# STA035B Final Prep, Winter 2024

You will be given this page for your final: two tables and formulas for SSG and SSE.

| | One sample | Two independent samples |
|---|---|---|
| Response variable | Numeric | Numeric |
| Standard error | $\frac{s}{\sqrt{n}}$ | $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ |
| Degrees of freedom | $n-1$ | $\min(n_1 - 1, n_2 - 1)$ |

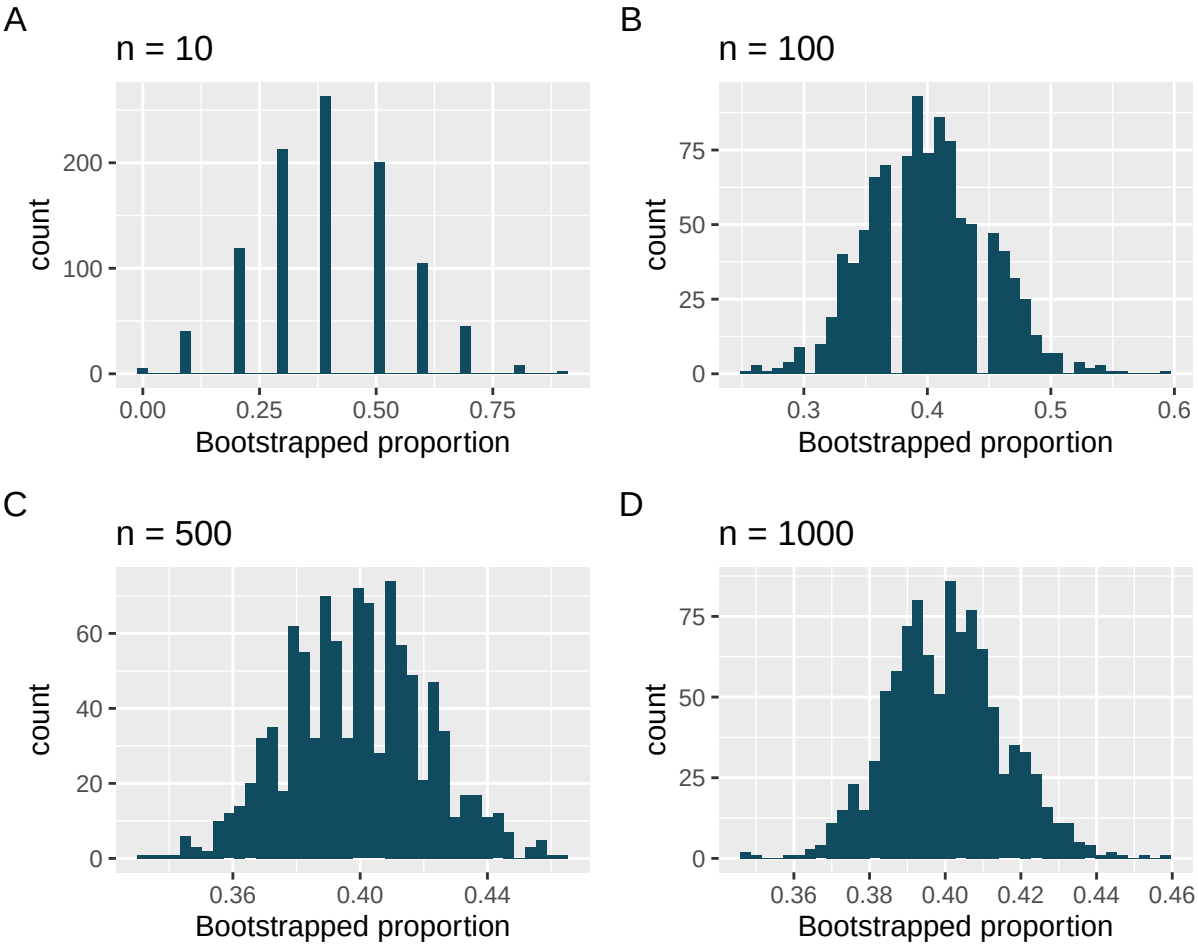
| | One sample | Two independent samples |
|---|---|---|
| Response variable | Binary | Binary |
| Standard error: HT | $\sqrt{\frac{p_0(1-p_0)}{n}}$ | $\sqrt{\widehat{p}_{pool}\left(1-\widehat{p}_{pool}\right)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$ |
| Standard error: CI | $\sqrt{\frac{\widehat{p}(1-\widehat{p})}{n}}$ | $\sqrt{\frac{\widehat{p}_1(1-\widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1-\widehat{p}_2)}{n_2}}$ |

For $k$ groups with sample means $\bar{x}_i$, $i = 1, \ldots, k$ per group (each with $n_i$ samples) and for $\bar{x}$ as the sample mean across all $n = \sum_{i=1}^{k} n_i$ samples,

$$SSG := \sum_{j=1}^{k} n_j(\bar{x}_j - \bar{x})^2, \qquad SSE := \sum_{i=1}^{n}(x_i - \bar{x})^2.$$

**Problem 1 (repeat from hw 4)**

Each of the following four distributions was created using a different dataset. Each dataset had the same proportion of successes ($\hat{p} = 0.4$) but a different sample size. The four datasets were given by $n = 10, 100, 500,$ and 1000.




Consider each of the following values for the true popluation $p$ (proportion of success). Datasets A, B, C, D were bootstrapped 1000 times, with bootstrap proportions as given in the histograms provided. For each parameter value, list the datasets which could plausibly have come from that population. (Hint: there may be more than one dataset for each parameter value.)

a. $p = 0.20$

b. $p = 0.45$

c. $p = 0.55$

**Problem 2**

Suppose that we have a dataset on car crashes, where we have information on whether or not the driver was over 75 years old as well as knowledge of whether or not the individual has had an accident in the last 3 years:

```
# A tibble: 2 x 2
  over75 crash_last3
   <dbl>       <dbl>
1     55          12
2     45           3
```

The way to read this table: 55 people over the age of 75, of whom 12 crashed in the last 3 years. 45 people under the age of 75, of whom 3 crashed in the last 3 years.

Suppose we want to perform a randomization test to see whether or not an individual being over 75 is independent of whether or not they have had a car crash in the last 3 years, where we randomize the car crash response. For each of the following, explain whether or not this outcome is likely, possible, or impossible as a randomization of the original dataset.

(i)

```
# A tibble: 2 x 2
  over75 crash_last3
   <dbl>       <dbl>
1     55           8
2     45           8
```

(ii)

```
# A tibble: 2 x 2
  over75 crash_last3
   <dbl>       <dbl>
1     55          13
2     45           2
```

(iii)

```
# A tibble: 2 x 2
  over75 crash_last3
   <dbl>       <dbl>
1     55           8
2     45           7
```

**Problem 3**

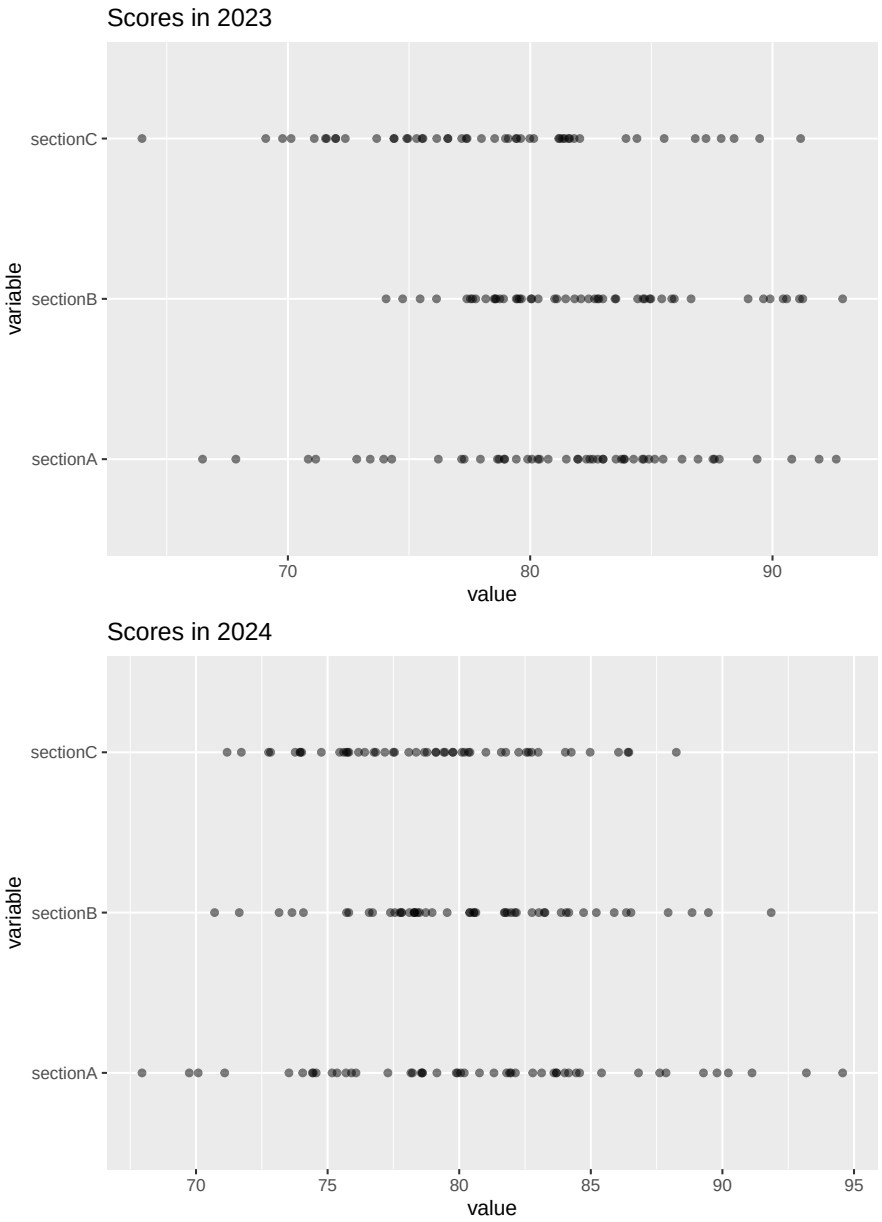
Suppose we are interested in developing a confidence interval for the difference of two population means. We are given two datasets consisting of independent samples, where the responses are approximately normal. Suppose in dataset A, we have 30 samples per group and the sample mean for group 1 is 10 and sample variance is 3, while the sample mean for group 2 is 12 and sample variance is 2. Meanwhile, in dataset B we have 40 samples per group but the sample mean and variance per group is the same as in dataset A: the sample mean for group 1 is 10 and sample variance is 3, while the sample mean for group 2 is 12 and sample variance is 2.

If we were to construct a 90% confidence interval for the difference in population means from group 1 to group 2, which dataset would give a smaller confidence interval? Explain. Does this change if we ask for a 99% confidence interval?

**Problem 4**

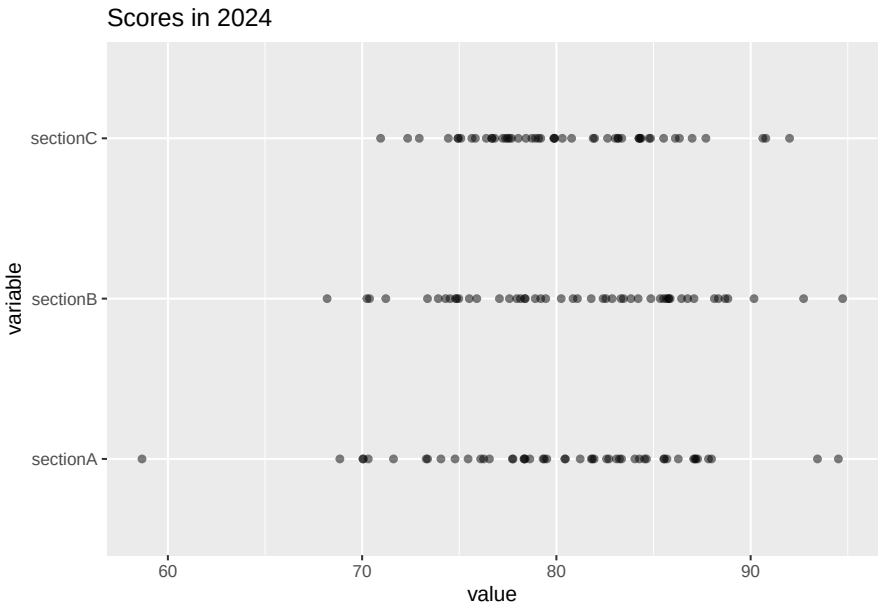
**Part 1** Suppose there is a class with 3 sections, A, B, C, each of which has 500 students, and suppose that we have the results of the class scores of a sample of 50 students in each section in 2023 and 2024. We are interested in formulating a hypothesis test to check whether or not the average score in each section is the same **in the year of 2023 only**. Assuming each sample independent and the scores are approximately normal, which test statistic would we use to evaluate this hypothesis test? Explain.

**Part 2** Suppose there is a class with 3 sections, A, B, C, each of which has 500 students, and suppose that we have the results of the class scores of a sample of 50 students in each section in 2023 and 2024. We are interested in formulating a hypothesis test to check whether or not the average score in each section is the same. Assuming each sample independent and the scores are approximately normal, and given the data visualized below, which year would have a larger value for the test statistic you described in Part 1? What does the larger value of the test statistic imply in the context of the problem?



**Part 3** The code used to make the plot below is as follows.

```
df <- tibble(
  obs = seq(1, 50),
  sectionA = 80 + 5*rnorm(50, sd = 1.1),
  sectionB = 81 + 5*rnorm(50, sd = 1.2),
  sectionC = 80 + 5*rnorm(50, sd = 1)
) %>%
  pivot_longer(cols = sectionA:sectionC) %>%
  rename(variable = name)
ggplot(df, aes(x = value, y = variable)) +
  geom_point(alpha = 0.5) +
  labs(title = "Scores in 2024")
```




Suppose we wanted to make it so that the sections were ordered so that section A appears on top and section C appears on bottom. What changes or additions to the code would we need to do this? Explain (no explicit code necessary)

**Problem 5**

Researchers collected data on heart and body weights of 144 domestic adult cats. The table below shows the output of a linear model predicting heat weight (measured in grams) from body weight (measured in kilograms) of these cats.

| term | estimate | std.error | statistic | p.value |
|---|---|---|---|---|
| `(Intercept)` | -0.3567 | 0.6923 | -0.5152 | 0.6072 |
| `Bwt` | 4.0341 | 0.2503 | 16.1194 | <0.0001 |

a. What are the hypotheses for evaluating whether body weight is positively associated with heart weight in cats?

b. State the conclusion of the hypothesis test from part (a) in context of the data.

c. Calculate a 95% confidence interval for the slope of body weight, and interpret it in context of the data.

d. Do your results from the hypothesis test and the confidence interval agree? Explain.