## Questions

1. P-values and decisions:
   (a) Suppose I do a test with $\alpha = 0.05$ and tell you that I rejected $H_0$. What does that tell you about my P-value?
   (b) Suppose that you wanted to do this test with $\alpha = 0.01$. What can you conclude from what I told you?
   (c) Suppose now I tell you that my P-value was 0.0234. What do you conclude now, at $\alpha = 0.01$?

a: less than 0.05
b: nothing: P-value might be less than 0.01 or might not be.
c: Do not reject null hypothesis. (Knowing that the P-value is 0.0234 tells you that the evidence against H0 is strong but not very strong; if all you know is that the P-value is less than 0.05, you don't know *how* strong the evidence is.)

2. Suppose we test $H_0 : \mu = 20$ vs. $H_a : \mu \neq 20$ using sample $n = 10,000$, $\sigma = 0.2$, $\alpha = 0.05$. The sample mean is $\bar{x} = 20.005$.
   (a) Would you guess that the sample mean is (statistically) significantly different from 20?
   (b) Calculate the $z$ test statistic and P-value for this test. What do you conclude?
   (c) The researchers conducting the study tell you that the difference between 20 and 20.005 is not important in practice. Is this consistent with the result of your test? Why or why not?

a: No; it seems very close to 20.

b: $z = \frac{20.005 - 20}{0.2/\sqrt{(10000)}} = 2.5$ , P-value 2x0.0057=0.0114.

c: No (would reject H0): statistical significance not same as practical significance (sample size "too big").

3. This is similar to the previous question. This time, $H_0 : \mu = 20$, $H_a : \mu \neq 20$, $\bar{x} = 26$, $n = 2$ and $\sigma = 5$.
   (a) This time the sample mean is far from the null hypothesis. Would you expect to reject the null hypothesis?
   (b) Calculate the $z$-statistic and P-value. What do you conclude?
   (c) This time, the researchers tell you that the difference between 20 and 26 is definitely important in practice. Is this consistent with the result of your test? Why or why not?
   (d) Extra: use StatCrunch to calculate 95% confidence intervals for the population mean in this question and the last one. Does this help to explain what is going on?

a: yes: H0 seems very wrong.

b: $z = \frac{26-20}{5/\sqrt{2}} = 1.70$ ; P-value 0.0897, do not reject H0.

c: no, as before; this time sample size too small.

d: this question: 19.1 to 32.9 (very wide), previous 20.001 to 20.009 (very narrow). Here, 26 is inside the CI (plausible given data); previously, 20 is outside the CI (not plausible given data).

4. Usually, we are searching for *small* P-values, so that we can reject our null hypothesis in favour of our alternative. Suppose we have some scientific theory that we test, based on an experiment with adequate sample size, and obtain a P-value that was not small. Can you think of two reasons why this finding might be interesting? Explain briefly.

– plausible theory that turns out to be wrong
– some problem with the experiment

Either way, the result seems interesting and would be worth following up on, except that it probably wouldn't be publishable -- journals like small P-values.

5. An industrial process is supposed to produce output with mean 60. So each day, a sample is taken from the output and a test is conducted of $H_0 : \mu = 60$ vs. $H_a : \mu \neq 60$. The process manager thinks that if the null hypothesis is rejected, there must be a fault in the process. Perform a simulation in StatCrunch to see whether the process manager is right, as follows:
   (a) Generate a simulated population of 1000 normal values with mean 60 and SD 10.
   (b) Take 100 random samples of size 50 from this population, using Data and Sample Columns. Store the samples Stacked with a Sample ID, as we did in §5.1.
   (c) Carry out a test of the hypotheses given above for each sample. This is done by Stat, Z-statistics, 1-sample Z, With Data. Select the column with the random-sample data in it, fill in the population SD of 10, and Group By the sample number. Set up the test, click Calculate, and you'll see the results of the 100 tests.
   (d) These tests were all based on data where the null hypothesis is correct. What would you expect to see in the P-values? What do you see?
   (e) Do you think the process manager is right? Explain.



You do the simulation.

Idea of answer: if you do a lot of tests, some of them (about 5% if $\alpha=0.05$ ) will be significant just by chance even if the null hypothesis is true. So it is perfectly reasonable to conclude that the process mean really is 60 if somewhere around 5% of the 100 tests came out significant.

A step further: if the null hypothesis is true, the number of significant tests is binomial with n=100 and p=0.05, and you can see whether

your simulation is out of line compared to that. (Extra extra: what if the process mean is actually 70, so that the null *is* wrong? What would you expect to happen then? What actually does happen?)

6. One of the concerns about SAT scores is that they are increasing over time. The historical mean is 450. Suppose we consider taking a sample of 500 students. What is the power of this test to detect an increase of 10 points in mean score? Use $\alpha = 0.05$ and assume $\sigma = 100$; proceed by simulation as below, using StatCrunch:
    1. Generate a large normally-distributed population with mean 460 (why?) and SD 100, using Data and Simulate. (A population of size 5000 is good.)
    2. Generate a large number (100 will do) of samples of size 500 from this population, using Data and Sample Columns. Save them "stacked with a sample ID".
    3. For each sample, test $H_0 : \mu = 450$ against $H_a : \mu > 450$ (using your different samples as "group by"). You'll get a table with 100 P-values in it. How many of those P-values are less than 0.05? This, out of 100, is the power of your test. (Alternatively, save the output as a Data Table, and then sort the P-values into order. This makes it easier to count how many of them are less than 0.05.)

This is about power, so you can skip it.

7. In the situation of the previous question, roughly how large a sample should be used to obtain a power of 0.60 (60%)? Assess this by trying a few different sample sizes until you find one with about the right power. (A larger sample size should give a larger power, all else being equal.)

This is also about power; skip this one.