Lecture 6 calculations

Lecture 6

1 Confidence Intervals

1.1 Confidence Interval Form

The confidence intervals in this chapter will follow the form:

$$estimate \pm reliability\ coefficient * standard\ error$$

1.2 Confidence Intervals for Population Means

We can estimate the population mean by using the mean of a representative sample from the population.

> **Warning**
>
> The method differs for cases with known and unknown population variance. Some approximations can be done when there is a large sample size.

1.3 CI for Mean: Known Population Variance

Cases in which population variances are known are rare. The best example of measurements with known population variances are standardized scales.

> **Note**
>
> The Wechsler Adult Intelligence Scale (WAIS) maintains a standard deviation of 15 points for its full scale IQ scores.

For a known population variance σ^2, sample mean $\bar{x}$, and sample size n, the 95% confidence limits for the $(1-\alpha)$ can be calculated using the following formula:

$$\bar{x} \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

1.4 CI for Mean: Notes

Confidence intervals are often written in parentheses notation as such:

$$\left(\bar{x} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$$

$z_{1-\alpha/2}$ is referred to as the reliability coefficient and defined as the following:

$$P(Z \leq z_{1-\alpha/2}) = 1 - \alpha/2$$

> **Tip**
>
> Recall that Z follows the standard normal distribution. $z_{1-\alpha/2}$ can be calculated using `qnorm(1-alpha/2,mean=0,sd=1)`. For a 95% confidence interval, $\alpha = 0.05$, then it follows that $1-\alpha/2 = 1-0.05/2 = 0.975$. Therefore, the reliability coefficient for a 95% confidence interval can be calculated as:

```
alpha=0.05
qnorm(1-alpha/2,mean=0,sd=1)
```

```
[1] 1.959964
```

1.5 Example

The file `lec6example.csv` contains simulated values from a normal distribution with a variance of 4.

1.5.1 Question

Create a 95% confidence interval for the population mean. Is 1 a plausible value for the population mean?

1.5.2 Answer

```
# setwd("where your file is")
example <- read.csv("datasets/lec6example.csv")
sdpop <- sqrt(4)
```

We need to calculate the sample mean of the `Sims` column, the sample size, and the reliability coefficient.

```
xbar <- mean(example$Sims)
sampsize <- nrow(example)                                    ①
alpha <- 1-0.95
relcoeff <- qnorm(1-alpha/2,mean = 0, sd=1)                  ②
```

① `nrow()` calculates the number of rows of the data frame. If dealing with a singular vector, you can use `length()`. `nrow(example)` will yield the same answer as `length(example$Sims)`.

We can now calculate the 95% confidence interval limits.

```
lower <- xbar - relcoeff*sdpop/sqrt(sampsize)
upper <- xbar + relcoeff*sdpop/sqrt(sampsize)
lower
```

```
[1] 0.6637448
```

```
upper
```

```
[1] 1.303866
```

The resulting 95% confidence interval is (0.66, 1.3). Since 1 is in the confidence interval, then we can claim that 1 is a plausible value for the population mean.

1.6 Exercise

Consider the sleep health data uploaded on Canvas as `SleepHealthData.csv`. Suppose that in the population that this sample represents, the variance of the sleep duration is 0.5.

1.6.1 Question

Calculate the 95% confidence interval for the average sleep duration (`sleep_duration`) in the population this sample represents. Is 7 hours a plausible value for the average sleep distribution of the population?

1.6.2 Answer

```
sleep <- read.csv("SleepHealthData.csv")
sdpop <- sqrt(0.5)
xbar <- mean(sleep$sleep_duration)
sampsize <- nrow(sleep)
alpha <- 1-0.95
relcoeff <- qnorm(1-alpha/2)
# not specifying the mean and sd in qnorm assumes standard normal

lower <- xbar - relcoeff*sdpop/sqrt(sampsize)
lower
```

```
[1] 7.060422
```

```
upper <- xbar + relcoeff*sdpop/sqrt(sampsize)
upper
```

```
[1] 7.203749
```

The 95% confidence interval is (7.06,7.2). 7 is not a plausible value for the average sleep distribution of the population.

1.7 Precision

The precision of an interval estimate is related to its width.

> **Note**
>
> The precision, also known as the **margin of error**, can be expressed as the product of the reliability coefficient and the standard error.

For the case of the population mean with known population variance,

$$precision = z_{1-\alpha/2} * \frac{\sigma}{\sqrt{n}} = \frac{upper - lower}{2}$$

> **Note**
>
> Because of the symmetry of the confidence interval, the precision is half the width of the confidence interval.

1.8 CI for Mean: Unknown Population Variance

It is more common to not know the population variance when estimating the population mean. Because of this, we are inclined to use the next best thing: an estimate of the population variance from the collected sample. This estimate is the **sample variance**

> **Warning**
>
> Using an approximate value for the population variance implies that our standardized statistic **might not** follow the standard normal distribution. We need a new distribution to calculate reliability coefficients for these confidence intervals.

1.9 The Student's t-distribution

The t distribution is symmetric and bell-shaped like the normal distribution, but has heavier tails.

> **Note**
>
> Heavier tails mean that the distribution is more likely to produce values that fall farther from the mean compared to the normal distribution. The heavier tails account for the extra uncertainty introduced by using the sample variance to estimate the population variance.

> **Important**
>
> The t-distribution is defined by the degrees of freedom denoted by ν or `df`, and specific t-distributions can be written as t_ν.

1.10 The t distribution: `R`

The `R` syntax for the t-distribution with degrees of freedom `df` consists of the following functions:

- PDF: `dt(x,df)`
- CDF: `pt(x,df)`
- Quantile: `qt(x,df)`
- Generate/Simulate `n` t-distribution points: `rt(n,df)`

1.11 CI for Mean: Unknown Population Variance

When the population variance is unknown OR the sample from a normally distributed population has a low sample size, the $(1-\alpha)*100$ confidence interval can be calculated using the following formula:

$$\bar{x} \pm t_{\nu,1-\alpha/2}\frac{s}{\sqrt{n}}$$

::: callout-important

s is the sample standard deviation $t_{\nu,1-\alpha/2}$ is defined as $P(t < t_{\nu,1-\alpha/2}) = 1-\alpha/2$. The degrees of freedom ν can be calculated as $\nu = n-1$.

:::

The 95% confidence interval can be written as:

$$\left(\bar{x} - t_{\nu,1-\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + t_{\nu,1-\alpha/2}\frac{s}{\sqrt{n}}\right)$$

Precision

The precision of an interval estimate is related to its width.

> **Note**
>
> The precision, also known as the **margin of error**, can be expressed as the product of the reliability coefficient and the standard error.

For the case of the population mean with unknown population variance,

$$precision = t_{\nu,1-\alpha/2}\frac{s}{\sqrt{n}} = \frac{upper - lower}{2}$$

> **Note**
>
> Because of the symmetry of the confidence interval, the precision is half the width of the confidence interval.

1.12 Example

The file `lec6example.csv` contains simulated values from a normal distribution.

1.12.1 Question

Supposed the population variance is unknown. Create a 95% confidence interval for the population mean. Is 1 a plausible value for the population mean?

1.12.2 Answer

```
# setwd("where your file is")
example <- read.csv("datasets/lec6example.csv")
```

We need to calculate the sample mean of the `Sims` column, the sample size, and the reliability coefficient.

```
xbar <- mean(example$Sims)
sampsize <- nrow(example)                                    ①
stdev <- sd(example$Sims)                                    ②
df <- sampsize-1                                             ③
alpha <- 1-0.95
relcoeff <- qt(p=1-alpha/2,df=df)
```

① `nrow()` calculates the number of rows of the data frame. If dealing with a singular vector, you can use `length()`. `nrow(example)` will yield the same answer as `length(example$Sims)`.
② We are calculating the sample variance using `sd()`.
③ Degrees of freedom `df = n-1`.

We can now calculate the 95% confidence interval limits.

```
lower <- xbar - relcoeff*stdev/sqrt(sampsize)
upper <- xbar + relcoeff*stdev/sqrt(sampsize)
lower
```

```
[1] 0.6642442
```

```
upper
```

```
[1] 1.303367
```

The resulting 95% confidence interval is (0.66, 1.3). Since 1 is in the confidence interval, then we can claim that 1 is a plausible value for the population mean.

1.13 Exercise

Consider the sleep health data uploaded on Canvas as `SleepHealthData.csv`.

1.13.1 Question

Calculate the **99%** confidence interval for the average heart rate (`heart_rate`) in bpm in the population this sample represents. Is 65 bpm a plausible value for the average sleep distribution of the population? Calculate the margin of error.

1.13.2 Answer

```
sleep <- read.csv("SleepHealthData.csv")
sdpop <- sd(sleep$heart_rate)
xbar <- mean(sleep$heart_rate)
sampsize <- nrow(sleep)
alpha <- 1-0.99
relcoeff <- qt(p=1-alpha/2,df=sampsize-1)
# not specifying the mean and sd in qnorm assumes standard normal

lower <- xbar - relcoeff*sdpop/sqrt(sampsize)
lower
```

```
[1] 69.6121
```

```
upper <- xbar + relcoeff*sdpop/sqrt(sampsize)
upper
```

```
[1] 70.71945
```

The 95% confidence interval is (69.61,70.72). 65 is not a plausible value for the average heart rate of the population.

The margin of error can be calculated as:

```
relcoeff*sdpop/sqrt(sampsize)
```

```
[1] 0.5536753
```

```
(upper-lower)/2
```

```
[1] 0.5536753
```