STAT 135 Lecture 20

Henry Liev

7 August 2025

Method 0.1 (Permutation Test)

Data from cholesterol-lowering study

Control	Treatment
103	100
115	105
123	121
131	

H_0: Treatment does nothing

H_1: Treatment does something Mann Whitney:

1. Under H_0, all the data is interchangeable
2. Throw away data, replace with ranks
3. Define test statistic T, sum of the ranks
4. Calculate t
5. Compare to the rank sum distribution

Permutation Test:

1. Under H_0, all the data is interchangeable
2. Define test statistic $\Delta = \bar{Y} - \bar{X}$
3. Calculate δ
4. Use the distribution of Δ across all equally likely assignments of data to treatment

We don't need to use all $\binom{m+n}{n}$

- Pick B random assignments of data to treatment
- Calculate $T_{(b)}$ for each assignment $(\bar{Y} - \bar{X})$
- Estimated p-value $\hat{P}$ = fraction of B assignments for which $t > T$
- If P is true p-value: $\mathbb{E}(\hat{p}) = p, SE(\hat{p}) = \sqrt{\frac{p(1-p)}{B}}$ Pick B large enough for desired accuracy
- End result: Totally nonparametric two sample test
- Obviously better than Mann-Whitney

Method 0.2 (Bootstrap (nonparametric))

Due to Brad Efron (1979) $X_1, \dots, X_n \overset{iid}{\sim} F$
Interested in some parameter $\theta(F)$
Have in mind an estimator $\hat{\theta} = T(X_1, \dots, X_n)$
Want to get $\mathrm{SE}(\hat{\theta})$
$\hat{\theta} = \bar{X}, SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$
$\hat{\theta} = f(\bar{X}), SE(f(\bar{X}))$ Delta Method
Big Idea of nonparametric (bootstrap)
Estimate the unknown F by $\hat{F}$
$\hat{F}$ is the empirical distribution which probability $\frac{1}{n}$ on each data point, "pretend the sample is the population"
$\hat{F}_n$ is a good approximation to F
Want $\mathrm{SE}_F(\hat{\theta})$, Randomness comes from F
$\mathrm{SE}_{\hat{F}}(\theta^*)$

Example 0.3

$X_1, \dots, X_n$ Assume data points are all distinct
Sampling from $\hat{F}$
There are $N = \binom{2n-1}{n}$ samples of size n drawn with replacement from $X_1, \dots, X_n$

1. In principle, we could draw all N possible samples of size n from the data
2. For $b = 1, \dots, N$ Get resampled data $X_{(b)}$, $\theta^*_{(b)} = T(X_(b))$

$$\mathrm{SE}_{\hat{F}} = \sqrt{\frac{1}{N}\sum_{b=1}^{N}(\hat{\theta}^*_{(b)} - \hat{\theta}^*_{(.)})^2}, \hat{\theta}^*_{(.)} = \frac{1}{N}\sum_{i=1}^{N}\hat{\theta}^*_{(b)}$$

Gives us an estimate of $SE_F(\theta)$ by replacing F with $\hat{F}$
$\hat{\theta} = T(x_1, \dots, x_n)$
Based on randomness in F
$\theta^* = T(Y_1, \dots, Y_n), Y_i \sim \hat{F}$
puts mass $\frac{1}{n}$ on each X_i
Choose a large number B (200, 1000)
For $b = 1, \dots, B$
Draw resamples $X_{(b)}$ from $X_1, \dots, X_n$
Draw a random sample of size n with replacement from $X_1, \dots, X_n$

$$\widehat{\mathrm{SE}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}\left[\hat{\theta}^*_{(b)} - \hat{\theta}^*_{(.)}\right]^2} \Rightarrow \hat{\theta}^*_{(.)} = \frac{1}{B}\sum_{b=1}^{B}\hat{\theta}^*_{(b)}$$

Nonparametric method for estimating $\mathrm{SE}(\hat{\theta})$ with no assumptions and no theory, $\hat{\theta} =$ median
Are using "large n", $\hat{F}_n \approx F$ asymptotic

Remark 0.4 (Bootstrap confidence intervals)

Many methods for using bootstrap to get CIs

	Seems to make sense	Automatically works well	Confusing
Basic Bootstrap	✓	×	A little
Percentile Interval	×	✓	No
Bootstrap-t	✓	?	Yes

Method 0.5 (Basic bootstrap)

Basic bootstraps works on the same principle as parametric bootstrap: Approximating $\hat{\theta} - \theta \Rightarrow \theta^* - \hat{\theta}$

$P(a \leq \hat{\theta} - \theta \leq b) = 0.95 \rightarrow (\hat{\theta} - b, \hat{\theta} - a)$

Generate θ^* parametrically via $f(x|\hat{\theta})$ (Parametric bootstrap)

- Basic bootstrap "makes" sense by same logic as parametric
- But doesn't work well becaues $\theta^* - \theta$ is not a good estimate of $\hat{\theta} - \theta$ in a nonparametric setting

Method 0.6 (Percentile Interval)

As before, generate many instance $\theta^*_{(b)}$ and take the percentiles to generate confidence intervals for θ

Method 0.7 (Bootstrap-t)

Normal data setting

Know that $Z = \frac{\bar{X} - \mu}{\mathrm{SE}(\bar{X})} \sim T$

$P(t_{0.025} \leq \frac{X - \mu}{\mathrm{SE}(\bar{X})} \leq t_{0.975}) = 95\% \rightarrow (\bar{X} - t_{0.975}\,\mathrm{SE}(\bar{X}), \bar{X} + t_{0.975})\,\mathrm{SE}(\bar{X})$

Build your own custom t-distribution from the data by bootstrapping $Z(b) = \frac{\hat{\theta}^*_{(b)} - \hat{\theta}}{\mathrm{SE}^*(b)}$

For $b = 1, \ldots, B$ generate resamples compute $Z(b)$ to generate the confidence intervals above

For each bootstrap need to conduct a bootstrap to find $\mathrm{SE}^*(b)$