Perm

June 8, 2023

1 STAT 207: Permutation Tests and the Bootstrap

- Both techniques involve random resampling of observed data
- Less reliant on model assumptions and large sample requirements
- Comes with higher computational demands
-

1.1 Permutation Tests

- Permutation tests differ from parametric tests in how the distribution of the test statistic is evaluated.
- Two-sample t-test as an example.
- $x_1, ..., x_m$ from population 1; $y_1, ..., y_n$ from population 2. Propose the test statistic $T = \overline{x} - \overline{y}$.
- Concatenate the observations $z_1, \dots, z_{m+n}$.
- Under the null, the permutations generate $\binom{m+n}{m}$ equally likely versions of the test statistic T, $T_1, ..., T_J$.
- The p-value attached to the observed value T_{obs} is just the fraction $p = |\{j : |T_j| \geq |T_{\text{obs}}|\}|/J$.
- Precision of the permutation p-value: k random permutations, then $\hat{p}$ has mean p and variance $\sigma_k^2 = p(1-p)/k$, which leads to the CI $(\hat{p} - 2\hat{\sigma}_k, \hat{p} + 2\hat{\sigma}_k)$.
- The set of possible p-values is discrete, more of an issue for small samples.
- The width of the CI is proportional to $1/\sqrt{k}$.

1.1.1 Example: The One-Way Layout and Reading Speeds

- k samples of sizes $n_1, .., n_k$ with unknown means $\mu_1, ..., \mu_k$ and a common unknown variance σ^2.
- Observations $y_{ij}, j = 1, .., n_i$ and total sample size n.
- $H_0 : \mu_1 = .. = \mu_k = \mu$. Assuming y_{ij}'s are normally distributed, the test statistic is the

difference between the two

$$\min_{\mu} \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \mu)^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}^2 - n\bar{y}^2,$$

$$\sum_{i=1}^{k} \min_{\mu_i} \sum_{j=1}^{n_i} (y_{ij} - \mu_i)^2 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{k} n_i \bar{y}_i^2.$$

- The permutation test statistic is

$$S = \sum_{i=1}^{k} n_i \bar{y}_i^2.$$

- Permuting all n samples and separate into k populations and recompute S.

TABLE 24.2. Reading Speeds and Typeface Styles

Typeface Style		
1	**2**	**3**
135	175	105
91	130	147
111	514	159
87	283	107
122		194

1.2 The Bootstrap

- We have i.i.d. samples $\mathbf{x} = (x_1, .., x_n)$ from an unknown distribution $F(x)$.
- Estimator $T(\mathbf{x})$ of a parameter (or functional) $t(F)$.
- Define bootstrap samples $\mathbf{x}^* = (x_1^*, .., x_n^*)$ and calculate the estimator $T(\mathbf{x}^*)$ provided $F_n^*(x)$ approximates $F(x)$.
  - Non-parametric bootstrap, $F_n^*(x)$ is the empirical distribution function
  - Parametric bootstrap, $F_n^*(x) = F_{\widehat{\alpha}}(x)$ for some parametric form $F_\alpha(x)$.
- $T(\mathbf{x}) - t(F)$ and $T(\mathbf{x}^*) - t(F_n^*)$ have similar distributions.
- In many examples, $t(F_n^*) = T(\mathbf{x})$ holds.
- Approximate the distribution and moments of $T(\mathbf{x})$ by independent Monte Carlo sampling Efron 1979.

1.2.1 Examples of bootstrap parameters

- The moments and central moments:

$$\mu_k(F) = \int x^k dF(x),$$
$$\omega_k(F) = \int [x - \mu_1(F)]^k \, dF(x).$$

- The pth quantile

$$\xi_p(F) = \inf\{x : F(x) \geq p\}.$$

- The natural estimators are

$$\hat{\mu}_k(\mathbf{x}) = \mu_k(F_n^*) = \frac{1}{n}\sum_{i=1}^{n} x_i^k$$
$$\hat{\omega}_k(\mathbf{x}) = \omega_k(F_n^*) = \frac{1}{n}\sum_{i=1}^{n} (x_k - \bar{x})^k$$
$$\hat{\xi}_p(\mathbf{x}) = \xi_p(F_n^*) = \inf\{x : F_n^*(x) \geq p\}.$$

- $t(F_n^*) = T(\mathbf{x})$?

1.2.2 Bias Reduction

- The bootstrap bias:

$$\text{bias}^* = E[T(\mathbf{x}^*)] - t(F_n^*)$$

can be estimated by the Monte Carlo average

$$\widehat{\text{bias}}_B^* = \frac{1}{B}\sum_b T(\mathbf{x}_b^*) - t(F_n^*).$$

- The revised estimator would have smaller bias:

$$T(\mathbf{x}) - \widehat{\text{bias}}_B^* = 2T(\mathbf{x}) - \frac{1}{B}\sum_b T(\mathbf{x}_b^*).$$

1.2.3 Confidence Intervals

- A set $C(\mathbf{x})$ is a $1-\alpha$ level confidence set for $t(F)$ if

$$P[t(F) \in C(\mathbf{x})] \geq 1 - \alpha.$$

- Bootstrap percentile interval:

$$\left[\xi_{\frac{\alpha}{2}}(G_n^*), \xi_{1-\frac{\alpha}{2}}(G_n^*)\right]$$

should approximate the α level CI. But in practice, the deviation can be substantial.

1.2.4 Applications in Regression

- Linear regression model, $y = X\beta + u$.
- Define the residuals $r_i = y_i - \hat{y}_i$.
- Bootstrapping residuals samples $\mathbf{r}^*$ from $\mathbf{r} = (r_1, .., r_n)$ with replacement. And let $\mathbf{y}^* = X\hat{\beta} + \mathbf{r}^*$.
- $E(\hat{\beta}^*) = \hat{\beta}$?
- Parametric bootstrap in GLM: using the MLE coefficients to resample responses.

1.3 Efficient Bootstrap Simulations

Methods to reduce Monte Carlo sampling error

1.3.1 The Balanced Bootstrap

- For bias estimation, for example, we estiamte the bias of the sample mean $\overline{x}$ by the Monte Carlo difference
$$\frac{1}{B}\sum_b \overline{x_b^*} - \overline{x}.$$
- Retain the randomness in the bootstrap resamples x_b^* while forcing each original x_i to appear exactly B times.

1.3.2 The Antithetic Bootstrap

- To reduce variance and improve efficiency.
- Antithetic resampling: think of both unbiased estimators V and W that are negatively correlated rather than independent.
$$Var(\frac{V+W}{2}) = \frac{1}{4}Var(V) + \frac{1}{4}Var(W) + \frac{1}{2}Cov(V, W)$$
- To generate antithetic bootstrap resamples: think of the order statistics $x_{(1)} \leq .. \leq x_{(n)}$ and the permutation $\pi(i) = n - i + 1$, then
$$x_{(\pi[1])} \geq .. \geq x_{(\pi[n])}.$$
- For any bootstrap sample $x^* = (x_1^*, ..., x_n^*)$, define x^{**} by substituting $x_{(\pi[i])}$ for every appearance of $x_{(i)}$ in x^*.
- T^* and T^{**} are negatively correlated.

1.3.3 Importance Resampling

- Assign different resampling probabilities p_i to the different observations x_i.
- The connection between the uniform expectation and the importance expectation

$$
\begin{aligned}
\mathrm{E}\left[T\left(\mathbf{x}^*\right)\right] &= \mathrm{E}_p\left[T\left(\mathbf{x}^*\right)\frac{\binom{n}{m_1^* \cdots m_n^*}\left(\frac{1}{n}\right)^n}{\binom{n}{m_1^* \cdots m_n^*}\prod_{i=1}^n p_i^{m_i^*}}\right] \\
&= \mathrm{E}_p\left[T\left(\mathbf{x}^*\right)\prod_{i=1}^n \left(np_i\right)^{-m_i^*}\right]
\end{aligned}
$$

- Estimate $\mathrm{E}\left[T\left(\mathbf{x}^*\right)\right]$ by the bootstrap average

$$
\frac{1}{B}\sum_b T\left(\mathbf{x}_b^*\right)\prod_{i=1}^n \left(np_i\right)^{-m_{bi}^*}.
$$

- Minimize the variance with respect to p.

1.3.4 Python implementations

[1]:
```python
# list all permutations
from itertools import permutations
l = list(permutations(range(1, 4)))
print(l)

# Generate one random permutation
import numpy as np

print(np.random.permutation(10))

np.random.permutation([1, 4, 9, 12, 15])
```

```
[(1, 2, 3), (1, 3, 2), (2, 1, 3), (2, 3, 1), (3, 1, 2), (3, 2, 1)]
[7 3 8 5 9 1 0 4 6 2]
```

[1]:
```
array([ 4,  1, 12,  9, 15])
```

[2]:
```python
import numpy as np

# Original data
data = np.array(range(1,11))

# Bootstrap resampling
bootstrap_sample = np.random.choice(data, size=len(data), replace=True)

print("Bootstrap sample:", bootstrap_sample)
```

```
Bootstrap sample: [ 5  2  3  2  5 10  2  2 10 10]
```

[ ]:

```

```

[ ]:

```

```