Two Sample Problem

Paweł Polak

March 22, 2016

STAT W4413: Nonparametric Statistics - Lecture 12

Introduction

So far, we have seen several instances of the one sample test:

- sign test,
- one-sample Kolmogorov-Smirnov test, and
- Lilliefors test.

In one sample test, the assumption is that we observe $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F$ and we would like to test some properties of F.

For instance, in the sign test we would like to know if the median of F has certain value μ or in Lilliefors test we would like to know whether F is Gaussian or not.

Another problem of major importance is known as two-sample test.

What is a two sample test?

In a two sample test we observe two sets of samples:

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} F \quad \text{and} \quad Y_1, Y_2, \ldots, Y_m \overset{iid}{\sim} G,$$

and we would like to test some relation between F and G.

$$H_0 : \mathcal{R}(F, G)$$

Example:
Suppose that you are working for a company and the researchers in the company claim that they have a better drug for a certain disease.

How would you evaluate the correctness of this claim?

What is a two sample test?

- We can first have several subjects that are suffering from the disease and give some of them the old drug and ask the rest to use the new drug.
- Then, you check the patients after a while and record their improvement.
- Let's call the improvement of the subjects that used the old drug $X_1, X_2, \ldots, X_n$, and the subjects that used the new drug, $Y_1, Y_2, \ldots, Y_m$.
- Given these numbers we would like to compare the performance of the new drug with the old one.
- This is going to be the discussion of this lecture.

Parametric two-sample tests

Recall that in the "parametric setting" we always assume that the distribution of the data is given in a parametric form and the test reduces to questions regarding the parameters of the distributions.

Of course, the most popular distribution that is considered in most cases is Gaussian.

Here we summarize a few popular two-sample parametric tests that are designed for Gaussian distribution.

Parametric two-sample tests

Example:

- A company is planning to change their instructions for the new employees.
- Since the new instruction is slightly more time consuming and hence more costly, they are worried that the new instruction will not meet their expectation and does not improve the performance of the employees.
- Therefore, they pick seven new employees and use the traditional method on 3 of them and the new method on 4 of them.
- After the training they took the same test from all the employees. The results are shown in Table 1.

Table : The grades of the employees after receiving different trainings.

New method	37, 49, 55, 57
Traditional method	23, 31,46

The main question the company would like to answer now is whether the new approach has improved the grades or not.

Two-sample Z-test

We can assume:

- that the results of the new training are called $X_1, \ldots, X_n$ and the results of the old training are called $Y_1, Y_2, \ldots, Y_n$;
- the Gaussianity of our dataset, i.e., $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu_x, \sigma_x^2)$ and $Y_1, Y_2, \ldots, Y_m \overset{iid}{\sim} N(\mu_y, \sigma_y^2)$, where σ^2 is known;
- that both samples have the same variances, i.e., $\sigma_x^2 = \sigma_y^2$;

Now, the improvement in the training can be translated to an increase in the mean and we can test for

$$H_0 : \mu_x = \mu_y \quad \text{vs.} \quad H_1 : \mu_x > \mu_y.$$

Two-sample Z-test

- The null hypothesis indicates that the new training is the same as the old training, while H_1 implies an improvement in the new training.
- The most natural test that we can have to check the validity of the null hypothesis is based on the Z statistic defined by

$$Z \triangleq \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}},$$

  where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and $\bar{Y} = \frac{1}{m}\sum_{i=1}^{m} Y_i$.
- When H_0 is true, we expect the Z statistic to be close to zero. I H_0 is violated we would expect Z to be large.
- As is clear under the null hypothesis the distribution of Z is

$$\frac{\bar{X} - \bar{Y}}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} \sim N(0,1).$$

- Therefore we can easily characterize the probability of type I error and/or the p-value.
- This test is called the Z-test.

Two-sample Z-test

In order to come up with the Z-test we made several assumptions on the data:

R1: The variance of the distribution is given.

R2: The variance of the two samples are the same.

R3: The two samples are distributed according to the Gaussian distribution.

As is clear all these assumptions are very restrictive. Therefore, in the next few slides we would like to remove these constraints one by one.

Two-sample T-test

- Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu_x, \sigma^2)$ and $Y_1, Y_2, \ldots, Y_m \overset{iid}{\sim} N(\mu_y, \sigma^2)$ where σ^2 is NOT known.
- This is the main difference between the assumptions of the Z-test and T-test. In other words, here we remove restriction R1.
- As before we are interested in checking the null hypothesis

$$H_0 : \mu_x = \mu_y \quad \text{versus} \quad H_1 : \mu_x > \mu_y.$$

- We define the T statistic as

$$T \triangleq \frac{\bar{X} - \bar{Y}}{S\sqrt{\frac{1}{n} + \frac{1}{m}}},$$

where

$$S \triangleq \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{m}(Y_i - \bar{Y})^2}{n + m - 2}},$$

and $\bar{X}$ and $\bar{Y}$ are the empirical averages.

Two-sample T-test

- As before we should reject the null hypothesis for large values of T and we accept it for small values.

- Note that under the null hypothesis the distribution of $\widehat{X} - \widehat{Y}$ is $N(0, \sigma^2/m + \sigma^2/n)$.

- Also $\bar{X}$ is independent of $\sum_{i=1}^{n}(X_i - \bar{X})^2$. If you do not know why, then check the Appendix in the next set of slides.

- $\bar{X}$ is also independent of $\sum_{i=1}^{m}(Y_i - \bar{Y})^2$.

- Similar statements hold for $\bar{Y}$. Therefore, $\bar{X} - \bar{Y}$ is independent of S.

- Under the null hypothesis $(n + m - 2)S^2/\sigma^2$ has χ^2 distribution with $m + n - 2$ degrees of freedom. If you do not understand this, study the appendix in the next set of slides.

Two-sample T-test

- Combining these results, we conclude that $\frac{\bar{X}-\bar{Y}}{S\sqrt{\frac{1}{n}+\frac{1}{m}}}$ is a T distribution with $m+n-2$ degrees of freedom. (If $Z \sim N(0,1)$ and V is $\chi^2(v)$, then $\frac{Z}{\sqrt{V/v}}$ has the T distribution with v degrees of freedom.)

- As you can see we removed one of the most unrealistic assumptions of the Z-test that assumes the variance is known (R1).

- Can we remove R2 as well? We address this question in the next slides.

Behrens-Fisher Problem

If we remove restriction $R2$, then we should model the problem in the following way.

- Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu_x, \sigma_x^2)$ and $Y_1, Y_2, \ldots, Y_m \overset{iid}{\sim} N(\mu_y, \sigma_y^2)$ where σ_x^2 and σ_y^2 are NOT known.
- As in the Z-test and T-test we are interested in checking the null hypothesis

$$H_0 : \mu_x = \mu_y \quad \text{versus} \quad H_1 : \mu_x > \mu_y.$$

This problem is known as the Behrens-Fisher problem.

Behrens-Fisher Problem

- Using similar heuristics one can come up with the following statistic for testing H_0:

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

where S_X^2 and S_Y^2 are unbiased estimates of the variance of X and variance of Y, e.g.,

$$S_X^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

$$S_Y^2 = \frac{1}{m-1} \sum_{i=1}^{m} (Y_i - \bar{Y})^2.$$

Behrens-Fisher Problem

- The main challenge here is that under the null hypothesis the distribution of T_{BF} may still depend on the actual values of σ_X and σ_Y that are not given. You will double-check this in the homework.

- However, this dependence is usually weak. A good approximation for the distribution of $\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{S_X^2}{n}+\frac{S_Y^2}{n}}}$ under the null hypothesis is given by Welch.

- According to Welch this distribution is still close to a T-distribution with certain number of degrees of freedom. You may check this formula in Wikipedia.

Two-Sample Tests

So far we have been able to remove the first two constraints of the Z test, i.e.,

- the assumption that the variances are equal and known.
- A take-home message of this example is that as soon as we remove some of the assumptions and make the model more realistic, then characterizing the distribution under null becomes very difficult.
- We will get back to this point and provide an easy way to resolve it.

The last assumption is about the Gaussianity of the data that we would like to remove next.

This will lead us to our discussion of the non-parametric tests.

But, before we move to the nonparametric tests we study one last famous parametric test, i.e., the F-test.

F-test

- Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu_x, \sigma_x^2)$ and $Y_1, Y_2, \ldots, Y_m \overset{iid}{\sim} N(\mu_y, \sigma_y^2)$ where σ_x^2.
- We would like to check the hypothesis:

$$H_0 : \sigma_X^2 = \sigma_Y^2 \quad \text{versus} \quad H_1 : \sigma_X^2 \neq \sigma_Y^2.$$

- The most natural statistic that can be used for testing H_0 is the F statistic as given by

$$F = \frac{S_X^2}{S_Y^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})^2/(n-1)}{\sum_{i=1}^m (Y_i - \bar{Y})^2/(m-1)}.$$

- Based on this statistic we reject H_0 if and only if $F \geq \kappa_2$ or $F \leq \kappa_1$.
- Under the null hypothesis one can prove that the F statistic has F distribution with parameters $n-1$ and $m-1$. [The ratio of two independent χ^2 distributions has F distribution.]
- Once we know this fact, we can easily characterize the probability of Type I error and also the p-value.

Weaknesses of parametric tests

- As is clear the main underlying assumption of all the discussions in the last section is the "Gaussian" assumption. However, Gaussian assumption is rarely met in practice.
- Remember that you know how to test if this assumption holds or not. You should use goodness of the fit tests.
- The argument that enables us to use parametric tests in practice is that some of these tests are, to some extent, robust to the non-Gaussianity of the data.
- For instance this argument is true for the T-test (specially if the sample size is reasonably large), but it is not true for the F-test.
- Nevertheless a better approach is to design tests that do not rely on any distributional assumption. Such tests are called *nonparametric*.

Weaknesses of parametric tests

- Another disadvantage of the parametric tests is that for any testing problem we should try to do some analytical work to characterize the distribution of the test statistic under the null hypothesis.

- This is most of the times difficult for practitioners and they usually look for some simpler tests that do not require any analytical work.

- As a result we set the following two goals for ourselves:

  G1: Providing general tests that do not rely on the Gaussian assumption.

  G2: Provide a simple approach that does not require much analytical work for characterizing the p-value.

Nonparametric test for the location (shift) problem

Modeling hypotheses

- Let's consider the example we considered in the last slides on the new training program of the company.
- So far, we have modeled the data as Gaussian random variables.
- The first challenge that we should address here is the following: how do you model the data if you want to avoid the Gaussianity assumption and
- once you model the data how do you cast the question of the company (whether the new training is better or not) as a testing problem?
- I suggest two different models and correspondingly two different hypotheses.

Nonparametric test for the location (shift) problem

Model 1:

1. Here we assume that $Y_1, Y_2, \ldots, Y_m \overset{iid}{\sim} F(y)$ and $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F(x - \Delta)$, and
2. that F is not known; (this is the main deviation from the Gaussian case).
3. Then if the new training has no advantage over the old training we can say that $\Delta = 0$, but if it has some positive impact then we can argue that $\Delta > 0$. Can you intuitively explain why positive Δ means positive impact of the new training?
4. Therefore, under this model H_0 and H_1 can be explained as:

$$H_0 : \Delta = 0 \quad \text{versus} \quad H_1 : \Delta > 0.$$

Nonparametric test for the location (shift) problem

Model 2:

1. In model 1, we assumed that the distribution of $X_1, X_2, \ldots, X_n$ is a shifted version of the distribution of $Y_1, Y_2, \ldots, Y_m$.
2. If we want to remove this constraint we can model the problem differently.
3. Here is how we can model it: $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F_1(x)$ and $Y_1, Y_2, \ldots, Y_m \overset{iid}{\sim} F_2(y)$.
4. Then we set H_0 and H_1 in the following way:
$$H_0 : F_1(x) = F_2(x) \;\; \forall x \quad \text{versus} \quad H_1 : F_1(x) < F_2(x) \;\; \forall x:$$
5. Can you convince yourself why H_1 means that the new training has been better than the old training?
6. As before, I assume that F_1 and F_2 are NOT known.

Nonparametric test for the location (shift) problem

Now that we could cast our problem as testing, the main remaining challenge is to propose test statistics whose distribution under H_0 does not depend on F in Model 1 and F_1 and F_2 in Model 2.

We first start with Model 1:

One of the main statistics that enable us to address non-parametric testing is called rank-statistic. We first review this statistic and will later use it to come up with good nonparametric test for Model 1.

Rank and order statistics

Given some observations $X_1, X_2, \ldots, X_n$ the rank of X_i is defined as

$$rank(X_i) = \sum_{j=1}^{n} \mathbb{I}(X_j \leq X_i).$$

Hence $rank(X_i)$ denotes the number of elements in the dataset that are less than or equal to X_i.

Example

Let $X_1 = 1, X_2 = 2, X_3 = 1.5, X_4 = 2.5, X_5 = 3$. Then we have $rank(X_1) = 1, rank(X_2) = 3, rank(X_3) = 2, rank(X_4) = 4, rank(X_5) = 5$.

The rank statistic is very popular in many nonparametric tests. The following theorem explains why.

Rank and order statistics

Lemma

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F$. Then $(rank(X_1), \ldots, rank(X_n))$ is uniformly distributed over the set of all permutations of $\{1, 2, \ldots, n\}$. For instance, $p(rank(X_1) = 1, rank(X_2) = 2, \ldots, rank(X_n) = n) = p(rank(X_1) = n, rank(X_2) = n - 1, \ldots, rank(X_n) = 1) = \frac{1}{n!}$

You will prove this lemma in the homework. This is a very intuitive lemma. So, while you are proving it, make sure you gain some intuition why this holds.

Rank and order statistics

According to this Lemma the distribution of the rank statistic is independent of F.

This is an appealing property for nonparametric statistic where F can be any distribution and is not known. Another interesting property of the rank statistic is given below.

Lemma

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F$. Then $(rank(X_1), \ldots, rank(X_n))$ is independent of the ordered statistic $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$.

Try to prove this result for yourself as well. Essentially you will prove it while you are proving the previous Lemma. Think about it.

Wilcoxon rank-sum test

Now that we have introduced the rank statistic, it is now time to see how we can use it for the following hypothesis testing problem:

- We have observed $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F(x - \Delta)$ and $Y_1, Y_2, \ldots, Y_m \overset{iid}{\sim} F(x)$, and we would like to test
$$H_0 : \Delta = 0 \quad \text{versus} \quad H_1 : \Delta > 0.$$
- Suppose that we combine all the samples $X_1, X_2, \ldots, X_n, Y_1, \ldots, Y_m$ together and calculate the rank of each data point.
- Wilcoxon rank-sum statistic is defined as
$$W = \sum_{i=1}^{m} rank(Y_i).$$
where the ranks of Y_i's are defined in the combined sample[1]

[1] There is a nice discussion of rank tests and their optimality properties in the book "Testing statistical hypothesis" by Lehman and Romano. Interested readers may refer to pages 239-242 of this book.

Wilcoxon rank-sum test

Suppose for a moment that H_0 holds. Then we expect $\sum_{i=1}^{m} rank(Y_i)$ to be around its expected value. Let's calculate its expected value. First, we simplify the W statistic.

$$W = \sum_{i=1}^{m} rank(Y_i) = \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{I}(X_j \leq Y_i) + \sum_{i=1}^{m} \sum_{j=1}^{\overset{\mathbf{m}}{n}} \mathbb{I}(Y_j \leq Y_i) \qquad (1)$$

The first expression on the right, i.e.

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{I}(X_j \leq Y_i),$$

is called Mann-Whitney statistic (U). It is straightforward to check that the second term in (1) satisfies:

$$\sum_{i=1}^{m} \sum_{j=1}^{n} \mathbb{I}(Y_j \leq Y_i) = \frac{m(m+1)}{2}.$$

In summary, we have

$$W = U + \frac{m(m+1)}{2}.$$

Wilcoxon rank-sum test

Now we can characterize the expected value of W under the null hypothesis. In other words,

$$\begin{aligned}\mathbb{E}_{H_0}(W) &= \mathbb{E}_{H_0}(U) + \frac{m(m+1)}{2} = \mathbb{E}_{H_0}(\sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{I}(X_j \leq Y_i)) + \frac{m(m+1)}{2} \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n}\mathbb{P}_{H_0}(X_j \leq Y_i) + \frac{m(m+1)}{2} = \frac{mn}{2} + \frac{m(m+1)}{2}. \qquad (2)\end{aligned}$$

Hence under the null hypothesis we expect W to be around $mn/2 + m(m+1)/2$.

What if H_1 is true?

Wilcoxon rank-sum test

- Intuitively speaking we expect the rank of Y_i's to decrease and hence W to decrease. This is intuitive since, the values of X_i's are larger in general.

- You can also see that the expected value of W will be smaller. In particular,

  $$\mathbb{E}_{H_1}(U) = mn\mathbb{P}(X_j \leq Y_i).$$

  Can you show this? Can you argue why $\mathbb{P}(X_j \leq Y_i)$ decreases as Δ increases? This argument leads us to Wilcoxon rank-sum test:

  $$\text{Reject } H_0 \text{ if } W < T.$$

The next question we should address now is the calculation of probability of Type I error. Specially note that we do not have access to F.

Wilcoxon rank-sum test

The following lemma indicates an important feature of this statistic.

Lemma

Under the null hypothesis H_0 the distribution of W is free of F.

The proof is a simple application of Lemma above.

According to this lemma we do not require any knowledge of the distribution F.

There are some analytical ways to characterize the distribution of W under H_0. We do not want to go in that direction.

Wilcoxon rank-sum test

Instead, similar to KS test you can imagine how for instance you can use Monte Carlo simulation to characterize the CDF of W under H_0.

Wilcoxon rank-sum test can be used for our second model as well. Let me review our second model. We assume that $X_1, X_2, \ldots, X_n \sim F_1(x)$ and $Y_1, Y_2, \ldots, Y_m \sim F_2(y)$ and we would like to test

$$H_0 : F_1(x) = F_2(x) \ \ \forall x \quad \text{vs.} \quad H_1 : F_1(x) < F_2(x) \ \ \forall x.$$

Convince yourself that Wilcoxon rank-sum statistic will show exactly the same behavior that it showed for our previous model.

Also, under H_0 the distribution of W is free of both F_1 and F_2. why?

Wilcoxon rank-sum test

As we have seen so far the rank statistic led to good tests for testing the alternate hypotheses of the form $F_1(x) < F_2(x)$.

Now, let's change H_1 a little bit:

- Consider again the training program example we have discussed so far and assume that the company would like to know whether the new training has had any effect on the grades of the new employees.

- We can model their question in the following way:

  - Given two samples $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F_1(x)$ and $Y_1, Y_2, \ldots, Y_m \overset{iid}{\sim} F_2(y)$ we can test for
  $$H_0 : F_1(x) = F_2(x) \;\; \forall x \quad \text{versus} \quad H_1 : F_1(x) \neq F_2(x) \;\; \exists x.$$

Wilcoxon rank-sum test

Can we use the rank statistics for this testing problem? The answer is NO.

This problem is know as the two-sample goodness-of-fit problem and people have derived such statistics for it.

The idea is very simple:

- first estimate $F_1(x)$ and $F_2(x)$, and call our estimates $\hat{F}_1(x)$ and $\hat{F}_2(x)$,
- then similar to Kolmogorov-Smirnov test we calculate the following distance between $\hat{F}_1(x)$ and $\hat{F}_2(x)$:

$$K_2 = \sum_x |\hat{F}_1(x) - \hat{F}_2(x)|.$$

Note that under the null hypothesis we expect K_2 to be small and under H_1 we expect it to be large. Hence, we will reject H_0 if $K_2 > T$, where T is some threshold.

Wilcoxon rank-sum test

It turns out that the distribution of K_2 is free of F_1 and F_2 under the null hypothesis.

Proposition

Under the null hypothesis for two sample tests, the distribution of K_2 is free of F_1 and F_2.

Proof.

The proof of this proposition is an optional part of the course. You can skip it if you are not interested. For notational simplicity I will use $\hat{F}$ and $\hat{G}$ instead of $\hat{F}_1(x)$ and $\hat{F}_2(x)$ respectively.
We focus on the CDF of K_2, $\mathbb{P}(K_2 \leq \gamma)$. Again similar to the derivations for the one sample Kolmogorov- Smirnov test we have

$$\mathbb{P}(K_2 \leq \gamma) = \mathbb{P}\left(-\gamma + \frac{i}{n} \leq \hat{G}(X_{(i)}) \leq \gamma + \frac{i-1}{n} \ \text{ for } i = 1, 2, \ldots, n\right) \tag{3}$$

□

Wilcoxon rank-sum test

Proof.

Note two things about this expression:

- $X_{(i)}$s are the ordered version of $X_1, X_2, \ldots, X_n$;
- unlike the one-sample case where the randomness was only on $X_1, X_2, \ldots, X_n$, here $\hat{G}$ is also a random variable and hence we should be more careful in calculating $\mathbb{P}(K_2 \leq \gamma)$.

In situations where we have two sources of randomness, one popular approach in calculating probabilities is to first condition on $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ and calculate the probability.
Then we will be able to calculate the entire probability through the following calculations:

$$\begin{aligned} \mathbb{P}(K_2 \leq \gamma) &= \mathbb{P}\left(-\gamma + \frac{i}{n} \leq \hat{G}(X_{(i)}) \leq \gamma + \frac{i-1}{n} \text{ for } i = 1, 2, \ldots, n\right) \\ &= \int \mathbb{P}(-\gamma + \frac{i}{n} \leq \hat{G}(X_{(i)}) \leq \gamma + \frac{i-1}{n} \text{ for } i = 1, \ldots, n | x_{(1)}, \ldots, x_{(n)}) \\ &\quad f_{X_{(1)},\ldots,X_{(n)}}(x_{(1)}, \ldots, x_{(n)}) dx_{(1)}, \ldots, dx_{(n)} \end{aligned} \tag{4}$$

Wilcoxon rank-sum test

Proof.

Therefore, the problem is simplified to the calculation of $\mathbb{P}(-\gamma + \frac{i}{n} \leq \hat{G}(X_{(i)}) \leq \gamma + \frac{i-1}{n}$ for $i = 1, \dots, n | x_{(1)}, \dots, x_{(n)})$.

Now we assume that the only randomness in this expression is the randomness of $Y_1, Y_2, \dots, Y_m$ that is reflected through $\hat{G}$.

Next we would like to characterize the joint distribution of $\hat{G}(X_{(1)}), \hat{G}(X_{(2)}), \dots, \hat{G}(X_{(n)})$ given the assumption that $X_{(1)}, \dots, X_{(n)}$ are fixed.

- note that the random variable $\hat{G}(X_{(i)})$ can only take values $\{0, \frac{1}{m}, \frac{2}{m}, \dots, \frac{m}{m} = 1\}$. Why?

□

Wilcoxon rank-sum test

Proof.

Therefore we can characterize the joint probability mass function of $\hat{G}(X_{(1)}), \hat{G}(X_{(2)}), \ldots, \hat{G}(X_{(n)})$. We have

$$\begin{aligned} &\mathbb{P}\left(\hat{G}(X_{(1)}) = \frac{\ell_1}{m}, \hat{G}(X_{(2)}) = \frac{\ell_2}{m}, \ldots, \hat{G}(X_{(n)}) = \frac{\ell_n}{m}\right) \\ &\overset{(a)}{=} \mathbb{P}\left(\hat{G}(X_{(1)}) = \frac{\ell_1}{m}, \hat{G}(X_{(2)}) - \hat{G}(X_{(1)})\frac{\ell_2 - \ell_1}{m}, \ldots, \hat{G}(X_{(n)}) - \hat{G}(X_{(n-1)}) = \frac{\ell_n - \ell_{n-1}}{m}\right) \quad (5) \end{aligned}$$

Note that $\hat{G}(X_{(1)}) = \frac{\ell_1}{m}$ means that exactly ℓ_1 of the $Y_1, \ldots, Y_n$ are below $X_{(1)}$. Then the event $\hat{G}(X_{(1)}) = \frac{\ell_1}{m}, \hat{G}(X_{(2)}) - \hat{G}(X_{(1)}) = \frac{\ell_2 - \ell_1}{m}$ is equivalent to the event that ℓ_1 of the $Y_1, \ldots, Y_n$ are below $X_{(1)}$ and $\ell_2 - \ell_1$ of them are in the interval $[X_{(1)}, X_{(2)})$. □

Wilcoxon rank-sum test

Proof.

Using the same argument we conclude that $\mathbb{P}(\hat{G}(X_{(1)}) = \frac{\ell_1}{m}, \hat{G}(X_{(2)}) - \hat{G}(X_{(1)}) = \frac{\ell_2 - \ell_1}{m}, \ldots, \hat{G}(X_{(n)}) - \hat{G}(X_{(n-1)}) = \frac{\ell_n - \ell_{n-1}}{m})$ is equal to the probability that ℓ_1 of the $Y_1, \ldots, Y_n$ are below $X_{(1)}$ and $\ell_2 - \ell_1$ of them are in the interval $[X_{(1)}, X_{(2)})$, and $\ldots, \ell_n - \ell_{n-1}$ of them are in the interval $[X_{n-1}, X_n)$.

We also know that

$$\begin{aligned} \mathbb{P}(Y_i \in (-\infty, X_{(1)})) &= G(X_{(1)}), \\ \mathbb{P}(Y_i \in (X_{(1)}, X_{(2)})) &= G(X_{(2)}) - G(X_{(1)}), \\ &\vdots \\ \mathbb{P}(Y_i \in (X_{(n-1)}, X_{(n)})) &= G(X_{(n)}) - G(X_{(n-1)}). \qquad (6) \end{aligned}$$

□

Wilcoxon rank-sum test

Proof.

Combining (5) and (6) we obtain

$$\mathbb{P}\left(\hat{G}(X_{(1)}) = \frac{\ell_1}{m}, \hat{G}(X_{(2)}) = \frac{\ell_2}{m}, \ldots, \hat{G}(X_{(n)}) = \frac{\ell_n}{m}\right) \tag{7}$$

$$\frac{m!}{\ell_1!(\ell_2-\ell_1)!\ldots(\ell_n-\ell_{n-1})!(m-\ell_n)!}$$

$$(G(X_{(1)}))^{\ell_1}[G(X_{(2)})-G(X_{(1)})]^{\ell_2-\ell_1}\ldots(G(X_{(n)})-G(X_{(n-1)}))^{\ell_n-\ell_{n-1}}(1-G(X_{(n)}))^{m-\ell_n} \tag{8}$$

Note the first important implication of the above result. If we assume the null hypothesis is true i.e. $F = G$, then the distribution of $G(X_{(1)}), \ldots, G(X_{(n)})$ is independent of F and G. Why?
The next step would be to calculate the probability specified in (4). However, instead of taking the integral with respect to $X_1, \ldots, X_n$ we will change the variables to $z_1 \triangleq G(X_{(1)}), \ldots, z_n \triangleq G(X_{(n)})$.

□

Wilcoxon rank-sum test

Proof.

Define $C^m_{\ell_1,\ell_2,\ldots,\ell_n} = \frac{m!}{\ell_1!(\ell_2-\ell_1)!\ldots(\ell_n-\ell_{n-1})!(m-\ell_n)!}$ we have

$$\begin{aligned}
&\int \mathbb{P}(-\gamma + \frac{i}{n} \leq \hat{G}(x_{(i)}) \leq \gamma + \frac{i-1}{n} \ \text{ for } i = 1,\ldots,n | x_{(1)},\ldots,x_{(n)}) \\
&f_{x_{(1)},\ldots,x_{(n)}}(x_{(1)},\ldots,x_{(n)}) dx_{(1)},\ldots,dx_{(n)} \\
&= \int \sum_{\ell_1,\ldots,\ell_n} C^m_{\ell_1,\ell_2,\ldots,\ell_n} z_1^{\ell_1}(z_2 - z_1)^{\ell_2-\ell_1}\ldots(z_n - z_{n-1})^{\ell_n-\ell_{n-1}}(1-z_n)^{m-\ell_n} \qquad (9) \\
&f_{z_1,z_2,\ldots,z_n}(z_1,\ldots,z_n) dz_1 \ldots dz_n.
\end{aligned}$$

Note that under the null hypothesis $f_{z_1,z_2,\ldots,z_n}(z_1,\ldots,z_n)$ is free of F or G and hence we conclude that this probability is independent of the null distributions. □

Wilcoxon rank-sum test

- Rank-statistics and some other statistics such as Kolmogorov-Smirnov are capable of providing nonparametric tests whose distributions under H_0 do not depend on our distributional assumptions.
- However, as soon as we change the alternate hypothesis we have to go through a usually tedious process of coming up with good test statistics whose distributions under H_0 is free of the actual distributions.
- While for the standard problems we have considered so far well-known tests exist, it may not be an easy task for the problems we may see in practice.
- Furthermore, even if we can come up with some statistic that has the following two properties:
  1. It can distinguish between H_0 and H_1.
  2. Its distribution under null is free of the distributions involved in the null hypothesis,

Wilcoxon rank-sum test

- It is not necessarily the best statistic to distinguish between H_0 and H_1.
- Usually coming up with statistics that have the first property, i.e., distinguishing between H_0 and H_1, is not difficult.
- The main issue in the design of the test is the second constraint we imposed above. Can we get rid of that condition? The answer is yes.
- Permutation tests have provided a simple approach for using test statistics without worrying about the second property we described above.
- While these tests were not very popular 20 years ago (due to their high computational complexity), they have become more and more popular these years.
- As you will see these tests provide very flexible alternative.