Lab 7: Confidence Intervals

For this lab, it will be helpful to have a copy of the knitted version of this document to answer the questions as much of it is written using mathematical notation that may be difficult to read when the document is not knitted. For your convenience, a pdf of this document is in the lab folder on blackboard.

Lab Goals

The purpose of this lab is to practice constructing 100(1 - α)% confidence intervals for the population mean, μ, given a random sample, $X_i, i = 1, \ldots, n$, of observations from the population of interest.

Constructing Confidence Intervals Using the Standard Normal Distribution

Assume you have obtained a SRS of size n from a population with mean, μ, and variance, σ^2. Using this sample, the sample mean and sample variance can be computed. Let

1) $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$ = the sample mean
2) $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$ = sample variance

Under the following conditions, quantiles from the standard normal distribution can be used to determine confidence intervals for μ:

1) σ is known and
    a) the sample is iid from $N(\mu, \sigma)$
    b) the sample is iid from any distribution, **and** n is "large enough" for the CLT to apply (usually when $n \geq 30$)
2) σ is unknown, the sample is iid from any distribution, **and** n is "large enough" for the CLT to apply (usually when $n \geq 30$, assuming distribution is not strongly skewed)

Note: Here "iid" stands for independent and identically distributed. (While technically a SRS refers to sampling without replacement, meaning there's dependence between observations, as long as n is much smaller than the population size, we can usually neglect this dependence.)

When the quantiles from the standard normal distribution can be used to determine a 100(1 - α)% confidence interval for μ, the confidence interval takes the following form:

($\bar{x}_n$ - $SE(\bar{X}_n) \times z_{\alpha/2}$, $\bar{x}_n$ + $SE(\bar{X}_n) \times z_{\alpha/2}$)

where

1) $\bar{x}_n$ is the sample mean
2) $SE(\bar{X}_n)$ is either the estimated or true standard error of $\bar{X}_n$

`a) if $\sigma$ is known, $SE(\bar X_n) = \sigma/\sqrt{n}$`

`b) if $\sigma$ is unknown, we replace $SE(\bar X_n)$ with an estimate of it $s/\sqrt{n}$`

3) $z_{\alpha/2}$ is the $100(1 - \alpha/2)$ percentile of the standard normal distribution. In R, $z_{\alpha/2}$ = `-qnorm(α/2)`.

Problem 1

Fuel economy information for cars is determined by the EPA (Environmental Protection Agency) by testing a sample of cars. A random sample of 12 Toyota Prius cars is selected to determine a 95% confidence interval for the mean fuel economy of all Toyota Prius cars. Fuel economy is known to be normally distributed with standard deviation $\sigma = 5$. The mean fuel economy of the 12 sampled cars was 48.3 mpg.

a) Can we use quantiles from the standard normal distribution to determine a 95% confidence interval for the mean fuel economy of Toyota Prius cars? If so, why?

Yes, we can use quantiles from the N(0,1) distribution because the population we are sampling from is normally distributed and the population variance is known.

b) What is the quantile needed for this confidence interval? Compute it using the `qnorm()` function in R.

The quantile is $z_{0.5/2}$.

```
-qnorm(.025)
```

```
## [1] 1.959964
```

c) What is $SE(\bar{X}_n)$ for this study?

$SE(\bar{X}_{12}) = 5/\sqrt{12}$.

d) Determine a 95% confidence interval for the mean fuel economy of Toyota Prius cars. Interpret this confidence interval in the context of the study.

($\bar{x}_n$ *-* $SE(\bar{X}_n) \times z_{\alpha/2}$, $\bar{x}_n$ *+* $SE(\bar{X}_n) \times z_{\alpha/2}$*) = (48.3 -* $5/\sqrt{12} \times z_{0.5/2}$, *48.3 +* $5/\sqrt{12} \times z_{0.5/2}$*)*

```
smean = 48.3
se = 5/sqrt(12)
quant = -qnorm(.025)

Lower = smean - se*quant
Upper = smean + se*quant
Lower
```

```
## [1] 45.47104
```

```
Upper
```

```
## [1] 51.12896
```

We are 95% confident that the mean fuel economy for Toyota Prius cars is between 45.47 mpg and 51.13 mpg.

Problem 2

Obesity is one of the most serious health risks of the 21st century. It has become the leading preventable cause of death worldwide. Obesity is defined as having a BMI (Body Mass Index) over 29.9. Researchers would like to determine a 99% confidence interval for the mean BMI of adults living in the US. A random sample of 450,000 adults from this population is taken and the BMI for each adult is recorded. The sample variance of BMI was 36. The sample mean of BMI was 27.7.

a) Can we use quantiles from the standard normal distribution to determine a 99% confidence interval for the mean BMI of adults living in the US? If so, why?

Yes, quantiles from the standard normal distribution can be used because the sample size is large.

b) What is the quantile needed for this confidence interval? Compute it using the `qnorm()` function in R.

The quantile is $z_{.005}$.

```
-qnorm(.005)
```

```
## [1] 2.575829
```

c) What is $SE(\bar{X}_n)$ for this study?

$SE(\bar{X}_{450000}) = 6/\sqrt{450000}$.

d) Determine a 99% confidence interval for the mean BMI of adults from the US. Interpret this confidence interval in the context of the study.

($\bar{x}_n$ - $SE(\bar{X}_n) \times z_{\alpha/2}$, $\bar{x}_n$ + $SE(\bar{X}_n) \times z_{\alpha/2}$) = (27.7 - $6/\sqrt{450000} \times z_{.01/2}$, 27.7 + $6/\sqrt{450000} \times z_{.01/2}$)

```
smean = 27.7
se = 6/sqrt(450000)
quant = -qnorm(.005)

Lower = smean - se*quant
Upper = smean + se*quant
Lower
```

```
## [1] 27.67696
```

```
Upper
```

```
## [1] 27.72304
```

We are 99% confident that the mean BMI for US adults lies in the interval (27.68,27.72).

Special Consideration for the Bernoulli Distribution

Suppose a random sample, $X_i, i = 1, \ldots, n$, is taken from a Bernoulli(p) distribution. Additionally, assume p, the mean of the Bernoulli distribution, is unknown. If n is large enough, by the CLT

$\bar{X}_n = \hat{p}_n \approx N(p, \sqrt{p(1-p)/n})$

What this means for us is that we can use quantiles from the standard normal distribution to construct confidence intervals for p, the unknown population mean, *if n is large enough.*

In this particular case, some guidelines are available to determine whether the CLT applies. If both

1) $np \geq 10$ and
2) $n(1-p) \geq 10$

then $\hat{p}_n \approx N(p, \sqrt{p(1-p)/n})$.

Fantastic! However, we don't know p. So, instead of using p to check the above conditions, we'll instead use $\hat{p}_n$ resulting in the following two criterion.

1) $n\hat{p}_n \geq 10$ and
2) $n(1-\hat{p}_n) \geq 10$

If (1) and (2) hold, quantiles from the standard normal distribution can be used to construct 100(1-α)% confidence intervals for p.

Using $\hat{p}_n$ to denote the sample mean of an iid Bernoulli(p) sample of size n, a 100(1-α)% confidence interval for p can be expressed as

($\hat{p}_n$ - $SE(\hat{p}_n) \times z_{\alpha/2}$, $\hat{p}_n$ + $SE(\hat{p}_n) \times z_{\alpha/2}$)

where

1) $\hat{p}_n$ is the sample proportion

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

3) $z_{\alpha/2}$ is the $100(1-\alpha/2)$ percentile of the standard normal distribution.

Problem 3

Quebec is a large province in Eastern Canada, and is the only province in Canada with a primarily French-speaking population. Historically there has been a debate over whether Quebec should secede from Canada. In a recent study of 800 Quebecois (residents of Quebec), 28% thought Quebec should secede from Canada.

a) Can we use quantiles from the standard normal distribution to determine a 95% confidence interval for the true proportion of Quebecois that want to secede from Canada? If so, why?

Yes. Checking the conditions for the CLT to hold for $\hat{p}_{800}$, we find 800 × .28 = 224 and 800 × (1-.28) = 576. Both are greater than 10 as required.

```
800*.28
```

```
## [1] 224
```

```
800*(1-.28)
```

```
## [1] 576
```

b) What is $SE(\hat{p}_n)$ for this study?

$SE(\hat{p}_n) = \sqrt{\frac{.28(1-.28)}{800}}$

c) Determine a 95% confidence interval for the true proportion of Quebecois that want to secede from Canada. Interpret this confidence interval in the context of the study.

$(\hat{p}_n - SE(\hat{p}_n) \times z_{\alpha/2}, \hat{p}_n + SE(\hat{p}_n) \times z_{\alpha/2}) = (.28 - \sqrt{\frac{.28(1-.28)}{800}} \times z_{.05/2}, .28 + \sqrt{\frac{.28(1-.28)}{800}} \times z_{.05/2})$

```
smean=.28
se=sqrt((.28*(1-.28))/800)
quant = -qnorm(.025)
Lower = smean - se*quant
Upper = smean + se*quant
Lower
```

```
## [1] 0.2488865
```

```
Upper
```

```
## [1] 0.3111135
```

We are 95% confident that the true proportion of Quebecois who want to secede from Canada is in the interval (0.25,0.31).

The t-distribution

In all of the guidelines given above for using quantiles from the N(0,1) distribution to construct confidence intervals for μ, at least one of the following two conditions was satisfied:

1) A sample of n iid observations from $N(\mu, \sigma)$ was obtained **and** σ was known

or

2) The sample size, n, was large enough to apply the CLT to determine the distribution of $\bar{X}_n$

How do we determine the appropriate quantiles to use to construct a confidence interval for μ when n is too small for the CLT to apply and σ is unknown?

Answer: If the sample is iid from a population that is approximately $N(\mu, \sigma)$, quantiles of a t-distribution with $n-1$ degrees of freedom can be used to construct confidence intervals for μ.

Define t_{df} as a t distribution with df degrees of freedom. Here are some properties of the distribution of t_{df}:

1) It is a continuous distribution characterized by a symmetric probability density function centered at 0.
2) For every value of df, a different t-distribution is defined.
3) As df increases, the distribution of t_{df} becomes more and more like a N(0,1) distribution.

If a t-distribution with $n-1$ degrees of freedom is appropriate to determine the quantiles for a 100(1-α)% confidence interval, the confidence interval takes the following form:

($\bar{x}_n$ - $SE(\bar{X}_n) \times t_{n-1,\alpha/2}$, $\bar{x}_n + SE(\bar{X}_n) \times t_{n-1,\alpha/2}$)

where

1) $\bar{x}_n$ is the sample mean
2) $SE(\bar{X}_n)$ is the estimated as $s/\sqrt{n}$
3) $t_{n-1,\alpha/2}$ is the $100(1-\alpha/2)$ percentile of a t-distribution with $n-1$ degrees of freedom. In R, $t_{n-1,\alpha/2}$ = `-qt(`α`/2,n-1)`.

Problem 4

Assume for the study described in Problem 1 that the sample standard deviation of the 12 sampled Toyota Prius cars is 5 mpg and the population standard deviation is unknown.

a) What distribution should be used to compute a confidence interval for the mean fuel econonmy of Toyota Prius cars under the new assumptions? Why?

A t_{11} distribution can be used to determine quantiles for the confidence interval since the population distribution of fuel economy for Toyota Prius cars is normal, the standard error is estimated, and the sample size of 12 is small.

b) Recalculate the 95% confidence interval for the mean fuel economy of Toyota Prius cars under the new assumptions.

($\bar{x}_n$ - $SE(\bar{X}_n) \times t_{n-1,\alpha/2}$, $\bar{x}_n + SE(\bar{X}_n) \times t_{n-1,\alpha/2}$) = (48.3 - $5/\sqrt{12} \times t_{11,.05/2}$, 48.3 + $5/\sqrt{12} \times t_{11,.05/2}$)

```
smean=48.3
se=5/sqrt(12)
quant = -qt(.025,11)
Lower = smean - se*quant
Upper = smean + se*quant
Lower
```

```
## [1] 45.12315
```

```
Upper
```

```
## [1] 51.47685
```

We are 95% confident that the mean fuel economy of Toyota Prius cars is between 45.12 mpg and 51.48 mpg.

c) How does this confidence interval compare to the one computed in 1(d)?

This interval is wider than the one calculated in 1(d) due to the extra variability associated with using the sample variance in place of the true population variance.