18.650: Fundamentals of Statistics October 30, 2023

Lecture 18 — Nonparametric tests

Prof. Philippe Rigollet *Scribe: Anya Katsevich*

So far we have looked at parametric tests, e,g

$$\begin{array}{ll} H_0 : \theta \leq 0 & \text{vs} \quad H_1 : \theta > 0 \\ \theta = 0 & \theta \neq 0 \\ \mu \leq 30 & \mu > 30 \end{array}$$

Non-parametric tests are tests for the *entire* distribution. For example,

$$\begin{array}{ll} H_0 : X \sim \mathcal{N}(0,1) & \text{vs} \quad H_1 : X \sim \text{any other dist.} \\ X \text{ is Gaussian} & X \text{ is non-Gaussian} \\ X \overset{d}{=} Y & X \overset{d}{\neq} Y \end{array}$$

Remark.

Consider the test $H_0 : X \sim \mathcal{N}(0,1)$ vs $H_1 : X \sim$ any other distribution. This is *not* the same as $H_0 : X \sim \mathcal{N}(0,1)$ vs $H_1 : X \sim \mathcal{N}(\theta, 1)$ for some $\theta \neq 0$. In the latter test, the alternative hypothesis is a much smaller class of distributions then the alternative hypothesis in the former test. See Figure 1 for a visualization of the difference.

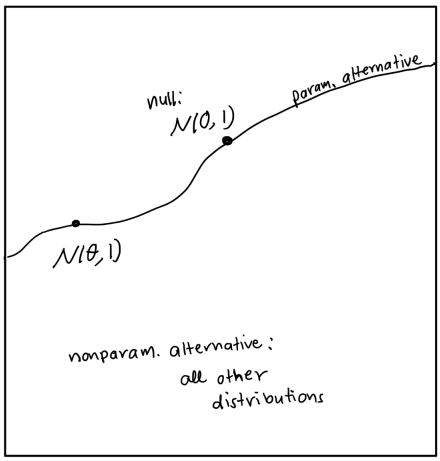

Figure 1: In the nonparametric case, the alternative to the null $H_0 : X \sim \mathcal{N}(0,1)$ is the set of all other distributions. In the parametric case, the alternative is the set of all Gaussian distributions $\mathcal{N}(\theta, 1)$ such that $\theta \neq 0$.

1 Kolmogorov-Smirnov test

Suppose $X_1, \ldots, X_n$ are iid with cdf F. We want to test whether $H_0 : F = F_0$ vs $H_1 : F \neq F_0$. (This is a goodness of fit test!) To do so, we use the *empirical cdf*:

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \leq t).$$

Properties of the empirical cdf.

1. $\mathbb{E}[\hat{F}_n(t)] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[\mathbb{1}(X_i \leq t)] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}(X_i \leq t) = F(t)$, so $\hat{F}_n$ is an unbiased estimator of F.

2. $\mathbb{1}(X_i \leq t) \sim \text{Ber}(F(t))$, and hence $n\hat{F}_n(t) \sim \text{Bin}(n, F(t))$.

3. The second property tells us $\hat{F}_n$ is an average of Bernoulli random variables. Therefore, $\sqrt{n}(\hat{F}_n(t) - F(t)) \rightsquigarrow \mathcal{N}(0, F(t)(1 - F(t))$ by the CLT.

Intuitively, we should reject the null if $|\hat{F}_n(t) - F_0(t)|$ is large for any t. So we look at the maximum discrepancy between the two over all possible t. See Figure 2 for a visualization of this maximum discrepancy.

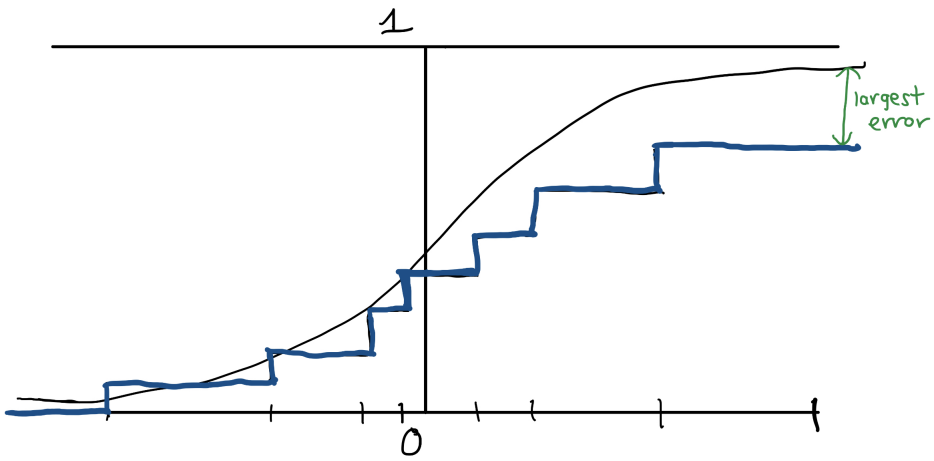

Figure 2: Empirical cdf compared to the true cdf; the dashes on the x-axis are the locations of the samples X_i.

This leads to the following rejection region:

$$R_\alpha = \{\sup_t |\hat{F}_n(t) - F_0(t)| > c_\alpha\}.$$

But how do we find c_α? It should satisfy

$$\mathbb{P}_{H_0}\left(\sup_t |\hat{F}_n(t) - F_0(t)| > c_\alpha\right) \approx \alpha,$$

where "under H_0" means that the data is drawn from F_0. So we need to understand the distribution of the test statistic $T := \sup_t |\hat{F}_n(t) - F_0(t)|$ when the $X_i \overset{\text{i.i.d.}}{\sim} F_0$, $i = 1, \ldots, n$.

Theorem 1.1: Distribution of T

If $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F_0$, then $T = \sup_t |\hat{F}_n(t) - F_0(t)|$ is distributed according to the Kolmogorov-Smirnov (KS) distribution, which *does not depend on* F_0!

The fact that the distribution of T does not depend on F_0 is powerful because it means we only need a single look-up table for the c_α. They are simply the quantiles of the KS distribution, regardless of what F_0 is. So whether F_0 is $\mathcal{N}(0, 1)$ or Poisson(22) or Exp(0.5), the choice of c_α is the same. Note that the KS distribution *does* depend on n.

1.1 Kolmogorov Lilliefors Test

Suppose we want to test

$$H_0\text{: } X \text{ is Gaussian vs } H_1\text{: } X \text{ is not Gaussian.}$$

This is a useful first hypothesis test to run on a dataset. If we confirm our data is Gaussian, we can then proceed to test some actual hypothesis of interest using one of the many tests that are based on the assumption of Gaussian data.

A natural test statistic to use is

$$T = \sup_t |\hat{F}_n(t) - \Phi_{\hat{\mu},\hat{\sigma}^2}(t)|, \tag{1}$$

where $\Phi_{\hat{\mu},\hat{\sigma}^2}$ is the cdf for the Gaussian $\mathcal{N}(\hat{\mu}, \hat{\sigma}^2)$, and $\hat{\mu}, \hat{\sigma}^2$ are the sample mean and variance, respectively.

However, it turns out that this test statistic does *not* have a KS distribution. This has to do with the fact that we've already matched the first two moments of the real distribution to that of the Gaussian we're testing against. Instead, T defined in (1) has a Kolmogorov-Lilliefors (KL) distribution, with corresponding modified quantiles c_α^{KL}. It turns out that $c_\alpha^{\text{KL}} < c_\alpha^{\text{KS}}$. This means that if we (incorrectly) used the KS quantiles, then the criterion to reject the null would be more stringent and we would end up retaining the null in cases when we shouldn't.

For example, when $n = 14$ and $\alpha = 5\%$, we have $c_\alpha^{\text{KL}} = 0.207$ and $c_\alpha^{\text{KS}} = 0.349$.

2 Permutation test

Often the setting is that we have two sets of samples $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} F_X$ and $Y_1, \ldots, Y_m \overset{\text{i.i.d.}}{\sim} F_Y$, where the X_i are also independent of the Y_j. We want to test $H_0 : F_X = F_Y$ vs $H_1 : F_X \neq F_Y$. This is called a 2-sample test.

Unlike the Gaussian vs non-Gaussian test (which is usually only preliminary), the 2-sample hypothesis test can itself be used to confirm or reject a scientific discovery. For example, consider testing whether or not the Higgs boson exists. The X_i are simulated data of a world without the Higgs boson, and the Y_i are real-world observations. Rejecting the null amounts to proving the Higgs boson must exist. (This hypothesis test was actually used in the discovery of the Higgs boson!)

So how should we test $H_0 : F_X = F_Y$ vs $H_1 : F_X \neq F_Y$?

Idea #1: use the test statistic $T := \sup_t |\hat{F}_X(t) - \hat{F}_Y(t)|$ and reject if it's larger than some c_α.
Issue: the distribution of T under the null, in which $F_X = F_Y = F$, depends on the actual cdf F. But we don't even have a candidate for this cdf! Therefore, we don't know how to choose c_α.

Idea #2: reject if $|\bar{X}_n - \bar{Y}_n| > c_\alpha$.
Issues: (1) this is a reasonable test from the perspective of type I error, but you could be committing a large type II error. Indeed, you would fail to reject if $F_X \neq F_Y$ but $\mathbb{E}[X] = \mathbb{E}[Y]$. (2) As in Idea #1, we again have the issue that the distribution of $|\bar{X}_n - \bar{Y}_n|$ depends on the true F. So we don't know how to choose c_α.

The **permutation test** is based on Idea #2. Assume for simplicity that $m = n$. We will reject if $T = |\bar{X}_n - \bar{Y}_n| > c_\alpha$. To find c_α, we will use a quantile of the T distribution. To approximate the T distribution, we draw new samples of T and construct a histogram. This is similar in spirit to the bootstrap.

How do we draw new samples of T? Let $Z_1, \ldots, Z_{2n}$ denote all $2n$ samples, which are iid under H_0. Suppose we reshuffle the samples and then split them up into two blocks. Under the null hypothesis, the absolute value of the difference between the sample means of these two blocks *has the same distribution as* $|\bar{X}_n - \bar{Y}_n|$!

Formally, let $\sigma : \{1, 2, \ldots, n\} \mapsto \{1, 2, \ldots, n\}$ denote any permutation of the indices. Split the data into the two blocks $\boxed{Z_{\sigma(1)}, \ldots, Z_{\sigma(n)}}, \boxed{Z_{\sigma(n+1)}, \ldots, Z_{\sigma(2n)}}$.

Now, let

$$\bar{Z}^{\sigma}_{\text{left}} = \frac{1}{n}\sum_{i=1}^{n} Z_{\sigma(i)}, \qquad \bar{Z}^{\sigma}_{\text{right}} = \frac{1}{n}\sum_{i=n+1}^{2n} Z_{\sigma(i)}, \quad T^{\sigma} = |\bar{Z}^{\sigma}_{\text{left}} - \bar{Z}^{\sigma}_{\text{right}}|. \tag{2}$$

Definition 2.1: Permutation test

Let $T = |\bar{X}_n - \bar{Y}_n|$ be the original test statistic, and $T^1, T^2, \ldots, T^B$ denote B test statistics T^{σ} constructed by reshuffling the data, splitting it into two blocks, and taking the absolute value of the difference between the sample means of the two blocks, as in (2). The permutation test at level α rejects the null if $T > t_\alpha$, where t_α is the αth quantitle of the T distribution, i.e. the point such that

$$\frac{1}{B}\sum_{b=1}^{B} \mathbb{1}(T^b \geq t_\alpha) = \alpha.$$

For a given observation T^{obs}, the p-value for the test is

$$\text{p-value} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}(T^b > T^{\text{obs}}).$$

Remarks.

- Note that T^{obs} in the above definition is the particular observed value of $T = |\bar{X}_n - \bar{Y}_n|$.
- The number of reshufflings B can be as large as $(2n)!$, though in practice it is infeasible to take B this large. Instead, we just use some subset of all possible reshufflings.
- The permutation test does not resolve issue #1 above: if two distinct distributions have the same mean, then the test will not reject the null.