

Practice Session 8

Introduction:

Distribution of a Sample Proportion:

When selecting random samples of size n from a population with proportion π , the distribution of the sample proportions is centered at the population proportion π , and has a standard error given by:

$$SE = \sqrt{\frac{\pi(1 - \pi)}{n}}$$

and is reasonably normally distributed if $n\pi \geq 10$ and $n(1 - \pi) \geq 10$.

Confidence Interval for a Proportion:

If z^* is a standard normal endpoint to give the desired level of confidence, and if the sample size is large enough so that $n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$, the confidence interval for a population proportion π is

$$\text{Sample statistic} \pm z^* \cdot SE$$

Where sample proportion based on a random sample of size n and the standard error are

$$\text{Sample statistic} = \hat{p}$$

$$SE = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$$

Hypothesis Test for a Proportion:

If the sample size is reasonably large so that $n\pi_0 \geq 10$, then we can test $H_0 : \pi = \pi_0$ vs $H_a : \pi \neq \pi_0$ (or a one-tail alternative), and the standardized test statistic is:

$$z = \frac{\text{Statistic} - \text{Null Value}}{SE} = \frac{\hat{p} - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}}$$

Practice with `rnorm`, `pnorm`, `qnorm`, and `mosaic::cnorm`

- a.) Using the `rnorm`, generate 100 random observations from a normal distribution. What are the default mean and standard deviation values? Plot a histogram of the data.
- b.) Now, change the mean and standard deviation to $\mu = 10$ and $\sigma = 5$. Plot a histogram of the data and compare.
- c.) Using the `pnorm` function with $\mu = 10$ and $\sigma = 5$, calculate the probability of observing a value greater than 13.
- d.) Using the `pnorm` function with $\mu = 10$ and $\sigma = 5$, calculate the probability of observing a value less than 7.
- e.) Using the `pnorm` function with $\mu = 10$ and $\sigma = 5$, calculate the probability of observing a value less 15 but greater than 11.
- f.) Using the `qnorm` function with $\mu = 10$ and $\sigma = 5$, find the population value that corresponds to the 90th percentile.
- g.) Using the `qnorm` function with $\mu = 10$ and $\sigma = 5$, find the population value that corresponds to the 20th percentile.
- h.) Using the `qnorm` function with $\mu = 10$ and $\sigma = 5$, find the population values corresponding to a two-sided interval that contains 95% of the population.
- i.) Using the `mosaic::cnorm` function with $\mu = 10$ and $\sigma = 5$, find the population values corresponding to a two-sided 95% interval.

Z-Scores and the Standard Normal Distribution

- a.) Generate 1000 random observations from a normal distribution with mean 10 and standard deviation 2. Save the result to the variable `norm_data`
- b.) Create a histogram of `norm_data`.
- c.) Convert `norm_data` into a vector of Z-scores using the following formula. Call this vector `z_data`.

- d.) Compute the mean and standard deviation of `z_data`. How do they compare to the parameter values you defined in part (a)?
- e.) Create a histogram of `z_data`. How does its center and spread compare to the histogram you created in part (b)?

Introduction to the Central Limit Theorem

- a.) Using the `rexp()` function, generate 1000 observations from an exponential distribution with `rate = 1` (the default). Save this vector as the variable `exp_data`.
- b.) Create a histogram of `exp_data`. How would you describe its shape?
- c.) Using the “do it” function, generate 10000 samples of size $n = 10$ from an exponential distribution with `rate = 1`. Compute the mean at each iteration. Save the output of “do it” to the variable `sampling_dist_exp_10`.
- d.) Create a histogram of `sampling_dist_exp_10`. How would you describe its shape?
- e.) Repeat part (c), except now increase the size of each sample to $n = 30$. Save the output of “do it” to the variable `sampling_dist_exp_30`.
- f.) Create a histogram of `sampling_dist_exp_30`. How does it compare to the histogram in part (b)? How does it compare to the histogram in part (d)?
- g.) Finally, repeat the sampling process for $n = 100$. Comment on the shape of the histogram.

Confidence Interval

Survival of ICU Patients in the dataset `ICUAdmissions` includes information on 200 patients admitted to an Intensive Care Unit. One of the variables, `Status`, indicates whether each patient lived (indicated with a 0) or died (indicated with a 1).

Construct and interpret a 95% confidence interval for the proportion of ICU patients who live. You can find the `ICUAdmissions` data in the `Lock5Data` library.

```
#your code here
```

Is B a Good Choice on a Multiple-Choice Exam?

Multiple-choice questions on Advanced Placement exams have five options: A, B, C, D, and E. A random sample of the correct choice on 400 multiple-choice questions on a variety of AP exams shows that B was the most common correct choice, with 90 of the 400 questions having B as the answer.

Does this provide evidence that B is more likely to be the correct choice than would be expected if all five options were equally likely? Show all details of the test. The data are available in `APMultipleChoice` from the `Lock5Data` library.

```
#your code here
```

Incentives for Quitting Smoking: Do They Work?

With no incentives, the proportion of smokers trying to quit who are still abstaining six months later is about 0.06. Participants in the study were randomly assigned to one of four different incentives, and the proportion successful was measured six months later. Of the 498 participants in the group with the least success, 47 were still abstaining from smoking six months later. We wish to test to see if this provides evidence that even the smallest incentive works better than the proportion of 0.06 with no incentive at all.

- 1.) State the null and alternative hypotheses, and give the notation and value of the sample statistic.
- 2.) Use a randomization distribution and the observed sample statistic to find the p-value.

```
#your code here
```

- 3.) Give the mean and standard error of the normal distribution that most closely matches the randomization distribution, and then use this normal distribution with the observed sample statistic to find the p-value.

```
#your code here
```

- 4.) Use the standard error found from the randomization distribution in part (2) to find the standardized test statistic, and then use that test statistic to find the p-value using a standard normal distribution.

```
#your code here
```

5.) Compare the p-values from parts (2), (3), and (4). Use any of these p-values to give the conclusion of the test.

```
#your code here
```