Two-Sample Numerical Data

paired and unpaired, t-test, bootstrap t-test, permutation test, rank-sum test, Kolmogorov-Smirnov test

University of California, San Diego
Instructor: Ery Arias-Castro
http://math.ucsd.edu/~eariasca/teaching.html

Two-sample paired data

□ Suppose we have numerical data of the form \((X_1, Y_1), \ldots, (X_n, Y_n)\), so that the variables \(X\) and \(Y\) are paired. We assume the pairs are iid from some distribution.

□ Example: Corn yield from split plots (indexed by \(i\)) of regular \((X_i)\) and kiln dried \((Y_i)\) seed.

□ We can analyze such data in at least two ways:

1. If we want to compare the magnitude of these variables, then typically one takes the difference

\[ Z_i = Y_i - X_i \]

and work on the sample \(Z_1, \ldots, Z_n\) (one-sample numerical data).

NOTE: This requires that taking the difference of \(X\) and \(Y\) makes sense!

2. If want to model how \(Y\) varies with \(X\), say, then we need to perform a correlation or regression analysis. (This will come later in the quarter.)

NOTE: This is applicable even when \(X\) and \(Y\) are different types of measurements.

(In this particular example, the first is more appropriate.)

Two-sample unpaired data

□ Suppose we have numerical data of the form \(X_1, \ldots, X_m\) and \(Y_1, \ldots, Y_n\), so that the variables \(X\) and \(Y\) are unpaired. We assume that the \(X\)'s are iid from \(F_X\) and the \(Y\)'s are iid from \(F_Y\), both distributions being unknown.

□ Example: Math SAT scores for some seniors at La Jolla High School and for some seniors at University City High School.

□ Typically, we want to compare the magnitude of these variables.

NOTE: Taking pairwise differences does not make sense here. In fact, the sample sizes could be different \((m \neq n)\). Even when they are equal, the pairing of \(X_i\) with \(Y_i\) would be completely arbitrary.
Testing for equal means

- Let $\mu_X$ and $\mu_Y$ be the means of $X$ and $Y$, respectively.
- Suppose we want to test $H_0 : \mu_X \leq \mu_Y$ vs $H_1 : \mu_X > \mu_Y$

Two-Sample $t$-Test with Equal Variances

- First, assume the variances are equal $\sigma_X^2 = \sigma_Y^2$.
- The two-sample $t$-test with equal variances rejects for large values of
  
  $$T = \frac{\bar{X} - \bar{Y}}{S \sqrt{\frac{1}{m} + \frac{1}{n}}}$$

  where $S^2$ is the pooled sample variance:

  $$S^2 = \frac{1}{m + n - 2} \left( \sum_{i=1}^{m} (X_i - \bar{X})^2 + \sum_{j=1}^{n} (Y_j - \bar{Y})^2 \right)$$

  $$= \frac{m - 1}{m + n - 2} S_X^2 + \frac{n - 1}{m + n - 2} S_Y^2$$

- **Theory.** If all the observations (both groups combined) are i.i.d. normal (same mean and variance), then $T$ has the $t$-distribution with $m + n - 2$ degrees of freedom.

Two-Sample $t$-Test with Unequal Variances

- Sometimes the equal-variance assumption is not reasonable.
  - **NOTE:** Testing for equal variances is typically not recommended as it is at least as hard (if not harder) than testing for the equal means.
- The two-sample $t$-test with unequal variances (aka the Welch t-test) rejects for large values of

  $$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}}$$

- **Theory.** If the observations within each sample are i.i.d. normal, then under the null, $T$ has approximately the $t$-distribution with degrees of freedom:

  $$\frac{(S_X^2/m + S_Y^2/n)^2}{(S_X^2/m)^2/(m-1) + (S_Y^2/n)^2/(n-1)}$$
Checking assumptions

- Homoscedasticity (equality of variances) is checked by comparing the spreads of the samples, for example, in side-by-side boxplots.

- When assuming equal variances, normality is checked with a combined Q-Q plot after removing the respective sample means.

  Otherwise, normality is checked with a separate Q-Q plot for each sample.

  NOTE: If the sample sizes are small to moderate, approximate normality is important. Otherwise, a simple check that the distributions are not too asymmetric or heavy-tailed may be enough.

  Although this may be standard practice, it is definitely ad hoc and unreliable, particularly in small samples.

Transformations

- One may apply a transformation to make the data look more ‘normal’. The same transformation is applied to both samples.

- Only a few transformations are used in practice:
  - Power functions \( x \rightarrow (x + \xi)^\eta \), in particular \( \eta = 1/2 \) or \( 1/3 \).
  
  - Logarithmic functions \( x \rightarrow \log(x + \xi) \)

  We then apply the t-test to the transformed data.

  NOTE: The t-test after transformation really addresses a different hypothesis because the mean after transformation may not agree even if they agree before transformation.

  Although this may be standard practice, it remains ad hoc. In particular, the transformation is often decided based on the data and the computation of the p-value does not take that into account.
Another option is to obtain a p-value using a form of nonparametric bootstrap. Perhaps the simplest way to do this is via a confidence interval for the difference in means. Say we do not assume the variances are equal and choose as pivot Welch t-ratio

\[ T = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}}} \]

Let \( B \) be a large integer. For \( b = 1, \ldots, B \), do the following:

1. Generate \( X_1^{(b)}, \ldots, X_m^{(b)} \) from \( \hat{F}_X \) and \( Y_1^{(b)}, \ldots, Y_n^{(b)} \) from \( \hat{F}_Y \). (Note that there is no pairing.)

2. Compute the t-ratio for that bootstrap sample

\[ T_b = \frac{\bar{X}_b - \bar{Y}_b - (\bar{X} - \bar{Y})}{\sqrt{\frac{S_{X,b}^2}{m} + \frac{S_{Y,b}^2}{n}}} \]

Let \( t_{\alpha}^{\text{boot}} \) denote the \( \alpha \)-quantile of the bootstrap pivots \( \{T_b : b = 1, \ldots, B\} \).

A (bootstrap) one-sided level \( 1 - \alpha \) confidence interval for \( \mu_X - \mu_Y \) is

\[ \left[ \bar{X} - \bar{Y} - t_{\alpha}^{1-\alpha} \widehat{SE}, \infty \right) \]

where

\[ \widehat{SE} = \sqrt{\frac{S_X^2}{m} + \frac{S_Y^2}{n}} \]

We then compute the p-value as seen before.

In words, in this particular case, it’s the largest \( \alpha \) such that \( (-\infty, 0] \) has a non-empty intersection with the \( (1 - \alpha) \)-confidence interval.

(Equivalently, it’s the largest \( \alpha \) such that 0 is in the \( (1 - \alpha) \)-confidence interval.)
Stochastic dominance

Consider two numerical random variables $X$ and $Y$. We say that $X$ stochastically dominates $Y$, sometimes denoted

$$X \overset{\text{sto}}{\geq} Y$$

if

$$P(X > t) \geq P(Y > t), \text{ for all } t \in \mathbb{R}$$

or, equivalently,

$$F_X(t) \leq F_Y(t), \text{ for all } t \in \mathbb{R}$$

(Note that the inequality is reversed.)

EXAMPLE: Suppose that $a X \sim Y + \Delta$, where $\Delta$ is deterministic. Then $X$ stochastically dominates $Y$ if and only if $\Delta \geq 0$.

Here $A \sim B$ reads ‘$A$ and $B$ have the same distribution’.

Permutation test

Suppose we want to test

$$H_0 : X \sim Y \quad \text{versus} \quad H_1 : X \overset{\text{sto}}{\geq} Y \text{ (strictly)}$$

Let $Z_1, \ldots, Z_{m+n}$ be the concatenated sample $X_1, \ldots, X_m, Y_1, \ldots, Y_n$.

Under the null, the $Z_k$’s are exchangeable because the variables are independent and have the same distribution.

Permutation test

A permutation test based on the difference in means works as follows.

First, compute the observed difference $D_{\text{id}} = \bar{X} - \bar{Y}$.

Then do the following:

1. For each permutation $\pi$ of $\{1, \ldots, m+n\}$, let

$$X_i^\pi = Z_{\pi(i)}, \quad i = 1, \ldots, m$$

$$Y_j^\pi = Z_{\pi(j+m)}, \quad j = 1, \ldots, n$$

Compute

$$D_\pi = \bar{X}^\pi - \bar{Y}^\pi$$

2. The (exact) permutation p-value is

$$\frac{\# \{\pi : D_\pi \geq D_{\text{id}}\}}{(m+n)!}$$

since there are $(m+n)!$ permutations of $\{1, \ldots, m+n\}$. 
Even for small \( m \) and \( n \), computing the exact p-value as done above may not be feasible. Indeed, assuming all the observations are distinct, there are \( \binom{m+n}{n} \) non-equivalent permutations.

For \( m = n = 20 \), \( \binom{m+n}{n} > 10^{11} \) already!

When this is the case, we estimate the p-value by Monte Carlo sampling. We sample \( B \) permutations (\( B \) is a large integer) uniformly at random instead of going over all permutations.

### Tests based on ranks

- A test based on ranks is based on a statistic of the form

  \[ h(R_1, \ldots, R_m) \]

  where \( R_i \) is the rank of \( X_i \) in the combined sample.

  NOTE: If we know the ranks of the \( X \)'s (up to permutation among the \( X \)'s), we know the ranks of the \( Y \)'s (up to permutation among the \( Y \)'s).

- Equivalently, a rank-based test statistic is one that can be computed based solely on the ordering of the \( X \)'s and \( Y \)'s, such as

  \[ XXXYYXYYXXX \]

Such tests are distribution-free, in that the distribution of the test statistic is the same under the null, regardless of the common distribution (as long as it is continuous). Indeed, when \( X \sim Y \), all the possible ordering patterns are equally likely.

This explains why

**Rank tests are special cases of permutation tests.**

Such tests are invariant with respect to any (strictly) increasing transformation \( g \), because the test statistic remains unchanged if \( X \) is changed to \( g(X) \) and \( Y \) to \( g(Y) \).

The tests that follow are famous examples of tests based on ranks.
The Wilcoxon Rank-Sum Test

- Recall that $R_i$ is the rank of $X_i$ in the combined sample. Ties are given the average rank among them or are broken at random.
- The Wilcoxon rank-sum test rejects for large values of

$$V = \sum_{i=1}^{m} R_i$$

- Theory. The distribution of $V$ can be computed exactly and efficiently using some recursion formulas. In the large-sample limit where $m \to \infty$ and $n \to \infty$, $V$ has a normal distribution.

The Mann-Whitney $U$-Test

- The Mann-Whitney $U$-test rejects for large values of

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{I}\{X_i > Y_j\}$$

- It is completely equivalent to the Wilcoxon rank-sum test!

Indeed

$$U = V - \frac{m(m+1)}{2}$$

The Kolmogorov-Smirnov Two-Sample Test

- The (one-sided) Kolmogorov-Smirnov test rejects for large values of

$$D_{m,n}^+ = \sup_{t \in \mathbb{R}} [\hat{F}_Y(t) - \hat{F}_X(t)]$$

- Theory. The distribution of $D_{m,n}^+$ can be computed exactly and efficiently using some recursion formulas. In the large-sample limit

$$\lim_{m,n \to \infty} \mathbb{P}\left( \sqrt{\frac{mn}{m+n}} D_{m,n}^+ \leq d \right) = 1 - e^{-2d^2}, \quad \forall d \geq 0$$

NOTE: This happens to be the same limiting distribution as in the one-sample case with sample size $\lfloor \frac{mn}{m+n} \rfloor$. 