ENGRD 2700, Basic Engineering Probability and Statistics, Fall 2019
Homework 8 Solutions

Due **Friday December 6** at 11:59 pm. Submit to Gradescope by clicking the name of the assignment. See `https://people.orie.cornell.edu/yudong.chen/engrd2700_2019fa.html#homework` for detailed submission instructions.

The same stipulations from Homework 1 (e.g., independent work, computer code, etc.) still apply.

1. For some reason, Harry has kept meticulous records of Harry's muffins, and found that they usually contain a number of blueberries that is normally distributed with mean $\mu_0 = 25$ and standard deviation $\sigma = 4$. However, recently Jimmy suspects that Harry has been adding more blueberries than usual to his muffins. The last 12 muffins Jimmy bought contained the following numbers of blueberries:

   33 21 28 25 24 31 17 31 29 30 29 31

   Jimmy is considerate, and decides that he doesn't want to alarm Harry unless his findings are significant at the $\alpha = 0.05$ level.

   (a) Write down the (one-sided) hypothesis test being conducted above.
   Jimmy is conducting the hypothesis test

   $$H_0 : \mu = 25 \qquad\qquad H_1 : \mu > 25.$$

   (b) Treating σ^2 as the true variance of the distribution, compute the test statistic associated with the above data.
   We obtain the test statistic

   $$t_{obs} = \frac{\bar{X}_{12} - 25}{\sigma/\sqrt{12}} = \frac{27.417 - 25}{4/\sqrt{12}} \approx 2.093.$$

   (c) Does Jimmy reject H_0, the null hypothesis?
   The test statistic from part (b) gives the p–value

   $$p = P(Z > 2.093) \approx 0.018.$$

   Since $p < \alpha$, Jimmy rejects the null hypothesis.

   (d) If your answer to part (c) is "yes", how small would the sample mean $\bar{x}_{12}$ need to be for Jimmy not to reject H_0? Alternatively, if your answer to part (c) is "no", how large would $\bar{x}_{12}$ need to be for Jimmy to reject H_0?
   Jimmy rejects the null hypothesis whenever

   $$T = \frac{\bar{X}_{12} - 25}{4/\sqrt{12}} > z_{0.05} = 1.645 \implies \bar{X}_{12} > 1.645\frac{4}{\sqrt{12}} + 25 = 26.90$$

   (e) Repeat parts (b) and (c), but assume this time that Jimmy has no idea what σ^2 is.
   This time, our test statistic becomes

   $$t_{obs} = \frac{\bar{X}_{12} - 25}{S_{12}/\sqrt{12}} = \frac{27.417 - 25}{4.757/\sqrt{12}} \approx 1.760.$$

   and we obtain the p–value

   $$p = P(T_{11} \geq 1.760) \approx 0.053.$$

   where we evaluate the probability using the R command `1 - pt(1.760, df=11)`. Here, interestingly, Jimmy would fail to reject H_0.

2. Harriet suspects that the quarter in her pocket may not be a fair coin. She flips it 50 times, and to conduct a two-sided hypothesis test at the $\alpha = 0.025$ significance level. Heads appears $h = 28$ times.

   (a) Does Harriet reject H_0, the null hypothesis that the coin is fair, for a 2-sided test?
   Let p be the probability that the coin comes up heads. Harriet is conducting the hypothesis test

   $$H_0 : p = 0.5 \qquad\qquad H_1 : p \neq 0.5.$$

   Here, $n = 50$ and $\bar{x}_{50} = 28/50 = 0.56$, from which we obtain the test statistic

   $$t_{obs} = \frac{0.56 - 0.5}{\sqrt{(0.5)(0.5)/50}} \approx 0.849$$

   and the p-value

   $$p = P(Z \leq -0.849) + P(Z \geq 0.849) \approx 0.395.$$

   Harriet fails to reject H_0.

   (b) For what values of h would Harriet reject H_0?
   Harriet would reject H_0 if either $T \geq z_{0.0125} = 2.241$ or $T \leq -z_{0.0125} = -2.241$. Both inequalities hold when

   $$|t_{obs}| = \left| \frac{\bar{X}_n - 0.5}{\sqrt{0.5(0.5)/50}} \right| \geq |2.241|$$

   Or, equivalently

   $$(\bar{X}_n - 0.5)^2 \geq 2.241^2 \frac{0.25}{50}$$

   Solving the quadratic equation

   $$(x - 0.5)^2 - \frac{2.241^2}{200} = 0$$

   gives the solutions $x = 0.342$ and $x = 0.658$. Thus, we must have that $t_{obs} \geq 2.241$ when $\bar{X}_n \geq 0.658$, and $t_{obs} \leq -2.241$ when $\bar{X}_n \leq 0.342$. Since sample means of 0.342 and 0.658 correspond to obtaining roughly 17.1 and 32.9 heads, respectively, Harriet would reject H_0 if either $h \leq 17$ or $h \geq 33$.

3. According to the CDC, 17% of school-age children in the United States are obese, while 33.8% of adults in the U.S. are obese (having a Body Mass Index, or BMI, of at least 30).

   (a) In 2005, the Health Department in Marion County, Indiana measured the heights and weights of 90,147 school-age children, allowing exact determination of their BMIs. Among the children participating in the study, 22% were considered obese. Does this indicate that the true obesity rate for children in Marion County is different from the national average? Conduct a two-sided hypothesis test.
   Following the problem description, we conduct the hypothesis test

   $$H_0 : p = 0.17 \qquad\qquad H_a : p \neq 0.17.$$

   Here, our sample mean is $\hat{p} = 0.22$, and our test statistic is

   $$\frac{\hat{p} - 0.17}{\sqrt{\dfrac{0.17(1 - 0.17)}{n}}} = \frac{0.22 - 0.17}{\sqrt{\dfrac{0.17(0.83)}{90147}}} = 39.965.$$

   This corresponds to a p-value of

   $$P(Z > 39.965) + P(Z < -39.965) = 2.94 \times 10^{-349}.$$

   We reject the null hypothesis, as we have strong evidence that the true children obesity rate in Marion County is different from 0.17.

(b) The Marion County Health Department simultaneously conducted a telephone survey of 4784 adults. 25% of participants reported as being obese. Does this indicate that the true adult obesity rate in Marion County is lower than the national average? Conduct a one-sided hypothesis test.

Here, we perform the hypothesis test

$$H_0 : p = 0.338 \qquad\qquad H_a : p < 0.338.$$

Our test statistic is

$$\frac{\hat{p} - 0.338}{\sqrt{\frac{0.338(1-0.338)}{n}}} = \frac{0.25 - 0.338}{\sqrt{\frac{0.338(0.662)}{4784}}} = -12.867,$$

which corresponds to a p-value of

$$P(Z < -12.867) = 3.45 \times 10^{-38},$$

We again reject the null hypothesis, as we have strong evidence that the true adult obesity rate in Marion County is lower than 0.338.

(c) What are the potential issues with the study above?

There are several possible issues, one of which is the following. Note that children had their heights and weights measured, while the adults were only surveyed over the phone. The Health Department may have obtained a biased sample (as they only obtained data from individuals who agreed to provide a response). Also, survey participants may not have been honest when reporting their weights.

4. Consider once again the temperature data `ithaca.csv` and `syracuse.csv` from Homework 7. We want to conduct the hypothesis test

$$H_0 : \mu_i = \mu_s \qquad\qquad H_1 : \mu_i \neq \mu_s$$

at the $\alpha = 0.05$ significance level, where μ_i and μ_s denote the mean temperatures in both cities during the month of March. Attach your code for the following questions.

(a) If we make the (unrealistic) assumption that the two samples are independent, do we reject H_0?

We know that dataset ithaca.csv has $\bar{X}_i = 51.452$ and $S_i^2 = (12.662)^2$, and dataset syracuse.csv has $\bar{X}_s = 50.742$ and $S_s^2 = (12.596)^2$. Each has 31 samples. Thus, the test statistic is:

$$t_{obs} = \frac{51.452 - 50.742}{\sqrt{\frac{(12.662)^2}{31} + \frac{(12.596)^2}{31}}} = 0.221$$

The p-value is $P(|Z| \geq 0.221) = P(Z \geq 0.221) + P(Z \leq -0.221) = 0.825 > \alpha$. Hence we fail to reject H_0.

The sample R code is as follows:

```
syracuse = read.csv("syracuse.csv");
ithaca = read.csv("ithaca.csv");
x_i = mean(ithaca$maxtemp);
s_i = sd(ithaca$maxtemp);
x_s = mean(syracuse$maxtemp);
s_s = sd(syracuse$maxtemp);
Z = (x_i - x_s)/sqrt(s_i^2/31 + s_s^2/31);
2*pnorm(-Z);
```

(b) Repeat part (a), but relax the assumption that the two cities are independent. That is, we build a hypothesis test for paired, dependent data. (Now that this is only possible when the two samples contain the same number of observations for the same dates.)

We can set D to be the difference between the corresponding data of Ithaca and Syracuse. Using R, we can calculate $\bar{D} = 0.710$, $S_D = 3.21$ and we know $n = 31$. Then the test statistic is:

$$t_{obs} = \frac{0.710}{3.21/\sqrt{31}} = 1.23$$

$p = P(Z \le -1.23) + P(Z \ge 1.23) = 0.219$. Therefore, we fail to reject H_0.
The sample R code is as follows:

```
diff = ithaca$maxtemp - syracuse$maxtemp;
mean(diff);
sd(diff);
n = length(diff);
Z = mean(diff)/(sd(diff)/sqrt(31));
2*pnorm(-Z);
```

5. A certain NBA player had a field goal percentage (i.e., probability of making a shot) of $p_0 = 60\%$ before needing to take a season off to recover from an injury.

   (a) Since returning to the game from injury, the player has made 13 out of $n = 20$ shots. Is the player's new, post-injury field goal percentage higher than his old percentage p_0? Perform a suitable one-sided hypothesis test and state your conclusion, taking $\alpha = 0.05$.
   We do a test for a proportion. Null hypothesis: $p = 0.6$, alternative hypothesis: $p > 0.6$.
   Since $\hat{p} = 13/20 = 0.65$, the test statistic takes value $t_{obs} = \frac{\hat{p}-0.6}{\sqrt{0.6\cdot(1-0.6)/n}} = 0.456$,
   (Note that in the denominator above you should not use $\sqrt{\hat{p}(1-\hat{p})/n}$, because we are computing the distribution of the test statistic assuming that the null hypothesis is correct. The null hypothesis says $p = 0.6$, so we use that value.) The p-value is $p = P(Z > 0.456) = 0.32 > 0.05$, so we fail to reject the null hypothesis. Therefore, we cannot conclude or do not have strong evidence that the post-injury field goal percentage has increased.

   (b) Suppose that the true new field goal percentage is p, where $p \in (0.6, 1)$. If we perform a one-sided test as above and want to achieve type-I error rate of 0.05 and type-II error rate of 0.025, what is the number of shots n needed since returning from injury? Provide an approximate formula as a function of p, and compute the values of n for each of $p = 0.8, 0.7, 0.61$ (Notice that if p is very close to 0.6 then you may need a very large number of shots.)
   We can either work directly with a formula from the notes or derive the result directly.

   <u>Method 1</u>: Let's use the formula
   $$n = \frac{\sigma^2(z_\alpha + z_\beta)^2}{(\mu_1 - \mu_0)^2}$$
   from the hypothesis testing notes. Here we have $z_\alpha = z_{0.05} = 1.645$ and $z_\beta = z_{0.025} = 1.96$. Also $\mu_1 = p$ and $\mu_0 = p_0 = 0.6$. Finally, σ^2 is the variance of a single observation. This is tricky because the variance is different under the null and alternative hypothesis depending on p_0 and p. We can approximate this variance by its value at $p_0 = 0.6$ as $p_0(1 - p_0) = 0.6(1 - 0.6) = 0.24$. Thus,
   $$n = 3.12/(p - 0.6)^2.$$
   For $p = 0.8, 0.7, 0.61$ we get (rounding up to integers) 78, 312, 31200 shots required. It is hard to detect small differences in field goal percentage!

   <u>Method 2</u>: We can also derive the result directly. We will reject the null hypothesis if
   $$t_{obs} = \frac{\hat{p} - 0.6}{\sqrt{0.6 \cdot (1 - 0.6)/n}} > z_{0.05}.$$
   The choice of $z_{0.05}$ above ensures a type-I error rate of 0.05. We want the type-II error rate to be 0.025; that is, when the true new field goal percentage is p, we want the probability of rejecting the null to be 0.975:
   $$\begin{aligned} 0.975 &= P\left(\frac{\hat{p} - 0.6}{\sqrt{0.6 \cdot (1 - 0.6)/n}} > z_{0.05}\right) \\ &= P\left(\hat{p} > \frac{\sqrt{0.24} \cdot z_{0.05}}{\sqrt{n}} + 0.6\right) \\ &= P\left(\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} > \frac{\frac{\sqrt{0.24}\cdot z_{0.05}}{\sqrt{n}} + 0.6 - p}{\sqrt{p(1-p)/n}}\right). \end{aligned}$$

The quantity $\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}$ is approximately distributed as $\mathcal{N}(0,1)$. To make the above probability equal 0.975 as desired, we need to set

$$\frac{\frac{\sqrt{0.24}\cdot z_{0.05}}{\sqrt{n}}+0.6-p}{\sqrt{p(1-p)/n}}=-z_{0.025}$$

$$\Longrightarrow n=\left(\frac{\sqrt{0.24}\cdot z_{0.05}+z_{0.025}\sqrt{p(1-p)}}{p-0.6}\right)^2$$

Plugging in the values $z_{0.05}=1.645$ and $z_{0.025}=1.96$, we obtain the formula

$$n=\left(\frac{0.81+1.96\sqrt{p(1-p)}}{p-0.6}\right)^2$$

The values of n for $p=0.8, 0.7, 0.61$ (rounding fractions up) are $64, 292, 31188$. (Note that we get a slightly different answer to the first method above because we didn't make the variance approximation in this calculation that we did above. Also, if the students round the answers at only at the final step, they will get $64, 292, 31188$, which are also correct.)