Chapter 10: Nonparametrics

October 21, 2018

1 Tests of goodness-of-fit

In some problems, before we collect data we have some specific distribution in mind for the data we will observe. We can test if the data indeed comes from the specified distribution or not.

1.1 The χ^2-test

Suppose that a large population consists of items of k different types, and let p_i denote the probability that an item selected at random will be of type i, $i = 1, \ldots, k$.

Example: There are 4 blood types: A, B, AB, and O. We may be interested in testing whether data are consistent with a theory that predicts a particular set of blood types.

Of course, $p_i \geq 0$ for $i = 1, \ldots, k$, and $\sum_{i=1}^k p_i = 1$.

Let $p_1^{(0)}, \ldots, p_k^{(0)}$ be specific numbers such that $p_i^{(0)} > 0$ for $i = 1, \ldots, k$, and $\sum_{i=1}^k p_i^{(0)} = 1$ and suppose that we want to test the following hypotheses:

$H_0 : p_i = p_i^{(0)},$ for $i = 1, \ldots, k,$ versus $H_0 : p_i \neq p_i^{(0)}$ for at least one value of i.

We shall assume that a random sample of size n is to be taken from the given population.

For $i = 1, \ldots, k$, we let N_i denote the number of observations in the random sample that are of type i. Thus, $N_1, \ldots, N_k$ are nonnegative integers such that $\sum_{i=1}^k N_i = n$ and

$$(N_1, \ldots, N_k) \sim \text{Multinomial}(n, \mathbf{p} := (p_1, \ldots, p_k)).$$

When H_0 is true, $\mathbb{E}(N_i) = np_i^{(0)}$, for $i = 1, \ldots, k$.

The difference between the actual number of observations N_i and the expected number $np_i^{(0)}$ will tend to be smaller when H_0 is true than when H_0 is not true.

In 1900, Karl Pearson proved the following result, whose proof will not be given here.

Theorem 1. *The following statistic*

$$Q = \sum_{i=1}^{k} \frac{(N_i - np_i^{(0)})^2}{np_i^{(0)}} \tag{1}$$

has the property that if H_0 is true and the sample size $n \to \infty$, then

$$Q \xrightarrow{d} \chi^2_{k-1}.$$

Thus, we will reject the null hypothesis if

$$Q \geq c_\alpha,$$

where c_α can be taken as the $(1-\alpha)$ quantile of the χ^2 distribution with $k-1$ degrees of freedom.

This test is called the χ^2-*test of goodness-of-fit.*

Whenever the value of each expected count, $np_i^{(0)}$, for $i = 1, \ldots, k$, is not too small, the χ^2-distribution will be a good approximation to the actual distribution of Q.

Specifically, the approximation will be very good if $np_i^{(0)} \geq 5$ and the approximation should still be satisfactory if $np_i^{(0)} \geq 1.5$, for $i = 1, \ldots, k$.

1.1.1 Testing hypothesis about a continuous distribution

Example: Consider a continuous random variable X taking values in $[0, 1]$. We observe a random sample of size n and want to test the null hypothesis that the distribution is Uniform$[0, 1]$.

This is a *nonparametric* problem since the distribution of X might be any continuous distribution on $[0, 1]$.

Suppose that $n = 100$. We can divide $[0, 1]$ into 20 subintervals of equal length, namely $[0, 0.05), [0.05, 0.10), \ldots, [0.95, 1.00]$.

If the actual distribution is uniform, then the probability that each observation will fall within the i-th subinterval is $1/20$.

Let N_i denote the number of observations in the sample that actually fall within the i-th subinterval. Then the statistic Q can be written as

$$Q = \frac{1}{5} \sum_{i=1}^{20} (N_i - 5)^2,$$

and under the null hypothesis, Q will be approximately distributed as χ^2 with 19 degrees of freedom.

Suppose that we want to test whether a random sample of observations comes from a particular distribution. The the following procedure can be adopted:

(i) Partition the entire real line, or any particular interval that has probability 1, into a finite number of k disjoint subintervals. Generally, k is chosen so that the expected number of observations in each subinterval is at least 5, if H_0 is true.

(ii) Determine the probability $p_i^{(0)}$ that the particular hypothesized distribution would assign to the i-th subinterval, and calculate the expected number $np_i^{(0)}$ of observations in the i-th subinterval, $i = 1, \ldots, k$.

(iii) Count the number N_i of observations in the sample that fall within the i-th subinterval.

(iv) Calculate the value of Q as defined in (1). If the hypothesized distribution is correct, then Q will approximately follow a χ^2 distribution with $k - 1$ degrees of freedom.

1.2 Likelihood ratio tests for proportions

We could use the multinomial distribution to test for

$$H_0 : \mathbf{p} = \mathbf{p}^{(0)} \qquad \text{versus} \qquad H_1 : H_0 \text{ is not true.}$$

The likelihood function for the multinomial vector $\mathbf{X} = (N_1, \ldots, N_k)$ is

$$L(\mathbf{p}) = \binom{n}{N_1, \ldots, N_k} p_1^{N_1} \cdots p_k^{N_k}.$$

Show that

$$-2 \log \Lambda(\mathbf{X}) = -2 \sum_{i=1}^{k} N_i \log \left(\frac{np_i^{(0)}}{N_i} \right).$$

In order to apply Wilk's theorem (Theorem 9.1.4 in the book), the parameter space must be an open set in k-dimensional space. This is not true for the multinomial distribution if we let $\mathbf{p}$ to be the parameter (as $\sum_{i=1}^{k} p_i = 1$).

The set of probability vectors lies on a $(k-1)$ dimensional set of $\mathbb{R}^k$. However, we can effectively treat the vector $\boldsymbol{\theta} = (p_1, \ldots, p_{k-1})$ as the parameter, as $p_k = 1 - p_1 - \ldots - p_{k-1}$ is a function of $\boldsymbol{\theta}$.

As along as we believe that all the coordinates of $\mathbf{p}$ are strictly between 0 and 1, the set of possible values of the $(k-1)$-dimensional parameter $\boldsymbol{\theta}$ is open.

Therefore, by the Wilk's theorem, $-2\log\Lambda(\mathbf{X})$ is approximately χ^2 with $k-1$ degrees of freedom.

Exercise: Suppose that $Y_1, \ldots, Y_n$ is a random sample from a population with density function given by

$$f(y|\mathbf{p}) = \begin{cases} p_j & \text{if } y = j, \text{ where } j = 1,2,3 \\ 0 & \text{otherwise,} \end{cases}$$

where $\mathbf{p} = (p_1, p_2, p_3)$ is the vector of parameters such that $p_1 + p_2 + p_3 = 1$ and $p_j \geq 0$ for $j = 1,2,3$. Use the likelihood ratio test for testing

$$H_0 : p_1 = p_2 = p_3 \qquad \text{versus}q \quad H_1 : H_0 \text{ is not true.}$$

Use the level $\alpha = 0.05$.

Solution: We use the likelihood ratio (LR) test with statistic

$$\Lambda(\mathbf{Y}) = \frac{L(\frac{1}{3})}{L(\widehat{p_1}_{mle}, \widehat{p_2}_{mle}, \widehat{p_3}_{mle})} \quad \text{and} \quad RR: \; -2\log\Lambda > \chi^2_{0.05,2} = 5.99.$$

Note that under the null hypothesis $p_1 = p_2 = p_3 = \frac{1}{3}$, because $p_1 + p_2 + p_3 = 1$. Also under the alternative hypothesis we have

$$\widehat{p_j}_{mle} = \frac{k_j}{n} \quad \text{where } k_j \text{ is the number of times that we observe } j \text{, for } j = 1,2,3.$$

Therefore, we have

$$\Lambda = \frac{\left(\frac{1}{3}\right)^n}{\left(\frac{k_1}{n}\right)^{k_1}\left(\frac{k_2}{n}\right)^{k_2}\left(\frac{k_3}{n}\right)^{k_3}}$$

Thus we reject H_0 if

$$k_1\log\left(\frac{k_1}{n}\right) + k_2\log\left(\frac{k_2}{n}\right) + k_3\log\left(\frac{k_3}{n}\right) - n\log\left(\frac{1}{3}\right) > 2.995.$$

1.3 Goodness-of-fit for composite hypothesis

We can extend the goodness-of-fit test to deal with the case in which the null hypothesis is that the distribution of our data belongs to a particular parametric family.

The alternative hypothesis is that the data have a distribution that is not a member of that parametric family.

Thus, in the statistic Q, the probabilities $p_i^{(0)}$ are replaced by estimated probabilities based on the parametric family, and the degrees of freedom are reduced by the number of parameters.

We are now interested in testing the hypothesis that for each $i = 1, \ldots, k$, each probability p_i can be represented as a particular function $\pi_i(\boldsymbol{\theta})$ of a vector of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s)$. It is assumed that $s < k - 1$ and no component of $\boldsymbol{\theta}$ can be expressed as a function of the other $s - 1$ components.

We shall denote by Ω the s-dimensional parameter space of all possible values of $\boldsymbol{\theta}$. Furthermore, we will assume that the functions $\pi_1(\boldsymbol{\theta}), \ldots, \pi_k(\boldsymbol{\theta})$ always form a feasible set of value of $p_1, \ldots, p_k$ in the sense that for every value of $\boldsymbol{\theta} \in \Omega$, $\pi_i(\boldsymbol{\theta}) > 0$, for $i = 1, \ldots, k$, and $\sum_{i=1}^{k} \pi_i(\boldsymbol{\theta}) = 1$.

The hypothesis to be tested can be written as follows:

$H_0 :$ There exists a value of $\boldsymbol{\theta} \in \Omega$ such that $p_i = \pi_i(\boldsymbol{\theta})$ for $i = 1, \ldots, k$,

versus

$H_1 :$ The hypothesis H_0 is not true.

If $\hat{\boldsymbol{\theta}}$ denotes the MLE of $\boldsymbol{\theta}$ based on observations $N_1, \ldots, N_k$, then we define Q as

$$Q = \sum_{i=1}^{k} \frac{[N_i - n\pi_i(\hat{\boldsymbol{\theta}})]^2}{n\pi_i(\hat{\boldsymbol{\theta}})}.$$

Theorem 2. *Suppose that H_0 holds and certain regularity conditions are satisfied. Then, as $n \to \infty$,*

$$Q \xrightarrow{d} \chi^2_{k-1-s}.$$

Thus, we will reject H_0 if $Q \geq c_\alpha$, where c_α can be taken as the $1 - \alpha$ quantile of the χ^2-distribution with $k - 1 - s$ degrees of freedom.

1.3.1 Testing whether a distribution is normal

Consider now a problem in which a random sample $X_1, \ldots, X_n$ is taken from some continuous distribution for which the p.d.f is unknown, and it is desired to test the null hypothesis

$H_0 :$ distribution is normal versus $H_1 :$ distribution is NOT normal.

To perform the goodness-of-fit test in this problem, we divide the real line $\mathbb{R}$ into k subintervals and count the number N_i of observations in the random sample that fall into the i-th subinterval, $i = 1, \ldots, k$.

If the i-th subinterval is the interval (a_i, b_i), then

$$\pi_i(\mu, \sigma^2) = \int_{a_i}^{b_i} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] = \Phi\left(\frac{b_i - \mu}{\sigma}\right) - \Phi\left(\frac{a_i - \mu}{\sigma}\right).$$

It is important to note that in order to calculate the value of Q, the (ML) estimates $\hat{\mu}$ and $\hat{\sigma}^2$ must be found using the numbers $N_1, \ldots, N_k$. In other words, $\hat{\mu}$ and $\hat{\sigma}^2$ will be the values of μ and σ that maximize the likelihood function

$$L(\mu, \sigma^2) = [\pi_1(\mu, \sigma^2)]^{N_1} \cdots [\pi_k(\mu, \sigma^2)]^{N_k}.$$

Because of the complicated nature of the function $\pi_i(\mu, \sigma^2)$, a lengthy numerical computation would usually be required to determine the MLEs that maximize $L(\mu, \sigma^2)$.

The MLEs found using the original observations $X_1, \ldots, X_n$ should NOT be used!

Theorem 3. *Let $X_1, \ldots, X_n$ be a random sample from a distribution with a s-dimensional parameter $\boldsymbol{\theta}$. Let $\hat{\boldsymbol{\theta}}_n$ be the MLE computed from the X_i's. Partition the real line into $k > s+1$ disjoint intervals $I_1, \ldots, I_k$. Let N_i be the number of observations that fall into I_i, for $i = 1, \ldots, k$. Let*

$$\pi_i(\boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\theta}}(X_1 \in I_i).$$

Let

$$Q' = \sum_{i1=}^{k} \frac{[N_i - n\pi(\hat{\boldsymbol{\theta}}_n)]^2}{n\pi(\hat{\boldsymbol{\theta}}_n)}.$$

Assume that regularity conditions needed for the asymptotic normality of the MLE hold. Then, as $n \to \infty$,

$$Q' \xrightarrow{d} F$$

where F is a c.d.f that lies between the c.d.f of a χ^2_{k-1-s} and the c.d.f of a χ^2_{k-1}.

2 The sample distribution function

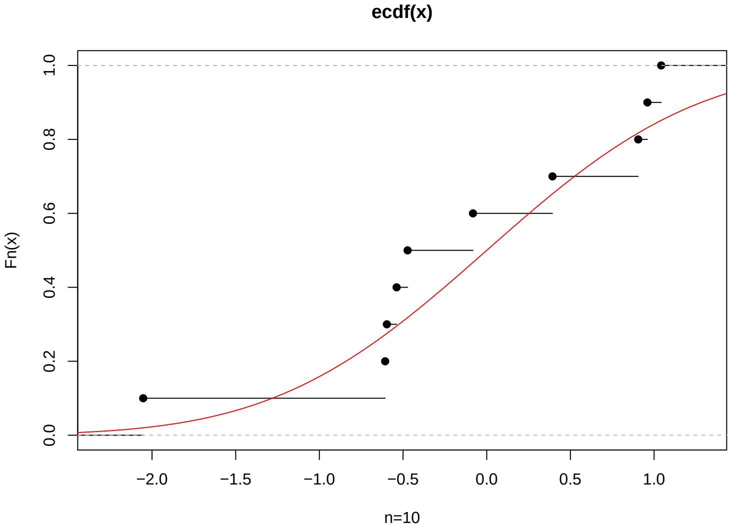
Let $X_1, \ldots, X_n$ be i.i.d F, where F is an unknown distribution function.

Question: We want to estimate F without assuming any specific parametric form for F.

Empirical distribution function (EDF): For each $x \in \mathbb{R}$, we define $F_n(x)$ as the proportion of observed values in the sample that are less than of equal to x, i.e.,

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}(X_i).$$

The function F_n defined in this way is called the *sample/empirical distribution function.*



Idea: Note that

$$F(x) = \mathbb{P}(X \le x) = \mathbb{E}[I_{(-\infty,x]}(X)].$$

Thus, given a random sample, we can find an *unbiased* estimator of $F(x)$ by looking at the proportion of times, among the X_i's, we observe a value $\le x$.

By the WLLN, we know that

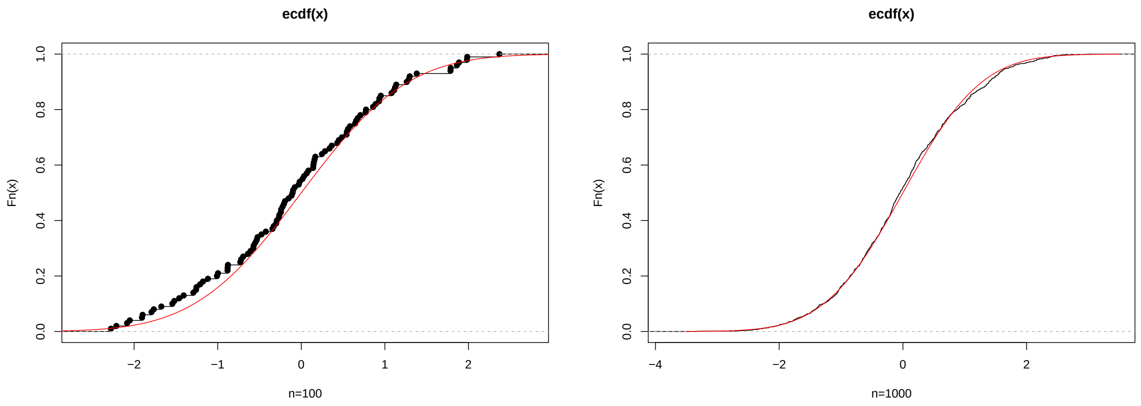
$$F_n(x) \xrightarrow{p} F(x), \qquad \text{for every } x \in \mathbb{R}.$$

Theorem 4. Glivenko-Cantelli Theorem. *Let F_n be the sample c.d.f from an i.i.d sample $X_1, \ldots, X_n$ from the c.d.f F. Then,*

$$D_n := \sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{p} 0.$$

By the CLT, we have

$$\sqrt{n}(F_n(x) - F(x)) \xrightarrow{d} N(0, F(x)(1 - F(x))), \qquad \text{for every } x \in \mathbb{R}.$$



As $F_n(x) \xrightarrow{p} F(x)$ for all $x \in \mathbb{R}$, we can also say that

$$\frac{\sqrt{n}(F_n(x) - F(x))}{\sqrt{F_n(x)(1-F_n(x))}} \xrightarrow{d} N(0,1), \qquad \text{for every } x \in \mathbb{R}.$$

Thus, an asymptotic $(1-\alpha)$ CI for $F(x)$ is

$$\left[F_n(x) - \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{F_n(x)(1-F_n(x))}, F_n(x) + \frac{z_{\alpha/2}}{\sqrt{n}}\sqrt{F_n(x)(1-F_n(x))}\right].$$

Likewise, we can also test the hypothesis $H_0 : F(x) = F_0(x)$ versus $H_1 : F(x) \neq F_0(x)$ for some known fixed c.d.f F_0, and $x \in \mathbb{R}$.

2.1 The Kolmogorov-Smirnov goodness-of-fit test

Suppose that we wish to test the simple null hypothesis that the unknown c.d.f F is actually a particular continuous c.d.f F^* against the alternative that the actual c.d.f is not F^*, i.e.,

$$H_0 : F(x) = F^*(x) \quad \text{for } x \in \mathbb{R}, \qquad H_0 : F(x) \neq F^*(x) \quad \text{for some } x \in \mathbb{R}.$$

This is a nonparametric (“infinite” dimensional) problem.

Let

$$D_n^* = \sup_{x \in \mathbb{R}} |F_n(x) - F^*(x)|.$$

D_n^* is the maximum difference between the sample c.d.f F_n and the hypothesized c.d.f F^*.

We should reject H_0 when

$$n^{1/2} D_n^* \geq c_\alpha.$$

This is called the **Kolmogorov-Smirnov** test.

How do we find c_α?

When H_0 is true, the distribution of D_n^* will have a certain distribution that is the same for every possible continuous c.d.f F. (Why?)

Note that, under H_0,

$$\begin{aligned}
D_n^* &= \sup_{x\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^n I(X_i \le x) - F^*(x) \right| \\
&= \sup_{x\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^n I(F^*(X_i) \le F^*(x)) - F^*(x) \right| \\
&= \sup_{F^*(x)\in\mathbb{R}} \left| \frac{1}{n}\sum_{i=1}^n I(U_i \le F^*(x)) - F^*(x) \right| = \sup_{t\in[0,1]} \left| \frac{1}{n}\sum_{i=1}^n I(U_i \le t) - t \right| \\
&= \sup_{t\in[0,1]} |F_{n,U}(t) - t|,
\end{aligned}$$

where $U_i := F^*(X_i) \sim \text{Uniform}(0,1)$ (i.i.d) and $F_{n,U}$ is the EDF of the U_i's. Thus, D_n^* is *distribution-free.*

Theorem 5. *(Distribution-free property) Under H_0, the distribution of D_n^* is the same for all continuous distribution functions F.*

We also have the following theorem.

Theorem 6. *Under H_0, as $n \to \infty$,*

$$n^{1/2} D_n^* \xrightarrow{d} H, \tag{2}$$

where H is a valid c.d.f.

In fact, the exact sampling distribution of the KS statistic, under H_0, can be approximated by *simulations*, i.e., we can draw *n data points from a Uniform(0,1) distribution and recompute the test statistic* multiple times.

2.1.1 The Kolmogorov-Smirnov test for two samples

Consider a problem in which a random sample of m observations $X_1, \dots, X_m$ is taken from the unknown c.d.f F, and an independent random sample of n observations $Y_1, \dots, Y_n$ is taken from another distribution with unknown c.d.f G.

It is desired to test the hypothesis that both these functions, F and G, are identical, without specifying their common form. Thus the hypotheses we want to test are:

$$H_0 : F(x) = G(x) \quad \text{for } x \in \mathbb{R}, \qquad H_0 : F(x) \neq G(x) \quad \text{for some } x \in \mathbb{R}.$$

We shall denote by F_m the EDF of the observed sample $X_1, \ldots, X_m$, and by G_n the EDF of the sample $Y_1, \ldots, Y_n$.

We consider the following statistic:

$$D_{m,n} = \sup_{x \in \mathbb{R}} |F_m(x) - G_n(x)|.$$

When H_0 holds, the sample EDFs F_m and G_n will tend to be close to each other. In fact, when H_0 is true, it follows from the Glivenko-Cantelli lemma that

$$D_{m,n} \xrightarrow{p} 0 \qquad \text{as } m, n \to \infty.$$

$D_{m,n}$ is also *distribution-free* (why?)

Theorem 7. *Under H_0,*

$$\left(\frac{mn}{m+n}\right)^{1/2} D_{m,n} \xrightarrow{d} H,$$

where H is a the same c.d.f as in (2)*.*

A test procedure that rejects H_0 when

$$\left(\frac{mn}{m+n}\right)^{1/2} D_{m,n} \geq c_\alpha,$$

where c_α (is the $(1-\alpha)$-quantile of H) is an appropriate constant, is called a *Kolmogorov-Smirnov two sample test.*

Exercise: Show that this test statistic is also distribution-free under H_0. Thus, the critical of the test can be obtained via simulations.

3 Bootstrap

Example 1: Suppose that we model our data $\mathbf{X} = (X_1, \ldots, X_n)$ as coming from some distribution with c.d.f F having median θ.

Suppose that we are interested in using the sample median M as an estimator of θ.

We would like to estimate the MSE (mean squared error) of M (as an estimator of θ), i.e., we would like to estimate

$$\mathbb{E}[(M - \theta)^2].$$

We may also be interested in finding a confidence interval for θ.

Example 2: Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ is a random sample from a distribution F. We are interested in the distribution of the sample correlation coefficient:

$$R = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\left[\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2\right]^{1/2}}.$$

We might be interested in the variance of R, or the bias of R, or the distribution of R as an estimator of the correlation ρ between X and Y.

Question: How do we get a handle on these problems?

How would we do it if an *oracle* told us F?

Bootstrap: The bootstrap is a method of replacing (plug-in) an unknown distribution function F with a known distribution in probability/expectation calculations.

If we have a sample of data from the distribution F, we first approximate F by $\hat{F}$ and then perform the desired calculation.

If $\hat{F}$ is a good approximation of F, then bootstrap can be successful.

3.1 Bootstrap in general

Let $\eta(\mathbf{X}, F)$ be a quantity of interest that possibly depends on both the distribution F and a sample $\mathbf{X}$ drawn from F.

In general, we might wish to estimate the mean or a quantile or some other probabilistic feature or the entire *distribution* of $\eta(\mathbf{X}, F)$.

The bootstrap estimates $\eta(\mathbf{X}, F)$ by $\eta(\mathbf{X}^*, \hat{F})$, where $\mathbf{X}^*$ is a random sample drawn from the distribution $\hat{F}$, where $\hat{F}$ is some distribution that we think is close to F.

How do we find the distribution of $\eta(\mathbf{X}^*, \hat{F})$?

In most cases, the distribution of $\eta(\mathbf{X}^*, \hat{F})$ is difficult to compute, but we can approximate it easily by simulation.

The bootstrap can be broken down in the following simple steps:

- Find a "good" estimator $\hat{F}$ of F.
- Draw a large number (say, v) of random samples $\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(v)}$ from the distribution $\hat{F}$ and then compute $T^{(i)} = \eta(\mathbf{X}^{*(i)}, \hat{F})$, for $i = 1, \ldots, v$.
- Finally, compute the desired feature of $\eta(\mathbf{X}^*, \hat{F})$ using the sample c.d.f of the values $T^{(1)}, \ldots, T^{(v)}$.

3.2 Parametric bootstrap

Example 1: (Estimating the standard deviation of a statistic)

Suppose that $X_1, \ldots, X_n$ is random sample from $N(\mu, \sigma^2)$.

Suppose that we are interested in the parameter

$$\theta = \mathbb{P}(X \leq c) = \Phi\left(\frac{c-\mu}{\sigma}\right),$$

where c is a given known constant.

What is the MLE of θ?

The MLE of θ is

$$\hat{\theta} = \Phi\left(\frac{c-\bar{X}}{\hat{\sigma}}\right).$$

Question: How do we calculate the standard deviation of $\hat{\theta}$? There is no easy closed form expression for this.

Solution: We can bootstrap!

Draw many (say v) bootstrap samples of size n from $N(\bar{X}, \hat{\sigma}^2)$. For the i-th sample we compute a sample average $\bar{X}^{*(i)}$, a sample standard deviation $\hat{\sigma}^{*(i)}$.

Finally, we compute

$$\hat{\theta}^{*(i)} = \Phi\left(\frac{c - \bar{X}^{*(i)}}{\hat{\sigma}^{*(i)}}\right).$$

We can estimate the mean of $\hat{\theta}$ by

$$\bar{\theta}^{*} = \frac{1}{v}\sum_{i=1}^{v}\hat{\theta}^{*(i)}.$$

The standard deviation of $\hat{\theta}$ can then be estimated by the sample standard deviation of the $\hat{\theta}^{*(i)}$ values, i.e.,

$$\left[\frac{1}{v}\sum_{i=1}^{v}(\hat{\theta}^{*(i)} - \bar{\theta}^{*})^2\right]^{1/2}.$$

Example 2: (Comparing means when variances are unequal) Suppose that we have two samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ from two possibly different normal populations. Suppose that

$$X_1, \ldots, X_m \text{ are i.i.d } N(\mu_1, \sigma_1^2) \qquad \text{and} \qquad Y_1, \ldots, Y_n \text{ are i.i.d } N(\mu_2, \sigma_2^2).$$

Suppose that we want to test

$$H_0 : \mu_1 = \mu_2 \qquad \text{versus} \qquad H_1 : \mu_1 \neq \mu_2.$$

We can use the test statistic

$$U = \frac{(m+n-2)^{1/2}(\bar{X}_m - \bar{Y}_n)}{\left(\frac{1}{m} + \frac{1}{n}\right)^{1/2}(S_X^2 + S_Y^2)^{1/2}}.$$

Note that as $\sigma_1^2 \neq \sigma_2^2$, U does not necessarily follow a t-distribution.

How do we find the cut-off value of the test?

The parametric bootstrap can proceed as follows:

First choose a large number v, and for $i = 1, \ldots, v$, simulate $(\bar{X}_m^{*(i)}, \bar{Y}_n^{*(i)}, S_X^{2*(i)}, S_Y^{2*(i)})$, where all four random variables are independent with the following distributions:

- $\bar{X}_m^{*(i)} \sim N(0, \hat{\sigma}_1^2/m)$.
- $\bar{Y}_n^{*(i)} \sim N(0, \hat{\sigma}_2^2/n)$.
- $S_X^{2*(i)} \sim \hat{\sigma}_1^2\, \chi_{m-1}^2$.

- $S_Y^{2*(i)} \sim \hat{\sigma}_2^2 \ \chi^2_{n-1}$.

Then we compute

$$U^{*(i)} = \frac{(m+n-2)^{1/2}(\bar{X}_m^{*(i)} - \bar{Y}_n^{*(i)})}{\left(\frac{1}{m}+\frac{1}{n}\right)^{1/2}(S_X^{2*(i)} + S_Y^{2*(i)})^{1/2}}$$

for each i.

We approximate the null distribution of U by the distribution of the $U^{*(i)}$'s.

Let c^* be the $\left(1-\frac{\alpha}{2}\right)$-quantile of the distribution of $U^{*(i)}$'s. Thus we reject H_0 if

$$|U| > c^*.$$

3.3 The nonparametric bootstrap

Back to Example 1: Let $X_1, \ldots, X_n$ be a random sample from a distribution F.

Suppose that we want a CI for the median θ of F.

We can base a CI on the sample median M.

We want the distribution of $M - \theta$!

Let $\eta(\mathbf{X}, F) = M - \theta$.

We approximate the $\alpha/2$ and the $1-\alpha/2$ quantiles of the distribution of $\eta(\mathbf{X}, F)$ by that of $\eta(\mathbf{X}^*, \hat{F})$.

We may choose $\hat{F} = F_n$, the empirical distribution function. Thus, our method can be broken in the following steps:

- Choose a large number v and simulate many samples $\mathbf{X}^{*(i)}$, for $i = 1, \ldots, n$, from F_n. This reduces to drawing **with replacement sampling** from $\mathbf{X}$.
- For each sample we compute the sample median $M^{*(i)}$ and then find the sample quantiles of $\{M^{*(i)} - M\}_{i=1}^{v}$.

Back to Example 2: Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ is a random sample from a distribution F. We are interested in the distribution of the sample correlation coefficient:

$$R = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\left[\sum_{i=1}^n (X_i - \bar{X})^2 \sum_{i=1}^n (Y_i - \bar{Y})^2\right]^{1/2}}.$$

We might be interested in the bias of R, i.e., $\eta(\mathbf{X}, \mathbf{Y}, F) = R - \rho$.

Let F_n be the discrete distribution that assigns probability $1/n$ to each of the n data points.

Thus, our method can be broken in the following steps:

- Choose a large number v and simulate many samples from F_n. This reduces to drawing **with replacement sampling** from the original paired data.
- For each sample we compute the sample correlation coefficient $R^{*(i)}$ and then find the sample quantiles of $\{T^{*(i)} = R^{*(i)} - R\}_{i=1}^{v}$.
- We estimate the mean of $R - \rho$ by the average $\frac{1}{n}\sum_{i=1}^{v} T^{*(i)}$.