Statistical Methods in Education Research
Vanderbilt University | Due September 24, 2024
Prof. Sean P. Corcoran | **52 total points**

Problem Set 4

Instructions: Answer the following questions in their entirety in a separate document. Submit your completed problem set as a PDF document via email to `sean.corcoran@vanderbilt.edu`. Use your last name and problem set number as the filename (e.g., *Bieber Problem Set 4.pdf*). Questions 1, 5 and 7 require the use of Stata, but the rest can be solved with or without Stata. Working together is encouraged, but it is expected that all submitted work be that of the individual student.

---

1. (**6 points—3 each**) On the midterm exam in introductory statistics, an instructor always gives a grade of B to students who score between 80 and 90. The scores tend to have a normal distribution with a mean $\mu = 83$ and a standard deviation $\sigma = 5$. About what fraction of the students get a B?

   (a) First, answer this question using what you know about the normal distribution.

   (b) Now use simulated data in Stata. Generate 1,000 student exam scores—this instructor has a big class!—from a normal distribution with the above parameters. Then answer the question based on the data you drew. Are there any differences between your two answers?

2. (**8 points—2 each**) Suppose the SAT math scores of high school seniors follow a normal distribution with mean $\mu = 550$ and standard deviation $\sigma = 100$. Now suppose you plan to take a sample of 25 seniors and calculate the sample mean ($\bar{x}$).

   (a) What is $E(\bar{x})$? What is $Var(\bar{x})$? What is the *standard error* of $\bar{x}$?

   (b) Suppose you would like to increase precision and cut your standard error in half. What sample size would you need to use?

   (c) Based on the original sample size of 25, what is the probability that you draw a random sample with a $\bar{x}$ of 590 or higher?

   (d) Based on the original sample size of 25, what is the probability what you draw a random sample with a $\bar{x}$ between 525 and 575?

3. (**4 points**) A national survey conducted in June 2021 by the University of Southern California as part of its Understanding Coronavirus in America program asked participants whether they had received at least one dose of the coronavirus vaccine. Of 1,626

adults interviewed, 67.58% said *yes*. Find the estimated standard error for the sample proportion ($\hat{\pi}$) reporting they had received at least one dose of the vaccine. Interpret this in words. Hint: use the sample proportion in place of the *population* proportion (π) where required.

4. (**5 points**) Mr. Grumpy and Mr. Happy are both running for Governor. Mr. Grumpy will eventually win the election with 51 percent of the vote. A day before the election, a state-wide newspaper surveys 100 people about their choice for governor. Assume the survey respondents accurately report who they will vote for. What is the probability *Mr. Happy* will be supported by 51 percent or more of the survey respondents?

5. (**18 points**) The Chi-squared (χ^2) distribution is a common probability distribution in statistics defined by one parameter k (the degrees of freedom). If $x \sim \chi^2(k)$, then $E(x) = k$ and $Var(x) = 2k$. The skewness of the distribution is $\sqrt{8/k}$. For large values of k, the distribution is symmetric. For smaller values of k, it is positively skewed. For this problem, let $k = 10$. Create a Stata do file that does the following:

   (a) (**2 points**) Draw 200 random values from this Chi-squared distribution. Note Stata has a random number function for the Chi-squared distribution. (Try `help functions` to view the statistical functions in Stata).

   (b) (**2 points**) Create a histogram for your sample data and find the sample mean $\bar{x}$ and standard deviation s. How do these compare to the (known) population mean and standard deviation? Describe the shape of your distribution: is it symmetric or skewed? What is the skewness statistic?

   (c) (**2 points**) Following one of the methods shown in class (and on the handout on simulations in Stata), conduct a simulation that repeatedly samples 10 observations from the Chi-squared distribution (with $k = 10$ as before), a total of 100 times. Your simulation should store the sample mean $\bar{x}$ on each iteration. When complete, use your data to answer the next questions.

   (d) (**4 points**) Create a histogram for these simulated sample means. What is the mean of these values? What is the standard deviation? How do these compare to what you would have predicted them to be, before you drew any samples? Explain.

   (e) (**2 points**) Based on your simulated sampling distribution, what is the probability of drawing a sample with a $\bar{x}$ greater than 11?

   (f) (**6 points**) Repeat parts (c)-(e) but increase the random sample size to 50. How does this change your results?

6. (**5 points**) Consider this probability distribution for student absences from the last problem set:

# of Days	0	1	2	3	4	5	6	7	8
Probability	0.2	0.14	0.25	0.11	0.1	0.09	0.05	0.03	0.03

   Now imagine you were to draw random samples of 50 students repeatedly from this population. What would the sampling distribution of $\bar{x}$ look like? What would be its mean? Its standard error? How do you know?

7. (**6 points—2 each**) In Stata, use the code below to create a "population" distribution that reflects the probability distribution in part (6):

```
clear all
set obs 20
gen abs = 0
insobs 14     /* insobs inserts additional observations */
replace abs = 1 if abs==.
insobs 25
replace abs = 2 if abs==.
insobs 11
replace abs = 3 if abs==.
insobs 10
replace abs = 4 if abs==.
insobs 9
replace abs = 5 if abs==.
insobs 5
replace abs = 6 if abs==.
insobs 3
replace abs = 7 if abs==.
insobs 3
replace abs = 8 if abs==.
tabulate abs
```

   (a) What is the mean, median, variance, and standard deviation of this population?

   (b) Use the `bootstrap` prefix to draw 1,000 repeated samples of size $n = 30$ from this population, each time retaining the mean, median, and standard deviation. (Hint: you will `bootstrap` the `summarize, detail` command). Save the results in a new dataset.

   (c) Provide histograms for your resulting means, medians, and standard deviations. Across your 1,000 samples, what the *average* mean, median, and standard deviation? For this simulation, what is the *standard error* of the mean, median, and standard deviation? Explain in words what this quantity represents.