Philip Ourso

ST516

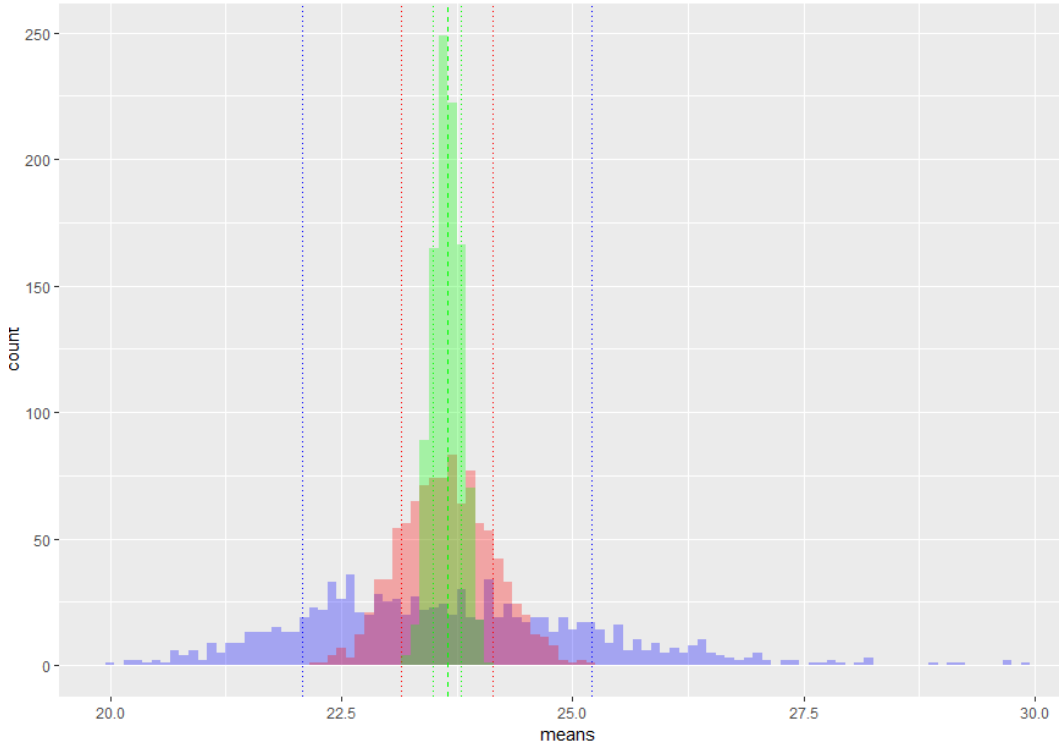
Module 9 Homework: Progress Report

1. Simulation study
    a. Describe the sampling distribution of the sampling mean of 2013 BMI data.
    Please see R code in accompanying .R script.

    Sample means and standard deviations for sample sizes n = 10, 100, 1000

```
       n_lo   n_md   n_hi
[1,] 23.642 23.642 23.643
[2,]  1.563  0.498  0.151
```

    As can be seen from the output of rbind(), the mean of the sampling distribution is fairly stable as sample size increases from 10 to 100 to 1000, but the standard deviation dramatically decreases with increasing sample size. This is illustrated in the histogram overlay, wherein the mean ± standard deviation is indicated by dashed and dotted lines of the corresponding color.



2. Translate the questions of interest into inferential questions.
    a. How has BMI changed? Are high-schoolers becoming more overweight?
    To answer this question, the data is already in acceptable form: a continuous numeric of type double. An appropriate analysis here would be to compare two population

means, those of the 2003 and 2013 data. A one-sided, two-sample t-test could be performed to determine if mean BMI had appreciably increased from 2003 to 2013, or a two-sided analysis could be performed to detect a potential difference.

b. Are male high-schoolers more likely to smoke than female high-schoolers, given 2013 data?

   For this question, the relevant data are the gender of the respondent, and their answer to the question regarding frequency of smoking within the last 30 days. The data would need to be transformed from the provided responses into something numeric that could be analyzed with a statistical procedure.

   One approach would be to transform the responses into a binomial condition: 0 corresponds to the "0 days" answer, and 1 to any other answer. From this, a difference in proportions could be investigated, using a 2-proportion hypothesis test. This approach would, however, discard information into the frequency of smoking, and hence more likely address the likelihood of being a non-smoker, not the likelihood of smoking.

   Instead, the responses could be converted to ordinal data, with increasing values indicating increasing frequency of smoking. From this, a comparison of medians could be performed, to understand the 50^{th} percentile of each gender.

c. How much TV do high-schoolers watch?

   The desired analysis here explores the q81 column data, an answer to the average television watched on a school day, to infer the amount of television watched by the average high-schooler. The question is somewhat vaguely worded, and so we interpret it to mean "how much TV do high-schoolers watch *on an average school day*?" Data can then be simply transformed into an integer value and a one-sample t-test performed to infer a population mean, however care must be taken when summarizing results, as the data is truncated at a level of "5 hours or more".