# Introduction to Econometrics II: Recitation 8*

Seung-hun Lee†

April 3rd, 2020 (Finally!)

## 1 Bootstrap Methods

### 1.1 Refresher on the properties of the estimators

Before we get to bootstrap, a reminder on the key properties of the estimators come in handy. Assume that we have data $(y_1, ..., y_n)$ distributed according to some probability measure $P$. The definition of a probability space is as follows.

**Definition 1.1** (Probability Measure)**.** Let $(\Omega, \mathcal{F})$ be a measurable space, with $\Omega$ referring to sets of states of the world and $\mathcal{F}$ referring to $\sigma-$algebra of subsets of $\Omega$ (or collection of events). The probability measure $P$ is a $[0, 1]$ valued function summarizing the likelihood of an event that satisfies

- $P(\Omega) = 1, P(\emptyset) = 0$
- $P\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$ for a countable disjoint sequence of elements of $\mathcal{F}$.

For instance, a normal distribution with mean $\mu$ and variance $\sigma^2$ can count.

Normally, we do not have an exact knowledge of what $P$ did the data $(y_1, ..., y_n)$ came from. We may know that the data is generated by a distribution $F_\theta$ for some $\theta \in \Theta$. However, there is a clear virtue in pinpointing a particular $\theta$ (denoted as $\theta_0$). Knowing this allows us

*This is based on the lecture notes of Professors Jushan Bai and Bernard Salanie. I was also greatly helped by previous recitation notes from Paul Sungwook Koh and Dong Woo Hahm. The remaining errors are mine.

†Contact me at sl4436@columbia.edu if you spot any errors or have suggestions on improving this note.

to discover the population from which our data is from. For instance, if $F_\theta = N(\theta, 1)$ and if we can pinpoint a particular $\theta_0$, we know the mean and the variance of our data.

Therefore, if we do not know the exact $\theta_0$ yet, we estimate what this might be. An **estimator** is a function of the observations $(y_1, ..., y_n)$, often denoted as $\hat{\theta}_n$. For instance, a mean, defined as $\frac{\sum_{i=1}^n y_i}{n}$, or the median is an estimator. While there are many possible estimators for a given data, not all of them are equal. Ideally, we would like the estimators to satisfy these properties

- **Unbiased**: $E(\hat{\theta}_n) = \theta_0$, where $E(\cdot)$ is the expectation under $F_{\theta_0}$
- **Consistent**: $\hat{\theta}_n \xrightarrow{p} \theta_0$ as $n \to \infty$
- **Asymptotically Normal**: $\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \Sigma_0)$ as $n \to \infty$. ($\Sigma_0$ is PD matrix)

In addition, we would like to carry out inference on our estimators reliably. To do this, we need to find a good way to approximate the finite-sample distribution of $\hat{\theta}_n$.

All of this is easy when we have a classical linear model, specified as

$$y = Xb_0 + u$$

where $u \sim N(0, \sigma_0^2)$. Assuming that the observations are IID conditional on $X = x$, the model has distribution $y \sim N(xb_0, \sigma_0^2)$. The OLS estimator has a finite-sample distribution that is characterized as follows

$$\begin{aligned} \hat{b}_n &= b_0 + (X'X)^{-1}X'u \\ \implies \sqrt{n}(\hat{b}_n - b_0) &= \left(\frac{X'X}{n}\right)^{-1} \frac{X'u}{\sqrt{n}} \\ &\sim N(0, \sigma_0^2(X'X)^{-1}) \end{aligned}$$

Since CLT implies that $\frac{X'u}{\sqrt{n}} \sim N(0, \sigma_0(X'X))$. If we are unsure of what $\sigma_0^2$ exactly is, we need to estimate for it using

$$\hat{\sigma}_n^2 = \frac{1}{n-p} \sum_{i=1}^n \hat{u}_i^2$$

where $p = \dim(X)$ and $\hat{u}_i = y_i - x_i\hat{b}_n$. This is an unbiased estimator for $\sigma_0^2$. We can see this

from the following.

$$\begin{aligned}\hat{u} &= y - X\hat{b}_n \\ &= y - X(X'X)^{-1}X'y \\ &= Xb_0 + u - Xb_0 - X(X'X)^{-1}X'u = Mu\end{aligned}$$

where $M = I - X(X'X)^{-1}X$ and is idempotent and symmetric. Then,

$$\begin{aligned}E[\hat{u}'\hat{u}] &= E[u'Mu] \\ &= E[tr(u'Mu)] \ (\because u'Mu \text{ is scalar.}) \\ &= E[tr(Muu')] \ (tr(AB) = tr(BA)) \\ &= tr[E(Muu')] \ (E(tr(A)) = E(A_{11} + ... + A_{nn}) = E(A_{11}) + ... + E(A_{nn}) = tr(E(A))) \\ &= tr[ME(uu')] \ (\text{If we condition } X = x,\ M \text{ is not stochastic}) \\ &= tr[M\sigma_0^2 I_n] = \sigma_0^2 tr(M) \\ &= \sigma_0^2 tr(I - X(X'X)^{-1}X') = \sigma_0^2[tr(I) - tr(X(X'X)^{-1}X')] \\ &= \sigma_0^2(n - tr((X'X)^{-1}X'X)) = \sigma_0^2(n-p)\end{aligned}$$

Therefore, $E\left[\frac{1}{n-p}\hat{u}'\hat{u}\right] = E\left[\frac{1}{n-p}\sum_{i=1}^{n}\hat{u}_i^2\right] = \sigma_0^2$.

For a more honest inference that captures how reliable the point estimates are, we would need a confidence interval. We can estimate the coverage probability for $b_0$ of a region $B \in \mathbb{R}^p$ by

$$\Pr(N(\hat{b}_n, \hat{\sigma}_n^2(X'X)^{-1}) \in B)$$

In words, it estimates the proportion of times that the region $B$ contains the distribution of $b_0$.

Suppose we are interested in finding the 95% confidence interval for component $k$ in $\hat{b}_n$, denoted as $\hat{b}_{kn}$. Applying above setup, the relevant standard error is $\hat{\sigma}_{kn} = \hat{\sigma}_n\sqrt{(X'X)_k^{-1}}$ and the interval would have upper and lower bounds of $\hat{b}_{kn} \pm 1.96\hat{\sigma}_{kn}$. We can test the null hypothesis that $b_{k0} = c$ with the rejection criterion

$$\frac{|\hat{b}_{kn} - c|}{\hat{\sigma}_{kn}} > 1.96$$

which gives a test with size (reject true null hypothesis wrongly by) 5% and power (correctly reject wrong hypothesis) that converges to 1.

**Comment 1.1** (What is a confidence interval?)**.** *There are some cautions when interpreting the confidence interval. Let $(l, u)$ be the 95% confidence interval after a sample is drawn. A 95% confidence interval can be interpreted as saying that* ***there is a 95% probability that the*** $(l, u)$ ***confidence interval encompasses the true value****. Notice that this is a probability statement about the confidence interval and not the true value. True value of $\beta_{k0}$ should not be treated as a random variable. Therefore, it is not necessarily true to say that there is a 95% probability that $b_{0k}$ is included in a particular $(l, u)$ confidence interval. Another way of seeing this is if the sample has been drawn N times, the fraction of calculated confidence intervals (that differs for each draw) which encompasses the true value is 95%, or that $N \times 0.95$ confidence intervals encompass true parameter.*

## 1.2 When we go beyond classical setup...

In most cases, we only know the asymptotics. Suppose that $\hat{\theta}_n$ is a CAN (consistent and asymptotically normal) estimator s.t.

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \sim N(0, \Sigma_0)$$

In a finite-sample context, we need to rely on an approximation of a following sense: Let $\widehat{\Sigma}_n$ be a consistent estimator for $\Sigma_0$. Then we can write the above equation as

$$\hat{\theta}_n = \theta_0 + \frac{1}{\sqrt{n}}\widehat{\Sigma}_n^{-1/2} N(0, I_n)$$

There are cases where above approximation can be very misleading. One example is from the White Misspecification Test.

To show this, we need to formulate Fisher's Information Matrix. Let $L$ be a likelihood function, which is a function of $y, x$, and parameter $\theta$. Define

$$I = E\left[\left(\frac{\partial \log L}{\partial \theta}\right) \cdot \left(\frac{\partial \log L}{\partial \theta'}\right)\right]$$

Which is the same as

$$J = -E\left[\frac{\partial^2 \log L}{\partial \theta \partial \theta'}\right]$$

By Cramer-Rao Lower Bound, we can estimate the variance of MLE as $I^{-1}$ or $J^{-1}$. Alterna-

tively, we can also use a Sandwich formula $J^{-1}IJ^{-1}$ for more robust estimate. White (1982) says that comparing the results gives an asymptotic test for misspecification. It turns out that the asymptotic approximation is unreliable to the extent that the test rejects well-specified models an order of magnitude more often than it should.

**Comment 1.2** (Why is $I = J$?)**.** *To show that expressions in I and J are equal, I define* $L(\theta; x) = f(x; \theta)$ *as a likelihood function. Notice that*

$$1 = \int f(x;\theta)dx$$

*Firstly, differentiate with respect to* $\theta$ *to get*

$$\begin{aligned} 0 &= \int \frac{\partial}{\partial\theta} f(x;\theta)dx \\ &= \int \frac{\partial}{\partial\theta} \log f(x;\theta) \times f(x;\theta)dx = E\left[\frac{\partial \log f(x;\theta)}{\partial\theta}\right] \end{aligned}$$

*Differentiate above with respect to* $\theta'$ *to get*

$$\begin{aligned} 0 &= \int \frac{\partial}{\partial\theta'}\left(\frac{\partial}{\partial\theta}\log f(x;\theta) \times f(x;\theta)\right)dx \\ &= \int \frac{\partial^2}{\partial\theta\partial\theta'}\log f(x;\theta)dx + \int \frac{\partial \log f(x;\theta)}{\partial\theta}\frac{\partial f(x;\theta)}{\partial\theta'}dx \\ &= \int \frac{\partial^2}{\partial\theta\partial\theta'}\log f(x;\theta)dx + \int \frac{\partial \log f(x;\theta)}{\partial\theta}\frac{\partial \log f(x;\theta)}{\partial\theta'}f(x;\theta)dx \\ &= E\left[\frac{\partial^2}{\partial\theta\partial\theta'}\log f(x;\theta)\right] + E\left[\frac{\partial}{\partial\theta}\log f(x;\theta)\cdot\frac{\partial}{\partial\theta'}\log f(x;\theta)\right] \end{aligned}$$

*Take the first term on RHS of the last line to LHS, and we get* $I = J$*. (Hogg et al. (2014))*

To overcome such problems, we can rely on resampling techniques. That is, we pretend that the sample is the entire population and that we use as much as we can by re-drawing within the samples. While this does not necessarily get us the better estimates, it may be possible to obtain a better confidence interval.

## 1.3 Bootstrap

In doing the bootstrap, we are looking for alternative ways to assign standard errors (and ultimately the test statistics and confidence intervals) that are *at least as good as* and ideally

*better than* the typical asymptotic approximation that we normally do. **Bootstrap** estimators are usually as good as asymptotic approximations in the sense that bootstrap estimators are also *consistent* and better in the sense that it has *smaller approximation error*.

The procedure works as follows for the $t$-statistic :

1. Start with an IID sample $(X_1, ..., X_n)$. Estimate the model and compute the $t$-statistic
2. Make $n$ draws *with replacement*. Re-estimate the model and compute the new $t$-statistic.
3. Repeat step 2 many times
4. Since there are different $t$-statistics for each sample, there will be a distribution of $t$-statistics. Use this $t$-distribution to get a reliable critical value

To generalize the discussion, I will use the following framework. We have a model and a data $(X_1, ..., X_n)$ generated from an unknown distribution $F_0$. We can characterize the test statistic as $T_n(X_1, ..., X_n)$. Define the finite sample distribution of the test statistic under $F_0$ as

$$G_n(t, F_0) = \Pr(T_n \leq t)$$

Two properties below are nice properties for $T_n$ to have:

**Definition 1.2** (Pivotal and Asymptotically Pivotal)**.**

- **Pivotal**: We say $T_n$ is pivotal if $G_n$ itself does not depend on the unknown parameters of $F_0$.
- **Asymptotically Pivotal**: We say $T_n$ is asymptotically pivotal if

$$G_\infty(t, F_0) = \lim_{n \to \infty} G_n(t, F_0)$$

does not depend on $F_0$. (so that $F_0$ can be omitted in the arguments.)

Since we do not know $F_0$, we need an estimator for it - an empirical distribution function which puts a probability mass $1/n$ on each observation

**Definition 1.3** (Empirical Distribution Function)**.** The **empirical distribution function** (or EDF) for the data $X_1, .., X_n$ is defined as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \leq x)$$

Using the EDF, we can estimate $G_n(t, F_n)$ instead of $G(t, F_0)$ in this way

1. Select a large $B$, say 10,000. (This depends on your computer's processing capacity)
2. For each $b \in \{1, .., B\}$ we do as follows
   - For each $i = 1, .., n$ pick $j = 1, ..., n$ randomly with replacement and let $X_i^b = X_j$. In other words, draw $n$ times with replacement from $(X_1, .., X_n)$ and redefine the 'shuffled' sample as $(X_1^b, ..., X_n^b)$.
   - Then, compute the new test statistic $T_n^b = T_n(X_1^b, ..., X_n^b)$.
3. Then, we will end up with $B$ numbers of $T_n^b$. Then $G_n(t, F_n)$ can be written as

$$G_n(t, F_n) = \frac{1}{B}\sum_{b=1}^{B} 1(T_n^b \leq t)$$

Once we are done with this, we find the bootstrapped critical values. Suppose that our test is of size 5% and we are doing a two-sided test. We find lower(upper) bounds for our critical value $c^-$ ($c^+$) s.t. $0.025B$ values of $T_n^b$ are smaller(larger) than that critical value. We then reject the null hypothesis if the $T_n$ from the original $(X_1, .., X_n)$ does not belong within the bounds set by this procedure. Bootstrapped confidence intervals are obtained by inverting the test - we take all values not rejected by the above procedure and construct an interval.

Now we show why bootstrapped estimators are *at least as good as* and sometimes ideally *better than* normal asymptotic approximations. Using the technic of Edgeworth Expansion, we can write

$$G_n(t, F_0) = G_\infty(t, F_0) + \frac{1}{\sqrt{n}}g_1(t, F_0) + \frac{1}{n}g_2(t, F_0) + \frac{1}{n\sqrt{n}}g_3(t, F_0) + O\left(\frac{1}{n^2}\right)$$
$$G_n(t, F_n) = G_\infty(t, F_n) + \frac{1}{\sqrt{n}}g_1(t, F_n) + \frac{1}{n}g_2(t, F_n) + \frac{1}{n\sqrt{n}}g_3(t, F_n) + O\left(\frac{1}{n^2}\right)$$

uniformly over $t$ almost surely. (From Horowitz, 2011). Suppose that $G_\infty(t,\cdot)$ is continuous at $F_0$. Then

$$G_n(t,F_n) - G_n(t,F_0) = G_\infty(t,F_n) - G_\infty(t,F_0) + \frac{1}{\sqrt{n}}(g_1(t,F_0) - g_1(t,F_n)) \\ + \frac{1}{n}(g_2(t,F_0) - g_2(t,F_n)) + \frac{1}{n\sqrt{n}}(g_3(t,F_0) - g_3(t,F_n)) + O\left(\frac{1}{n^2}\right)$$

Notice that $G_\infty(t,F_n) - G_\infty(t,F_0)$ is $O\left(\frac{1}{\sqrt{n}}\right)$ since $F_n - F_0 = O\left(\frac{1}{\sqrt{n}}\right)$.[1] Therefore, the size of the bootstrap approximation error is $O\left(\frac{1}{\sqrt{n}}\right)$ and the bootstrap error is consistent. Note that normal asymptotic approximation also has the same size of approximation error.

If $T_n$ is asymptotically pivotal, $G_\infty(t,F_n) - G_\infty(t,F_0)$ term vanishes. The leading term is $\frac{1}{\sqrt{n}}(g_1(t,F_0) - g_1(t,F_n))$ and the size of the approximation error is $O\left(\frac{1}{n}\right)$ - improvement.

## 1.4 Types of Bootstraps

- **Parametric Bootstrap** : Suppose that there is a parametric model for $y = g(x,\theta_0,u)$ and that you have faith in this form. Then 1) hold $x_i$'s fixed and bootstrap on the residuals $\hat{u}_i$, which gets you the bootstrapped $y_i = x\hat{\theta}_0 + \hat{u}_i$. Another way to do it is to draw from the estimated parametric distribution of $u$. Note that if you want to know the distribution of $T_n$ under $H_0$, then assume $H_0$ when resampling. Each step contributes to making bootstrap powerful.

- **Bias Corrected Bootstrap** : To get the idea, suppose that $X$ is a random vector where $\mu = E(X)$. Assume that the we are interested in the true value of $\theta$, which is $\theta_0 = g(\mu)$ ($g$ is known and continuous). We can consistently estimate $\theta_0$ by $\theta_n = g(\bar{X})$. However, when we take a Taylor expansion of $\theta_n = g(\bar{X})$ about $\bar{X} = \mu$, [2]

  $$\theta_n = g(\mu) + g'(\mu)(\bar{X} - \mu) + \frac{1}{2}g''(\mu)(\bar{X} - \mu)^2 + O(n^{-2})$$

  As such, the bias term is

  $$B_n = E[\theta_n] - \theta_0 = g'(\mu)E(\bar{X} - \mu) + \frac{1}{2}E(g''(\mu)(\bar{X} - \mu)^2) + O(n^{-2}) = \frac{1}{2}E(g''(\mu)(\bar{X} - \mu)^2)$$

  which is nonzero and $O(n^{-1})$ unless $g$ is a linear function.

[1] Refer to Dvoretzky-Kiefer-Wolfowitz inequality
[2] For simplicity, I discuss the scalar case

In the bootstrap setup, let $\bar{X}^*$ denote the mean of the 'shuffled' sample. The analogue for the above is

$$\theta_n^* = \theta_n + g'(\bar{X})(\bar{X}^* - \bar{X}) + \frac{1}{2}g''(\bar{X})(\bar{X}^* - \bar{X})^2 + O(n^{-2})$$

then, take the empirical expectation to get

$$B_n^* = E^*[\theta_n^* - \theta_n] = \frac{1}{2}E^*(g''(\bar{X})(\bar{X}^* - \bar{X})^2) + O(n^{-2})$$

Here, we have $E[B_n^*] = B_n + O(n^{-2})$. Thus, the bias corrected estimator of $\theta_0$ is now $\theta_n - B_n^* = 2\theta_n - E^*[\theta_n^*]$. The bias is

$$E[\theta_n - B_n^*] - \theta_0 = O(n^{-2})$$

- **Variance of an Estimator** : This is the most common use for the bootstrap. The bootstrap variance of the estimator $\hat{\theta}_n$ is defined as

$$\tilde{V}_n = \frac{1}{B-1}\sum_{b=1}^{B}\left(\hat{\theta}_n^b - \frac{1}{B}\sum_{i=1}^{B}\hat{\theta}_n^i\right)^2$$

Once you have a computer, it is not too difficult to compute. However, the resulting standard error depends on the choice of $B$.

- **Block Bootstrap** : This is used when we suspect that the data is serially correlated. What we do is to use $m$-size blocks of consecutive observations. Then we can use $B = n + 1 - m$ blocks in total. This works when $m/n \to 0, m \to \infty$ as $n \to \infty$

- **Wild Boostrap** : This is useful when we have a heteroskedastic error terms. Suppose we have a sample $b$, and there is an $i$'th observation in that sample. The trick is to use $u_i^* = \epsilon_i \hat{u}_i$ where $E[\epsilon_i] = 0, E[\epsilon_i^2] = 1$. As in class, you can use $\epsilon_i = \pm 1$. Mammen (1993) shows that choosing

$$\epsilon_i = \begin{cases} \frac{1+\sqrt{5}}{2} \text{ Prob:} \frac{\sqrt{5}-1}{2\sqrt{5}} \\ \frac{1-\sqrt{5}}{2} \text{ Prob:} \frac{\sqrt{5}+1}{2\sqrt{5}} \end{cases}$$

leads to a valid and very efficient bootstrap inference for linear models.

- **Subsampling**[3] : Consider the same setup as bootstrap - we are still interested in the distribution of $T_n(X_1, ..., X_n)$. Instead of resampling with the size $n$, we resample with size $m$ and do a bootstrap for $B$ times. So there are a total of $N_{n,m} = \binom{n}{m}$ total number of subsets that can be found. I will denote one of those subsample as $W_k, k \in \{1, ...., N_{n,m}\}$. Then we are going to make use of $T_m(W_k)$. We are trying to estimate $G_n(t, F_n)$ with $T_m$, but on a rescaled version. Define

$$J_n(t, F_0) = \Pr(r_n(T_n - T_\infty) \leq t) \text{ (You can define } J_m(t, F_0) \text{ similarly)}$$

where $r_n$ is a scaling parameter independent of $F_0$ and chosen such that $J_n(t, F_0)$ converges in distribution to a nondegenerate distribution $J_\infty(t, F_0)$ (say, $\sqrt{n}$). We estimate this using

$$L_{n,m} = \frac{1}{N_{n,m}} \sum_{i=1}^{N_{n,m}} 1[r_m(T_m - T_n) \leq t]$$

The $L_{n,m}$ is intended to be an estimator of $J_m(t, F_0)$. Then, as $m/n \to 0, r_m/r_n \to 0$ and $m \to \infty$ as $n \to \infty$, we have

$$r_n(T_n - T_\infty) \simeq r_m(T_m - T_n)$$

which works better than bootstrap as long as $J_\infty(t, F_0)$ is continuous with respect to $t$

# 2 Multiple Hypothesis Testing

## 2.1 Motivation

Suppose that we are testing $S$ hypothesis where $S$ is some large number and each hypothesis is mutually independent. For instance, one might run multiple regressions on several investment strategies while comparing to some benchmark and wishes to find which strategies are outperforming the said benchmark. Let $S = 1000$. Then, the probability of rejecting any one of the 1000 hypothesis is calculated by $1 - (0.95)^{1000} \simeq 1$[4], in other words, you get something significant just out of pure luck. In doing so, the most dangerous consequence is

[3] Usually carried without replacements. For the similar version that executes this with replacement, search $m$ out of $n$ bootstrap

[4] I did this on a Google calculator.

running in to a **false discovery** - a rejection of a true null hypothesis[5].

The best approach that prevents this mistake is finding either an empirical or theoretical motivations for the hypothesis. In a context with larger datasets, we need to come up with a technical method for testing multiple hypothesis.

## 2.2 Setting up

We are running $S$ simultaneous tests. For a particular test $s \leq S$, we are testing the null $H_{0s}$ versus the alternative $H_{1s}$, we know that the $p$-value is

$$p_s = \Pr(\text{Reject } H_{0s} | H_{0s} \text{ is true})$$

This is also the mathematical definition of the false discovery. Some other key terms are

- **False discovery proportion** (FDP): It is the proportion of the false discoveries out of all discoveries, or
- **False discovery rate** (FDR): $E[FDP]$
- **Familywise error rate** (FWE): It is the probability of one or more false discoveries. A $k$-FWE calculates probability of $k$ or more false discoveries.

There are two strands of methods. One controls for the FWE. The other controls for the FDR. We go over each strands

## 2.3 Controlling FWE: Bonferroni Method and Holm Method

**Bonferroni method** is a classical method based on Bonferroni inequalities. To see how this works, let $A_s$ be the event that $p_s < w$ (and for which $H_{0s}$ is rejected) for some value $w$ associated with the significance level $\alpha$, where $s = 1, ..., S$. Let $I_0$ be indices for the true hypothesis, with $n(I_s) = S_0 \leq S$. The goal is to determine the value of $w$ that makes the $FWE \leq \alpha$. Then the following logic works

$$FWE = \Pr(\cup_{s \in I_0} A_s) \leq \sum_{s \in I_0} \Pr(A_s) = S_0 w \leq S w$$

So for the FWE, we can let $w = \alpha / S$.

For the general case of $k$-FWE, we take this approach:

[5]Discovery refers to any rejection of a null hypothesis.

**Theorem 2.1** (From Lehmann and Romano (2005)[a]). *For testing $H_{0s}$ where $s = 1, .., S$, if $p_s$ is $U[0,1]$, then following the procedure where $H_{0s}$ is rejected if $p_s \leq \frac{k\alpha}{S}$ controls k-FWE*

*Proof.* Let $I_0$ be the collection of indices of the true null hypotheses and $n(I_0) = S_0$. Let $N$ be the number of rejections of true null hypothesis. Then we can write

$$\begin{aligned}
\Pr(N \geq k) &\leq \frac{E(N)}{k} \ (\because \text{ Markov Inequality}) \\
&= \frac{E\left[\sum_{s \in I_0} 1\left(p_s \leq \frac{k\alpha}{S}\right)\right]}{k} \\
&= \sum_{i \in I_0} \frac{\Pr\left(p_s \leq \frac{k\alpha}{S}\right)}{k} \\
&\leq \sum_{i \in I_0} \frac{k\alpha/S}{k} = S_0 \frac{\alpha}{S} \leq \alpha
\end{aligned}$$

□

[a] Lehmann, E. L. and J. P. Romano (2005) Generalizations of the Familywise Error Rate, *The Annals of Statistics*, 33(3), 1138-1154

While we do not need to make any assumptions about distribution of the $p$-values, the problem is that this method tends to be conservative - there are too few rejections.

**Holm method** is a step-down method in the sense that ordering of the different $p$-values are involved. The procedure is as follows

1. Order the $p$-values s.t. $p_1 < ... < p_S$, and denote each null hypothesis as $H_{01}, ..., H_{0S}$.
2. Reject $H_{0j}$ if $p_j \leq \frac{\alpha}{S-j+1}$ for $j = 1, 2, ..., S$

That is, we start with the hypothesis with lowest $p$-value and compare its $p$-value against $\alpha/S$. If this is rejected, move on to the next hypothesis and compare its $p$-value against $\alpha/(S-1)$. In the last hypothesis, we would be comparing against $\alpha$.

The way this works is as follows: Let $k$ be the index of the first true null hypothesis that is rejected. Since we ordered hypothesis according to the size of the $p$-value, the first $k-1$ rejected hypotheses are false null hypothesis. Since there are total of $S - S_0$ false hypotheses, we can claim $k - 1 \leq S - S_0$. This equation implies $\frac{1}{S-k+1} \leq \frac{1}{S_0}$. Since $k$th hypothesis is still rejected, $p_k \leq \frac{\alpha}{S-k+1} \leq \frac{\alpha}{S_0}$. So, there exists a true hypothesis whose $p$-value is bounded

above by $\alpha/S_0$. Then

$$\begin{aligned} FWE &= \Pr(\text{Reject at least one true hypothesis}) \\ &\leq \Pr(p_s \leq \alpha/S_0 \text{ for some true hypothesis}) \\ &\leq \sum_{s \in I_0} \Pr(p_s \leq \alpha/S_0) = S_0 \times (\alpha/S_0) = \alpha \end{aligned}$$

where I use the fact that if null hypothesis is true and the statistic is continuous, $p$-value has $U[0,1]$ distribution. As for the general case, we can do as follows

**Theorem 2.2** (From Lehmann and Romano (2005)[a]). *For testing $H_{0s}$ where $s = 1, .., S$, if $p_s$ is $U[0,1]$, then following the procedure where $H_{0s}$ is rejected if*

$$\alpha_s = \begin{cases} \frac{k\alpha}{S} & (s \leq k) \\ \frac{k\alpha}{S-s+k} & (s \geq k) \end{cases} \quad (\alpha_s \text{ is the cutoff p-value for sth null})$$

*controls k-FWE. This is equivalent to using $\frac{k\alpha}{\min\{S-s+k,S\}}$ as cutoff.*

*Proof.* Assume that $n(I_0) = S_0 \geq k$. Let $j$ be index of the $k$th true null hypothesis rejected. Then, $j-k$ of the hypotheses so far ar false, thus,

$$j - k \leq S - S_0 \implies k \leq j \leq S - S_0 + k$$

The Holm procedure would reject $k$ true null hypothesis iff $p_s \leq \alpha_s$ for $s = 1, ..., j$. Since we have $j \geq k$, it must be that

$$p_j \leq \alpha_j = \frac{k\alpha}{S - j + k} \leq \frac{k\alpha}{S_0}$$

This implies that when at least $k$ false rejections are made, there exists null hypotheses whose $p$-values is bounded above by $\frac{k\alpha}{S_0}$. We can then apply Theorem 2.1 to show that the $k$-FWE is bounded above by $\alpha$. □

[a] Lehmann, E. L. and J. P. Romano (2005) Generalizations of the Familywise Error Rate, *The Annals of Statistics*, 33(3), 1138-1154

Note that Holm method is not as conservative as Bonferroni method in that more rejections are likely to happen. They share the common feature in that no assumption about the (in)dependence structure of the hypotheses are required.

## 2.4 Controlling FDR: Benjamini & Hochberg Method

The goal of this approach is to set the false discovery rate to be below some $\gamma \in [0, 1]$. The procedure is as follows

1. Order $p$-values from the lowest to the highest: $p_1 < ..., < p_S$
2. Make the following adjustment:

$$\tilde{p}_s = \min \left\{ \frac{S p_j}{j} | j \geq s \right\}$$

3. We move from the $S$th hypothesis. If $\tilde{p}_S \leq \gamma$, then all $S$ hypothesis are rejected and the process ends. Otherwise, move to $S-1$th hypothesis.
4. At some point, you will find $i$ s.t. $\tilde{p}_i \leq \gamma$. Then reject the first $i$ hypothesis.

So how can we do this? We do not reject $H_{0s}$ iff for all $j \geq s, p_j \geq \frac{j\gamma}{S} \left(\geq \frac{s\gamma}{S}\right)$. So what we can do is to plot the ordered $p$-values and the threshold line $p(s) = \frac{s\gamma}{S}$. If a $p$-value of the $s$th hypothesis is larger than $\frac{s\gamma}{S}$, then that hypothesis can not be rejected. In this manner, we can find $s^*$, the largest index s.t. the $p$-value is below the $p(s)$ plot. Then, we reject all $H_{0s}$ s.t. $s \leq s^*$.

We can show that in suitable conditions, this method yields $FDR = \frac{S_0}{S}\gamma$. This is still conservative method in that $S_0 \leq S$. We can do better by estimating what $S_0$ would be and using $\frac{S}{S_0}\gamma$ as cutoff instead.