are genuine permutation replications of $\hat{\phi}$ (not obvious, but true..), and therefor

$$\widehat{ASL}_{perm} = \#\left\{ \hat{\phi}^*(b) \leq \hat{\phi} \right\} / B$$

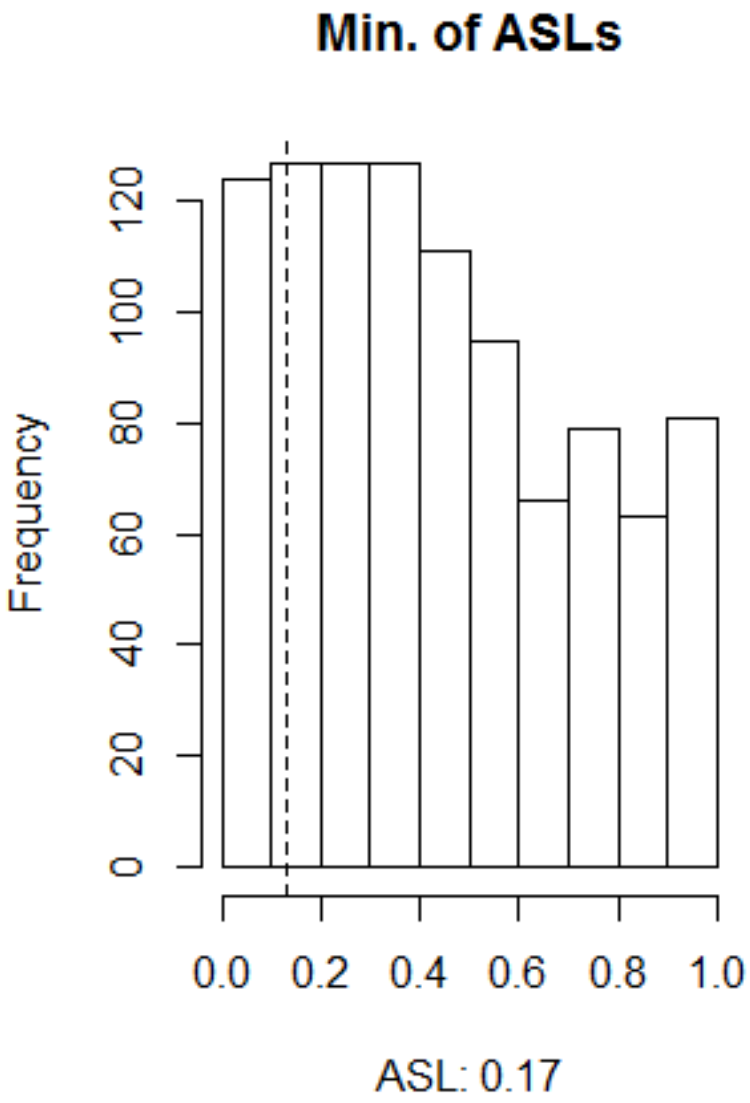
Other Test Statistics

In the mouse data, we get

$$\hat{\phi} = \min_k \{\widehat{ASL}_k\} = \widehat{ASL}_1 = .131$$

And 170 of the 1000 values of $\hat{\phi}^*(b)$ are less than .131, giving

$$\widehat{ASL}_{perm} = .17$$



Relationship to the Bootstrap

Permutation methods tend to apply to only a narrow range of problems. However when they apply, as in testing $F = G$ in a two-sample problem, they give gratifyingly exact answers without parametric assumptions.

The bootstrap distribution was originally called the "combination distribution". It was designed to extend the virtues of permutation testing to the great majority of statistical problems where there is nothing to permute.

Relationship of Hypothesis Tests to CI and the Bootstrap

We can use confidence intervals to calculate ASLs. For

$$H_0: F = G; \hat{\theta} = \bar{z} - \bar{y}$$

If we choose α so that $\hat{\theta}_{lo}$, the lower end of the $1 - 2\alpha$ CI for θ, exactly equals 0. Then

$$Prob_{\theta=0}\{\hat{\theta}^* \geq \hat{\theta}\} = \alpha^{***}$$

In other words, values of θ that are smaller than $\hat{\theta}_{lo}$ are implausible according to our data.

So instead we can ask ourselves what value of α will make the lower end of the bootstrap confidence interval equal to zero?

For the percentile method, the answer is

$$\alpha_0 = \#\{\hat{\theta}^*(b) < 0\}/B = \widehat{ASL}_{\%}$$

* Note that these are $\hat{\theta}^*(b)$ based on the bootstrap non-parametric sampling procedure for constructing CI's and not for estimating $\hat{F}_0$

Relationship of Hypothesis Tests to CI and the Bootstrap

- The permutation ASL is exact, while the bootstrap ASL is approximated.
- The bootstrap histogram is centered near $\hat{\theta}$, while the permutation histograms are centered near 0.
- The bootstrap ASL tests the null hypothesis $\theta = 0$, while the permutation ASL test $F = G$. The first is more realistic/practical. Still, permutation test usually perform reasonably well even when $F = G$ is far from reasonable.
- The combination of a point estimate and a confidence interval is usually more informative than just a hypothesis test by itself.