ExpEcon Methods: Resampling Methods Permutation Tests & Bootstrapping

ECON 8877

P.J. Healy

First version thanks to Sam Stelnicki

Updated 2023-11-17

Permutation Tests

Example: The Permutation Test

How it works:

`https://www.jwilber.me/permutationtest/`

Permutation Test

- Fisher (1935), Pitman (1937,1938)
- Resampling method where we use our data in different orders (without replacement) to test for differences between populations
- Example was for sample means
- Could do exact same for sample medians, modes, variances...
    - *Any* statistic of a sample!

Assumptions & Properties

- Only assumption: observations are exchangeable
  - Joint dist'n: $F(Y_1, Y_2, Y_3) = F(Y_3, Y_1, Y_2)$
  - Same marginals, "symmetric" correlation
  - True if treatments are randomly assigned!
- Permutation test is always **valid**
- The issues are**power** and **exactness**
  - e.g., outliers can affect resampled distributions

Two-Sample Framework

Chung and Romano (2013)

- Sample 1: $X_1, \ldots, X_m$ i.i.d. from P
- Sample 2: $Y_1, \ldots, Y_n$ i.i.d. from Q
- Let $Z = (Z_1, \ldots, Z_N) = (X_1, \ldots, X_m, Y_1, \ldots, Y_n)$, $N = m + n$
- Model/hypothesis: $(P, Q) \in \mathcal{P}$
- Important example: $\bar{\mathcal{P}} = \{(P, Q) : \ P = Q\}$
- Permutations: $\pi : \{1, \ldots, N\} \to \{1, \ldots, N\}$, $\pi \in \mathbf{G}_N$
- Let $Z_\pi = (Z_{\pi(1)}, \ldots, Z_{\pi(N)})$
- Test statistic: $T_{m,n}(Z)$ (eg, $T_{m,n}(Z) = \frac{1}{m}\sum_{i=1}^{m} X_i - \frac{1}{n}\sum_{i=1}^{n} Y_i$)
- $T_{m,n}(Z_\pi)$ calculated after permuting via π
- Order all $T_{m,n}(Z_\pi)$: $T_{m,n}^{(1)} \leq T_{m,n}^{(2)} \leq \cdots \leq T_{m,n}^{(N!)}$
- Given α, threshold ranking is $k^* = (1 - \alpha)N!$
- Permutation test function:

$$\phi(Z) = \begin{cases} 1 & T_{m,n}(Z) > T_{m,n}^{(k^*)} \\ 0 & T_{m,n}(Z) < T_{m,n}^{(k^*)} \end{cases}$$

Two-Sample Framework

- If $(P, Q) \in \bar{\mathcal{P}} = \{(P, Q) : \; P = Q\}$ then the test is exact:

$$E_{P,Q}[\phi(X_1, \ldots, X_m, Y_1, \ldots, Y_n)] = \alpha$$

- But what if we assume $(P, Q) \in \mathcal{P}_0 \supset \bar{\mathcal{P}}$?
    - Permuted data no longer has the same distribution as original
- Test may not even be asymptotically exact
- Example: $\mathcal{P}_0 = \{(P, Q) : \; \mu(P) = \mu(Q)\}$, $T_{m,n}(Z) = \sqrt{N}(\bar{X}_m - \bar{Y}_n)$
- Romano (1990): Rejection rate higher than α even with $N \to \infty$ unless
    1. $m/n \to 1$ as $N \to \infty$, or
    2. variances of P and Q are equal
- Unbalanced samples: Rejecting the null might actually be due to different variances, not different means

Chung and Romano (2013) Correction

Chung and Romano (2013) offer a correction:

$$S_{m,n}(Z) = \frac{T_{m,n}(Z)}{V_{m,n}}$$

where

$$V_{m,n}(Z) = \sqrt{\frac{N}{m}\hat{\sigma}^2_m(X_1, \ldots, X_m) + \frac{N}{n}\hat{\sigma}^2_n(Y_1, \ldots, Y_n)}$$

For testing difference in means:

$$S_{m,n} = \frac{\sqrt{N}(\bar{X}_m - \bar{Y}_n)}{\frac{N}{m}S^2_X + \frac{N}{n}S^2_Y}$$

where

$$S^2_X = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X}_m)^2$$

Why? Distribution of $T_{m,n}(Z_\pi)$ is asymptotically normal with mean 0

Chung and Romano (2013)

- Different adjustments for different statistics
- Based on variance of the large-sample distribution, which is approximately normal
- Always divide $T_{m,n}$ by an estimator of that asymptotic variance
- Paper gives guidance for testing medians

Randomization Tests

- You run ALL possible combinations of the data, rather than just a random subset
- Permutation test is a subset of randomization test

Young (2019)

- Runs permutation tests for 53 different published AEA papers
- Finds 13-22% fewer significant results than the methods used in the papers
- This increases to 33-49% for multiple effects

Young (2019)

- Runs permutation tests on regression coefficients of the previous paper
- Uses the Wald statistic and t-statistics
- Also runs bootstrap and jackknifes for all of these papers

Young (2019)

- Key Takeaways: the design of experiments is really important for whether the p-values of resampling vs. statistical testing are similar
- Lots of treatments and interactions allows for more sensitivity to outliers and creates more volatility causing these methods to vary greatly

Bootstrapping

Bootstrapping - Efrom (1979)

Goal: estimate a parameter of a distribution. e.g. median

- Resample your collected *n*-sized data *with replacement* to produce M samples of n-sized data
- Each data point in your original sample has $\frac{1}{n}$ chance of being chosen
- Plot the distribution of observed parameter values.
- Estimate: mean of bootstrap dist'n
- Standard error: std. deviation of boostrap dist'n
- Confidence interval: 5th to 95th quantile
- Completely non-parametric

Bootstrapping Assumptions

- For the standard bootstrap method, observations are assumed to be independent
  - Block bootstrapping was developed to deal with correlated data
- Sample data needs to resemble the population its drawn from and sufficiently large
- Do not need to know the real distribution
- Sufficiently large: enough data to get to around 200 samples, but it's better to run as many bootstrap samples until your statistics converge

Bootstrapping Consistency

- As long as the bootstrap variance converges, we have convergence of the entire distribution
  - It seems like the only case where it won't is if the variance is infinite

Using Bootstrapped Estimates

- Bootstrapping itself is not a statistical test, but rather just estimating different parts of a distribution
- You can then use these estimates in a hypothesis test

Bootstrapping Mean

- Compute sample mean. Is this truly the population mean?
- Step 1: Bootstrap some samples of the data
- Step 2: Find the mean of each of these samples
- Step 3: Plot these from smallest to largest to get a distribution of the bootstrapped mean
- Step 4: Find the confidence interval of these means and that's your estimate

Bootstrapping Standard Errors

- Sample standard error may not be enough to give you insight to a statistical test - Monte Carlo Simulation
- Step 1: Bootstrap some samples of data
- Step 2: Calculate the statistic of interest that you want standard errors for i.e. mean
- Step 3: Calculate the standard deviation of each of these statistics
- Step 4: As the number of bootstrapped samples grows, this will become the bootstrapped standard error

Standard Deviation of Bootstrapped Statistics Calculation

$$\hat{\sigma}_B = \left(\frac{\sum_{b=1}^{B}[\hat{\theta}^*(b) - \hat{\theta}^*(\cdot)]}{B-1}\right)^{1/2} \text{ where } \hat{\theta}^*(\cdot) = \frac{\sum_{b=1}^{B}\hat{\theta}^*(b)}{B}$$

As B $\rightarrow \infty, \hat{\sigma}_B \rightarrow \sigma$

Bootstrapping Difference of Two Sample Means

- You have two samples: X and Y, with n and m observations. You want to know if the means are the same.
- Step 1: Compute $t = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}}$
- Step 2: Compute x' and y', where $x_i' = x_i - \bar{x} + \bar{z}$ and $y_i' = y_i - \bar{y} + \bar{z}$ where $\bar{z}$ is the mean of the joint sample
- Step 3: Draw bootstrapped sample of x' and y' and use those to compute test statistic
- Step 4: p-value = $\frac{\sum_{i=1}^{B} I[t_i > t]}{B}$
- Basically a permutation test, but sampling with replacement

When to Use?

- When the sample size is too small for clear analysis
- When you do not have a clear understanding of the underlying population distribution

Example

Original Data:	13	8	1	11	7	4	15	12

Mean: 8.875, std err: 1.6844

Example

Original Data:	13	8	1	11	7	4	15	12
Bootstrap 1:	11	11	15	15	8	11	11	4
Bootstrap 2:	4	15	1	4	4	8	13	11
Bootstrap 3:	12	1	7	8	15	1	7	4
Bootstrap 4:	12	12	7	8	8	1	15	1
Bootstrap 5:	15	8	12	1	8	1	7	11

Means: 10.75, 7.5, 6.875, 8, 7.875 SE: 1.4911

Code

- Stata: bootstrap, reps(N): X Y Z
- Matlab: bootstrp(N,@stat,X,Y)

```
a = [13 8 1 11 7 4 15 12];
a_mean = mean(a);
a_stderror = std(a)/sqrt(length(a));

[a_bootstrap,a_bootstrapdata] = bootstrp(5,@mean,a);
a_bootstrapSE = std(a_bootstrap);
```

Jackknifing

- Earlier resampling method than bootstrapping, where we no longer use the whole sample for resampling
- Rather, we remove one datapoint and calculate whatever statistic we want using the n-1 observations
    - We do this for all possible samples of n-1, so we find a statistic removing every possible observation once

Jackknifing Assumptions

- Normally distributed data
  - Small sample sizes may not be normal
- Our resampled values are necessarily correlated

Bootstrap vs. Jackknife

- Jackknife gives a more conservative estimate of standard error, but usually it's not as accurate as the bootstrap
- Jackknife gives same results every time, whereas bootstraps can change

Literatur

Chung, E., Romano, J.P., 2013. Exact and asymptotically robust permutation tests. The Annals of Statistics 41, 484–507. doi:10.1214/13-AOS1090.

Romano, J.P., 1990. On the Behavior of Randomization Tests without a Group Invariance Assumption. Journal of the American Statistical Association 85, 686–692. doi:10.1080/01621459.1990.10474928.