

choose a

MATH 3339
Statistics for the Sciences
Sec 7.1-7.3

Wendy Wang
wwang60@central.uh.edu

x

Lecture 13 - 3339

Outline

- 1 Estimating Parameters
- 2 Introduction to Confidence Intervals
- 3 T-distribution
- 4 Confidence Interval for Population Mean
- 5 Examples of Confidence Intervals

Statistical Inference


- **Statistical inference** is the science of inferring characteristics of an entire population from the information contained in a sample from that population.
- **Statistical inference** is divided into two broad categories, **estimation** and **hypothesis testing**.
Chapter 7 *Chapter 8*
- They are not mutually exclusive.

Estimation

use sample information to determine population

- **Estimation** is the process of determining the value of a population parameter from evidence made available by a sample **statistic**.
- A point estimate is a single value that has been calculated from sample data to estimate the unknown population parameter.
- We like to use unbiased estimates to predict the parameter.

Unbiased Estimators

- If we desired to estimate a parameter, we want to know that we are using a good estimator. The way we know that is if the statistics is an unbiased estimate of the parameter.
- A point estimator $\hat{\theta}$ is said to be an unbiased estimator of θ if $E(\hat{\theta}) = \theta$ for every possible value of θ .
- If $\hat{\theta}$ is not unbiased, the difference $E(\hat{\theta}) - \theta$ is called the bias of $\hat{\theta}$.


Notation of Parameters and Statistics

Name	Sample Statistic	Population Parameter
mean	\bar{x}	μ mu
standard deviation	s	σ sigma
correlation	r	ρ rho
regression coefficient	b	β beta
proportion	\hat{p}	p

We are going to use the sample statistics to **estimate** the population parameters.

Standard Error

- The **standard error** of an estimator $\hat{\theta}$ is its standard deviation $SE(\hat{\theta}) = \sqrt{Var(\hat{\theta})}$.
- Examples of standard errors
 - ▶ $SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$.
 - ▶ $SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$.
- The problem is that often we do not know σ or p for example, we can use the estimators for these parameters. Then we have the **estimated standard error**.
- Again we need to know how these estimators $\hat{\theta}$ are being distributed.

What do we use for estimating?

- point estimation : one single value
- confidence interval

- A **confidence interval** is a range of possible values that is likely to contain the unknown population parameter we are seeking.
- First, we must have a **level of confidence**. 90%
- Then based on this level we will compute a **margin of error**. = $f(c.l.)$
- Last, we can say that we are $\alpha\%$ confident that the true population parameter falls within our confidence interval.



Example

Suppose the heights of the population of basketball players at a certain college are in question. A sample of size 16 is randomly selected from this population of basketball players and their heights are measured. The average heights is found to be 6.2 feet and the margin of error is found to be ± 0.4 feet. If this margin of error was determined with a 95% confidence level, find and interpret the confidence interval.

95% confidence interval for height for players in this college
is between $\frac{6.2 - 0.4}{= 5.8}$ and $\frac{6.2 + 0.4 = 6.6}{}$

$$\bar{X} \pm M.E. = (L.L., U.L.)$$

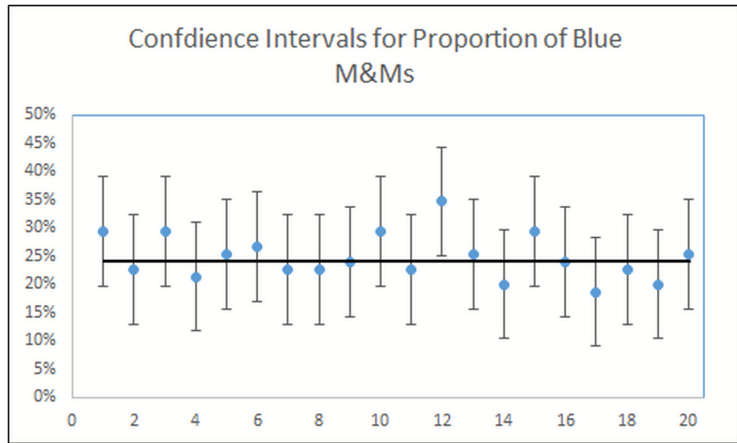
We are 95% confident that the height of the basketball players in this college is between 5.8 and 6.6 feet.

• •

So what does this mean?

- The confidence level is saying if we repeat this process several different times we would expect to include the population proportion \pm % of the time.
- For example, the producers of M&Ms say that 24% of the Milk Chocolate M&Ms are blue. There was a study to see if that was true. A sample of 75 milk chocolate M&Ms were randomly selected and then the process was repeated 20 times. The following graph is the percent we got that were blue M&Ms and the confidence intervals for a 95% confidence.

How many times should the true parameter be included in the confidence interval?



.24

What is the average cell phone bill per month?

- A survey taken in 2010 polled 400 randomly chosen cell phone users. They answered the question: "What is your average monthly cell phone bill?"
- The following are the characteristics of the sample:
 - ▶ The sample mean is $\bar{x} = \$71$.
 - ▶ Assume the population standard deviation to be $\sigma = \$20$.
 - ▶ The sample size is $n = 400$.

What is the population mean monthly cell phone bill?

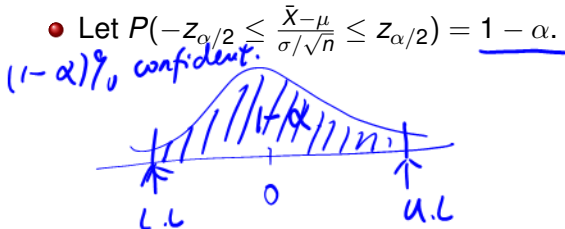
- From the sample of 400 people we found that the sample mean was \$71.
- We can use this number to *estimate* the population mean monthly cell phone bill.
- However, if we take another random sample of 400 cell phone users, would we get the same sample mean?

Confidence Interval

- To answer the previous questions we create what is called a **confidence interval**, a range of values to estimate the true parameter.
- That is, from a sample we get a point estimate. Then we say how *confident* we are that the true parameter is within the range of values in the interval.
- Today we will look at estimating the population mean, μ .
- To find this estimation we will use the sampling distributions of \bar{X} .

From Probability to Confidence

- Let X be a random variable with **unknown** mean μ and standard deviation σ .
- Suppose we are looking at the sample means \bar{X} for a sample of size n .
- Assume $\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$. C.L.T.
- Let $P(-z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}) = \underline{1 - \alpha}$.



$$\frac{\bar{X} - \mu}{(\frac{\sigma}{\sqrt{n}})} = Z \sim N(0,1)$$

The confidence interval

The $1 - \alpha$ confidence interval for μ , given that we know the population standard deviation is:

$$\bar{x} \pm z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$$

↑
critical value

\bar{x} : sample mean,

$z_{\frac{\alpha}{2}}$: critical value (it is determined by confidence level, C)

$$\alpha = 1 - C = 1 - \text{confidence level}$$

σ : population st. dev.

n : sample size

Assumptions for Estimating the Population Mean

1. The sample has to be as a result of a simple random sample (SRS).
2. The distribution of the population has to be Normal. By the Central Limit Theorem if our sample size is larger than 30 then the sample means have a Normal distribution.

$\bar{x} \overset{\text{C.L.T}}{\sim} \text{Normal}$
($n \geq 30$)

Confidence Level $C = 100(1 - \alpha)\%$

- The confidence interval is associated with a degree of confidence.
- This degree of confidence is the percentage of times that the confidence interval actually does contain the parameter.
- Most common choices of level of confidence: 80%, 90%, 95% and 99%.
- The confidence level is predetermined which will be given in the problem. If not assume $C = 95\%$.

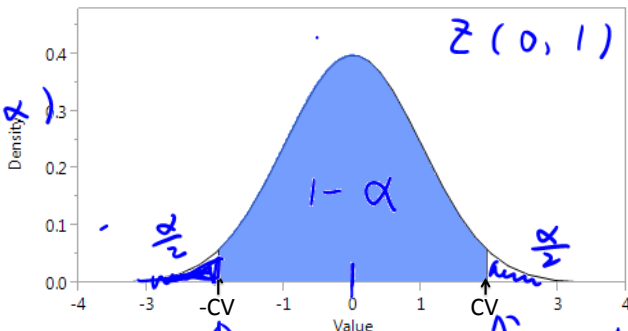
Critical value

$$Z_{\frac{\alpha}{2}}$$

- The critical value depends on the confidence level $1 - \alpha$.
- The critical value (CV) is the cut off point where $1 - \alpha$ is the middle area of the density curve.

$$\frac{\alpha}{2} + (1 - \alpha) + \frac{\alpha}{2} = 1$$

Distribution



$$qnorm\left(\frac{\alpha}{2}, 0, 1\right)$$

$$qnorm\left(1 - \frac{\alpha}{2}, 0, 1\right)$$

Critical Value for Means

(95% C.I for mean μ .)

$$C = 0.95$$

$$\alpha = 1 - C = 0.05$$

- We are assuming we know the population standard deviation.
- For means if the population standard deviation is known, the critical value is z^* . where the area under the Normal curve is between $-z_{\alpha/2}$ and $+z_{\alpha/2}$ is the confidence level $C = 1 - \alpha$.
- This can be found using the z-table or $qnorm((1 + C)/2)$ in R.
- The following table is the common confidence levels with their z-score

C	80%	90%	95%	99%
$z_{\alpha/2}$	$z_{0.10} = 1.28$	$z_{0.05} = 1.645$	$z_{0.025} = 1.96$	$z_{0.005} = 2.576$

Margin of Error

The margin of error is

$$m = \text{critical value} \times \text{standard error}$$

For means (given the population standard deviation is known), the margin of error is:

$$m = \underline{z_{\alpha/2}} \left(\frac{\sigma}{\underline{\sqrt{n}}} \right)$$

What is the margin of error for the mean monthly cell phone bill?

$$\begin{aligned} m &= z_{\frac{\alpha}{2}} * \left(\frac{\sigma}{\sqrt{n}} \right) \\ &= 1.96 * \left(\frac{\$20}{\sqrt{400}} \right) \\ &= \$1.96 \end{aligned}$$

$$\begin{aligned} \sigma &= 20 \\ n &= 400 \\ C &= 95\% \\ C.V. &= z_{\frac{\alpha}{2}} = \\ &= qnorm\left(\frac{1+0.95}{2}\right) \end{aligned}$$

```
> qnorm(1.95/2)
[1] 1.959964
```

Rcode

95% C.I for the monthly cell phone bill
is $(\bar{x} - 1.96, \bar{x} + 1.96)$ $\bar{x} = 71$

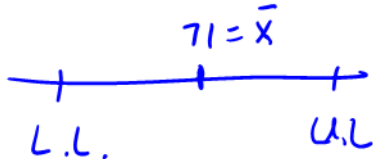
\bar{x}
↑
> 71 + c(-1, 1) * qnorm(1.95/2) * 20/sqrt(400)
[1] 69.04004 72.95996

C.V = $z_{\alpha/2}$

σ

$\sqrt{400}$

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



Gas Prices

The following is a random sample of the price of regular unleaded gasoline from 10 gas stations in the area

1.85 1.99 1.94 2.25 2.39 2.19 2.19 1.95 2.18 2.09

Suppose that the population standard deviation of gasoline prices is $\sigma = 0.1$. Give a 97% confidence interval for the mean gasoline prices.

$$\bar{x} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

$$\bar{x} = \frac{1.85 + \dots + 2.09}{10}, \quad \sigma = 0.1, \quad n = 10$$

$$z_{\frac{\alpha}{2}} = q_{\text{norm}}\left(\frac{1+0.97}{2}\right)$$

R code

```
gas=c(1.85,1.99,1.94,2.25,2.39,2.19,2.19,1.95,2.18,2.09)
> mean(gas)+c(-1,1)*qnorm(1.97/2)*0.1/sqrt(10)
[1] 2.033376 2.170624
```

Recap: The confidence interval for population mean

The $1 - \alpha$ confidence interval for μ , given that we know the population standard deviation is:

$$\bar{x} \pm z_{\alpha/2} \left(\frac{\sigma}{\sqrt{n}} \right)$$

A diagram illustrating the components of the confidence interval formula. Three blue arrows point upwards from below the formula to the terms \bar{x} , $z_{\alpha/2}$, and $\frac{\sigma}{\sqrt{n}}$. A blue letter 'C' is positioned below the arrows, centered under the $z_{\alpha/2}$ term.

what if we don't know σ ?

Recap: Critical Value when σ is known

- We are assuming we know the population standard deviation.
- For means if the population standard deviation is known, the critical value is z^* . where the area under the Normal curve is between $-z_{\alpha/2}$ and $+z_{\alpha/2}$ is the confidence level $C = 1 - \alpha$.
- This can be found using the z-table or `qnorm((1 + C)/2)` in R.
- The following table is the common confidence levels with their z-score

C	80%	90%	95%	99%
$z_{\alpha/2}$	$z_{0.10} = 1.28$	$z_{0.05} = 1.645$	$z_{0.025} = 1.96$	$z_{0.005} = 2.576$

Mean Amount of Coffee Dispensed

$$s = \text{sd}(\text{data})$$

A coffee machine dispenses coffee into paper cups. Here are the amounts measured in a random sample of 20 cups.

9.9, 9.7, 10.0, 10.1, 9.9, 9.6, 9.8, 9.8, 10.0, 9.5,
9.7, 10.1, 9.9, 9.6, 10.2, 9.8, 10.0, 9.9, 9.5, 9.9

Determine a 90% confidence interval for the mean amount of coffee dispensed from this machine.

$$\bar{x} \pm \underbrace{CV}_{\text{CV}} \cdot \frac{\cancel{X}^?}{\sqrt{n}} \quad \frac{s}{\sqrt{n}}$$

Standard Error when σ is unknown

- When the standard deviation of a statistic is estimated from the data, the result is called the **standard error** of the statistic.
- The standard error of the sample mean is

$$SE_{\bar{X}} = \frac{s}{\sqrt{n}}$$

where s is the computed **sample** standard deviation from the data.

- From our example: $SE_{\bar{X}} = \frac{0.1986}{\sqrt{20}} = \underline{\underline{0.0444}}$.

The T-distribution

- The problem is that the sample standard deviation s varies from sample to sample.
- William Gosset, (a quality control engineer for the Guinness Brewery) discovered this problem and figured out a new distribution that changes the critical value based on the sample size.
- This new distribution is called *Students T* distribution, because Guinness would not allow Gosset to publish his findings since he was their employee.
- The shape of this distribution changes with different sample sizes. So it depends on a parameter called the **degrees of freedom** (df)
- The degrees of freedom for the T-distribution of the sample mean is the sample size minus one: $(n - 1)$. Because we are using the sample standard deviation $s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$.

T distribution

if σ is known, \rightarrow z-dist.
if σ is unknown \rightarrow t-dist.

- Used for the inference of the population mean. When population standard deviation σ is unknown.
- The distribution of the population is basically bell-shape.

- Formula for t :

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

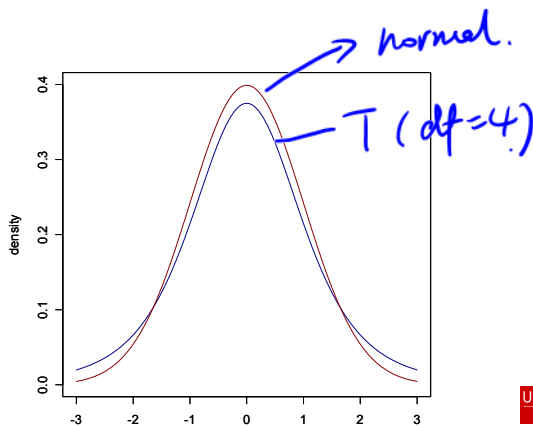
s : sample st. dev.
 $\sim T(n-1)$

- Use t-table, or `qt(probability,df)` in R.
- Degrees of freedom: $df = n - 1$.

\uparrow
sample size

Normal Distribution vs T distribution

The red graph is the Normal density curve and the blue graph is the T density curve with a degrees of freedom of 4.



Using T-table

- <https://www.math.uh.edu/~wwang/t-table.pdf>
- The top margin is the area in the right tail.
- The left margin is the degrees of freedom $n - 1$.
- The values inside the table are the t values.

Critical value when σ unknown

- When σ is **unknown** we use t -distribution.
- With degrees of freedom, $df = n - 1$.
- The critical value is $t_{\alpha/2}$, where the area between $-t_{\alpha/2}$ and $+t_{\alpha/2}$ under the T-curve is the confidence level $C = 1 - \alpha$.
- $t_{\alpha/2}$ is found in T-table using the row according to the degrees of freedom and the column according to the confidence level at the bottom of the table.
- In R use $qt((1 + C)/2, df)$.

$$\begin{array}{l} n=30 \quad C=96\% \\ df=n-1=29 \\ \underline{qt\left(\frac{1+0.96}{2}, 29\right)} \end{array}$$

Mean Amount of Coffee Dispensed

A coffee machine dispenses coffee into paper cups. Here are the amounts measured in a random sample of 20 cups.

9.9, 9.7, 10.0, 10.1, 9.9, 9.6, 9.8, 9.8, 10.0, 9.5,
9.7, 10.1, 9.9, 9.6, 10.2, 9.8, 10.0, 9.9, 9.5, 9.9

Determine a 90% confidence interval for the mean amount of coffee dispensed from this machine.

$$\bar{X} = \frac{9.9 + \dots + 9.9}{20} = 9.845$$

in R: `mean(data)`

$$S = \frac{1}{n-1} \sum (x_i - \bar{x})^2 = 0.1986$$

in R: `sd(data)`

$$\bar{X} \pm t(\alpha, df) \cdot \frac{S}{\sqrt{n}}$$

Determine the 90% Confidence Interval

$$\underbrace{9.845 \pm \underbrace{qt\left(\frac{1+0.90}{2}, 19\right) * \frac{0.1986}{\sqrt{20}}}}_{+ c(-1, 1) * \downarrow}$$

R code

```
coffee=c(9.9,9.7,10,10.1,9.9,9.6,9.8,9.8,10,9.5,9.7,  
10.1,9.9,9.6,10.2,9.8,10,9.9,9.5,9.9)  
t.test(coffee,conf.level = 0.9)
```

One Sample t-test

data: coffee

t = 221.68, df = 19, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 0

90 percent confidence interval:

9.768207 9.921793

sample estimates:

mean of x

9.845

Z-interval ^{for μ :} if σ is given,

$$\bar{X} \pm Z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

T-interval ^{for μ :} if σ is NOT given,

$$\bar{X} \pm t_{(\frac{\alpha}{2}, df)} \cdot \frac{s}{\sqrt{n}}$$

Example 1

$$CV: q_{\text{norm}}\left(\frac{1+C}{2}\right)$$

A soft-drink machine is regulated so that the amount of drink dispensed is approximately normally distributed with a standard deviation equal to 0.53 ounces. Find a 99% confidence interval for the mean of all drinks dispensed by this machine if a random sample of 36 drinks has an average content of 7.94 ounces.

$$n = 36$$

$$7.94 = \bar{x}$$

$$0.53 = \sigma$$

$$0.99 = C$$

$$\alpha = 1 - C = 1 - 0.99 = 0.01$$

σ is given, Z-interval

$$\bar{x} \pm z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

Annotations: $\bar{x} \rightarrow 7.94$, $z_{\frac{\alpha}{2}} \rightarrow q_{\text{norm}}\left(\frac{1+0.99}{2}\right)$, $\sigma \leftarrow 0.53$, $\sqrt{n} \leftarrow 36$

$$C + \alpha = 1$$



$$\alpha + C = 1$$

$$C = 1 - \alpha$$

$$1 - \frac{\alpha}{2}$$

$$= 1 - \frac{1 - C}{2}$$

$$= \frac{2 - (1 - C)}{2}$$

$$= \frac{1 + C}{2}$$

Example 2

The heights of a random sample of 50 college students showed a mean of 174.5 centimeters and a standard deviation of 6.9 centimeters. Construct a 98% confidence interval for the mean height of all college students.

$$n = 50 \quad \bar{X} = 174.5 \quad S = 6.9$$

$$\text{Confidence level} = 0.98 = C$$

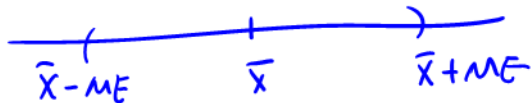
σ is NOT given. t -interval

$$\bar{X} \pm t_{(\frac{\alpha}{2}, df)} * \frac{S}{\sqrt{n}}$$

Annotations: $\bar{X} \rightarrow 174.5$, $t_{(\frac{\alpha}{2}, df)} \rightarrow t_{(\frac{1-C}{2}, 49)}$, $S \leftarrow 6.9$, $n \leftarrow 50$, $df = n - 1 = 49$

What will happen if changing Confidence Levels **C**

Suppose we have a 99% confidence interval for the population mean, μ of (2.272, 17.728). If we change the confidence level to 92% what is the confidence interval?



```
> qnorm(1.99/2)
[1] 2.575829
> qnorm(1.92/2)
[1] 1.750686
```

$$\text{width} = 2 * ME,$$

$$ME = qnorm\left(\frac{1+C}{2}\right) \cdot \frac{\sigma}{\sqrt{n}}.$$

changing C from 99% to 92%

when $C = 0.99$, $qnorm\left(\frac{1.99}{2}\right) = 2.58$

$C = 0.92$, $qnorm\left(\frac{1.92}{2}\right) = 1.75$

Changing Sample Size

The mean of a random sample of n measurements is equal to $\bar{x} = 33.9$. Assume $\sigma = 3.3$. Determine the margin of error for a 95% confidence interval for the population mean when the sample size is

1. $n = 100$

2. $n = 400$

$$q_{\text{norm}}\left(\frac{1.95}{2}\right) * \frac{3.3}{\sqrt{n}}$$

as $n \nearrow$, $M.E \downarrow$

Behavior of Confidence Intervals with different n

Notice that as the sample size increases the width of the interval decreases. Or the confidence interval becomes thinner.

- First: Mathematically, notice that we are dividing by a larger number, so that will decrease the quotient.
- Second: Intuitively, as the sample size increases the accuracy of the estimation becomes better, thus the point estimate is getting closer to the population mean and the interval does not need to be as wide.

Choosing Sample Size

You can have both a high confidence while at the same time a small margin of error by taking enough observations.

- Sample size for confidence intervals of means.

$$n > \left(\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{ME} \right)^2$$

Starting Salary

- We want to estimate annual starting salaries for college graduates with degrees in business administration. To determine this we need a sample.
- Assume that a 95% confidence interval estimate of the population mean annual starting salary is desired.
- Assume the standard deviation is $\sigma = \$7,500$.
- How large a sample should be taken if the desired margin of error is $m = \$500$?

$$n > \left(\frac{z_{\frac{\alpha}{2}} \cdot \sigma}{ME} \right)^2 = \left(\frac{z_{\text{norm}(\frac{1-0.95}{2})} \times \$7500}{\$500} \right)^2$$
$$= 864.3 \dots$$

n at least 865