Data analysis practice # 6. Stat 200H

Load again the data set Miami.mtw from the data directory Student1, containing several variables collected on a sample of n=83 female college students.

A. Confidence intervals for the mean of a quantitative variable

Consider Wgt=weight in pounds. Use Graph > Histogram and Graph > Probability plot (which we explained a few times ago) to evaluate the shape of the empirical (i.e. sample data) distribution for Wgt. Do you think these plots support or give evidence against a normal (or almost normal) underlying population shape?

Now, use Stat > Basic statistics > 1-sample Z (specifying confidence level 95% in “Options”) to produce confidence intervals for the population mean weight – ignore all the parts that have to do with testing. This command employs a standard normal distribution to derive the multiplier in a confidence interval (z*). We are using it under the assumption that n=83 is large enough for the normal approximation to work. You are asked to specify a “sigma” which we will take to be the sample standard deviation of Wgt (you need to find this using “Display descriptive statistics”, or “Column statistics”) – again, if we consider this as a large sample, using s instead of the unknown population sigma will not do much harm.

Next, use Stat > Basic statistics > 1-sample t (specifying confidence level 95% in “Options”) to produce the corresponding confidence interval. This command employs a t distribution to derive the multiplier in a confidence interval (t*). You don’t need to specify the degrees of freedom for the T distribution (Minitab takes them to be the number of cells in the column for Wgt, i.e. n, minus 1).

Compare the two 95% confidence intervals. Are they similar/different? Why? Interpret such an interval using non-statistical terms.

B. Confidence intervals for comparing the means of a quantitative variable in two populations

Now we will use the dichotomous variable Diet=currently dieting (Yes=1, No=2) to divide our sample into two sub-samples. To do this, use Manip > Split worksheet, and specify you want the split “By variable” Diet. This produces two additional worksheets: one contains the n1=30 students that are dieting, and one contains the n2=53 students that are not. You can now use any command on either of the worksheets separately. We will think of the two sub-samples as two independent samples (if all observations are independent, the two sub-samples will obviously be) obtained from the population of dieters and non-dieters, respectively.

Start by obtaining Stat > Basic Statistics > Display descriptive statistics, for Wgt in the two samples. What are the sample means, standard deviations, and standard errors?
Which sample has the lower average Wgt? Are the standard deviations in the two samples very different? Just based on this numbers, with an informal reasoning, do you think there is enough evidence to infer a difference in mean weight between the two populations of dieters and non-dieters?

Next, we will construct a 95% confidence interval for the difference \( d = \mu_1 - \mu_2 \). If this interval is entirely positive, we have strong evidence to back the claim that the population of dieters is on average heavier than that of non-dieters. However, if the interval contains 0 (so that our range of likely values for the difference contains both positive and negative values) the available data does not provide enough information to draw a strong conclusion. To build this interval, go back to the worksheet with the entire data set, and use Stat > Basic Statistics > 2-sample t. Our samples are here in one column (check this option), put Wgt into “Samples”, and Diet into “Subscripts” (the latter tells Minitab what other column to use to split the values in Wgt into two groups). In “Options” specify 95%. Also, go in “Graphs” and ask for a box-plot of the data (this is going to be interesting as a visual complement to our confidence interval).

From the descriptive statistics you displayed before, it should be clear that Wgt has different standard deviations in the two samples, so we should NOT assume equal variance between the two populations. Do NOT check “Assume equal variances”.

Comment on the resulting interval, and articulate some more based on the box-plots (the red dot displayed on top of each is the sample mean). Note that the confidence interval produces an inference on the difference between the population means/locations, while the box plots give you a snap-shot (albeit subject to sampling variability) of how the two populations may look like, besides their location. What are the most striking features here?

The two-sample confidence interval you just constructed employs a multiplier obtained from a T distribution (t*). The theory besides it is as follows (here I am using the language of the book): if the two populations are normal (or almost), or if the two samples are large, then the statistic

\[
\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim (\text{approx}) \quad t_{df}
\]

with a complicated formula for the degrees of freedom, which Minitab computes automatically – you can find it on page 411 of the book. The denominator is the standard error of the sample difference that appears at the numerator.
C. Is weight normal (or almost)?

Our overall sample is quite large (83 > 30), so regardless of whether weight is normal (or almost) in the overall population, we can use almost indistinguishably a z-based or a t-based confidence interval for the overall population mean weight (part A).

Our sub-samples are fairly large, too (30, 53 both ≥ 30). So the confidence interval we computed for the difference between mean weight of dieters and non-dieters is likely to be quite all right, regardless of whether weight is normal in the two populations (part B).

However, I would like you to comment on whether you think that the empirical distributions we observe for weight (in the overall sample, and within each sub-sample) present some evidence against normality – look back at the histogram, normality plot, and box plots you have produced, and produce more plots if you want. Can you see extreme high values? Could we actually suspect a heavy/long right tail?