Chapter 5: Inference about a population mean/median

1. Point estimation of a mean
2. Standard error of the estimated mean
3. Confidence interval for a mean
4. Statistical tests for the population mean
5. Type I and Type II errors and power of a test
6. The level of significance (p-values)
7. One-sided tests
8. Checking the assumption of normality
9. Sample size and power calculations
10. Robustness of t-tests and t-intervals
11. Bootstrap confidence interval
12. Inference about the population median

Chapter 5 Examples:

1. One-sample t-test and Confidence Interval
2. Confidence Interval Simulation
3. Power for a One-sample t-test
4. Bootstrap CI for Single Mean
5. Sign test

Cattle Example:
We are interested in the level of a particular hormone in the meat for a large herd of cattle. We are specifically interested in estimating the population mean (μ).

Let Y be the measured hormone level in the meat of a randomly sampled animal. Y is a random variable.

Sample n = 20 values: $y_1, y_2, \ldots, y_n$

Let: $\bar{y}$ = sample mean = 14.62 units
s = sample std. dev.= 2.73 units
μ = pop. mean (unknown)
σ = pop. std. dev. (unknown)

Recall that the sample mean $\bar{Y}$ is also a random variable. Based on the sampling distribution results (from Ch 4), we know that $\bar{Y}$ has mean = μ and standard deviation = $\sigma/\sqrt{n}$.

Objectives in the motivating example:

1. Estimate of μ

2. Standard Error (SE)

3. Confidence Interval

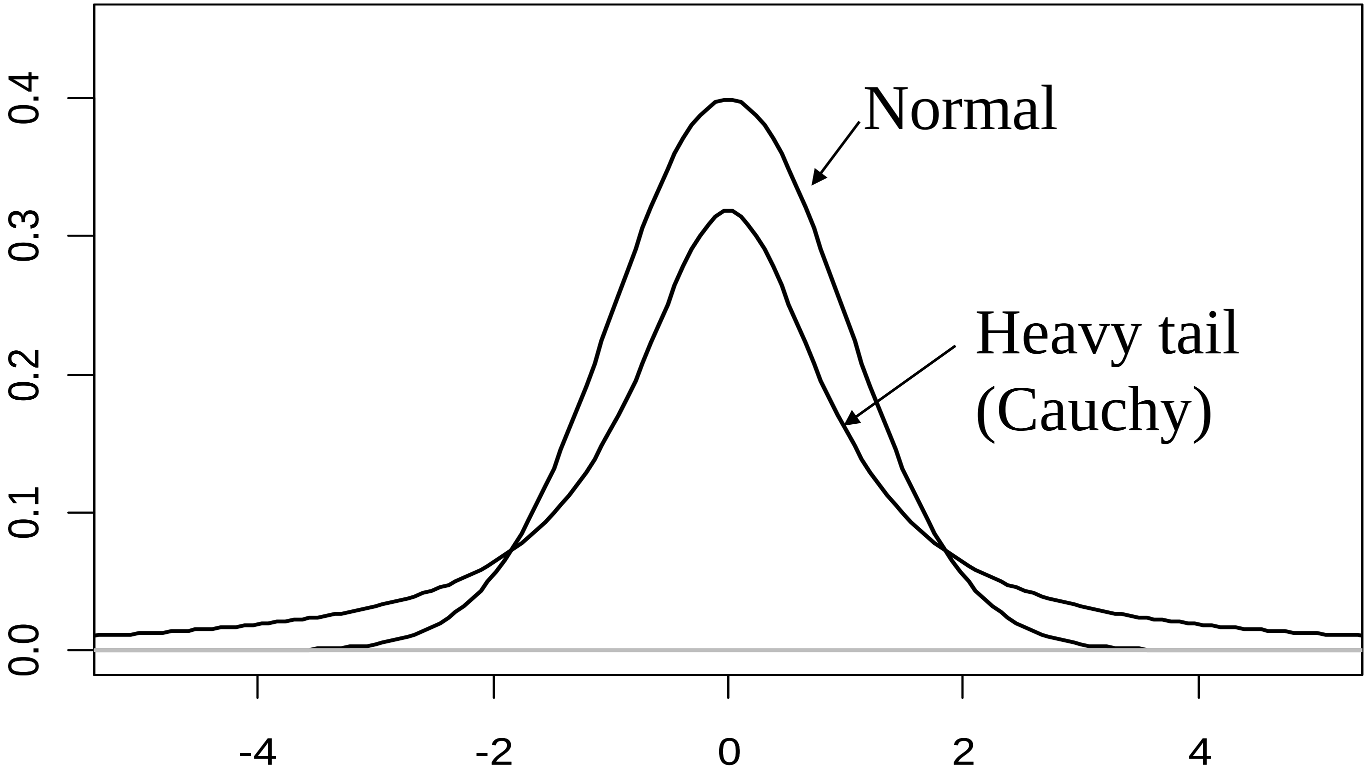
4. Hypothesis Testing

5. Plan for sampling: Power and Sample Size calculations

1. Estimate of population mean (μ)

$\bar{y}$ **(sample mean)** is almost always used.

When is it not used?



Use medians (or trimmed means) to estimate the mean of **heavy-tailed symmetric** distributions.

2. Standard Error (SE)

Standard Error (SE) is an indication of the precision of a (point) estimate. Often, an estimate (for example a sample mean) is presented along with a corresponding SE. Later in these notes we will see that the SE is used in the calculation of test statistics and confidence intervals.

The standard error is an estimate of the standard deviation of the sampling distribution of the estimate. Recall that the standard deviation of the sample mean is ${}^{\sigma}/_{\sqrt{n}}$.

The standard error of the mean is:
SE = SEM = ${}^{s}/_{\sqrt{n}}$

Roughly 68% of the time the sample mean will be within one SE of the true population mean μ (based on Empirical Rule).

Cattle Example:

In the cattle herd example a sample of n = 20 is taken. Recall:

$$\bar{y} = 14.62 \quad \text{and} \quad s = 2.73$$

$$\text{Then SE} = {}^{s}\!\!\diagup\!{}_{\sqrt{n}} = {}^{2.73}\!\!\diagup\!{}_{\sqrt{20}} = 0.61$$

$$14.62 \pm 0.61$$

would be quoted as estimate of μ and an informal statement about its precision.

You will often see the same information (mean and SE) summarized graphically. For example, a bar chart with error bars.

In my opinion, bar charts have a high “ink to information” ratio. Boxplots give more information in the same amount of space!

A common question: "**Should I use the standard deviation (*s*) or standard error (SE) in my summary tables (or graphs)**?"

- Use standard deviation (*s*) when you want to describe the sample.
- Use standard error ($SE = s/\sqrt{n}$)) when you want to describe the accuracy of $\bar{y}$ as an estimator of μ.
- Occasionally, people give the margin of error (ME), described later in these notes.
- In my opinion, it is most important to clearly state which value is presented. (Some articles are unclear whether they have reported *s* or SE.)
- As long as the sample size is given, you can calculate s from SE or vice versa.

3. Confidence Interval for Population Mean (μ)

We will often want to make a conclusion or inference about a population parameter based on a single sample.

One of the most common types of inference is to construct what is called a **confidence interval**, which is defined as
"*an interval of values computed from sample data that is almost sure to cover the true population value*."

During this course, we will be looking at confidence intervals for different population parameters.

It helps to remember that many confidence intervals will follow the same general form:

Estimate ± Table Value x Standard Error
OR Estimate ± Margin of Error

Deriving a Confidence Interval for Population Mean (μ)

Assume (temporarily) that σ is known; then from the Empirical Rule:

$$P\left(\mu - 2\frac{\sigma}{\sqrt{n}} \le \bar{y} \le \mu + 2\frac{\sigma}{\sqrt{n}} \right) \approx 0.95$$

After rearranging the terms using algebra we get

$$P\left(\bar{y} - 2\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{y} + 2\frac{\sigma}{\sqrt{n}} \right) \approx 0.95.$$

This gives,

$$\left(\bar{y} - 2\frac{\sigma}{\sqrt{n}}, \bar{y} + 2\frac{\sigma}{\sqrt{n}} \right)$$

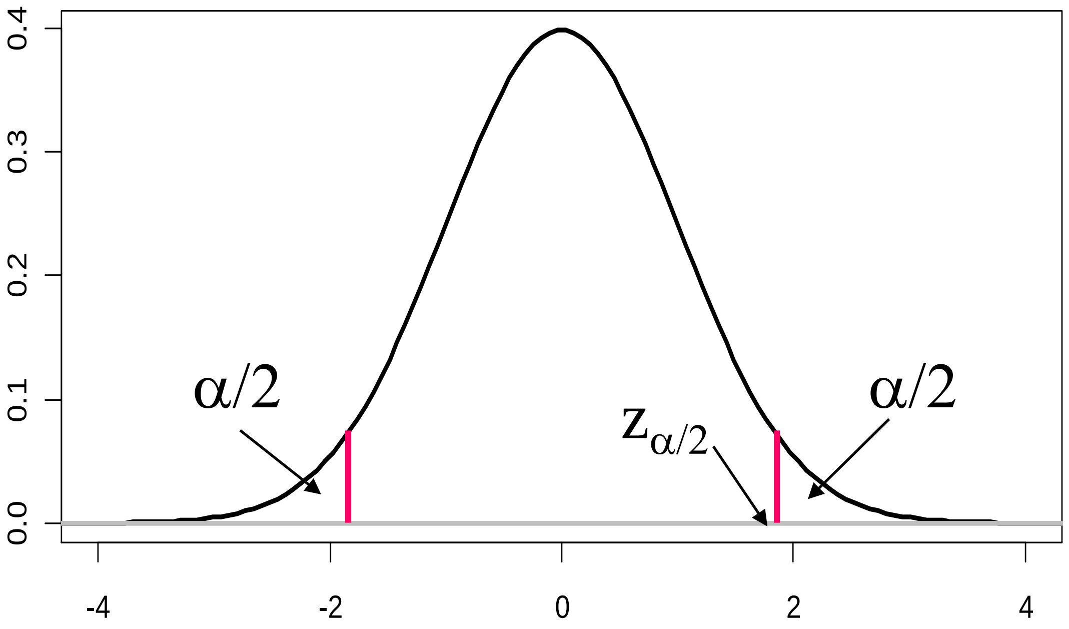
a random interval that will contain the true μ with probability approximately 0.95.

Such an interval is called a **95% confidence interval**.

Note #1: The multiplier 2 in the formula (taken from the Empirical Rule) is only approximate.

Let z_α be the value such that the probability of being greater than z_α is α. For Confidence Intervals, we want the total area on both tails to be α, so we use $z_{\alpha/2}$ as the multiplier.
Some Common values are shown below, additional values in O&L Table 1.

CI%	**α**	**α/2**	**$Z_{\alpha/2}$**	**R code**
99%	0.01	0.005	2.58	qnorm(0.995)
95%	0.05	0.025	1.96	qnorm(0.975)
90%	0.10	0.05	1.645	qnorm(0.950)



Note #2: The previous interval cannot be used for data analysis because it contains σ which cannot be determined from the data.

When we replace σ with its estimate (s) we use Student's t distribution (instead of the normal distribution) to construct the confidence interval.

The exact shape of a Student's t distribution depends on a quantity called degrees of freedom (df) which is related to sample size.

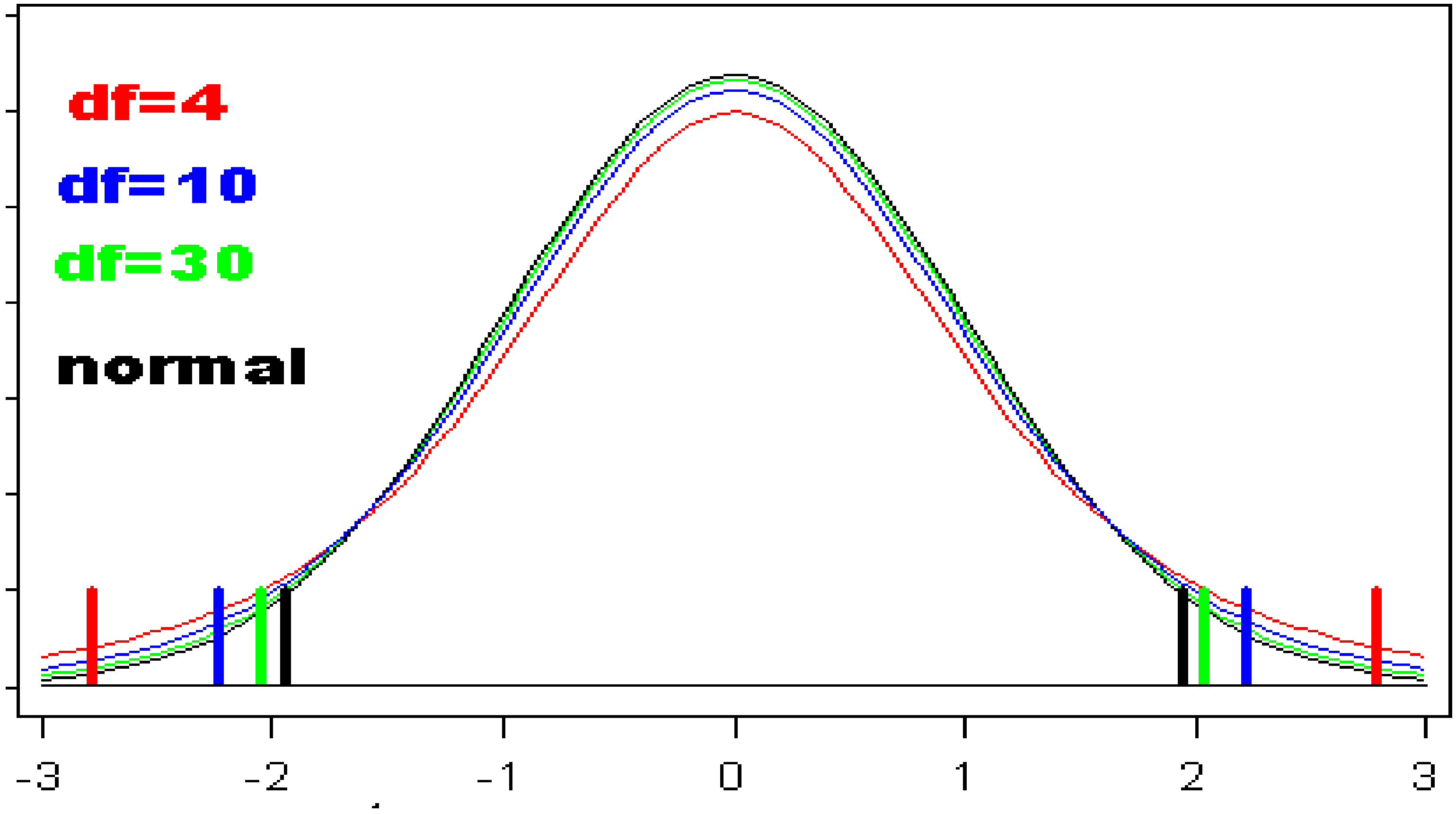
The quantity $t_{\alpha/2}$ value such that the interval between -$t_{\alpha/2}$ and +$t_{\alpha/2}$ contains 100(1-α)% of the area under the curve.

For table values see O&L Table 2.

In R: $t_{\alpha/2}$ is computed using qt((1-alpha/2) ,df)

The *t* distribution is symmetric and bell shaped like the normal curve, but has a larger standard deviation.

As the df increase, the *t*-curves approach the normal curve; thus the normal curve can be regarded as a t curve with infinite df (df=∞).



The lines are mark the edge of a 95% interval for each distribution.

Confidence Interval for μ

The (1-α)100% confidence interval for μ is:

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

where the table value $t_{\alpha/2,n-1}$ is determined from the Student's t-distribution with df = n-1.

Note: Margin of Error = ME = $t_{\alpha/2,n-1} \frac{s}{\sqrt{n}}$

Assumptions: Random sample, independent observations, normally distributed data and/or large sample size.

Interpretation: We can be 95% confident that μ is contained in the 95% confidence interval.

More Detail / **"Frequency" interpretation**: We have approximately 95% confidence that μ is in that interval, because it is an interval calculated by a method such that under repeated sampling approximately 95% of such intervals would include μ.

Return to the **Cattle Example:**

$$\bar{y} = 14.62\ ,\ s{=}2.73,\ \ n = 20$$

$$df\ = n - 1 = 19 \rightarrow t_{0.025} = \mathrm{qt}\left(0.975,\mathrm{df}\ =\ 19\right) = 2.093$$

Then the 95% interval is

$$\left(14.62 - 2.093\frac{2.73}{\sqrt{20}},14.62 + 2.093\frac{2.73}{\sqrt{20}}\right)$$

$$= \left(13.34\ ,\ 15.90\right)$$

Interpretation:

We are 95% confident that the true population mean hormone concentration is between 13.34 and 15.90.

Note #3: All of the formulas assume that the distribution of individual observations is normal. If that is not true, it causes a problem because the distribution of the sample mean is not quite normal, and using Student's t distribution is not quite right.

However, the confidence intervals, even under non-normality, are generally satisfactory if the distribution of Y is not too skewed and/or the sample size is large.

There are statistical procedures (histogram, qqplots, tests of normality) that are used to assess the validity of the normality assumption. These will be discussed later in these notes.

Note #4: The CI formula given here assumes that the sample size is much smaller than the population size (so the population size can be considered to be effectively infinite). A "finite population correction" is available, but not discussed in this class.

Note #5: Confidence vs Prediction vs Tolerance Intervals

This discussion is motivated by a blog from Jim Frost at Minitab.

The **confidence interval** gives a range that is likely to contain the unknown population mean. It does not tell us anything about the distribution of individual values!

A **prediction interval** is a range that is likely to contain the response value of a single new observation. The prediction interval is always wider than the corresponding confidence interval because of the added uncertainty involved in predicting a single response versus mean response.

$$(1-\alpha)100\% \text{ Prediction Interval}: \quad \bar{y} \pm t_{\alpha/2} s\sqrt{1+(1/n)}$$

A **tolerance interval** is a range that is likely to contain a specified proportion of the population. To generate tolerance intervals, you must specify both the proportion of the population and a confidence level.

4. Statistical Hypothesis Tests for μ

A hypothesis test is a formal procedure for comparing observed data with a hypothesis whose truth we want to assess.

Return to the problem of the hormone levels in cattle.
Say the owner of the herd claims that the average value for the herd is 12. That is, he is claiming that $\mu=12$. Do we have evidence against this claim?

The approach taken is to assume the claim is true (called the "**null hypothesis")**, and see if the data are consistent with that assumption, or consistent with some other value for μ (the "**alternative hypothesis"**)

Null Hypothesis H_0: $\mu=12$ (=μ_0 "mu naught")
Alternative Hypothesis H_A: $\mu\neq12$

Notes about Hypotheses

- Hypotheses are statements about population parameters (ex: μ = population mean).
- H_0, H_A, μ_0 are motivated by a specific research question. These can (should!) be specified before looking at the data.
- The **null hypothesis (H_0)** is the claim that is initially assumed to be true. Typically corresponds to the status quo or accepted beliefs.
- The **alternative hypothesis (H_A)** is the assertion that is contradictory to H_0. This hypothesis typically corresponds to discovery or new information or need to take action.
- H_0, H_A cover all possible outcomes.
- Note about the Cattle example: The storyline/motivation for the formal test is weak here. In practice, I think the confidence interval is a better fit for this scenario.

Statistical Hypothesis Test for μ using the confidence interval

If the claim is really true, then we would expect the hypothesized mean (μ_0 = 12) to be within a confidence interval for μ (with some high level of confidence).

We calculated the 95% confidence interval to be (13.34, 15.90)

Because 12 is well outside this interval, we conclude that the data are inconsistent with the owner's claim. We have evidence that $\mu \neq 12$.

Hence, we will **reject** the owners claim, but since our interval will not contain μ about 5% of the time, we must acknowledge that there is a 5% chance, an α=0.05 probability, that our interval will cause us reject by mistake.

Overview of Formal Hypothesis Testing

1. State the null and alternative hypotheses and choose α. This can (should) be done before any data is collected. Hypotheses are based on specific research questions.
2. Collect data, check assumptions, calculate summary statistics and Test Statistic (TS).
3. A. Define the Rejection Region (RR) based on a table value (TV)
   OR
   B. Calculate the p-value
4. Make a decision (Reject H_0 or Fail to Reject H_0) by
   A. Comparing Test Statistic to the Rejection Region
   OR
   B. Comparing p-value to α.

   Then draw conclusions.

NOTE: In STAT511 by “default”, we will use α = 0.05.

Conclusions from a test

The null hypothesis will be rejected in favor of the alternative only if sample evidence suggests that H_0 is false. If the sample does not strongly contradict H_0, we will continue to believe in the truth of the null hypothesis.

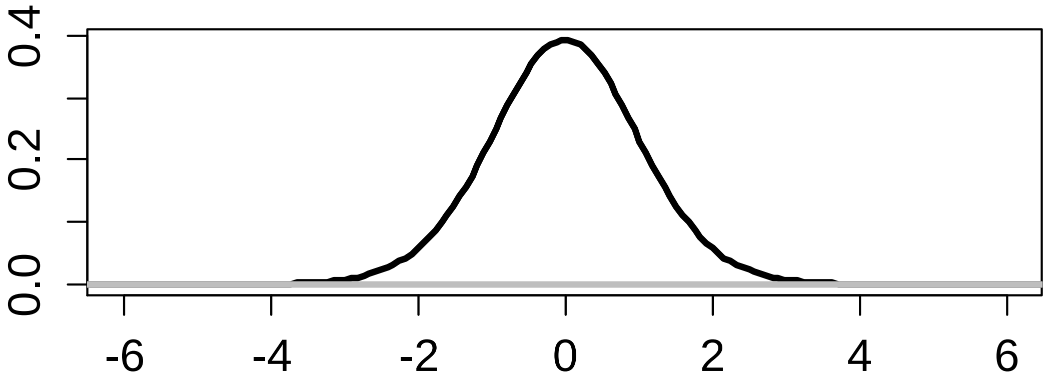
The two possible conclusions from a hypothesis-testing analysis are **Reject $\mathbf{H_0}$** or **Fail to reject $\mathbf{H_0}$**.

A formal hypothesis test for μ ("two-sided" alternative)

$$H_0: \mu = \mu_0 \qquad H_A: \mu \neq \mu_0$$

$$\text{TS}: t = \frac{\bar{y} - \mu_0}{(s/\sqrt{n})}$$

$$\text{RR}: \text{Reject } H_0 \text{ if } |t| > t_{\alpha/2, n-1}$$



For the **Cattle example**:

1. **State Hypotheses**: H_0: μ=12 (=μ_0) vs H_A: μ≠12

2. **Test Statistic (TS)**:
$$t = \frac{\bar{y} - \mu_0}{(s/\sqrt{n})} = \frac{14.62 - 12}{(2.73/\sqrt{20})} = 4.29$$

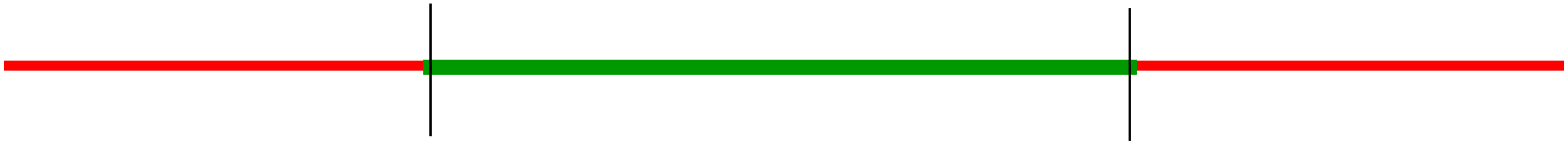
Interpretation: $\bar{y}$ is 4.29 SE's away from the hypothesized value.

3. **Define Rejection Region (RR)**: n=20 → $t_{\alpha/2,n-1}$= 2.093. So, reject H_0 if |t| > 2.093.

4. **Conclusion**: Since t = 4.29 > 2.093, we Reject H_0. We conclude that the true population mean hormone concentration is different from 12.

Comments about hypothesis testing

1. Usually hypothesis tests are set up so that the thing we want to prove is the alternative hypothesis.
2. Important: Failure to reject H_0 does not mean we really believe it is true.
3. Most research problems do not really require a decision. They require an estimate and an indication of its accuracy. **A Confidence Interval may be a better answer than a hypothesis test**.
4. The values of μ that are outside the 95% CI are exactly those values that would be rejected in the α=0.05 (2-sided) hypothesis test. Thus, the Confidence Interval can be used to learn the result of all hypothesis tests simultaneously (fixed α, any μ_0).



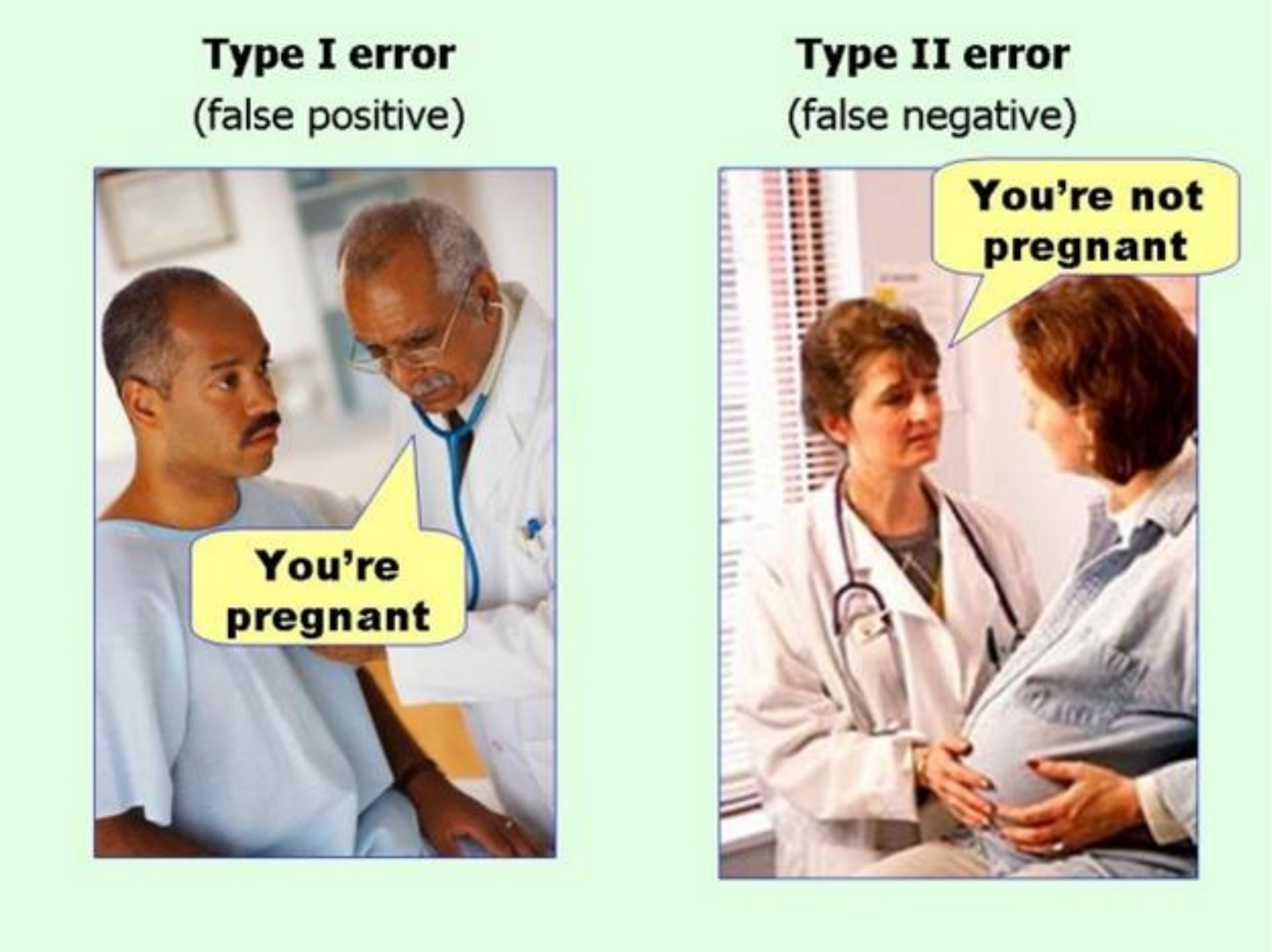
5. Classic example: The courtroom

   H_0: defendant is innocent, H_A: defendant is guilty

5. Type I and Type II Errors and Power

	Truth	
Decision	H_0 True	H_0 False
Reject H_0	P(Type 1 Error) = α	P(Correct) = 1- β = Power
Fail to Reject H_0	P(Correct) = 1-α	P(Type II Error) = β

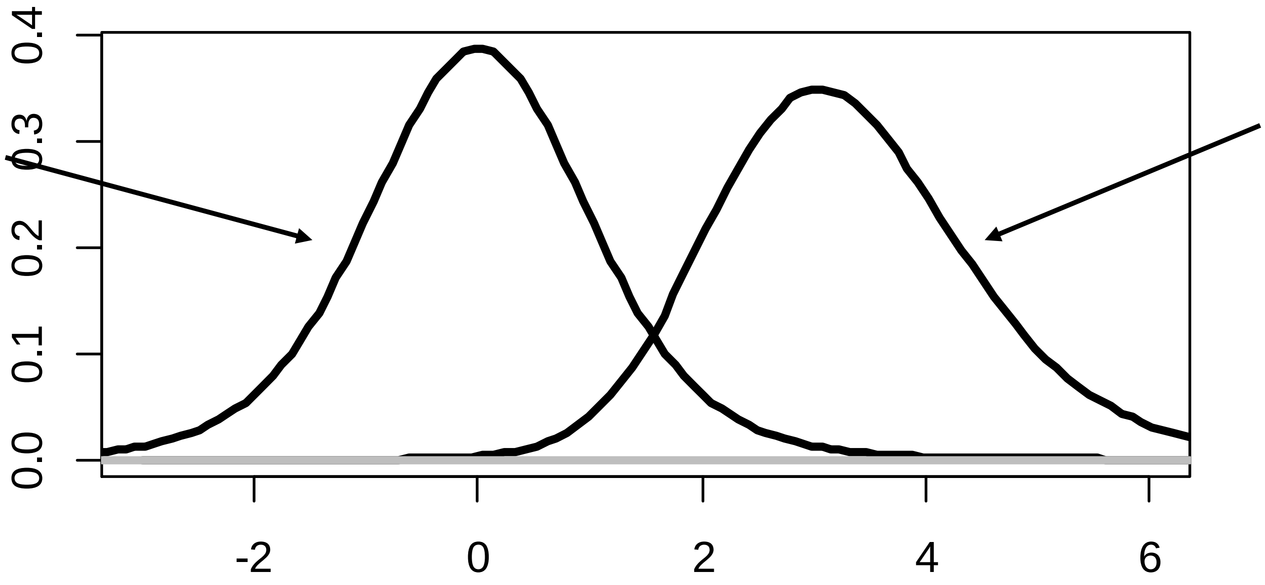
- **Type I error** is the error of rejecting H_0, when H_0 is true. We denote P(Type I error) = α. False positive.
- **Type II error** is the error of not rejecting H_0, when H_0 is false. We denote P(Type II error) = β. False negative.
- **Power** is the probability of rejecting H_0, when H_0 is in fact false. Power = 1-β.
- Common Procedure: Control α at a very small value (0.05 is the typical, but 0.01 and 0.10 are also used), and let β fall where it may.



H_0: Not Pregnant vs H_A: Pregnant

Graphical representation of Type I Error and Power

Distbn of t
when H_0
is true



Distbn of t
when H_A
is true

For the **Cattle example**:

H_0: μ=12 vs H_A: μ≠12

Reject H_0 if |t| > 2.093.

1. P(Type I Error) = α is the probability rejecting H_0, when H_0 is true.
2. Power (= 1 – β) is the probability of rejecting H_0, when H_0 is in false (hence H_A is true). In order to calculate power we need to conjecture a specific alternative. For example: μ = 15. More on power later in these notes.

6. The level of significance (P-values)

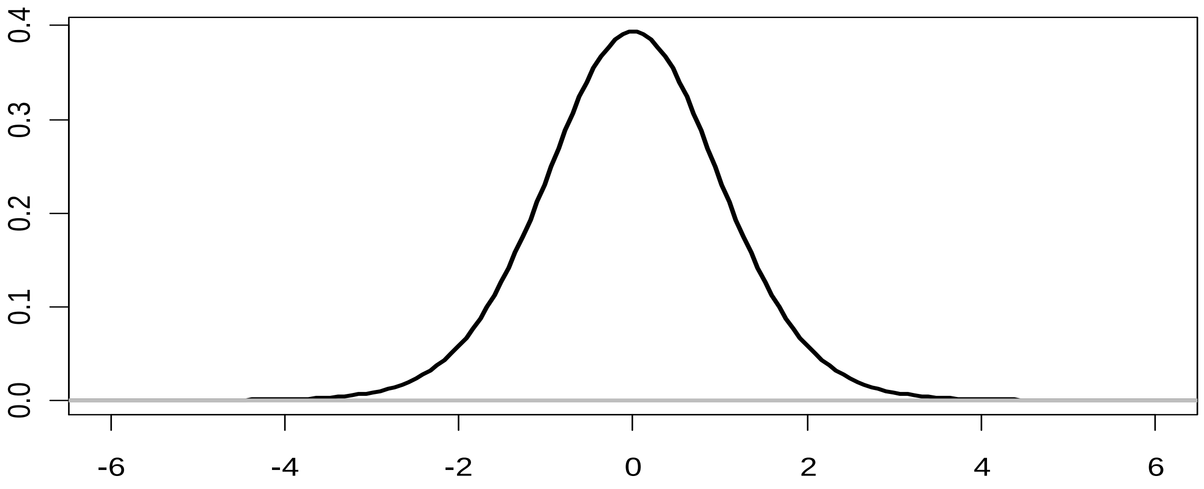
Rather than set α, better to summarize the evidence against H_0 with a "p-value" (significance probability).

p-value is probability of observing a value of the test statistic **as or more supportive of H_A** than the actual observed value**, given H_0 is true.**

If p-value is small enough, we conclude that it would have been very unusual to observe such an extreme value of the test statistic if the null hypothesis was true.

Hence, small p-values support the alterative hypothesis.

To make a decision, just compare p-value to α.
Reject H_0 if p-value $< \alpha$
Fail to Reject H_0 if p-value $\geq \alpha$



Return to the **Cattle example** ($\alpha = 0.05$, n = 20, df = 19):

1. **Hypotheses:** H_0: $\mu=12$ vs H_A: $\mu\neq12$
2. **Test Statistic:** t = +4.29
3. **P-value:**

   area under curve outside the interval (–t, t) (for 2-sided tests)

   `=2*(1-pt(abs(t),n-1))`

   `=2*(1-pt(4.29,df = 19)) = 0.0004`
4. **Conclusion:**

   Since p-value = 0.0004 < α = 0.05, we Reject H_0. We conclude that the true population mean hormone concentration is different from 12.

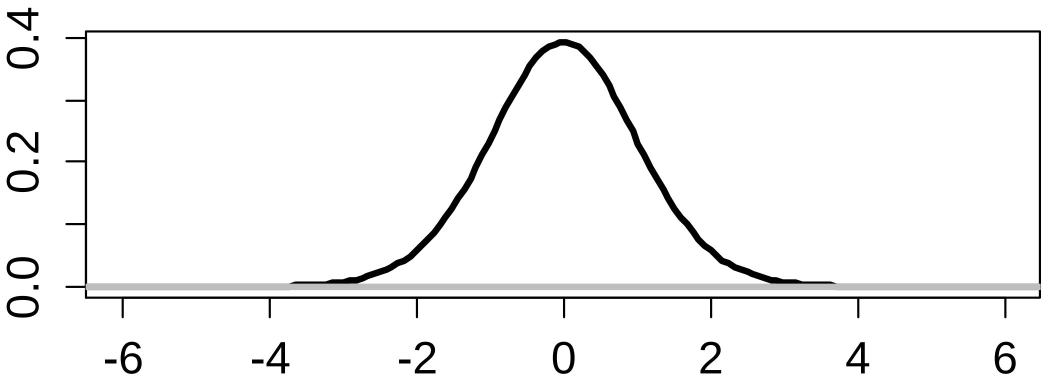
Notes about p-values

1. Since p-values are probabilities, they will take values between 0 and 1!
2. p-values are generally preferred over a rejection rule. By presenting a p-value (instead of using the rejection rule), we allow the reader to consider their own α. α depends on the seriousness of making a Type I error, which may depend on the user of the results.
3. In many scenarios, an appropriate confidence interval will provide the needed information for making a decision. Confidence intervals are an alternative to hypothesis testing.
4. NEVER report just a p-value! Estimate, SE and sample size are required for the reader to make sense of the results.
5. We will discuss ASA guidance about p-values in the CH6 notes.

7. One-sided tests

Return to **Cattle example**:
Say our research hypothesis is that μ > 12.

$$H_0 : \mu \leq 12$$
$$H_A : \mu > 12$$



Then the **Rejection Region is only on the side supporting H_A**.
The P-value takes **area only in the direction that supports H_A**.

Test Statistic t= +4.29 (same as two-sided case!).
Rejection Region: Reject H_o if $t > t_\alpha$ (Don't divide α by 2.)
Conclusion: Since 4.29 > 1.73, Reject H_o.

Critical value in R: `qt(1-alpha,df = 19)` (result is t_α= 1.73)
P-value in R: `1-pt(4.29,df = 19)` (result is p = 0.0002)

Note: Be careful if sample mean is in the opposite direction of the alternative.

One-Sample t-test for Single Population Mean (μ)

Assumptions: Random sample, independent observations**,** normally distributed data and/or large sample size.

Test Statistic:

$$t = \frac{\bar{y} - \mu_0}{s / \sqrt{n}}$$

Alternative Hypothesis:	**Rejection Region:**
(1) H_A: $\mu > \mu_0$	$t \geq +t_{\alpha,\, n-1}$
(2) H_A: $\mu < \mu_0$	$t \leq -t_{\alpha,\, n-1}$
(3) H_A: $\mu \neq \mu_0$	$\lvert t \rvert \geq +t_{\alpha/2,\, n-1}$

P-values: (1) area to the right of t (test statistic), (2) area to the left, (3) double the area to the right of |t|.

p-values for one and two-sided tests

For this example, the test statistic t = +2.

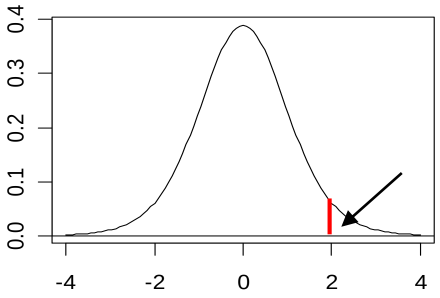
Alternative Hypothesis:

P-value:

Greater Than

$(H_A : \mu > \mu_0)$

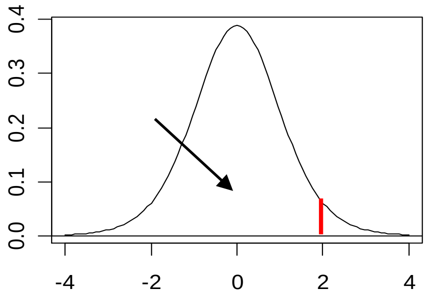
Area to the right of t



Less Than

$(H_A : \mu < \mu_0)$

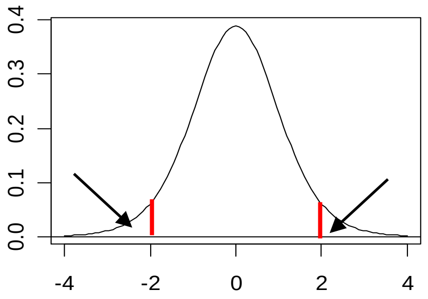
Area to the left of t



Not Equal

$(H_A : \mu \neq \mu_0)$

Area beyond –t and +t



One-sample t-test and CI using R/Rcmdr

- In R, use the function t.test().
- In Rcmdr, choose Statistics -> Means -> Single-Sample t-test.

- For the Exercise Example: (See “**One-Sample t-test**”)

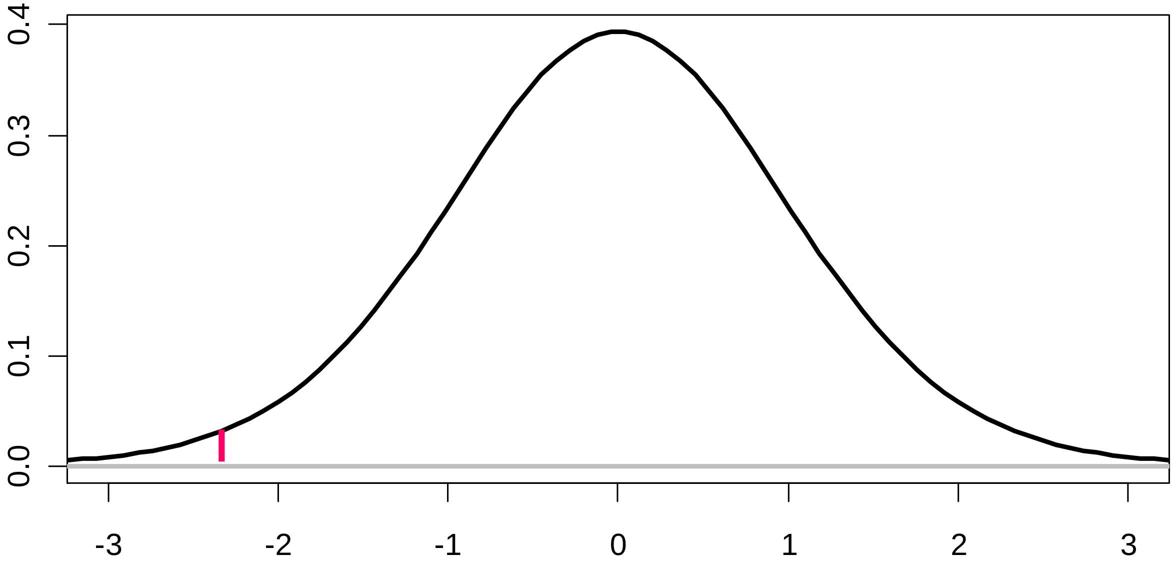
```
> t.test(exercise$y ,mu=30)

        One Sample t-test
data:  exercise$y
t = 0.2462, df = 34, p-value = 0.807
alternative hypothesis: true mean is not
equal to 30
95 percent confidence interval:
 26.26906 34.75951
sample estimates:
mean of x
 30.51429
```

One sided test Example: An EPA inspector took samples from a stream to determine if the mean level of a contaminant is above the allowed maximum of 10. If the mean level can be shown to be above 10, then they are required to start remediation procedures. Hence,

$$H_0 : \mu \leq 10$$
$$H_A : \mu > 10$$

Then a rejection would establish that the mean level is too high. The t-value and a two-sided p-value given by R were t = –2.45 and p-value = 0.0321. What is the appropriate one-sided p-value? Hint: Was $\bar{y}$ > or < 10?



P = 1-(0.0321)/2 = 1-0.0165 = 0.9835

Fail to reject H_0.

Notes about One-sided tests

1. **Two-sided tests are the “default”, one-sided tests should not be used unless there is some compelling reason.**
2. By default, programs (including R) will return two-sided p-values, but will give you the option to calculate a one-sided p-value. (See “**One-Sample T-test**”.)
3. Given the two-sided p-value and test statistic, you can also calculate the appropriate one-sided p-value.
   - It will always work to sketch the picture and mark the test statistic and p-value.
   - If the test statistic (or just estimated value) supports H_A, then one-sided p-value = (two-sided p-value)/2
   - If the test statistic (or just the estimated value) is on the opposite side relative to the side that supports H_A, then
     one-sided p-value = 1 – (two-sided p-value)/2

Note about one-sided Confidence "Intervals"

- In R (and some other software packages), if you ask for a one-sided test you will get a one-sided CI.
- For this class, when I ask for a confidence interval I mean a "standard" two-sided CI (unless specifically stated otherwise).
- Want both a one-sided test and a "standard" two-sided CI? Just run the t.test() function twice!
- Interpretation of one-sided lower limit (from "**One-Sample t-test**" Example): We can be 95% confident that the population mean is greater than or equal to 26.98.

One or Two SAMPLE Tests vs One or Two SIDED Tests

Some clarification…..

- In CH5, One-SAMPLE t-test refers to inference on a single mean (μ).

- In CH6, we will discuss the two-SAMPLE t-test for comparing two means (μ_1, μ_2) from independent samples.

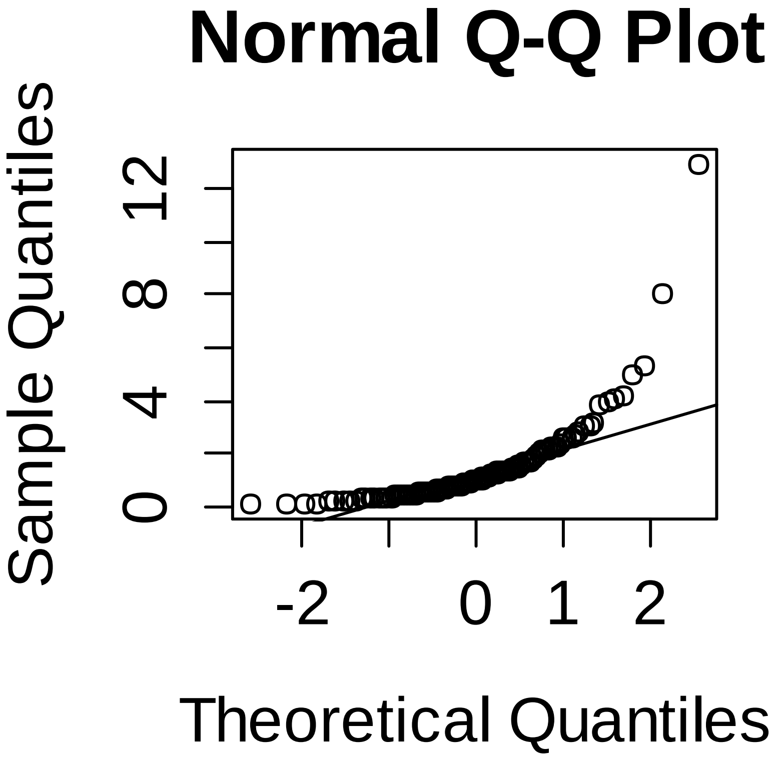
- One or Two SIDED test refers to the form of H_A.

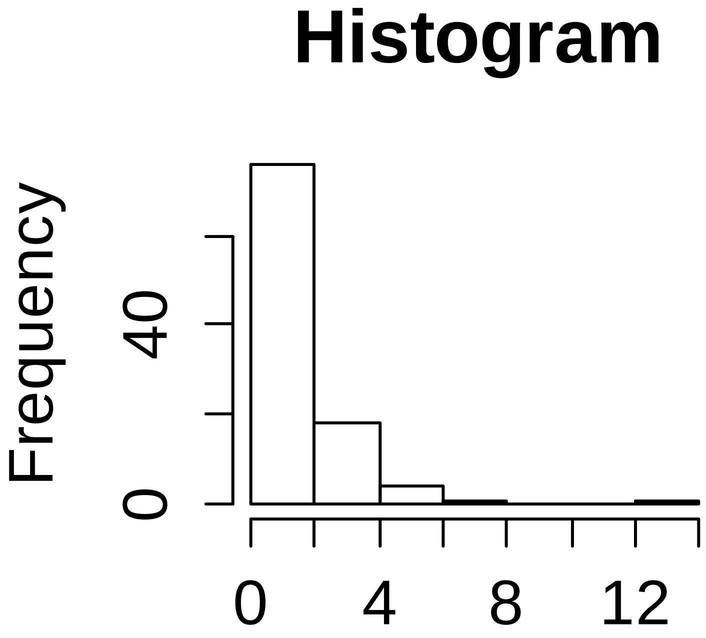
- We can have any combination of a 1 or 2 SAMPLE and 1 or 2 SIDED tests!

8. Checking Normality

A **Q-Q plot or Quantile-Quantile plot** is a common graphical way to check data for non-normality. Quantile is another term for percentile. A Q-Q plot is a plot of the quantiles of a data set versus the quantiles of a reference theoretical distribution.

If the plot is a straight line, it supports the idea that the data came from that distribution.





Testing for Normality in R

QQplots can be generated in R using qqnorm() function.

Two options are available in R for testing for normality:

1. Shapiro-Wilks Test - shapiro.test()
2. Kolmogorov-Smirnoff Test - ks.test()

See "**One Sample t-test**" example.

Notes about tests of normality:

1. **Both tests (SW, KS) use the H_0: data are normally distributed.** So, small p-values indicate that we should reject H_0 and conclude that the data are not normally distributed.
2. Histograms and QQ plots are usually more informative than the tests, because small sample sizes generally "pass" the test (high p-value, no evidence against normality), and large sample sizes generally "fail" (small p-value, evidence against normality).

CI Simulation "Example"

This is not a basic data analysis example! However, use of do (from the dplyr package) can be useful for practical data analysis.

In practice, we don't know the value of the true population mean (μ). So, we don't know whether (1) our CI has captured the true value or (2) whether we have made the correct decision in the hypothesis test.

Simulation gives us the ability to generate data where the truth is known and see how a method(s) performs.

For this simulation, we use rnorm() to generate random observations from the standard normal distribution (with mean = 0 and standard deviation = 1). Specifically, we work with 1000 samples (SampleID goes from 1 to 1000), each of size n = 25 where we know that μ = 0.

Then for each SampleID (with n=25 observations), we use t.test to (1) calculate the 95% CI and (2) test H0: μ = 0. Using α = 0.05, any p-value < 0.05 represents a false rejection (type I error).

CI Simulation Results

1. $952/1000 = 0.952 \approx 95\%$ of the confidence intervals include $\mu = 0$ (the true population mean). This is expected for 95% confidence intervals!
2. We find that that $48/1000 = 0.048 \approx 5\%$ of the p-values are less than 0.05. This is expected when using $\alpha = 0.05$. (Still a 5% chance of a type I error or false positive.)
3. The Sample IDs that yield a CI not including $\mu = 0$ are the same as the samples that have a p-value < 0.05. This is expected because the CI and hypothesis test will give the same conclusion for the two-sided test.
4. We find that the observed t test statistics follow a t-distribution with $df = 24$. This is the expected distribution “under the null hypothesis”.

9. Sample Size and Power Calculations

- These calculations are done during experiment planning before any data has been collected. We want to determine a reasonable sample size for the study so that we can achieve our research goals.
- Calculations are based on conjectures. Coming up with reasonable conjectures is often the hardest part!
- Sample size justification should match your planned analysis (and hence your research goals).
- Power corresponds to a hypothesis test. Recall that power is the probability of rejecting H_0 , given H_A is true.

We will consider several cases:

1. Find the n (sample size) required so that the expected width of a $100(1-\alpha)\%$ CI is approximately 2E (or the $100(1-\alpha)\%$ ME = E).
2. Use R to compute power for one-sided t-test.
3. Use R to compute power for two-sided t-test.
4. Use Lenth’s on-line power calculator for any of the calculations above.

“Power” Case 1: Find the n required so that the expected width of a 100(1-α)% Confidence Interval is approximately 2E (or the 100(1-α)% ME = E)

A 100(1-α)% CI is of the form:

$$\bar{y} \pm ME \quad \text{where} \quad \text{ME} = t_{\alpha/2,n-1} \frac{s}{\sqrt{n}}$$

So the formula for ME depends on n and s.
We will use a conjectured value for s, then try different values of n.

Two Approaches:
A. Use R (or Lenth!) to calculate ME for a range of n values.
B. Iteratively solve for n: Since $t_{\alpha/2}$ depends on n, we need to (1) use starting value for $t_{\alpha/2.}$ and plug into the calculation then (2) update/repeat the calculation with the updated value of $t_{\alpha/2}$ to find n.

$$n = \frac{\left(t_{\alpha/2,n-1}\right)^2 s^2}{E^2}$$

“Power” Case 1A Example

You want a 95% confidence interval of width less than 6 mm (and hence ME < 3 mm), and you conjecture that σ = 4mm.

Use R to try values of n between 5 and 15 with s=4 (see **“Power One-Sample t-test”**).

Based on these results, a value of n=10 will result in a 95% ME < 3
(or a total CI width < 6)

n	ME
5	4.967
6	4.198
7	3.699
8	3.344
9	3.075
10	2.861
11	2.687
12	2.541
13	2.417
14	2.31
15	2.215

“Power” Case 1B Example

You want a 95% confidence interval of width less than 6 mm (E = 3 mm), and you conjecture that σ = 4mm.

1. Take **initially** $t_{\alpha/2}$ = 2. This gives,

$$n = \frac{(2)^2 4^2}{3^2} = 8 \text{ (rounded up)}$$

2. With a ballpark estimate of n, we can now update $t_{\alpha/2.}$
df = n-1= 8-1 = 7. $t_{\alpha/2}$ =2.365.

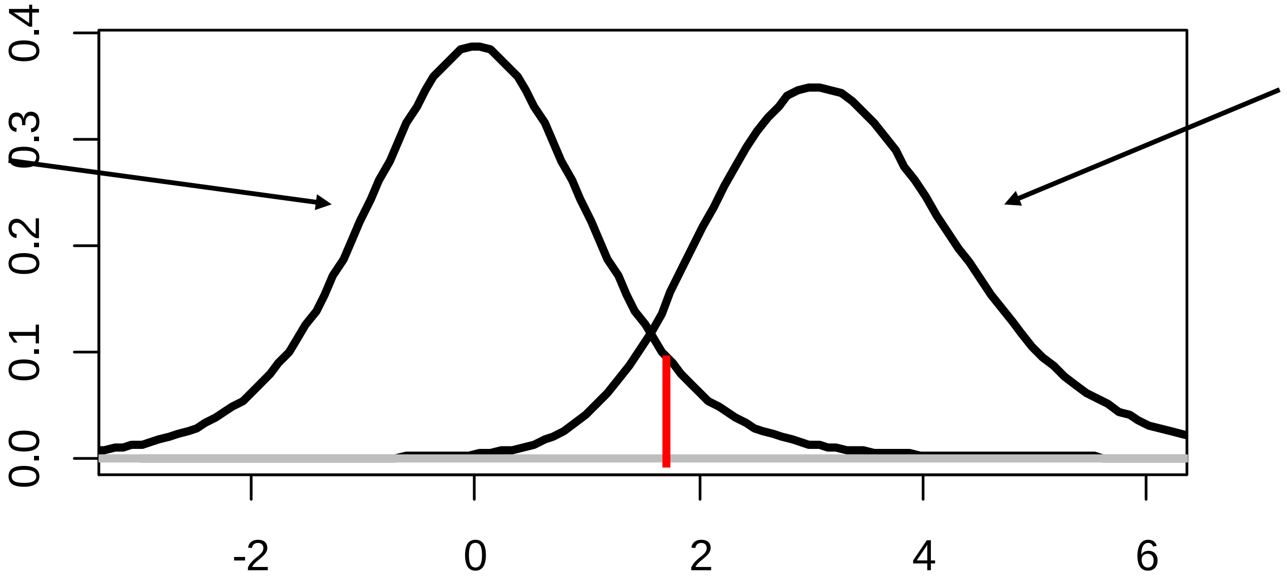
$$n = \frac{(2.365)^2 4^2}{3^2} = 9.9 \cong 10$$

3. Check the resulting value based on n=10. In this case, when n=10 E= 2.861 < 3 -> OK!

Power Case 2: Use R to compute power for one-sided t-tests.
Recall that power is the probability of rejecting H_0 , given H_A is true.

We need to make conjectures about true **μ** (under H_A) and **σ.**
(Coming up with reasonable conjectures is often the hardest part!)

Example: α=0.05, n=10, df = 9
H_0: $\mu \leq 12$ $(=\mu_0)$ vs H_A: $\mu > 12$
Rejection Region: We will reject H0 if $t > t_\alpha = 1.8333$
Conjectures: $\mu_A = 16$, $\sigma = 4$
How do we compute power for this test?



Distbn of t when H_0 is true (μ=12)

Distbn of t when H_A is true (μ=16)

Power Case 2 (one-sided tests) continued

If HA is true, then the distribution of t is not centered at zero; it is "non-central" with "noncentrality" parameter given by

$$\lambda = \frac{\mu_{\mathrm{A}} - \mu_0}{\sigma / \sqrt{n}}$$

(**Note:** Some authors define this with an additional 2 in the denominator of λ. We follow R notation which does not have a 2 in the denominator)

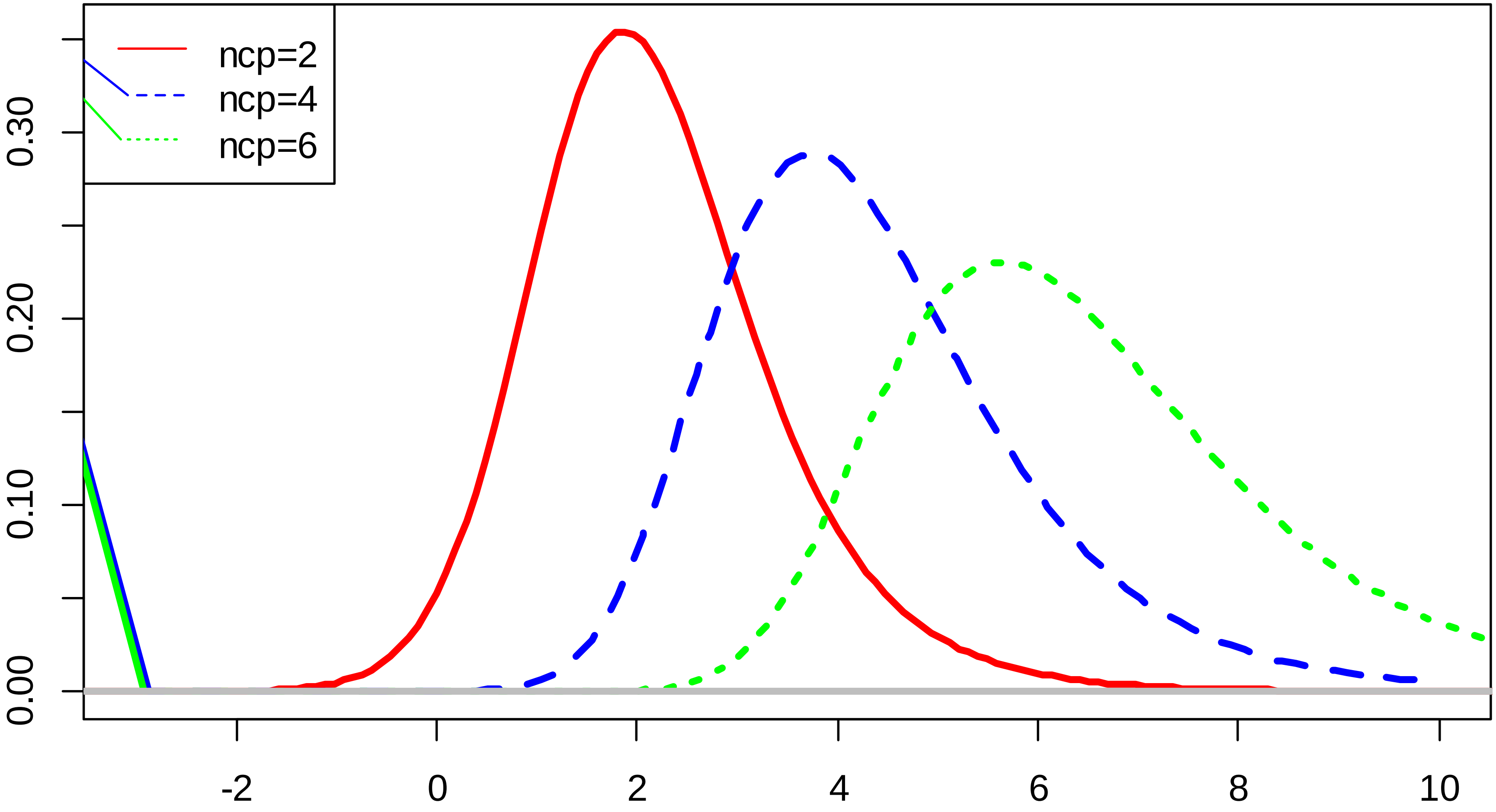
For the example, the value of λ is $\lambda = \frac{16-12}{4.0/\sqrt{10}} = 3.16$

Power can then be computed in R using the function **pt**

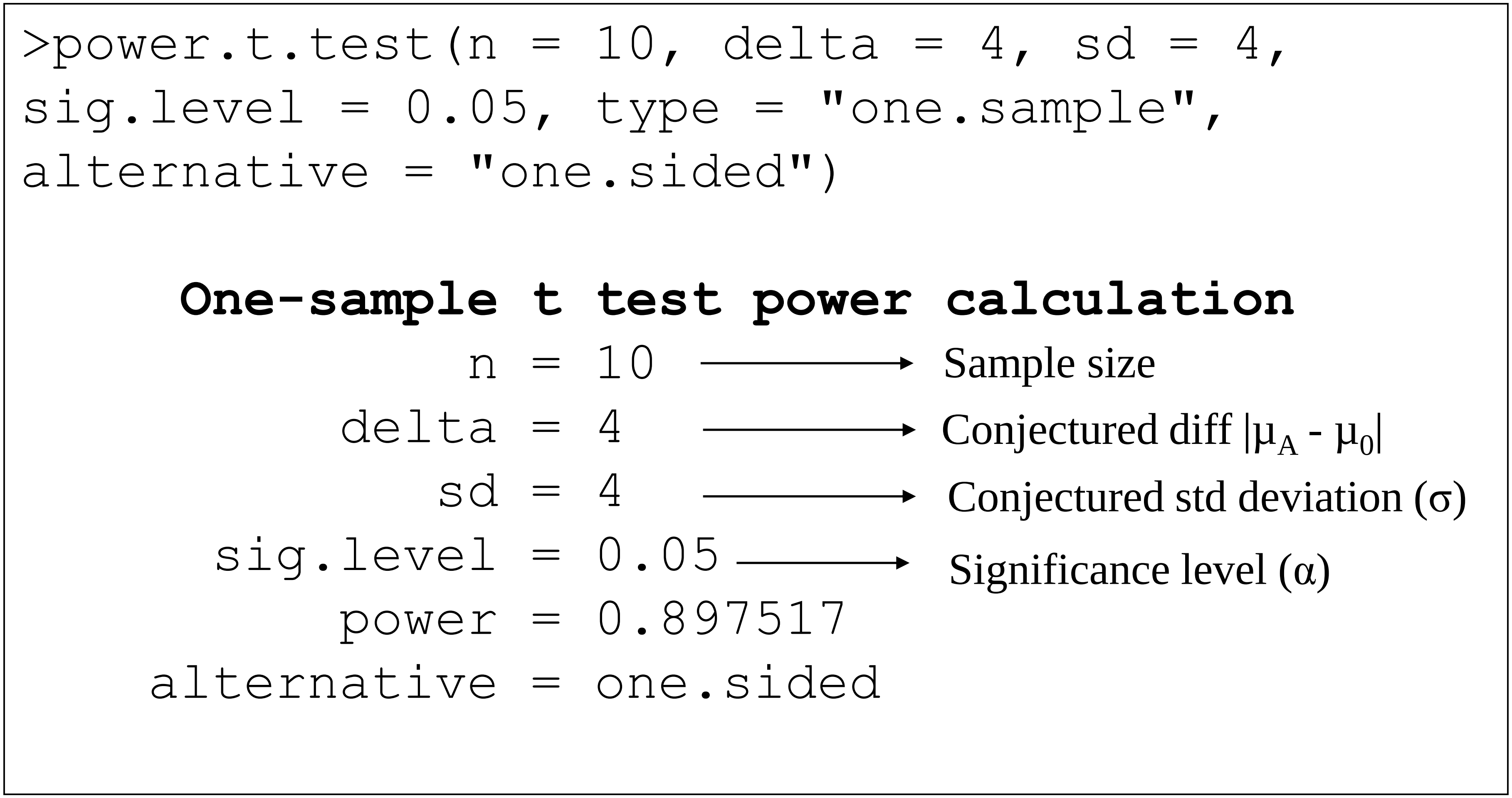
```
power= 1-pt(1.833, df = 9, ncp = 3.16)
     = 0.898
```

In practice, we can use **power.t.test()** to compute power (for fixed n) or n (to achieve a certain level of power). (See example: "**Power for a One-Sample t-test**".)

Example non-central t-distributions



Power for One-Sample t-test in R



```
>power.t.test(n = 10, delta = 4, sd = 4,
sig.level = 0.05, type = "one.sample",
alternative = "one.sided")

     One-sample t test power calculation
              n = 10
          delta = 4
             sd = 4
      sig.level = 0.05
          power = 0.897517
    alternative = one.sided
```

Sample size

Conjectured diff $|\mu_A - \mu_0|$

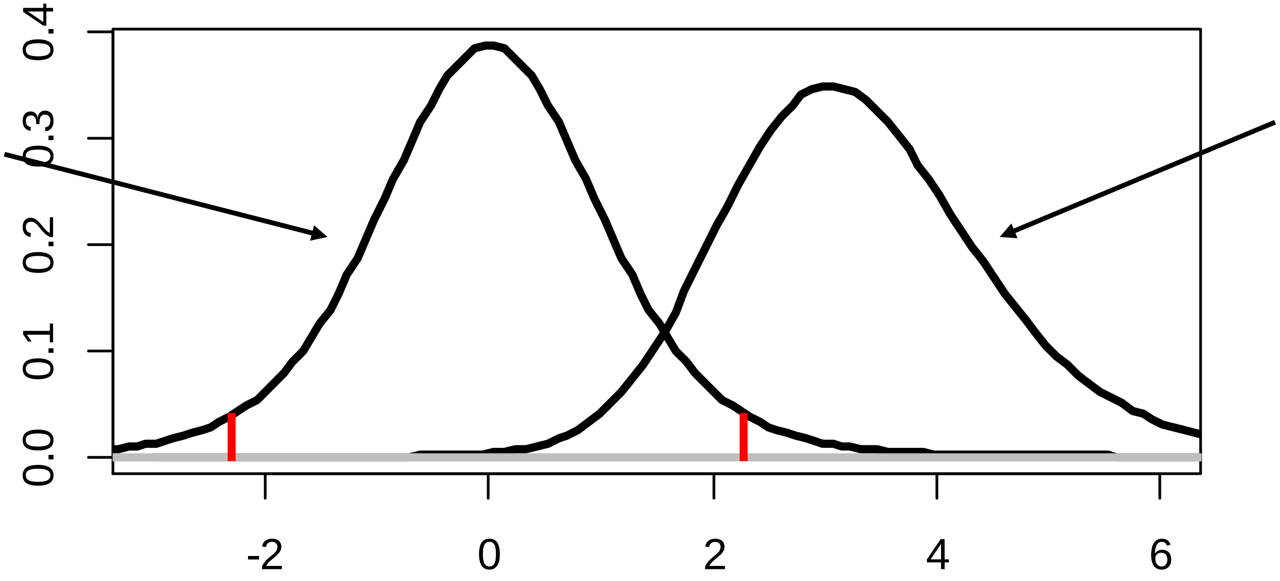
Conjectured std deviation (σ)

Significance level (α)

Power Case 3: Use R to compute power for two-sided t-tests.

Power for a two-sided t-test can be computed with a minor modification of the one-sided case. H0: μ = 12 vs HA: μ ≠ 12. n=10 (so df=9).

Reject if |t| > $t_{\alpha/2}$=2.262. Sum area under the non-central curve from both tails.



Distbn of t when H_0 is true

Distbn of t when H_A is true

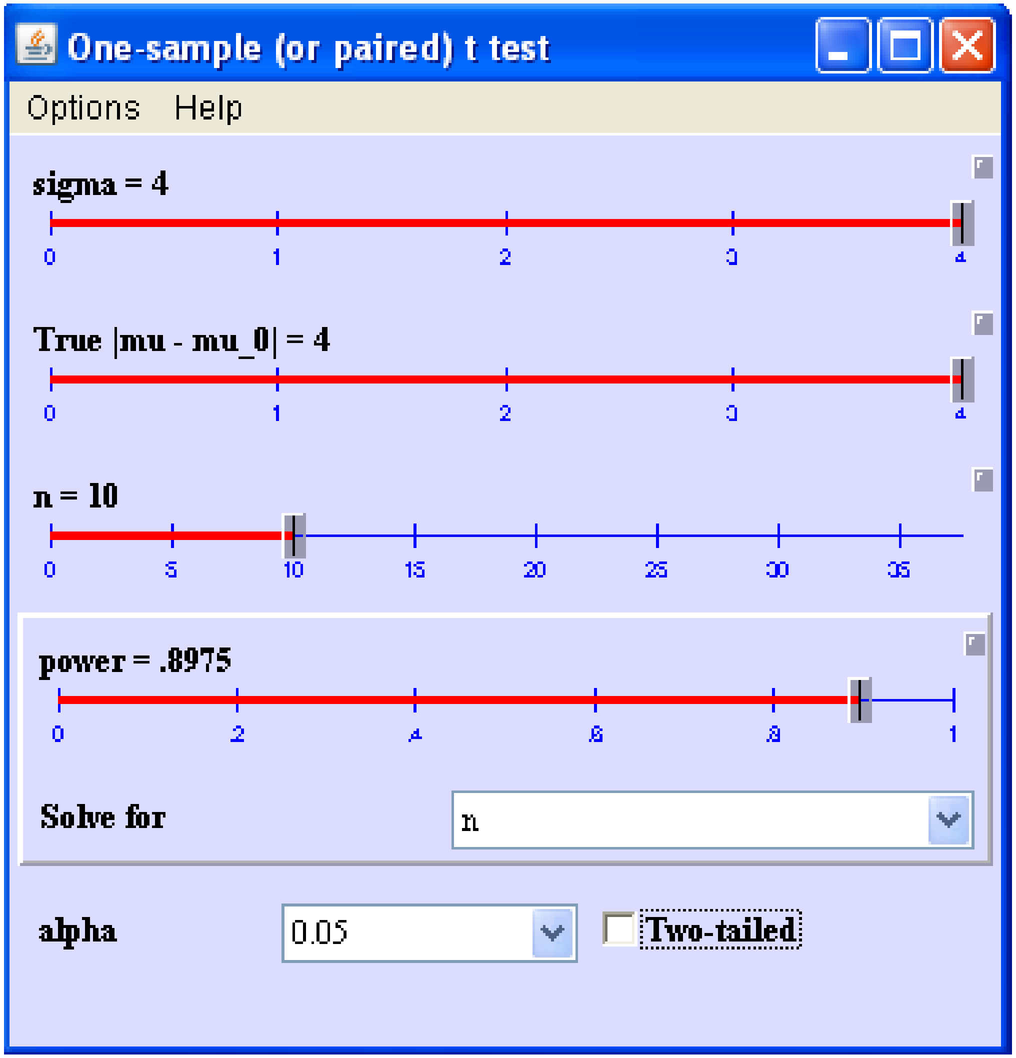
Power can then be computed in R using the function **pt**

```
power= 1-pt(2.262, 9, 3.16)+pt(-2.262,9,3.16)
     = 0.803
```

We use power.t.test() to compute the power or sample size. (See "**Power for a One-Sample t-test**")

Power Case 4: Use Lenth's online power calculator to compute power.
http://homepage.stat.uiowa.edu/~rlenth/Power/

* Choose One-sample t-test



sigma = 4

True |mu - mu_0| = 4

n = 10

power = .8975

Solve for n

alpha 0.05 □ Two-tailed

Conjectured Standard Deviation (σ)

Conjectured difference $|\mu_A - \mu_0|$

Sample Size

Power

Significance Level (α) and one or two-sided test

“Factors” that effect Power

- **n (sample size):** As n increases, so does power.
- **Magnitude of difference ($|\mu_A$ - $\mu_0|$)**: As magnitude of difference increases, so does power.
- **σ (standard deviation):** As σ increases, power decreases.
- Also α and whether test is one or two-sided.

How to see this?

1. λ (non-centrality parameter): As λ increases, so does power.

$$\lambda = \frac{\mu_{\mathrm{A}} - \mu_0}{\sigma / \sqrt{n}}$$

2. Graphs
3. Intuition

Comments about Power

1. Sample size justification is often required for grant proposals, Animal Care protocols or Human Subjects committees.
2. **Typically, people strive to achieve 80% or 90% power.**
3. Always round sample size up (to next integer value)!
4. The power calculation is based on conjectured values! If possible, use pilot data or published articles to come up with conjectures for μ_A and σ.
5. Another way to think about the conjecture for μ_A is to think about what would be a "meaningful difference" (between μ_A - μ_0). We want our study to be powerful enough to detect a meaningful difference.
6. Here is another approach for coming up with a conjecture for σ. Sometimes people have an expected range of values (min, max). The empirical rule (based on the normal distribution) tells us that almost all (>99%) of the data should fall within 3 standard deviations of the mean. Then σ = (max – min)/6.
   O&L use a more conservative denominator of 4.

7. We can do “what if?” calculations. Programs can be altered so that n is fixed and σ varies. Graphs can be added.
8. Formulas in O&L using z_α are only for larger n. We will use the R or Lenth programs which are based on the t distribution.
9. Using power.t.test(), we can find the power corresponding to a certain sample size by specifying n. We can find the sample size needed to achieve a certain level of power by specifying power. Ex: power.t.test(n=10,…) vs power.t.test(power=0.90,….)
10. The R package `pwr` contains some additional power calculations.
11. Sometimes power calculations are done to show lack of effect. In this case failure to reject H_0 is used to argue that H_0 is true. Power calculations are used to argue that if there had been an effect of a specified size, we probably would have rejected the null hypothesis. This argument is faulty. It is better to use a confidence interval to argue lack of an important effect.
12. For more discussion about power and sample size justification see the article “Some Practical Guidelines for Effective Sample Size Determination” (2001) by Lenth.

10. “Robustness” of the t-test and t- confidence intervals

A test is **robust** if the observed rate of type I errors is close to the claimed rate (α), despite failure of the assumptions of the procedure.

A confidence interval method is **robust** if the observed rate of coverage is close to the claimed rate of coverage (e.g. 95%), despite failure of the assumptions of the procedure.

O&L perform a simulation (where H_0: $\mu=\mu_0$ is true) and empirically estimate the Type I error rate (proportion of times that the H_0 is falsely rejected). See full table for power results (not shown here).

Observed Type I Error rates from O&L Table 5.6 (also see Figure 5.19) Stated α=0.05.

Distribution	n=10	n=15	n=20
Normal	0.05	0.05	0.05
Heavy Tailed	0.035	0.049	0.045
Light Skewness	0.025	0.037	0.041
Heavy Skewness	0.007	0.006	0.011

Results of “Robustness” Simulation

Based on statistical simulation studies using non-normal populations, we arrive at the following conclusions about the robustness of the t-test :

1. When the population is symmetric with heavy tails (“moderately outlier prone”) the error rates for the t-tests are slightly below α, but not unacceptably so. CI coverage will be slightly high.

2. When the population is highly skewed (long right tail), the error rates for the t-tests are consistently below α. This problem persists, even for relatively large sample sizes.

Note: When observed Type I Error Rate is below the stated α level, we say the method is “conservative”.

11. Bootstrap Confidence Interval

The bootstrap is a resampling based method originally suggested by Brad Efron in 1979. This approach can be used to construct a CI when the assumption of normality is not satisfied.

The bootstrap method involves using the <u>distribution of the sample data</u> to estimate the correct percentiles for our non-normal population.

The usual t interval for a 95% CI is:

$$\left(\bar{y} - t_{0.025} \frac{s}{\sqrt{n}}, \bar{y} + t_{0.025} \frac{s}{\sqrt{n}} \right)$$

where the $t_{0.025}$ is the upper 0.025 point is from the t table. A modified from of the interval for non-symmetric distributions is:

$$\left(\bar{y} - t_{0.025} \frac{s}{\sqrt{n}}, \bar{y} - t_{0.975} \frac{s}{\sqrt{n}} \right)$$

Bootstrap Method: Find what the percentiles of the t-statistic would be if the true population had the same shape as the sample. We do this by re-sampling from the data, as if it were the population.

If the data are: $y_1, y_2, ..., y_n$

Randomly sample (with replacement) to get: $y_1^*, y_2^*, ..., y_n^*$

Compute $\bar{y}^*$ and s^* from the new sample.

Compute:
$$t^* = \frac{\bar{y}^* - \bar{y}}{(s^* / \sqrt{n})}$$

Repeat many times (thousands), and use the percentiles of the observed t^* as if they were the percentiles of *t*.

The “studentized” bootstrap t interval is then:

$$\left(\bar{y} - t^*_{0.025} \frac{s}{\sqrt{n}}, \bar{y} - t^*_{0.975} \frac{s}{\sqrt{n}} \right)$$

Example: Largemouth Bass sampled from n=53 lakes in Florida (approx. 10 fish/lake). Y=Average mercury in fish at each lake. Ref: Lange, Royals, & Connor. (1993). *Transactions of the American Fisheries Society* .

See: "**Bootstrap CI Single Mean"** for computations.

1. For Y: histogram, normal plot, and hypothesis tests suggest non-normality.

2. For log(Y): histogram, normal plots, and hypothesis tests also suggest non-normality.

3. Perhaps the individual fish values are lognormal, but after averaging the fish in each lake, the result is not normal nor lognormal, but clearly skewed. Looks like a good case for the bootstrap!

$$\text{Bootstrap CI}: \quad \left(0.5272 - 1.878\frac{0.3410}{\sqrt{53}}, 0.5272 - (-2.176)\frac{0.3410}{\sqrt{53}}\right)$$

$$(0.439, 0.629)$$

$$\text{Standard CI}: \quad (0.433, 0.621)$$

Bootstrap CIs using the `boot` package in R

1. Define the statistic function.
2. The **`boot(data= , statistic= , R=, ...)`** function calls the statistic function "R" times. The results can be examined using `print()` and/or `plot()`.
3. The **`boot.ci()`** function can be used to generate confidence intervals. Several types of confidence intervals are available: normal, basic, student, percent, bca (bias corrected accelerated) or all. Note that variance estimates are required for studentized intervals.

See: "**Bootstrap CI Single Mean"** for examples.
For more information see the book "Bootstrap Methods and their Application" (1997) by Davison and Hinkley or the Rnews article "Resampling Methods in R: The boot Package" (2002) by Canty.

12. Inference about the population median

When the population parameter of interest is the median (M), one may test the null hypothesis that the median equals a specific value.

Example: (O&L example 5.20) Sample of household recyclable material for n=25 households.

Hypothesis test: $H_0 : M = 5\,\text{lbs./wk.}$

$H_A : M \neq 5\,\text{lbs./wk.}$

B = test statistic = number observations greater than M_0 (null hyp val).

Reject H_0 if $B \le C_{\alpha(2),n}$ or $B \ge n - C_{\alpha(2),n}$

where $C_{\alpha(2),n}$ is taken from Table 4.

In the Example $B = 13, C_{\alpha(2),n} = 7$, for $\alpha = 0.05$.

B is not less than or equal to 7.

B is not greater than or equal to $25 - 7 = 18$.

Therefore, Fail to Reject H_0.

Sign Test for median using R

```
> library(BSDA)
> SIGN.test(y, md = 5)

        One-sample Sign-Test

data:  y
s = 13, p-value = 1
alternative hypothesis: true median is not equal to 5
95 percent confidence interval:
 3.931247 6.700000
sample estimates:
median of x
        5.3

                  Conf.Level L.E.pt U.E.pt
Lower Achieved CI     0.8922 4.2000    6.7
Interpolated CI       0.9500 3.9312    6.7
Upper Achieved CI     0.9567 3.9000    6.7
```

Non-parametric Statistical Methods

The sign test and CI for the median are examples of non-parametric methods.

Non-parametric methods do not require distributional assumptions (or at least they require fewer assumptions). For example, the CI for the median does not require that the data comes from a normal distribution.

In many cases, non-parametric methods will be based on ranks or order statistics.

Do you want to make inference about the mean or the median?

Whether you are interested in the mean or the median depends on the purpose of the study, not the shape of the distribution:

1. If you are interested in the average value, cumulative cost, or cumulative exposure, then you are interested in the mean, whether the distribution is skewed or not. (That is why we are interested in the mean of skewed pollution distributions).
2. If you are interested in a "typical" response (half above, half below), then you are interested in the median.

Note: In the Recycling example, you can make the case that we are probably more interested in the mean, because we are interested in the accumulation of material in the landfill, and the cost of materials. If we were interested in staffing of the pick-up trucks we may also be interested in the median.