Homework 6: Mathematical Statistics (MATH-UA 234)

Due 11/17 at the beginning of class on Gradescope

Reminder. Don't forget the project proposals are due on 11/22. A good proposal is the key to doing well on the project, so if you have to choose between spending time on the homework or on the proposal, choose the proposal (although hopefully you can put sufficient time into both).

Problem 1. *Pick at least one of the following articles to read. Provide a one paragraph summary of what you think the most important points of the article were.*

- *An unhealthy obsession with p-values is ruining science*
- *Using Effect Size—or Why the P Value Is Not Enough*
- *The Extent and Consequences of P-Hacking in Science*

Problem 2. *Suppose that the size α test is of the form*

$$\textit{reject } H_0 = \{\theta \in \Theta_0\} \textit{ if and only if } T(X_1, \dots, X_n) > c_\alpha.$$

(a) Given data $X_1, \dots, X_n$, prove that,

$$p\text{-}\textit{value} = \sup_{\theta \in \Theta_0} \mathbb{P}[T(X_1', \dots, X_n') > T(X_1, \dots, X_n) | X_1', \dots, X_n' \sim F_\theta].$$

(b) Explain, in simple words, what this result says about the meaning of a p-value for this type of test.

Solution.

(a) We write the power function

$$\beta(\theta) = \mathbb{P}[T(X_1', \dots, X_n') > c_\alpha | X_1', \dots, X_n' \sim F_\theta].$$

By definition,

$$\alpha = \sup_{\theta \in \Theta_0} \beta(\theta) = \sup_{\theta \in \Theta_0} \mathbb{P}[T(X_1', \dots, X_n') > c_\alpha | X_1', \dots, X_n' \sim F_\theta].$$

Note that as α goes from 0 to 1, then c_α must go from large to small. In particular, c_α is a non-increasing function of α. Recall also that

$$p\text{-value} = \inf\{\alpha : T(X_1, \dots, X_n) > c_\alpha\}$$

Assuming c_α is continuous, then from drawing a picture it is clear that $c_p = T(X_1, \dots, X_n)$.

If we do not assume continuity, we can note that the above facts imply that the p-value is such that, for all $\epsilon > 0$, $c_{p-\epsilon} \geq T(X_1, \dots X_n) > c_{p+\epsilon}$. Observing that

$$p = \lim_{\epsilon \to 0} p + \epsilon = \lim_{\epsilon \to 0} \sup_{\theta \in \Theta_0} \mathbb{P}[T(X_1', \dots, X_n') > c_{p+\epsilon} | X_1', \dots, X_n' \sim F_\theta].$$

$$p = \lim_{\epsilon \to 0} p - \epsilon = \lim_{\epsilon \to 0} \sup_{\theta \in \Theta_0} \mathbb{P}[T(X_1', \dots, X_n') > c_{p-\epsilon} | X_1', \dots, X_n' \sim F_\theta].$$

we conclude

$$p\text{-value} = \sup_{\theta \in \Theta_0} \mathbb{P}[T(X_1', \dots, X_n') > T(X_1, \dots, X_n) | X_1', \dots, X_n' \sim F_\theta].$$

problems with a textbook reference are based on, but not identical to, the given reference

(b) This says that if we sampled new data according to some distribution from our statistical model for which the null hypothesis is true, then the p-value is the probability of observing a statistic larger than that of our original data.

Problem 3 (Wasserman 10.5). *Let $X_1, \dots, X_n \sim \mathrm{Unif}(0,\theta)$ and let $Y = \max\{X_1, \dots, X_n\}$. We want to test $H_0 = \{\theta \leq 1/2\}$ vs $H_1 = \{\theta > 1/2\}$ and we will reject if $Y > c$ for some constant c.*

(a) *Find the power function $\beta(\theta) = \mathbb{P}[\max\{X_1, \dots, X_n\} > c | X_1, \dots, X_n \sim \mathrm{Unif}(0,\theta)]$.*

(b) *What choice of c will make the size of the test 0.05?*

(c) *In a sample of size $n = 20$ with $Y = 0.48$, what is the p-value? What is the conclusion about H_0 that you would make?*

(d) *In a sample of size $n = 5$ with $Y = 0.45$, what is the p-value? What is the conclusion about H_0 that you would make? Since $Y = 0.45$ is smaller than the previous case, we might expect it would be "harder to reject" H_0, yet the p-value is smaller than the previous case. Explain this observation.*

(e) *In a sample of size $n = 20$ with $Y = 0.52$, what is the p-value? What is the conclusion about H_0 that you would make?*

Solution.

(a) In homework 3 we computed the distribution of the maximum of n uniform random variables on $(0,\theta)$. Using this, we can compute

$$
\begin{aligned}
\beta(\theta) &= \mathbb{P}[\max\{X_1, \dots, X_n\} > c | X_1, \dots, X_n \sim \mathrm{Unif}(0,\theta)] \\
&= 1 - \mathbb{P}[\max\{X_1, \dots, X_n\} \leq c | X_1, \dots, X_n \sim \mathrm{Unif}(0,\theta)] \\
&= 1 - \begin{cases} c^n/\theta^n & c \in [0,\theta] \\ 1 & c > \theta \\ 0 & c < 0. \end{cases} \\
&= \begin{cases} 1 - c^n/\theta^n & c \in [0,\theta] \\ 0 & c > \theta \\ 1 & c < 0. \end{cases}
\end{aligned}
$$

(b) The size is

$$
\sup_{\theta \leq 1/2} \beta(\theta) = \begin{cases} 1 - 2^n c^n & 0 \leq c \leq 1/2 \\ 0 & c > 1/2 \\ 1 & c < 0. \end{cases}
$$

We can make the size equal to 0.05 if $0.05 = 1 - 2^n c^n$. I.e. if $c = (1 - .05)^{1/n}/2$.

(c) Here we will reject at the size α corresponding to $c < Y = 0.52$. This threshold is $\alpha = 1 - 2^{20}(0.48)^{20} \approx 0.558$, so $p \approx 0.558$.

(d) We now have $p = 1 - 2^5(0.45)^5 \approx 0.41$.

Since there is much less data, assuming the data was drawn from $\mathrm{Unif}(0, 1/2)$, it is more probable that we see a maximum of 0.45 when we draw only 5 samples than we see a maximum of 0.48 if we draw 20 samples.

This is why p-values can be meaningless if the sample sizes aren't sufficiently large.

(e) In this case we reject even at the size zero test with $c = 1/2$. So $p = 0$.

Note that if you try to compute the p-value from the formula $1-2^{20}(0.52)$ you will get a negative number which doesn't make sense for a p-value. This is because you must use the correct case when solving for the p-value. In particular, we should use the case for the size when $c > 1/2$.

Problem 4 (Wasserman 10.6). *There is a theory that people can postpone their death until after an important event. To test the theory, Phillips and King (1988) collected data on deaths around the Jewish holiday Passover. Of 1919 deaths, 922 died the week before the holiday and 997 died the week after.*

We can model this situation by thinking of whether each person died after or before the holiday as an iid Bernoulli random variable with parameter p. Then, the number of people who died after the holiday is a Binomial random variable with parameter (n, p), where $n = 1919$.

(a) for each $\alpha \in (0, 1)$, devise a size-α test for the null hypothesis that $p = 1/2$.

(b) Report and interpret the p-value.

(c) Construct a confidence interval for p.

Solution.

(a) Because we need a separate test for each α, we will want to come up with a class of tests. A simple form is what we've been using: reject if and only if $T(X_1, \dots, X_n) > c$, where c is a parameter we can choose to control the size.

We know $\bar{X}_n = (X_1 + \cdots + X_n)/n$ is an unbiased estimator for p. Thus, if we define $T(X_1, \dots, X_n) = |\bar{X}_n - 1/2|$, then if $T(X_1, \dots, X_n)$ is big, then it would be reasonable to reject.

So, we will use the test of the form reject if and only if $|\bar{X}_n - 1/2| \geq c$. The power function for this test is

$$\beta(p) = \mathbb{P}[|\bar{X}'_n - 1/2| \geq c | X'_1, \dots, X'_n \sim \text{Ber}(p)].$$

The size of the test is α if

$$\alpha = \sup_{p=1/2} \beta(p) = \mathbb{P}[|\bar{X}'_n - 1/2| \geq c | X'_1, \dots, X'_n \sim \text{Ber}(1/2)],$$

so we must solve for c in this equation.

Note that $n\bar{X}'_n$ is a binomial random variable with parameters $1919, 1/2$.

Let $F(x)$ be the PDF for a binomial random variable with the above parameters; i.e. $F(x) = \mathbb{P}[n\bar{X}'_n \leq x]$. Then, since the distribution is symmetric about $n/2$, conditioned on $\bar{X}_n \sim \text{Bin}(1919, 1/2)$ we have that

$$\mathbb{P}[|\bar{X}'_n - 1/2| \geq c] = \mathbb{P}[|n\bar{X}'_n - n/2| \geq nc] = 2\mathbb{P}[n\bar{X}'_n \leq n/2 - nc] = 2F(n(1/2 - c)).$$

Thus, the test has size α when $c = 1/2 - n^{-1}F^{-1}(\alpha/2)$.

(b) In our case, we have $\bar{X}_n = 997/1919$. Using problem 2, we simply need to compute

$$\begin{aligned} p\text{-value} &= \mathbb{P}[|\bar{X}'_n - 1/2| \geq |997/1919 - 1/2| | X'_1, \dots, X'_n \sim \text{Ber}(1/2)] \\ &= 2\mathbb{P}[\bar{X}_n \leq 1919/2 - |997 - 1919/2|] = 2F(1912/2 - 997) \approx 0.0911. \end{aligned}$$

(c) You can just reuse the confidence interval for Bernoulli that we've done before.

For instance, from Chebyshev's inequality we had

$$\mathbb{P}[p \in (\bar{X}_n - \epsilon, \bar{X}_n + \epsilon)] \le \frac{1}{4n\epsilon^2}$$

So if we construct a confidence interval so that $\bar{X}_n - \epsilon = 1/2$, then our interval is $(1/2, 0.539083)$ and our value of α is about 0.34.

On the other hand, if we want $\alpha = 0.1$, then our interval is $997/1919 \pm (0.483, 0.556)$.

Problem 5. *Suppose you work for a pharmaceutical company. You are trying to develop drugs to reduce the recovery time from infection by a particular virus.*

To model the effectiveness of a given drug, you could do the following: Let X_i the recovery time of a patient given a given drug, and suppose $X_i \sim N(\mu, 1)$ some unknown μ. The mean recovery time for patients who did not receive any treatment is μ_0. Thus, you would like to test the null hypothesis $H_0 = \{\mu \ge \mu_0\}$ by analyzing the outcomes $X_1, \ldots, X_n$ of a drug trial. If you are able to reject the null hypothesis, then you will be able to market your drug.

In reality, you studied statistics in college instead of chemistry and every drug you make is ineffective ($\mu = \mu_0$). However, you are greedy and want to make money by selling drugs. Your boss doesn't know statistics, and in your area there are no governmental regulation agencies. Thus, if you perform an experiment where you reject the null hypothesis at a p-value of 0.05, then you will be able to sell the drug and get rich.

(a) Explain how you could design an experiment (or series of experiments) in order to reject the null hypothesis for one of your drugs at a p-value of 0.05 (or at least give you a decent shot at this). Your experiments should be scientifically legitimate (i.e. you cannot just lie about the data collected in the trial).

(b) Now, suppose you are a governmental regulator. Explain what regulations you might introduce which would have prevented your alter-ego from getting away with such tricks.

Solution.

(a) Theorem 10.14 says that the probability of a given sample of data having p-value p or less is p. For this particular case, you can work out this fact exactly using an approach similar to the worksheets we did in class.

Thus, *regardless of n*, a given drug trial will have a 5% change of having a p-value of at most 0.05. If we do k trials, then we will have a $1-(1-0.05)^k$ change of at least one of them being successful. This means for $k = 90$, we would have over 99% change of finding a drug with a small p-value. Moreover, over k-trials. The expected number of "successful" drugs would be $0.05k = k/20$.

(b) There are many possibilities here.

One is to require drug manufacturers to report every test that they do. That way you could identify if they are trying to "get lucky" by doing a ton of tests. To strengthen this, you might also require that they have significant evidence based on science to suggest why the drug should work.

Another is to use something besides p-values which is better able to indicate the scientific significance of the test. For instance, even if n is large, it's still relatively easy to get a small p-value, but it would be much harder to construct a confidence interval which was scientifically relevant.

Finally, one could change the null hypothesis from $\mu \ge \mu_0$ to $\mu \ge \mu_0 - c$, where c is some number indicating how much of a decrease in recovery time is necessary to justify potential side-effects. In conjunction with a requirement on large sample size, this would make it very difficult to get a drug to pass the tests if $\mu = \mu_0$.