

Midterm 1

QTM 285-1

About the Midterm

The exam is open-book and untimed, and is designed to take approximately 75 minutes. You'll have one week to complete and submit it. The longer window is intended to give you flexibility in scheduling, not because the exam is expected to take a week.

You may consult the course book, your notes, and other course materials, and me, but you must complete the exam independently. No collaboration with other students or AI assistants, and no discussion of exam questions or answers until the submission deadline has passed.

Every plot you see below is faithful to the described scenario. You may be able to use them to check your work.

Problem 1

You have a sample of 100 observations drawn with replacement from a population and use their mean to estimate the population mean. Your coworker makes a new dataset containing four copies of each observation, for 400 entries in all, and uses the mean of those 400 entries.

Your coworker claims that the new mean has half the standard error of the original mean. Is your coworker right? Explain.

Problem 2

Turnout in a county's midterm elections has usually been about 25% of registered voters. This year's election is getting much more attention than usual, and nobody knows what turnout will be.

You and a coworker are designing a survey for a news organization to estimate the proportion θ of the county's registered voters who will turn out. You will sample with replacement and report an interval of the form $\hat{\theta} \pm .01$, where $\hat{\theta}$ is the sample proportion. You want this interval estimator to have at least roughly 95% coverage, whatever this year's turnout turns out to be.

Relying on the county's usual 25% turnout, your coworker proposes sampling 7,500 voters.

Part A

Does your coworker's choice of sample size guarantee roughly 95% coverage for the proposed interval? Explain your answer. If not, what sample size would you recommend?

For Parts B and C, suppose you conduct the survey using a sample size that gives your interval roughly 95% coverage.

Part B

After the election, the actual turnout falls outside your interval. An editor at the news organization that hired you to run the poll calls you angrily: “Everyone I’ve talked to says your intervals cover. How could you get it wrong this time?”

Write a short reply to the editor, who has not studied statistics.

Part C

The editor hasn't told you the actual turnout yet. You're trying to figure out how far off your estimate could be. Would you be surprised if it were off by more than 1.5 percentage points?

Explain your answer. Use the reference picture below, and assume the normal approximation is exact. A rough explanation is enough; you don't need to calculate a precise probability.



Problem 3

A population of $m = 200$ people is arranged in a circle, with 100 zeros followed by 100 ones as you go around it. We'll compare three ways of drawing a sample of $n = 100$ people from it. In each case, we estimate θ by the proportion of 1s in the sample, $\hat{\theta} = \frac{1}{100} \sum_{i=1}^{100} Y_i$, and report the interval estimate $\hat{\theta} \pm 1.96\sqrt{\hat{\theta}(1-\hat{\theta})/n}$.

1. Draw 100 people uniformly at random *with* replacement.
2. Draw 100 people uniformly at random *without* replacement.
3. Choose one starting person uniformly at random, and take that person and the next 99 people around the circle.

One of these intervals has coverage of roughly 95%, one has noticeably more, and one has noticeably less. Which is which? Explain. You may want to try to include a sketch or two.

Problem 4

A coworker has a sample of $n = 100$ household incomes $Y_1 \dots Y_{100}$, drawn with replacement from a population, and wants a 95% interval estimate for the population's mean income. They bootstrap: they draw 10,000 samples of size 100 with replacement *from their sample*, take the mean $\bar{Y}^*$ of each, and report $\bar{Y} \pm 1.96 \text{ se}$, where se is the standard deviation of these bootstrap means.

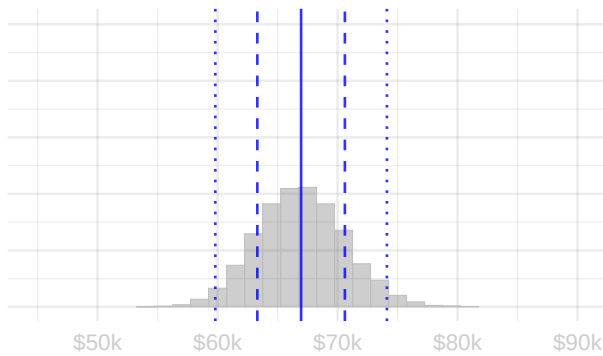
The plot labelled Original shows the distribution of the bootstrap means in this procedure.

Match each scenario to the plot of the resulting sampling distribution below. Use each of A, B, and C once. Explain.

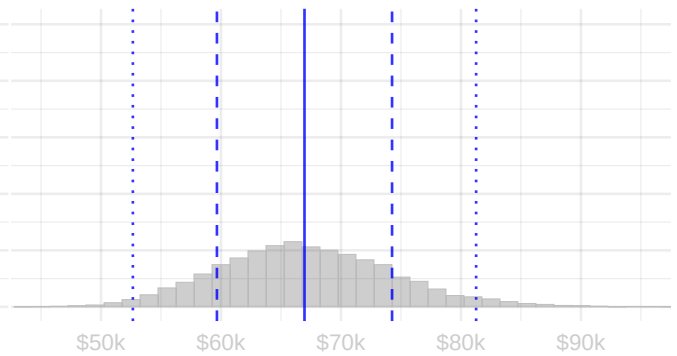
1. Your coworker proposes increasing the number of bootstrap resamples from 10,000 to 40,000, using the same 100 incomes and keeping each resample's size at 100.
2. Instead, suppose your coworker collects a new sample of 400 household incomes, drawn with replacement from the same population, and uses the mean of those 400 observations as the estimate.
3. Your coworker uses the same 100 incomes and 10,000 resamples, but draws only 25 observations in each resample instead of 100.

As usual, the solid, dashed, and dotted lines show the mean, one standard error to either side, and roughly two standard errors to either side, respectively.

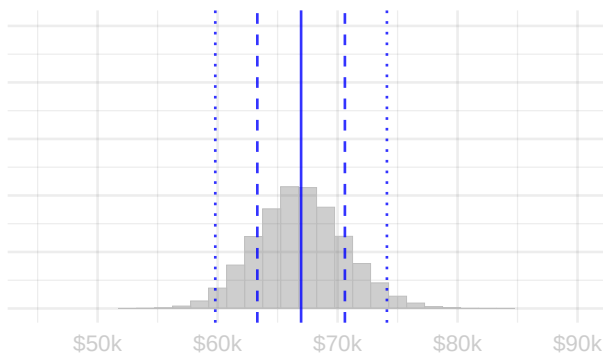
Original



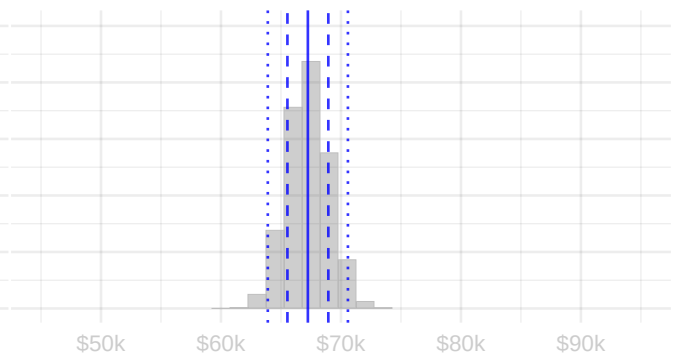
A



B



C



Problem 5

You have a sample $Y_1 \dots Y_n$ drawn with replacement from a population with mean μ and standard deviation σ . Your coworker proposes estimating μ by dividing the sum of the observations by $n + 100$ instead of n :

$$\bar{Y}_{100} = \frac{1}{n + 100} \sum_{i=1}^n Y_i.$$

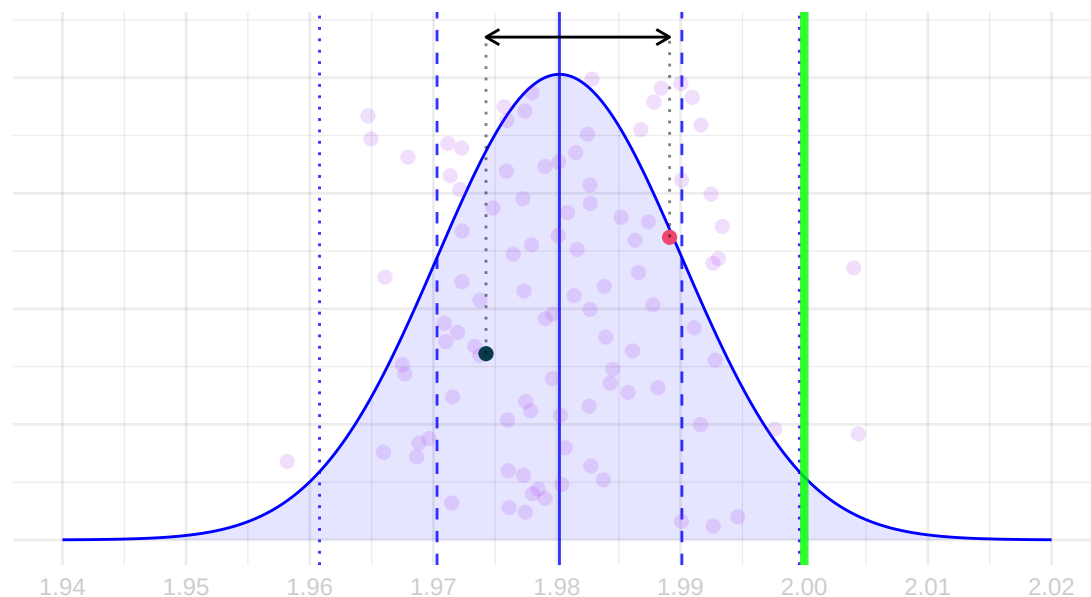
Your coworker has calculated formulas for the bias and standard deviation of the ‘divide by $n+k$ estimator’

$$\bar{Y}_k = \frac{1}{n+k} \sum_{i=1}^n Y_i.$$

$$\begin{aligned} E[\bar{Y}_k] - \mu &= -\frac{k}{n+k} \mu \\ \text{sd}[\bar{Y}_k] &= \frac{\sigma}{\sqrt{n}} \times \frac{n}{n+k}. \end{aligned}$$

Suppose $n = 10,000$, $\mu = 2$ and $\sigma = 1$.

The plot below shows an approximation to the sampling distribution of $\bar{Y}_{100}$. The green line marks the population mean μ . The solid blue line marks the estimator’s expected value, the dashed blue lines are one standard error to either side, and the dotted blue lines are 1.96 standard errors to either side.



Part A

1. Calculate the bias and standard error of $\bar{Y}_{100}$.
2. Your coworker argues that with so many observations, adding 100 to the denominator should make little difference, and proposes the interval $\bar{Y}_{100} \pm 1.96 \text{ sd}(\bar{Y}_{100})$.

Is your coworker's interval calibrated to have roughly 95% coverage? Give an approximate coverage probability and explain your answer.

Part B

Apparently you're running this season's popular survey design. 1,000 other pollsters independently ran a survey exactly like yours and reported their estimates of the population mean. What is the typical distance between two of these estimates?

Give a numerical answer for the root-mean-squared distance. Explain your answer.