S&DS 242/542: Theory of Statistics

Lecture 7: Nonparametric tests, permutation tests

Two-sample tests

Two-sample tests

Given two independent data samples

$$X_1, \ldots, X_n \qquad \text{and} \qquad Y_1, \ldots, Y_m$$

are their distributions different? Is one distribution "larger than" the other?

Two-sample z-test

Suppose

$$X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\mu_X, \sigma^2), \qquad Y_1, \ldots, Y_m \overset{IID}{\sim} \mathcal{N}(\mu_Y, \sigma^2)$$

The two samples are assumed independent, with a common variance $\sigma^2 > 0$. We wish to test

$$H_0 : \mu_X = \mu_Y \quad \text{vs.} \quad H_1 : \mu_X > \mu_Y$$

Assuming that σ^2 is known, the **two-sample z-statistic** is

$$Z = \frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

$\bar{X} \sim \mathcal{N}(\mu_X, \frac{\sigma^2}{n})$ is independent of $-\bar{Y} \sim \mathcal{N}(-\mu_Y, \frac{\sigma^2}{m})$. Then under H_0, $\bar{X} - \bar{Y} \sim \mathcal{N}(0, \sigma^2(\frac{1}{n} + \frac{1}{m}))$, so $Z \sim \mathcal{N}(0, 1)$.

The **two-sample z-test** at level-α rejects H_0 when $Z > z^{(\alpha)}$.

Two-sample t-test

When σ^2 is unknown, we may estimate it by the **pooled sample variance**

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

This estimate is reasonable assuming that the X_i's and Y_j's have the same variance.

The **two-sample t-statistic** is

$$T = \frac{\bar{X} - \bar{Y}}{S_{\text{pooled}}\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

and a test of H_0 based on T is called a **two-sample t-test**.

Distribution of the pooled sample variance

Theorem

If $X_1, \dots, X_n \overset{IID}{\sim} \mathcal{N}(\mu_X, \sigma^2)$ and $Y_1, \dots, Y_m \overset{IID}{\sim} \mathcal{N}(\mu_Y, \sigma^2)$ are independent, then S^2_{pooled} is independent of $\bar{X} - \bar{Y}$, with

$$S^2_{pooled} \sim \frac{\sigma^2}{m+n-2} \cdot \chi^2_{m+n-2}$$

Proof:

- Let $\bar{X}, S_X^2$ be sample mean/variance of X_i's, similarly $\bar{Y}, S_Y^2$
- Applying Thm from last lecture, $\bar{X}, \bar{Y}, S_X^2, S_Y^2$ are all independent
- $S^2_{pooled} = \frac{1}{m+n-2}\Big(\underbrace{\sum_{i=1}^n (X_i - \bar{X})^2}_{=(n-1)S_X^2} + \underbrace{\sum_{j=1}^m (Y_j - \bar{Y})^2}_{=(m-1)S_Y^2}\Big)$

  $\Rightarrow S^2_{pooled}$ is independent of $\bar{X} - \bar{Y}$.
- From last class: $S_X^2 \sim \frac{\sigma^2}{n-1}\chi^2_{n-1}$, $S_Y^2 \sim \frac{\sigma^2}{m-1}\chi^2_{m-1}$

  $\Rightarrow S^2_{pooled} = \frac{1}{m+n-2}\left(\sigma^2 \chi^2_{n-1} + \sigma^2 \chi^2_{m-1}\right) \sim \frac{\sigma^2}{m+n-2}\chi^2_{m+n-2}$.

Two-sample t-test

The two-sample t-statistic may be written as

$$T = \frac{\bar{X} - \bar{Y}}{S_{\text{pooled}}\sqrt{\frac{1}{n} + \frac{1}{m}}} = \underbrace{\frac{\bar{X} - \bar{Y}}{\sigma\sqrt{\frac{1}{n} + \frac{1}{m}}}}_{=Z} \Bigg/ \sqrt{\underbrace{\frac{S^2_{\text{pooled}}}{\sigma^2}}_{=U}}$$

Under H_0, $Z \sim \mathcal{N}(0,1)$, $U \sim \frac{1}{m+n-2} \cdot \chi^2_{m+n-2}$, and these are independent. So by definition of the t-distribution,

$$T \sim t_{m+n-2}$$

The two-sample t-test at significance level α would reject H_0 when $T > t^{(\alpha)}_{m+n-2}$, the upper-α point of the t_{m+n-2} distribution.

Note that T has the same distribution for any $\sigma^2 > 0$ and also any $\mu_X = \mu_Y$, so it is pivotal under H_0.

Welch's t-test

If instead

$$X_1, \dots, X_n \overset{IID}{\sim} \mathcal{N}(\mu_X, \sigma_X^2), \qquad Y_1, \dots, Y_m \overset{IID}{\sim} \mathcal{N}(\mu_Y, \sigma_Y^2)$$

with different variances σ_X^2, σ_Y^2, then

$$\bar{X} - \bar{Y} \sim \mathcal{N}(\mu_X - \mu_Y, \tfrac{\sigma_X^2}{n} + \tfrac{\sigma_Y^2}{m})$$

We may estimate this variance by $\frac{S_X^2}{n} + \frac{S_Y^2}{m}$ where S_X^2, S_Y^2 are the individual sample variances, and test H_0 using **Welch's t-statistic**

$$T_{\text{welch}} = \bar{X} - \bar{Y} \Bigg/ \sqrt{\tfrac{S_X^2}{n} + \tfrac{S_Y^2}{m}}.$$

This is called **Welch's *t*-test** or the **unequal variances *t*-test**. Welch showed that the null distribution of T_{welch} is approximately (but not exactly) a t-distribution with degrees-of-freedom

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

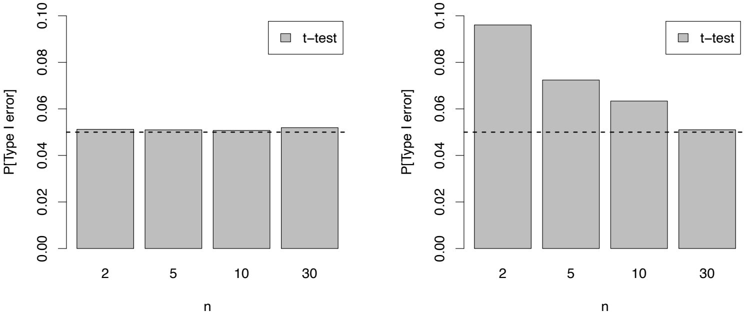
Robustness in large samples

The reason why the t-test is widely used is not because our data are usually normally distributed. Instead, as long as each sample is IID with mean 0 and (finite) variance σ^2:

- The z-statistic
$$Z = \bar{X} - \bar{Y} \Big/ \sigma\sqrt{\tfrac{1}{n} + \tfrac{1}{m}}$$
converges in distribution to $\mathcal{N}(0,1)$ as $m, n \to \infty$, by the CLT.
- The pooled variance $S^2_{\text{pooled}} \to \sigma^2$ in probability.
- Then also the t-statistic
$$T = Z \Big/ \sqrt{S^2_{\text{pooled}}/\sigma^2}$$
converges in distribution to $\mathcal{N}(0,1)$. [This is formalized by a result known as Slutsky's Lemma.]

Thus a level-α t-test will have Type I error probability $\approx \alpha$ when m, n are large, even when the data are not normally distributed.

Robustness in small samples?



Data: Uniform([0, 1]) Data: 10% $\mathcal{N}(10, 1)$, 90% $\mathcal{N}(0, 1)$

Sample sizes: control $m = 30$, experiment $n \in \{2, 5, 10, 30\}$

Wilcoxon rank-sum statistic

The **Wilcoxon (a.k.a. Mann-Whitney) rank-sum test** is a two-sample test that is valid for non-normally-distributed data:

1. Sort the *pooled sample* $X_1, \ldots, X_n, Y_1, \ldots, Y_m$, and assign the smallest a rank of 1, the next smallest a rank of 2, etc., and the largest a rank of $m + n$.[1]
2. The test statistic T is the sum of ranks of the values $Y_1, \ldots, Y_m$ of the second sample.

Example: Consider sample sizes $m = n = 2$,

$$(X_1, X_2) = (1.8, -0.5), \quad (Y_1, Y_2) = (0.4, -2.3)$$

In sorted order, the pooled observations and their ranks are

Observation	Y_2	X_2	Y_1	X_1
Rank	1	2	3	4

So the rank-sum statistic is $T = 1 + 3 = 4$.

[1] For simplicity, let us assume that there are no ties in the data values.

Null hypothesis of the rank-sum test

If $X_1, \ldots, X_n \overset{IID}{\sim} F$ and $Y_1, \ldots, Y_m \overset{IID}{\sim} G$ for two continuous distributions F and G, this tests the *nonparametric* null hypothesis

$$H_0 : F = G$$

Under H_0, each permutation of the ranks $1, 2, \ldots, m+n$ is equally likely for $X_1, \ldots, X_n, Y_1, \ldots, Y_m$, e.g. for $m = n = 2$:

Ranks of X_1, X_2, Y_1, Y_2	Value of T	Probability
$1, 2, 3, 4$	7	$\frac{1}{4!}$
$1, 2, 4, 3$	7	$\frac{1}{4!}$
$1, 3, 2, 4$	6	$\frac{1}{4!}$
$\vdots$	$\vdots$	$\vdots$
$4, 3, 2, 1$	3	$\frac{1}{4!}$

This gives the null distribution of T, and T is pivotal under H_0.

Wilcoxon rank-sum test

Theorem

Let T be the rank-sum statistic. Under H_0,

$$\mathbb{E}[T] = \frac{m(m+n+1)}{2}, \qquad \mathrm{Var}[T] = \frac{mn(m+n+1)}{12}.$$

We may compute/approximate the null distribution of T by:

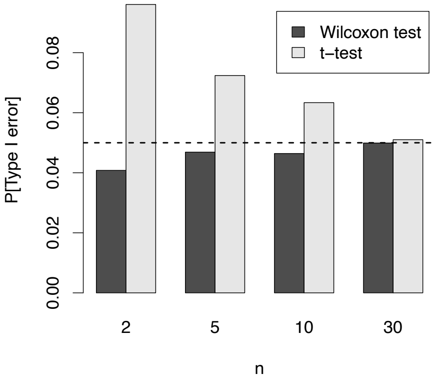
- Exhaustive enumeration of all permutations, if m, n are small.
- Applying a normal approximation if m, n are large:

$$T \sim \mathcal{N}\left(\tfrac{m(m+n+1)}{2}, \tfrac{mn(m+n+1)}{12}\right)$$

- Simulating permutations of $1, 2, \ldots, m+n$ uniformly at random, and computing T for these simulations.

Testing against a one-sided alternative H_1 that the X_i's "tend to be larger" than the Y_j's, the ranks of the Y_j's should be smaller under H_1, so we would reject $H_0 : F = G$ for small values of T.

Type I error probabilities in small samples



Data: 10% $\mathcal{N}(10, 1)$, 90% $\mathcal{N}(0, 1)$

Sample sizes: control $m = 30$, experiment $n \in \{2, 5, 10, 30\}$

Statistical power

The **power** of a test is its ability to successfully distinguish an alternative H_1 from the null H_0. It is defined as

$$\text{Power} = \mathbb{P}_{H_1}[\text{reject } H_0]$$

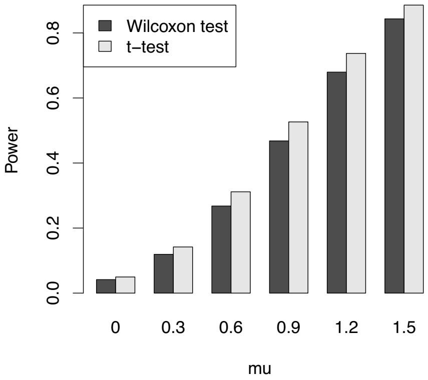
where $\mathbb{P}_{H_1}$ means that this probability is computed assuming the alternative hypothesis is true.

[The complement of power is the probability of **Type II error**, i.e. the probability that we do not reject H_0 when H_1 is in fact true:

$$\mathbb{P}[\text{Type II error}] = 1 - \text{Power} = \mathbb{P}_{H_1}[\text{accept } H_0]$$

We will stick to thinking about power instead of Type II error.]

Simulated power under the normal model



Data: $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 1)$ and $Y_1, \ldots, Y_m \sim \mathcal{N}(0, 1)$
Sample sizes $m = n = 8$

Permutation testing

A second view of the rank-sum test

Consider two independent samples

$$X_1, \ldots, X_n \overset{IID}{\sim} F, \qquad Y_1, \ldots, Y_m \overset{IID}{\sim} G$$

and the problem of testing equality of distribution

$$H_0 : F = G$$

Another way to understand the rank-sum test is: Let

$$\{Z_1, \ldots, Z_{m+n}\} = \{X_1, \ldots, X_n, Y_1, \ldots, Y_m\}$$

denote the set of all observations, discarding their ordering.[2] Under H_0, given only $\{Z_1, \ldots, Z_{m+n}\}$, each of the $(m+n)!$ assignments of these values to $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ is equally probable.

So each assignment of ranks to $Y_1, \ldots, Y_m$ is also equally probable.

[2] Again let us assume that there are no ties in the data values.

The permutation null distribution

For the same testing problem, consider *any* test statistic $T(X_1, \ldots, X_n, Y_1, \ldots, Y_m)$, not necessarily the rank-sum.

The **permutation null distribution** of T is the distribution of

$$T(X_1^*, \ldots, X_n^*, Y_1^*, \ldots, Y_m^*)$$

when we fix the set of values

$$\{Z_1, \ldots, Z_{m+n}\} = \{X_1, \ldots, X_n, Y_1, \ldots, Y_m\}$$

and let $X_1^*, \ldots, X_n^*, Y_1^*, \ldots, Y_m^*$ be a permutation of these values chosen uniformly at random.

Equivalently, it is the *conditional* distribution of T under H_0 given the pooled sample $\{Z_1, \ldots, Z_{m+n}\}$.

The permutation null distribution

Example: Consider data $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$, and the t-statistic

$$T = \bar{X} - \bar{Y} \Big/ S_{\text{pooled}} \sqrt{\tfrac{1}{n} + \tfrac{1}{m}}$$

Its permutation null distribution is the distribution of

$$\bar{X}^* - \bar{Y}^* \Big/ S^*_{\text{pooled}} \sqrt{\tfrac{1}{n} + \tfrac{1}{m}}$$

when $X_1^*, \ldots, X_n^*, Y_1^*, \ldots, Y_m^*$ is a random permutation of the pooled sample $X_1, \ldots, X_n, Y_1, \ldots, Y_m$, and

$$\bar{X}^* = \frac{1}{n} \sum_{i=1}^{n} X_i^*, \qquad \bar{Y}^* = \frac{1}{m} \sum_{j=1}^{m} Y_j^*$$

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

The permutation test

Suppose large values of T provide evidence against H_0 in favor of an alternative H_1.

A level-α **permutation test** based on T rejects H_0 if the observed value of T exceeds the upper-α point of its permutation null distribution.

This ensures the conditional Type I error guarantee

$$\mathbb{P}[\text{ Type I error} \mid \{Z_1, \ldots, Z_{m+n}\}\,] \leq \alpha$$

for any possible observed values of $\{Z_1, \ldots, Z_{m+n}\}$.

Hence, averaging over all possible values of $\{Z_1, \ldots, Z_{m+n}\}$, this also ensures $\mathbb{P}[\text{Type I error}] \leq \alpha$ unconditionally.

The permutation test

Example: Suppose H_1 specifies that the mean of F (distribution of X_i's) is larger than the mean of G (distribution of Y_j's). A level-α permutation test of H_0 vs. H_1 based on the t-statistic

$$T = \bar{X} - \bar{Y} \Big/ S_{\text{pooled}} \sqrt{\tfrac{1}{n} + \tfrac{1}{m}}$$

would reject H_0 when T exceeds the upper-α point of the distribution of

$$\bar{X}^* - \bar{Y}^* \Big/ S^*_{\text{pooled}} \sqrt{\tfrac{1}{n} + \tfrac{1}{m}}$$

over random permutations $X_1^*, \ldots, X_n^*, Y_1^*, \ldots, Y_m^*$ of the data.

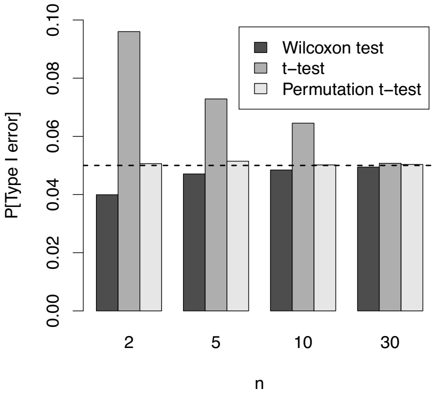
We may *simulate* this permutation null distribution by computing T on randomly generated permutations of the data, and compare T for the original (unpermuted) data with these simulated values.

Advantages of permutation testing

Why might we compare T to its permutation null distribution, rather than its actual (unconditional) null distribution under H_0?

- The permutation null distribution does not rely on parametric modeling assumptions, and is robust to misspecifications of the data model.
- Permutation testing is easy to apply for test statistics T where we may not know its theoretical null distribution.
- We do not need T to be pivotal under H_0: Even if T has different sampling distributions for different data distributions $F = G$, its conditional distribution given $\{Z_1, \ldots, Z_{m+n}\}$ no longer depends on the data distribution, and is always given by uniform sampling of a permutation of these observed values.

Robustness of the permutation t-test

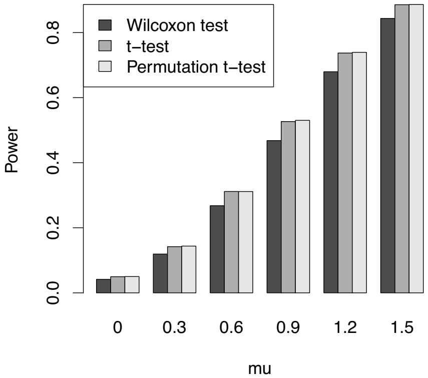


Data: 10% $\mathcal{N}(10,1)$, 90% $\mathcal{N}(0,1)$

Sample sizes: control $m = 30$, experiment $n \in \{2, 5, 10, 30\}$

(1000 permutations per simulation)

Power of the permutation t-test



Data: $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 1)$ and $Y_1, \ldots, Y_m \sim \mathcal{N}(0, 1)$
Sample sizes $m = n = 8$ (1000 permutations per simulation)

Two-sample testing in higher dimensions

Suppose

$$X_1, \ldots, X_n \overset{IID}{\sim} F, \qquad Y_1, \ldots, Y_m \overset{IID}{\sim} G$$

are data in a general metric space, e.g. images or documents represented in a feature space $\mathbb{R}^p$. We wish to test

$$H_0 : F = G \qquad \text{vs.} \qquad H_1 : F \neq G$$

There may not be a reasonable notion of "ordering" or "rank" for the data. Instead, many test statistics have been proposed:

- Compute the average distances $d_{XY} = \frac{1}{nm} \sum_{i,j} d(X_i, Y_j)$, $d_{XX} = \frac{1}{\binom{n}{2}} \sum_{i<i'} d(X_i, X_{i'})$, $d_{YY} = \frac{1}{\binom{m}{2}} \sum_{j<j'} d(Y_j, Y_{j'})$. Set

$$T = 2d_{XY} - d_{XX} - d_{YY}$$

Two-sample testing in higher dimensions

- For each observation X_i and Y_j, count how many of its k nearest neighbors come from the same sample as itself. Take

  $$T = \text{ average of this count across all } m+n \text{ observations}$$

- Construct a minimal spanning tree of

  $$\{X_1, \ldots, X_n, Y_1, \ldots, Y_m\}$$

  (This is the tree connecting all $m+n$ observations and having smallest total edge length.) Delete those edges whose endpoints do not belong to the same sample. Take

  $$T = \text{number of remaining connected components}$$

These statistics have complex distributions, and also may not be exactly pivotal under H_0, but one may use them to test H_0 in a permutation testing framework.