Stat 104: Quantitative Methods

Class 21: Point Estimation and Confidence Intervals

Recap-What is our goal?

- To get on the Owl's invite list?
- Explore new space frontiers?
- Get a job?
- Make new friends?

- No! our goal is to make **statistical inference**

We want to draw conclusions from **sample** data about the larger **populations** from which the samples are drawn.

Recap-Terminology

- A **parameter** is a characteristic of a population. A **statistic** is a characteristic of a sample
- Inferential statistics enables you to make an educated guess about a population parameter based on a statistic computed from a random sample drawn from that population.

	Sample **S**tatistic	**P**opulation **P**arameter
Mean	$\bar{X}$	μ
Proportion	$\hat{p}$	π
Variance	s^2	σ^2
Correlation	r	ρ

greek letters

Examples

- The sample mean is a statistic used to estimate μ: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- There could be many possible estimators!
- For example, the sample median is another statistic that could be used to estimate μ.

Examples

- The sample variance is a statistics used to estimate σ^2.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

- But one could also use the range, IQR or even the above divided by n to estimate σ^2.

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

What kind of estimators do we want?

- Statisticians spend lots of time trying to develop estimators of parameters.
- In particular, there are two properties we want our estimators to have:
  - ❑ They should be unbiased
  - ❑ They should have minimum variance
- We'll discuss these concepts in turn.

Unbiased Estimates

- What kind of estimator do we want ? Probably one that always gives a good approximation to the truth.
- Can we be right all the time ? Probably not. What about being correct **on average** ?
- Generically, let $\hat{\theta}$ denote the estimate of some parameter θ.
- The bias of $\hat{\theta}$ is defined to be

$$bias(\hat{\theta}) = E(\hat{\theta}) - \theta$$

- As estimator is called **unbiased** if its bias=0.

Example

- The sample mean is an unbiased estimate of the population mean.
- That is $E(\bar{X})=\mu$ $\cancel{\bar{X}=\mu}$
- But there are many other unbiased estimates of the population mean.
- We also want our estimators to have minimum variance.

The Sample Variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

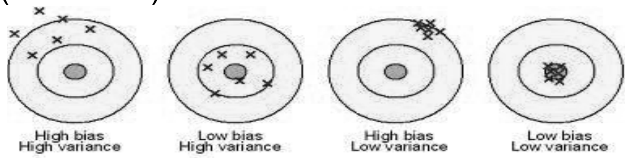
- Why do we divide by n-1? So it is unbiased!
- It is a bit of work but it can be shown that

$$E(s^2) = \sigma^2.$$

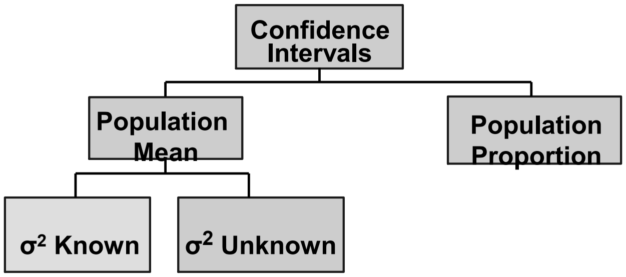
- So in short, we divide by “n-1” to make the sample variance an unbiased estimator.

Bias-Variance Trade-off

- Given a choice, we want estimators with low (or no bias) and low variance.



Confidence Intervals



Introduction

Say we have a population of interest and we want to determine what its mean is or the proportion with some characteristic.

Population

$$\mu = ?$$

$$p = ?$$

We know the procedure is to generate a random sample $X_1 X_2 ,..., X_n$ and form the estimates

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \ or\ \hat{P} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

How Wrong Are We ?

It would be **grossly misleading** to claim that μ is precisely equal to the observed $\bar{x}$ or p to the observed $\hat{p}$.

To detail our uncertainty about our estimate for μ, we can construct a **confidence interval** or **interval estimate** for μ.

Example: what is the average weight of graduate students' at Harvard ?

Gender	Sample Mean	Std. Err.	DF	L. Limit	U. Limit
0	168.52577	2.663393	96	163.23898	173.81256
1	126.34821	3.4025612	55	119.52933	133.1671

interval estimates

point estimates

The same idea works for proportions.

What proportion of students are right handed ?

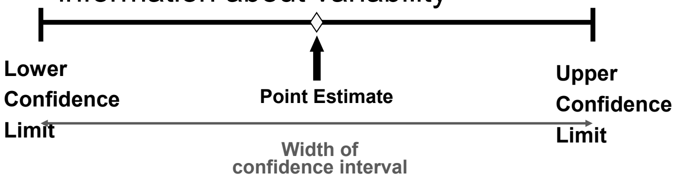
Gender	Count	Total	Sample Prop.	Std. Err.	L. Limit	U. Limit
0	90	98	0.9183673	0.027658405	0.86415786	0.9725768
1	56	58	0.9655172	0.023958908	0.91855866	1.0124758

(silly?)

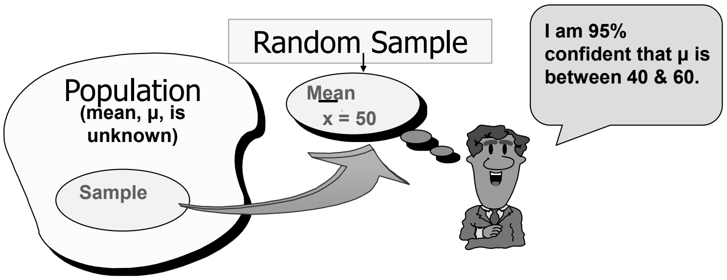
Confidence intervals are a vital aspect to statistics since an estimate is useless without some concept of its precision-that is exactly what a confidence interval tells us; **how good is our estimate**.

Point and Interval Estimates

- A point estimate is a single number,
- a confidence interval provides additional information about variability



The General Estimation Process



We construct the interval estimate using the CLT.

Recall that the CLT says that (**for large *n***)

$$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$$

Then by the **Standardization Rule**:

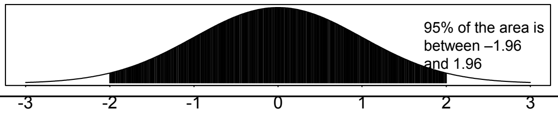
$$\frac{\bar{X}-\mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

We have

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

Then from what we know about the standard normal distribution:

$$P(-1.96 < \frac{\bar{X}-\mu}{\sigma / \sqrt{n}} < 1.96) = 95\%$$



We have

$$P(-1.96 < \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} < 1.96) = 95\%$$

By doing some **algebra**, we may rearrange stuff so that

$$P(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}) = 95\%$$

lower bound

\

truth

upper bound

We are thus 95% confident that the true mean μ is in the interval

$$(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}})$$

What does it mean that we are 95% confident ?

It means that if we had 100 different samples and created 100 different intervals, 95 out of 100 would contain the true (but unknown) mean.

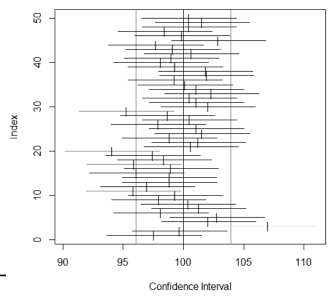
Interpretation

- Confidence intervals are often wrongly interpreted. We are confident in the *process*, not in any one interval.
- The confidence level is ***not*** the probability that a specific confidence interval contains the population parameter.
- "The parameter is an unknown constant and no probability statement concerning its value may be made."
  —Jerzy Neyman, original developer of confidence intervals.

Interpretation

- ***Confidence level****:* To say that we are 95% *confident* is shorthand for "95*% of* all possible samples of a given size from this population will result in an interval that captures the unknown parameter."
- ***Confidence interval****:* To interpret a *C*% confidence interval for an unknown parameter, say, "We are *C%* confident that the interval from _____ to _____ captures the actual value of the [population parameter in context]."

Graph of Many Confidence Intervals



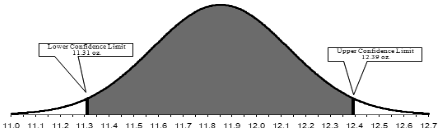
Example: Mechanics of the Interval

- As part of their quality-control program, Whole Foods wants to know the average weight of the grapefruits they purchase from Florida growers.
- They take a sample of 40 Florida grapefruits with sample mean 11.85 ounces, and know from past experience that the weight of Florida grapefruits is normally distributed with a standard deviation of 1.75 ounces.

Example (cont)

- The 95% confidence interval for the true mean weight of Florida grapefruits, is then

$$(\overline{x}-1.96\frac{\sigma}{\sqrt{n}},\overline{x}+1.96\frac{\sigma}{\sqrt{n}})=11.85\pm1.96\left(\frac{1.75}{\sqrt{40}}\right)=(11.31,12.39)$$



The confidence interval for the mean is given by

$$(\overline{x}-1.96\frac{\sigma}{\sqrt{n}},\overline{x}+1.96\frac{\sigma}{\sqrt{n}})$$ or $$\overline{x}\pm1.96\frac{\sigma}{\sqrt{n}}$$

There is a slight problem here. Do we know the true value of σ ? Probably not. Do we panic ? If n>30 we can **substitute s for σ** so that the interval estimate of μ is then given by

$$\overline{x}\pm1.96\frac{s}{\sqrt{n}}$$

Note-this is an approximation but ok to do if n>30

Stay tuned for what to do for small n.

Assumptions Review

- IF we know our data is normally distributed AND we know σ (big IF) the confidence interval for any sample size is

$$\overline{x}\pm1.96\frac{\sigma}{\sqrt{n}}$$

- IF our sample size is 30 or larger, we don't need to know σ and the CLT kicks in

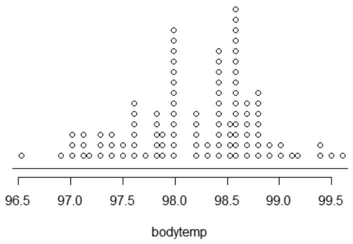
$$\overline{x}\pm1.96\frac{s}{\sqrt{n}}$$

Is it hot in here or is it just you?

- What are we always told "normal" body temperature is ?
- Where did this number come from ? It turns out that a while back some German researchers ran extensive studies and discovered that average body temperature is around 36.83 Celsius.
- It was easier to just say, heck, average temp is 37 Celsius, which works out nicely to 98.6.

Some University of Maryland researchers studied 105 healthy undergrads to get a handle on what the usual body temperature should be. Here is a dotplot of their data:

```
> mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat100/bodytemp.csv")
> names(mydata)
[1] "fahrenheit"
> bodytemp=mydata$fahrenheit
> library(BHH2)
> dotPlot(bodytemp)
```



Confidence Intervals in R

```
> t.test(bodytemp)

        One Sample t-test

data:  bodytemp
t = 1608.5, df = 104, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 98.07703 98.31916
sample estimates:
mean of x
  98.1981
```

```
> mean(bodytemp)
[1] 98.1981
> sd(bodytemp)
[1] 0.6255737
```

$$\overline{x}\pm1.96\frac{s}{\sqrt{n}}=98.2\pm1.96(.06)$$

In Celsius, this interval is (36.71,36.85)

Review of s versus $s/\sqrt{n}$

```
> describe(bodytemp,skew=F)
   vars   n mean   sd  min  max range   se
X1    1 105 98.2 0.63 96.5 99.6   3.1 0.06
```

s tells us how variable the sample is

$s/\sqrt{n}$ tells us how variable the sample mean is

Some More R Examples

- Sometimes we just have the summary statistics and not the full data set.
- In this case there is a command in the `BSDA` package called `tsum.test` we can use.

Example

- A random sample of 121 automobiles traveling on an interstate showed an average speed of 65 mph with a standard deviation of 22 mph. Find the 95% CI.

```
> library(BSDA)
> tsum.test(n.x=121,mean.x=65,s.x=22)

        One-sample t-Test

data:  Summarized x
t = 32.5, df = 120, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 61.04014 68.95986
```

Finding any size confidence interval

A 95% Confidence interval is given by

$$\bar{X} \pm 1.96\frac{s}{\sqrt{n}}$$

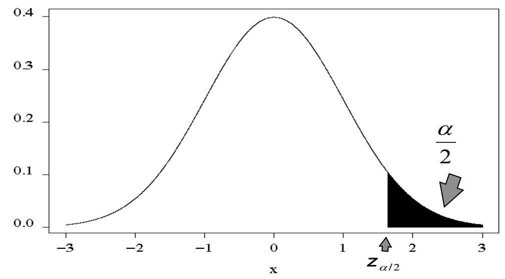
What if you want a 99% interval, or an 80% interval, or something else ?

What changes is the number in the box:

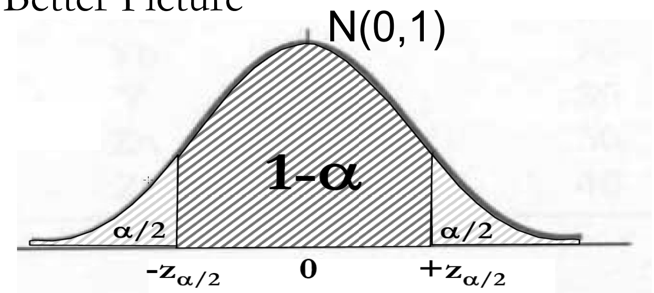
$$\bar{X} \pm [\quad]\frac{s}{\sqrt{n}}$$

Some strange notation

- We define $z_{\alpha/2}$ to be the point on the normal curve as follows

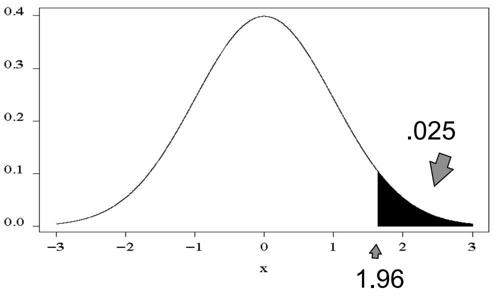


A Better Picture



Example of Strange Notation

For example, when α=5%, $z_{\alpha/2}$=1.96 and we have



Confidence Levels

- Any size confidence interval is then given by

The (1-α)$\bullet$100% C.I. $\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$

- Here are the most common values

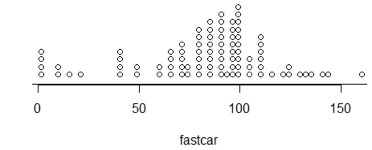
Confidence Level	*Confidence Coefficient,* $1-\alpha$	*z value,* $z_{\alpha/2}$
80%	**.80**	**1.28**
90%	.90	1.645
95%	.95	1.96
98%	**.98**	**2.33**
99%	.99	2.58
99.8%	**.998**	**3.08**
99.9%	**.999**	**3.27**

Confidence and Width

- Example : What is the fastest you have ever driven a car ?

```
> mydata=read.csv("http://people.fas.harvard.edu/~mparzen/stat100/stat111_survey.csv")
> names(mydata)
 [1] "id"           "stat104"      "stat110"      "stat139"
 [5] "cs50"         "r"            "year"         "female"
 [9] "height"       "weight"       "primarycell"  "looks"
[13] "sleep"        "haircut"      "smoke"        "snap"
[17] "righthanded"  "manual"       "vegetarian"   "glasses"
[21] "coffee"       "hair"         "shower"       "fastestdriven"
[25] "relationship" "siblings"     "texts"        "exercise"
[29] "random"       "ttest"        "heartrate"
> fastcar=mydata$fastestdriven
```

Example: Fastest you've driven a car



```
> describe(fastcar,skew=FALSE)
   vars  n  mean    sd min max range   se
X1    1 89 83.11 33.11   0 160   160 3.51
```

Example:

```
> t.test(fastcar,conf.level=.80)$conf.int
[1] 78.58051 87.64421
```

$\bar{X} \pm 1.28\left(\frac{s}{\sqrt{n}}\right)$

```
> t.test(fastcar,conf.level=.90)$conf.int
[1] 77.27805 88.94667
```

$\bar{X} \pm 1.64\left(\frac{s}{\sqrt{n}}\right)$

```
> t.test(fastcar,conf.level=.95)$conf.int
[1] 76.13763 90.08709
```

$\bar{X} \pm 1.96\left(\frac{s}{\sqrt{n}}\right)$

```
> t.test(fastcar,conf.level=.99)$conf.int
[1] 73.87190 92.35282
```

$\bar{X} \pm 2.58\left(\frac{s}{\sqrt{n}}\right)$

```
> t.test(fastcar,conf.level=.999)$conf.int
[1] 71.16353 95.06119
```

What happens to the width of the C.I. as confidence increases ?

Margin of Error

- Margin of Error (e): the amount added and subtracted to the point estimate to form the confidence interval

Example: Margin of error for estimating μ, σ known:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \qquad e = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Factors Affecting Margin of Error

$$\boxed{e = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}}$$

- Data variation, σ : e ⬇ as σ ⬇
- Sample size, n : e ⬇ as n ⇧
- Level of confidence, 1 - α : e ⬇ if 1 - α ⬇

What About Small *n*?

- From the Central Limit Theorem and the Standardization Rule we always have (for large n or normal observations)

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

- If *n* is large (> 30) we can **replace σ with s** and still have

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim N(0,1).$$

Technically this is an approximation that gets better as *n* increases.

However, if *n* is small…..

- The CLT works if *n* is large.
- If n is small, replacing σ with s results in **more uncertainty** so the **CLT doesn't exactly hold** and we have a new result:

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$$

The t distribution

Looks like the normal but fatter tails

Formally..

- When σ is not known
- The sample size is less than 30
- The data is approximately normally distributed
- THEN

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$$

This result is only important for small samples…..so don't have small samples ☹

Small sample modification

- Strange but true story…William Gosset was in charge of quality control for Guiness Brewing Company.
- He used very small samples, and constructed CI's using our formula.
- But he discovered he was making mistakes (saying good batches were bad) more like 15% of the time instead of 5%.
- He discovered the "*t*" distribution.

If n is small, replacing σ with s results in **more uncertainty** so

$$\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}.$$

The t distribution

Looks like the normal but fatter tails

The $(1\text{-}\alpha)\bullet 100\%$ C.I. for μ is then given by

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

defined similarly to $z_{\alpha/2}$ for the N(0,1)

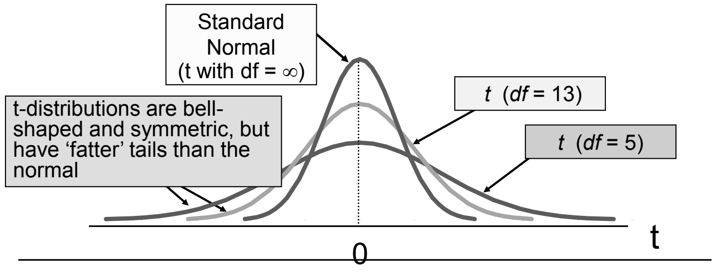
The *t* Distribution

- The t distribution looks like the N(0,1) distribution except it has **fatter tails**.
- It is centered at zero and **defined by its *degrees of freedom* which equal n-1.**
- As the sample size n gets large, the t distribution looks like the N(0,1) distribution.

$$t_{n-1} \xrightarrow{n\to\infty} N(0,1)$$

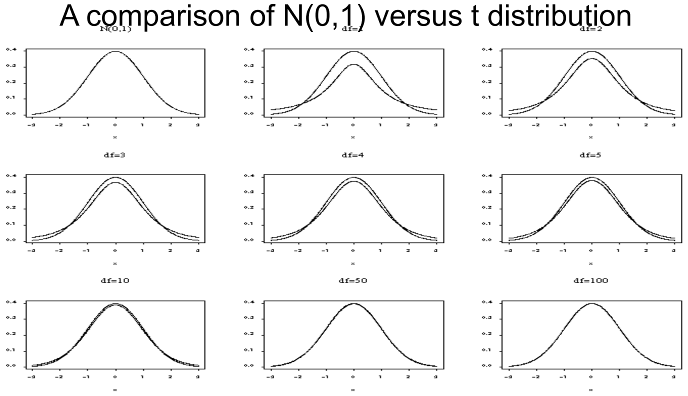
Student's t Distribution

Note: t ⟶ z as n increases



t-distributions are bell-shaped and symmetric, but have 'fatter' tails than the normal

A comparison of N(0,1) versus t distribution



Summary for CI for mean, small n

Recall : If you have a **"small" sample**...

Replace the 1.96 value with a **t value** to get:

$$\bar{x} \pm t\left(\frac{s}{\sqrt{n}}\right)$$

All we are doing is pumping up the volume

where "**t**" comes from **Student's t distribution**, and depends on the sample size through the **degrees of freedom "n-1".**

Critical Values of Student's *t*

For 90% ci's

For 99% ci's

	$\alpha/2$				
d. f.	0.100	0.050	0.025	0.010	0.005
1	3.078	6.314	12.706	31.821	63.656
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763

For the 95% Confidence Interval

n-1

In R
`qt(.975,df)`

Example

So if n=2 and you want to compute a 95% confidence interval (you doofus), it would be:

$$\bar{x} \pm 12.706\left(\frac{s}{\sqrt{n}}\right)$$

But if n=29 and you want to compute a 95% confidence interval it would be:

$$\bar{x} \pm 2.048\left(\frac{s}{\sqrt{n}}\right)$$

Example: CI for mean, small sample

- Random sample of 15 students slept an average of 6.4 hours last night with standard deviation of 1 hour.
- What is the average amount all students slept last night?

Need t with n-1 = 15-1 = 14 d.f.

For 95% confidence, t_{14} = 2.145

```
> qt(.975,14)
[1] 2.144787
```

$$\bar{x} \pm t\left(\frac{s}{\sqrt{n}}\right) = 6.4 \pm 2.145\left(\frac{1}{\sqrt{15}}\right) = 6.4 \pm 0.55$$

This is exactly what the tsum.test() or t.test() command is doing

Note

- Since we are usually always replacing σ by s, it turns out that for any sample size
$$\frac{\overline{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}.$$
- BUT we usually don't care about this as long a *n* is large (bigger than 30).

Note about the "*t*"

- **R always uses the t-distribution when computing confidence intervals**.
- However, **we will always use "1.96"** for 95% CI's when *n* is large in the notes and HW.
- For large *n*, there wont be too much of a difference.

Caution

- One requirement (often overlooked in practice) is that to construct the t confidence interval your data needs to be (approximately) normally distributed.
- So one should do some sort of check of their data to ensure it is somewhat normally distributed before the confidence interval is constructed (and this can be difficult to ascertain in small samples).

However, this method is robust

- How problematic is it if we use the *t*- confidence interval even if the population distribution is not normal?
- For large random samples, it's not problematic
- The Central Limit Theorem applies: for large *n*, the sampling distribution is bell-shaped even when the population is not

The *t* method is robust

- What about a confidence interval using the *t*-distribution when n is small?
- Even if the population distribution is not normal, confidence intervals using *t*-scores usually work quite well
- We say the *t*-distribution is a *robust method* in terms of the normality assumption

When does the *t* method not work?

- With binary data
- With data that contain extreme outliers
- You need to be extra cautious to look for extreme outliers or great departures from the normal population assumption **when *n* is small.**

Example

❑: A simple random sample of 35 men yields a mean pulse rate of 72.5 beats per minute (bpm) and a standard deviation of 10.2 bpm.

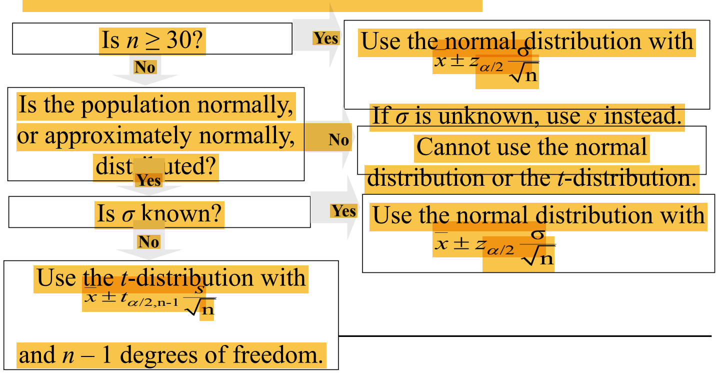
❑ Find the 95% confidence interval estimate for the mean pulse rate of all men.

```
> tsum.test(n.x=35,mean.x=72.5,s.x=10.2)

        One-sample t-Test

data:  Summarized x
t = 42.051, df = 34, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 68.99618 76.00382
```

Assumptions

- As usual, we assume our data is independent and we took a random sample.
- Ideally the sample size should be no more than 10% of the population.
- If the sample size is small, the data needs to be mound shaped'ish, or at least not extremely skewed.
- For larger data sets the skewness doesn't matter because of the Central Limit Theorem.

Normal or t-Distribution?



Is $n \geq 30$?

Yes

Use the normal distribution with $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

If σ is unknown, use s instead.

No

Is the population normally, or approximately normally, distributed?

No

Cannot use the normal distribution or the t-distribution.

Yes

Is σ known?

Yes

Use the normal distribution with $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

No

Use the t-distribution with $\bar{x} \pm t_{\alpha/2,\text{n-1}} \frac{s}{\sqrt{n}}$

and $n - 1$ degrees of freedom.

Things you should know

❑How to construct a confidence interval for the mean

Correct formula

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

Approximation for large n

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

For HW and Exams

$$\bar{x} \pm 1.96 \frac{s}{\sqrt{n}}$$