
Problem Set 6 *Solutions*

1. **(2 points)** When a recent survey asked whether the government should impose strict laws to make industry do less damage to the environment, a 95% confidence interval for the population proportion responding *yes* was (0.87, 0.90). Would a 99% confidence interval be wider or narrower? Why?

A 99% confidence interval would be **wider**. By definition, a 99% confidence interval contains the true population mean in 99% of random samples, while a 95% confidence interval contains the true population mean in 95% of random samples. For a given sampling distribution of $\hat{\pi}$, greater confidence requires a wider range of possible values.

2. **(4 points)** An education school wants to estimate the mean annual salaries of the school's former students 5 years after graduation. A random sample of 25 such graduates found a sample mean of \$51,288 and a sample standard deviation of \$5,736. Assuming that the population distribution is normal, find a 90% confidence interval for the population mean. Show your work.

Here $n=25$, $\bar{x} = 51288$ and $s = 5736$. Although the population distribution is normal, its standard deviation σ is unknown and estimated using the sample standard deviation s . Thus the confidence interval requires a t statistic. With $df = 24$ and $\alpha/2 = 0.05$ (for a 90% confidence interval), the t value used is 1.711 (`display invttail(24,0.05)`), resulting in the confidence interval:

$$51288 \pm 1.711 \left(\frac{5736}{\sqrt{25}} \right) = (49325, 53251)$$

3. **(6 points)** A graduate school admissions officer has determined that historically, applicants have undergraduate grade point averages that are normally distributed with standard deviation 0.45. From a random sample of 25 applications for the current year, the sample mean grade point average is 3.20.

(a) Find a 95% confidence interval for the population mean.

Here $n=25$, $\bar{x} = 3.20$ and $\sigma = 0.45$. Using the z value of 1.96 associated with a 95% confidence interval, the interval estimate is:

$$3.20 \pm 1.96 \left(\frac{0.45}{\sqrt{25}} \right) = (3.0236, 3.3764)$$

- (b) Based on this sample, a statistician computes a confidence interval for the population mean extending from 3.06 to 3.34. What is the *confidence level* associated with this interval? Show your work or explain how you found your answer.

In this confidence interval the upper bound was calculated as:

$$3.34 = 3.20 + z \left(\frac{0.45}{\sqrt{25}} \right)$$

Solving for z we get 1.556. A z value of 1.556 corresponds to a confidence level of about 88%. $P(z > 1.556) \approx 0.06$ multiplied by 2 is an $\alpha \approx 0.12$.

4. (**3 points**) Find the t value that would be multiplied by the standard error to form a:
- (a) 95% confidence interval with a sample size of 5
 - (b) 95% confidence interval with a sample size of 15
 - (c) 95% confidence interval with a sample size of 25
 - (d) 95% confidence interval with degrees of freedom of 25
 - (e) 99% confidence interval with degrees of freedom of 25
 - (f) 90% confidence interval with a sample size of 500

t -values can be obtained in Stata using `display invttail(df,p)` where df is the degrees of freedom, and p is the right tail probability. (You can also use online lookup tables or the table in the textbook). If the interval has a confidence level $(1 - \alpha)$, then the probability p in each tail is $\alpha/2$. For example, for a 95% confidence interval we find the t -value for which there is an 0.025 probability in the right tail. We can denote this $t_{\alpha/2}$.

- (a) for $n=5$ and $df=4$, $t_{0.025} = \mathbf{2.777}$
- (b) for $n=15$ and $df=14$, $t_{0.025} = \mathbf{2.145}$
- (c) for $n=25$ and $df=24$, $t_{0.025} = \mathbf{2.064}$
- (d) for $df=25$, $t_{0.025} = \mathbf{2.060}$

- (e) for $df=25$, $t_{0.005} = \mathbf{2.787}$. Notice $\alpha/2 = 0.005$ here.
- (f) for $n=500$ and $df=499$, $t_{0.05} = \mathbf{1.648}$. Notice $\alpha/2 = 0.05$ here and that this t value is quite close to the z value used when $p=0.05$. As the sample size gets large, the t distribution gets closer to the standard normal (z).

5. **(6 points)** An estimate is needed of the mean travel time from home to school in a large urban school district. The estimate should be correct to within 2 minutes in 95% of random samples. A previous study of school commuting time suggests that 15 minutes is a reasonable approximation for the standard deviation (σ) of commuting time.

- (a) How large of a sample of families is needed to meet this requirement?

We would like the margin of error for the confidence interval to be 2 minutes:

$$me = 2 = 1.96 * \frac{15}{\sqrt{n}}$$

Solving for n we get a minimum sample size of about 216.

- (b) A random sample is selected of the size you reported in part (a). The sample has a standard deviation of 9 minutes, rather than 15. What is the margin of error for a 95% confidence interval for the mean travel time to school?

Replacing the guessed standard error of 15 with $s = 9$, the margin of error is:

$$me = 1.96 * \frac{9}{\sqrt{216}} = 1.20$$

I am using a z of 1.96 instead of t only under the assumption that my n is large enough that $t \approx z$

6. **(6 points)** A question in the General Social Survey asks whether respondents agree or disagree with the following statement: "It is much better for everyone involved if the man is the achiever outside the home and the woman takes care of the home and family." The sample proportion agreeing was 0.36 in 2004 ($n=883$).

- (a) Show that the estimated standard error for the sample proportion was 0.016.

$$\sqrt{\frac{\hat{\pi}(1 - \hat{\pi})}{n}} = \sqrt{\frac{0.36(1 - 0.36)}{883}} = 0.016$$

- (b) Show that the margin of error for a 95% confidence interval was 0.03.

$$1.96 * 0.016 = 0.030 \text{ (that is, three percentage points)}$$

- (c) Construct the 95% confidence interval and interpret it.

$0.36 \pm 1.96 * 0.016 = \mathbf{(0.329, 0.391)}$. Over repeated samples, a confidence interval calculated in this way will contain the true population mean 95% of the time. Thus, we have 95% confidence that this interval contains μ .

7. **(8 points)** For this question, use the Stata dataset *Grade4_classrooms.dta* on Github. This file represents a random sample of 4th grade classrooms located in urban school districts in Texas. Each observation is a classroom, and the variables either describe the teacher (e.g., total teaching experience, teacher race/ethnicity), or are an average for the classroom (e.g., reading z-scores, % of students who are economically disadvantaged). All data are from 2006.

use https://github.com/spcorcor18/LP0-8800/raw/main/data/Grade4_classrooms.dta, clear

- (a) Inspect the histograms for the variables *totexp* (the teacher's total years of experience), *maplus* (an indicator of whether the teacher holds a master's degree or higher), and *readz_class* (the mean z-score in reading for the students in the class). How would you describe the shape of these distributions? Note: the mean z-score is based on students' position in the statewide distribution of test-takers. **See attached log**
- (b) Provide 90% confidence intervals for each of the variables listed in part (a), and interpret each in words. **See attached log**
- (c) For purposes of constructing confidence intervals, are you concerned about the normality (or lack thereof) in the distributions viewed in part (a)? Explain why or why not. **See attached log**
- (d) Consider your confidence interval for reading z-scores. Is it consistent with the hypothesis that students in urban school districts perform about the same as the statewide average in reading? Explain your reasoning.

By construction the statewide average reading score is zero. (The z -score was calculated such that the mean is 0). The 95% interval estimate is (-0.06, 0.009). Since the confidence interval is a “range of likely values” and it includes zero, it is consistent with the hypothesis that students in urban districts perform about the same, on average, in reading as the state.

```

.
. // *****
. // LP0.8800 Problem Set 6 - Solutions to Question 7
. // Last updated: October 18, 2021
. // *****

. cd "$db\_TEACHING\Statistics I - PhD\Problem sets\Problem set 6"
C:\Users\corcorssp\Dropbox\_TEACHING\Statistics I - PhD\Problem sets\Problem set 6

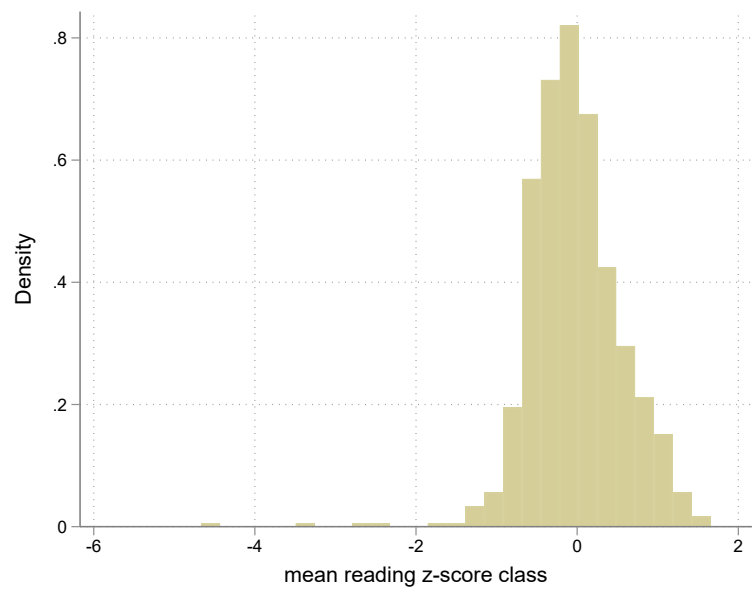
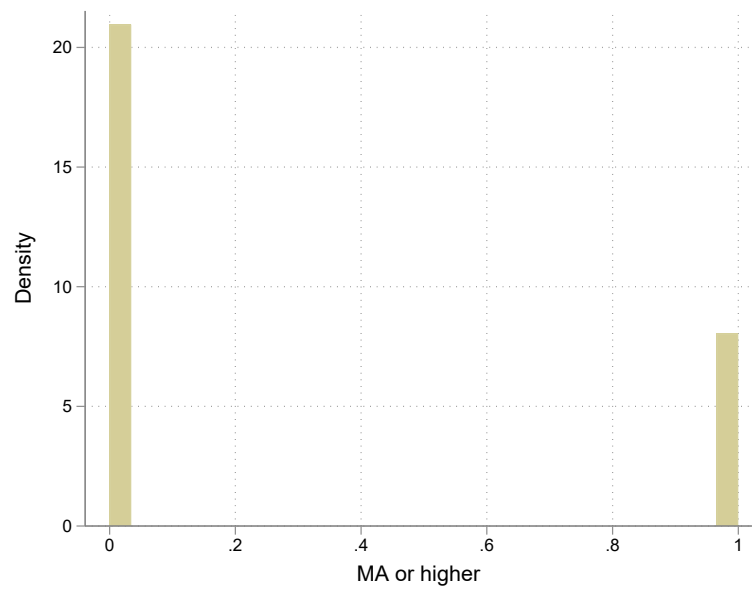
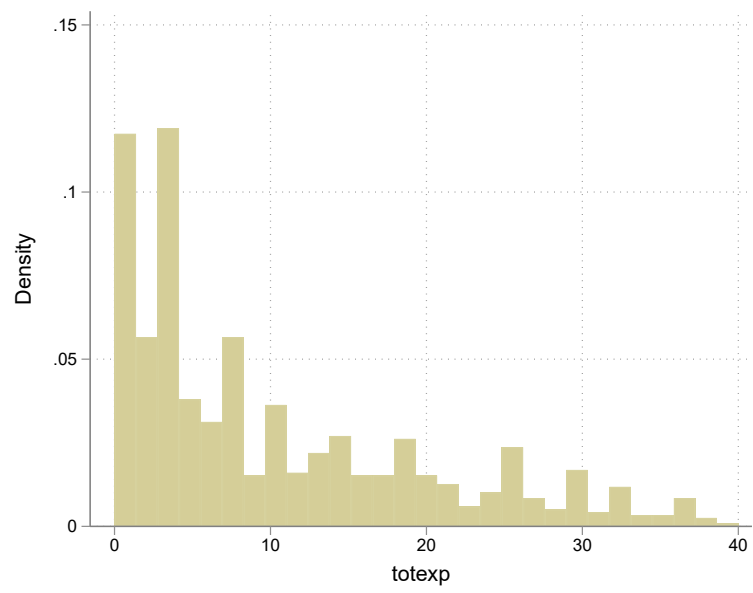
.
. // *****
. // Question 6
. // *****

. use https://github.com/spcorcor18/LP0-8800/raw/main/data/Grade4_classrooms.dta, clear

.
. // *****
. // Part a
. // *****
. histogram totexp, name(hist1, replace) nodraw
(bin=29, start=0, width=1.3793103)
. histogram maplus, name(hist2, replace) nodraw
(bin=29, start=0, width=.03448276)
. histogram readz_class, name(hist3, replace) nodraw
(bin=27, start=-4.6651998, width=.23412534)

.
. graph combine hist1 hist2 hist3, xsize(3) ysize(7) col(1)
. graph export hists.pdf, as(pdf) replace
(file hists.pdf written in PDF format)

```



```
.
. // The distribution of teacher experience is right skewed, with a somewhat
. // large mass of inexperienced teachers and long right tail of more experienced
. // teachers. MAPplus is a binary variable, so its distribution is concentrated
. // on the values of 0 and 1. Classroom average reading scores are more
. // symmetrical, albeit a bit left skewed.
```

```
.
. // *****
. // Parts b-c
. // *****
. // 90% confidence intervals
. mean totexp, level(90)
```

```
Mean estimation                Number of obs   =           858
```

	Mean	Std. Err.	[90% Conf. Interval]	
totexp	10.169	.3315576	9.623044	10.71495

```
. mean maplus, level(90)
```

```
Mean estimation                Number of obs   =           858
```

	Mean	Std. Err.	[90% Conf. Interval]	
maplus	.2773893	.0152935	.2522065	.3025721

```
. mean readz_class, level(90)
```

```
Mean estimation                Number of obs   =           765
```

	Mean	Std. Err.	[90% Conf. Interval]	
readz_class	-.0255273	.0208613	-.0598827	.0088282

```
.
. // The confidence intervals are estimated above. If we can assume the sampling
. // distribution of the sample mean is normal, the 90% confidence interval will
. // contain the population mean in 90% of random samples. Per the CLT, the
. // normality assumption is reasonable if the sample size is large enough. In
. // this case with a sample size of 858 we can be reasonably confident that the
. // sampling distributions are normal, even in cases where the original
. // distribution is very non-normal (like maplus).
```

```
.
.
. capture log close
```