Lecture 9:
Confidence Interval for One Mean

STAT 310, Spring 2023

Warm-up Question

As part of the General Social Survey, a random sample of 300 Americans were asked how many hours per day they spend watching television. The average for these respondents was 2.9 hours.

(a) What is the sample size?

(b) What is the sample mean?

(c) Describe the population mean.

(d) Suppose another survey is conducted with a different random sample of 300 Americans. Would you expect the sample mean to the same or slightly different?

Review of Key Terms

- A **parameter** is a numerical characteristic of the population (fixed number, that is usually unknown).
- A **statistic** is a numerical characteristic of the sample (varies depending on sample). The statistic is also sometimes referred to as a **point estimate**, since it is our best guess at the value of a population parameter.
- The **standard error** (SE) measures the variability of a statistic from sample to sample.
- A **confidence interval** gives a range of plausible values for the population parameter. The half-width of a confidence interval is called the **margin of error**.

Notation

Some notation for common parameters and statistics:

	parameter	statistic
proportion	p	$\hat{p}$
mean	μ	$\bar{x}$
standard deviation	σ	s

Central Limit Theorem (CLT) for the Mean

Suppose we repeatedly take random samples of the same size n from a population, and then compute the sample mean from each. Then if the sample size n is sufficiently large (usually $n > 30$), the distribution of the sample means follows an approximate normal distribution centered around the population mean μ, and with standard error $SE = \sigma/\sqrt{n}$. We can express this succinctly using the following notation:

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Here the symbol $\sim$ translates to the word "follows."

Using Simulation to Verify the CLT

We can conduct a simulation in R to verify the CLT. As our population we'll consider the salaries (in thousands of dollars) of all San Francisco (SF) city employees from the year 2017. This data is publicly available on the internet.[1]

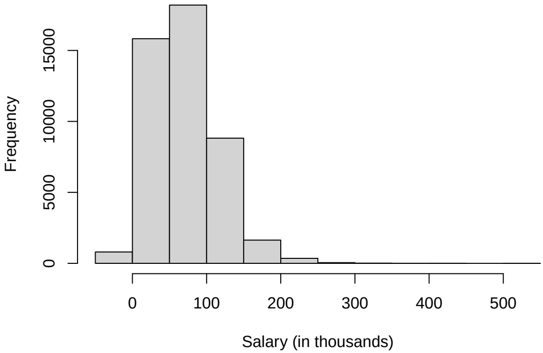
```
# load salary data
> Salaries = readRDS(url("https://ericwfox.github.io/data/Salaries.rds"))
```

```
# population size (number of SF city employees)
> length(Salaries)
[1] 45693

# population mean
> mu = mean(Salaries)
> mu
[1] 67.06238

# population standard deviation
> sigma = sd(Salaries)
> sigma
[1] 47.58665
```

[1] https://data.sfgov.org/City-Management-and-Ethics/Employee-Compensation/88g8-5mnd

Below is a histogram of the salaries of all SF city employees. This is the *population distribution*, which we can see is right skewed.

```
> hist(Salaries, xlab = "Salary (in thousands)", main = "")
```



Below is code for taking a random sample of $n = 50$ salaries of city employees, and then comping mean.

```
> samp1 = sample(Salaries, size = 50)
> mean(samp1)
[1] 73.92442
```

If we take another random sample of $n = 50$ salaries, we get a different sample mean.

```
> samp2 = sample(Salaries, size = 50)
> mean(samp2)
[1] 48.20304
```

Also note that the sample means estimate the population mean ($\mu = 67.062$ thousand dollars) with some error. For instance, here the first sample mean is an overestimate, and the second sample mean is an underestimate.

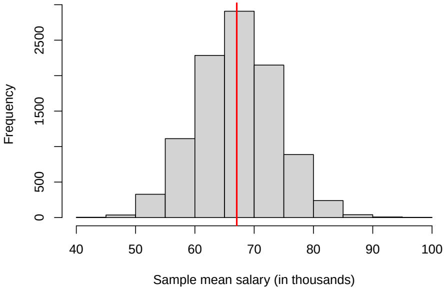
To construct a *sampling distribution* we repeatedly take random samples from the population, and compute the sample mean from each. To do this in R, we can use the `replicate()` function:

```
xbars = replicate(10000, {
  samp = sample(Salaries, size = 50)
  mean(samp)
})
```

The code above takes 10,000 random samples of $n = 50$ salaries from the population, and computes the mean salary of each sample. The 10,000 sample means are stored in the vector `xbars`.

A histogram of the 10,000 sample means is plotted below (this is our sampling distribution for $\bar{x}$). Remarkably, we see that the histogram looks normally distributed, and centered around the population mean (red vertical line). This is exactly what the CLT says the sampling distribution should look like.

```
> hist(xbars, xlab = "Sample mean salary (in thousands)", main = "")
# plot vertical line at population mean
> abline(v = mu, col = "red", lwd = 2)
```



The standard error (SE) is given by the standard deviation of the 10,000 sample means.

```
> sd(xbars)
[1] 6.737655
```

This is approximately the same value we get when we use the formula $SE = \sigma/\sqrt{n}$ from the CLT.

```
> sigma / sqrt(50)
[1] 6.729768
```

Remarks:

- Note that even though the population distribution for salary is skewed right, the distribution of the sample means is normal. This is true in general - according to the CLT, as long as the sample size is large (usually $n > 30$), the sampling distribution will be approximately normal, regardless of the shape of the population distribution.
- In applications, we usually work with a single random sample of data, and therefore can't plot a histogram of the sampling distribution like we did in the simulation. However, because of the CLT we can still *assume* that the sampling distribution is approximately normal when certain conditions are met.
- The primary application of the CLT is to make inferences for population parameters by constructing confidence intervals or performing hypothesis tests. These methods only require a single random sample of data.

Confidence Interval for μ

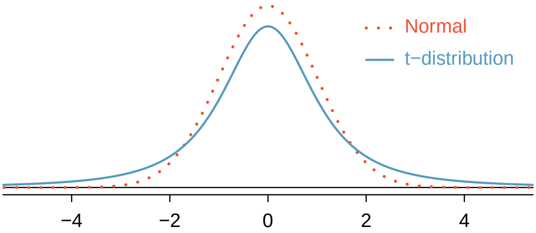
- Based on the CLT, we can construct a 95% confidence interval for the population mean μ as

$$\bar{x} \pm z^* SE \implies \bar{x} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

- However, one issue with this confidence interval formula is that the standard error is in terms of the population standard deviation σ, which is unknown.
- We can resolve this by plugging in the sample standard deviation s as the estimate of σ. That is, use $SE \approx s/\sqrt{n}$.
- This is a sensible approach when the sample size is large. But when the sample size is small, the confidence interval needs to be adjusted to account for additional uncertainty in estimating σ with s.
- It turns out that we can get a more accurate confidence interval, which accounts for this additional uncertainty, by using the t-distribution to calculate the critical value instead of the z-distribution.

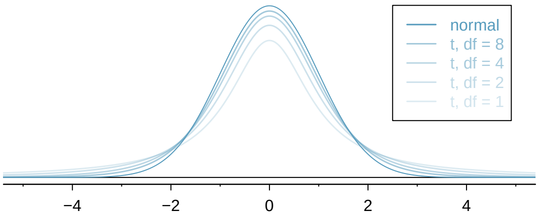
t-disribution

- A *t*-distribution is bell-curve shaped distribution that is centered around zero.
- It looks similar to a standard normal distribution, but it has wider tails.



t-disribution

- The shape of the t-distribution depends on the degrees of freedom, which is defined as $df = n - 1$
- When the sample size is small the t-distrubiton has noticeably wider tails than a normal distribution.
- When the sample size is large (about 30 or more), the t-distribution is close to a normal distribution.



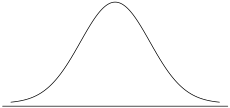
Confidence Interval for μ

A confidence interval for the the population mean μ is given by

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$$

The critical value t^* is found using t-distribution. It depends on the confidence level and degrees of freedom, and can be computed using the R function `qt()`.

Example: Calculate the critical value t^* when the confidence level is 95% and $n = 15$.



Conditions

The confidence interval for μ is valid if the following conditions are satisfied:

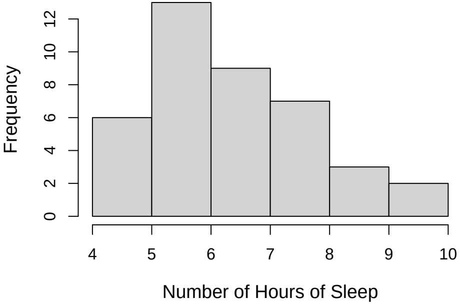
- The data come from a random sample. (This is called the **independence condition** in the textbook.)
- The sample size n is large ($n \geq 30$). Otherwise, if the sample size is small ($n < 30$), the data should have an approximate normal distribution. (This is called the **normality condition** in the textbook.)

These conditions ensure that the CLT holds. For small sample sizes ($n < 30$), look at histogram of the data to check normality.

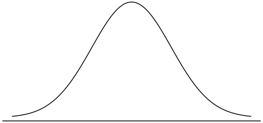
Example

As part of a survey conducted by the US National Center for Health Statistics, a random sample of $n = 40$ Americans were asked how many hours of sleep they get on a typical weekday. Some summary statistics and a histogram of this data are shown below.

n	$\bar{x}$	s	min	max
40	6.8	1.4	4	10



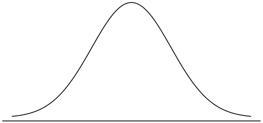
(a) Calculate and interpret a 95% confidence interval for the population mean.



(b) Comment on whether the conditions for the confidence interval appear satisfied.

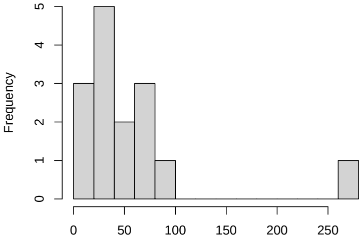
Example

Calculate a 90% confidence interval for the population mean number of hours of sleep Americans get on a typical weekday.



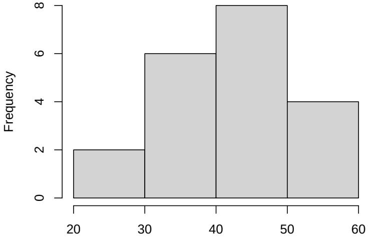
Example: Checking the Conditions

Consider the following histogram of a random sample of size $n = 15$. We want to use this data to calculate a confidence interval for the population mean. Comment on whether you think the conditions for the confidence interval are satisfied.



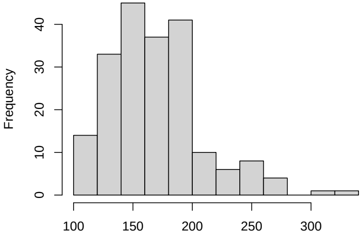
Example: Checking the Conditions

Consider the following histogram of a random sample of size $n = 20$. We want to use this data to calculate a confidence interval for the population mean. Comment on whether you think the conditions for the confidence interval are satisfied.



Example: Checking the Conditions

Consider the following histogram of a random sample of size $n = 200$. We want to use this data to calculate a confidence interval for the population mean. Comment on whether you think the conditions fo the confidence interval are satisfied.


Frequency