STAT/CS 94 Fall 2015 Adhikari HW08, Due: 10/28/15

NAME: SID:

This week's homework is a bit longer than the previous weeks' and has two pages: A question sheet and an answer sheet. Both are two-sided. In the published PDF document, the answer sheet pages come after the questions. *Please write your answers on the (double-sided) printed answer sheet, in the space provided.* (There will be a small penalty for not following this instruction; it makes the grader's job more difficult.)

Some problems include numerical output of Python code. You are welcome to round the output to two decimal places in your calculations.

Problem 1 Normal Newborns

The distribution of the birth weights of babies at a hospital follows the normal curve quite closely, with an average of 120 ounces and an SD of 15 ounces. In each part below, write one line of Python code that evaluates to the value that should be used to fill in the blank. You can assume that the module `stats` has been imported from `scipy`.

(a) The proportion of birth weights that are more than 110 ounces is approximately _____.

(b) The 80th percentile of the birth weights is approximately _____ ounces.

(c) The proportion of birth weights that are in the range 125 ounces to 130 ounces is approximately _____ ounces.

(d) About 50% of the birth weights are in the range 120 ounces plus or minus _____ ounces.

(e) About 50% of the birth weights are in the range 100 ounces to _____ ounces.

Answer:

(a) `1 - stats.norm.cdf(110, 120, 15)`

(b) `stats.norm.ppf(.8, 120, 15)`

(c) `stats.norm.cdf(130, 120, 15) - stats.norm.cdf(125, 120, 15)`

(d) `stats.norm.ppf(.75, 120, 15) - 120`

(e) `stats.norm.ppf(.5 + stats.norm.cdf(100, 120, 15), 120, 15)`

Note: You can also convert to standard units by hand if you prefer; for example, another correct answer to part (a) is `1 - stats.norm.cdf((110 - 120)/15)` and to part (d) is `15*stats.norm.ppf(.75)`.

Problem 2 Pressure Probabilities

Here is some code with its output. Use this (not other code) to find the following values. You might need to use some of the given output more than once, and some not at all.

(i) `stats.norm.ppf(0.7)`: 0.52440051270804067

(ii) `stats.norm.cdf(0.3)`: 0.61791142218895256

(iii) `stats.norm.cdf(1.5)`: 0.93319279873114191

(iv) `stats.norm.cdf(2.5)`: 0.99379033467422384

(a) the approximate proportion of diastolic blood pressures that are above 86 mm, assuming a distribution of diastolic blood pressures that is approximately normal with average 76 mm and SD 4 mm

(b) the approximate proportion of diastolic blood pressures that are between 70 mm and 86 mm, assuming a distribution of diastolic blood pressures that is approximately normal with average 76 mm and SD 4 mm

(c) the approximate 30th percentile of heights, assuming a distribution of heights that is approximately normal with average 69 inches and SD 3 inches

Answer:

(a) `1 - stats.norm.cdf(2.5)` (since 86mm is 2.5 standard units above 76mm for this Normal distribution)

(b) `stats.norm.cdf(2.5) - (1 - stats.norm.cdf(1.5))`

(c) `69 - 3*stats.norm.ppf(0.7)` (since `-1*stats.norm.ppf(0.7)` gives the number of standard units whose CDF value is 0.3, and multiplying by 3 and adding 69 converts that to the given Normal distribution)

Problem 3 Example Exam

(a) Describe a situation in which you would use the Table method `.join` when analyzing data. Don't use an example you've seen in existing course materials.

(b) Describe a situation in which you would use the Table method `.group` when analyzing data. Pick an example that *does not use* `collect=sum`, `collect=len`, or `collect=max`. Don't use an example you've seen in existing course materials.

Answer: Of course, there are many correct answers to this problem.

(a) I have one dataset of pictures from one telescope at a certain location on Earth along with the time at which each picture was taken, and another dataset giving the spatial orientation of the telescope's location at different times. I want to associate each picture with the spatial orientation of the telescope when the picture was taken, so I can figure out what part of the sky I am seeing in each picture. So I join the two tables on the time column.

(b) I have a dataset of income records for many people around the world, along with the country in which each person lives. I would like to compute the Gini coefficient of each country. (The Gini coefficient is a rough but popular measure of income inequality within a group of people; see `https://en.wikipedia.org/wiki/Gini_coeffic` I write a function called `gini` that takes a single array of incomes and computes its Gini coefficient. Then I group my table of income records by country, with `collect=gini`, producing a table with the name and Gini coefficient of each country.

Problem 4 Dividend Deviations

Incomes in a large population have an average of \$63,000 and an SD of \$40,000. Fill in each blank with the best choice from the list below, and explain.

(i) \$100 (ii) \$2,000 (iii) \$40,000

(a) If you draw one income at random from the population, it is expected to be \$63,000 give or take _____.

(b) If you draw a sample of 400 incomes at random with replacement from the population, the average of the sample is expected to be \$63,000 give or take _____.

Answer:

(a) 40,000. (Just the population standard deviation.)

(b) 2,000. (Imagine repeatedly taking samples of 400 things and calculating the mean of each one. The standard deviation of that distribution of means, a.k.a. the standard error of the mean, is the population standard deviation divided by $\sqrt{400}$. Note that this *always* true of the standard error of the *mean*, even if, as in this case, the population is not approximately a Normal distribution.)

Problem 5 Expected Excursion

A random sample of 1,000 adults is taken from all the adults in a city of over 2 million inhabitants. You can assume that the method of sampling is essentially indistinguishable from random sampling with replacement.

The average commute distance of the sampled people is 5.6 miles and the SD is 4.1 miles.

(a) True or false (explain): Because the sample is large, the distribution of the commute distances in the sample is approximately normal.

(b) Pick one of the two options to complete the sentence: The interval "5.6 miles $\pm 2 \times 4.1/\sqrt{1000} = 5.34$ to 5.86" is an approximate 95%-confidence interval for the average commute distance of all the adults in the [sample, city].

(c) Pick one option and explain your choice. In the calculation in part (b), the normal curve

(i) is not used at all.

(ii) is an approximation to the histogram of all the commute distances in the sample.

(iii) is an approximation to the histogram of all the commute distances in the population.

(iv) is an approximation to the probability histogram of the sample average.

Answer:

(a) False. Because the sample is large, the distribution of the commute distances in the sample is approximately *the distribution of the commute distances in the population*. The problem did not say that the commute distances in the population follow a Normal distribution.

(b) City. (We already know the average commute distance of all the adults in the sample – 5.6 miles. So you should be suspicious of any attempt to estimate it.)

(c) (iv). The standard deviation of the means of repeated samples of size 1000 (a.k.a. the standard error of the mean) is $4.1/\sqrt{1000}$ (just as in the previous problem). The Normal approximation comes in when we go from the standard deviation of that distribution to a 95% confidence interval. 2 standard deviations to the left and right of the mean of a Normal distribution cover about 95% of that distribution, and that is the number we multiplied $4.1/\sqrt{1000}$ by to get our confidence interval. Any time you compute a confidence interval that way, you are implicitly using a Normal approximation to the probability distribution of the sample average. Note that the Central Limit Theorem is a reasonable justification for this approximation, since we have a large sample (1000) and the statistic we're computing on the sample is its mean.

Problem 6 Hot Hominids

A "normal" human body temperature has long been considered to be 98.6 degrees Fahrenheit. In a sample of 100 people taken from a large population, the average temperature was 98.4 degrees Fahrenheit with an SD of 0.2 degrees. You can assume that the method of sampling was essentially the same as random sampling with replacement.

Use these data to test the following hypotheses, in the steps laid out in parts (a) through (c).

Null. The average temperature in the population is 98.6 degrees Fahrenheit; the average in the sample is different due to chance.

Alternative. The average temperature in the population is not 98.6 degrees Fahrenheit.

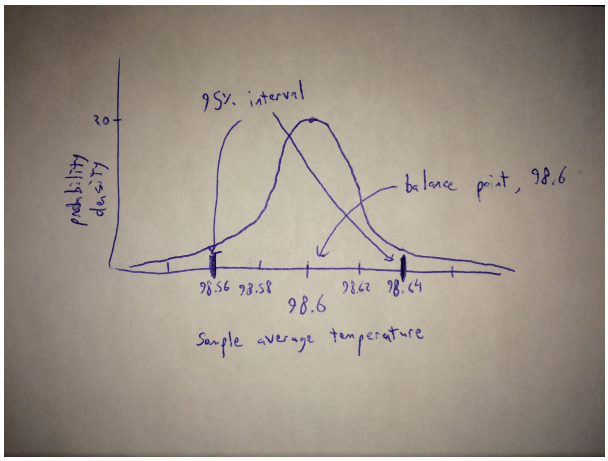
(a) Sketch (no computer; just draw a freehand sketch) the probability histogram of the sample average, assuming that the null hypothesis is true. Show where the histogram balances and provide a numerical value for the balance point. Mark an interval on the horizontal axis that is symmetric about the balance point and has probability about 95%.

(b) Give numerical values for the two ends of the interval that you marked in (a), and explain whether your values are exact or approximate.

(c) Complete the test and make a conclusion. You can use any reasonable cutoff for the P-value.

Answer:

(a)



Note: The null hypothesis does not specify the *standard deviation* of the temperature in the population, only the *mean* and *shape* (Normal). So we have made our best estimate of the population standard deviation under the null hypothesis by plugging in the standard deviation of the sample. This is a typical move, but it's important to remember that what we have drawn is only an approximation to the probability histogram of the sample average.

Note 2: We drew in actual numbers for the density, but you didn't need to do that for full credit.

(b) $98.6 \pm 2 \times \frac{0.2}{\sqrt{100}}$, or $[98.56, 98.64]$. Our values are roughly a 95% interval for the Normal distribution with mean 98.6 and SD 0.02. They are approximations to the histogram of the sample average (across repeated sampling under the null hypothesis), for three reasons:

(i) As noted above, we have simply plugged in the sample standard deviation as our best guess at the population standard deviation under the null.

(ii) The Central Limit Theorem applies, but it only guarantees that the distribution of the sample average is *approximately* Normal, not exactly Normal.

(iii) 2 is only an approximation to the 97.5th percentile of the standard Normal distribution; $98.6 \pm$ `stats.norm.ppf(.975)` $\times \frac{0.2}{\sqrt{100}}$ is technically more accurate.

(c) We will use the sample mean minus 98.6 as our test statistic. This necessitates a two-sided test, since large negative and large positive values are evidence against the null. The P-value is therefore approximately $2 \times$ `stats.norm.cdf((98.4 - 98.6)/0.02)`, or $2 \times$ `stats.norm.cdf(-10)`. Python reports that this is roughly 1.5×10^{-23}, which is a very, very small number. (You should have been able to guess that without calculating it.) So we can reject the null hypothesis at any cutoff, say .0000001.

Problem 7 Vigilant Voters

A poll of voters in a large city is based on a sampling method that is essentially the same as random sampling with replacement a large number of times. The methods of this class have been used to construct confidence intervals for various proportions in the voting population. For example, an approximate 95%-confidence interval for the proportion of voters who will vote for Proposition X is (0.36, 0.42).

Find an approximate 99%-confidence interval for the proportion of voters that will vote for Proposition X, in the following steps. Some Python output is provided for you, in case you want to use it.

`stats.norm.ppf(0.95)`: 1.6448536269514722
`stats.norm.ppf(0.975)`: 1.959963984540054
`stats.norm.ppf(0.99)`: 2.3263478740408408
`stats.norm.ppf(0.995)`: 2.5758293035489004

(a) Pick one of the two options *and explain*: The center of the interval (0.36, 0.42) is the proportion of "Yes on X" supporters in the [city, sample].

(b) The SE of the sample proportion is approximately _____.

(c) Complete the problem: find an approximate 99%-confidence interval for the proportion of voters that will vote for Proposition X.

Answer:

(a) Sample. The method we've seen for computing confidence intervals for population proportions computes a confidence interval that is *centered* on the sample proportion. That means that the confidence interval is of the form [sample mean $- A$, sample mean $+ A$] for some number A. If we knew that the center of the confidence interval were exactly the population proportion, we would just report that number and forget about the confidence interval!

(b) .03/`stats.norm.ppf(0.975)`, or roughly 0.015. (From part (a), $A = .42 - .39 = .03$. A 95% confidence interval, using a Normal approximation to the sample proportion, is [sample mean+`stats.norm.ppf(.025)`$\times$ SE, sample mean+`stats.norm.ppf(.975)`$\times$SE].) Therefore .03 = `stats.norm.ppf(.975)`$\times$SE, and solving for the SE we get SE = .03/`stats.norm.ppf(.975)`.

(c) Now we know the (approximate) SE and sample mean, so we can just use our formula to get different-sized confidence intervals. As in the previous part, our 99% confidence interval will be [sample mean + `stats.norm.ppf(.005)`$\times$SE, sample mean+`stats.norm.ppf(.995)`$\times$SE]. We can replace `stats.norm.ppf(.005)` with `-stats.norm.ppf(.995)` and substitute in the values we've calculated to get [.39$-$`stats.norm.ppf(.995)`$\times$ .03/`stats.norm.ppf(0.975)`, .39 + `stats.norm.ppf(.995)` $\times$ .03/`stats.norm.ppf(0.975)`].