Chapter 6: Comparing two population means or medians

1. Two sample t-test and CI (equal variances)
2. Welch-Satterthwaite t-test and CI (unequal variances)
3. Practical considerations: p-values, effect sizes and independence
4. Sample size and power for two-sample t-test
5. Wilcoxon Rank-sum (Two-sample) test
6. Paired samples: t-test and CI
7. Wilcoxon Signed Rank (Paired sample) test
8. Bootstrap two-sample t confidence interval

Examples:

1. Two-sample t-test (and Wilcoxon Rank-Sum test)
2. Power for a Two-sample t-test
3. Wilcoxon (Two-Sample) Rank Sum test
4. Paired t-test (and Wilcoxon Signed Rank test)
5. Bootstrap Two-sample t confidence interval

1. Two sample t-test and Confidence Interval

Rat Lead Example: Twenty rats were randomly assigned to two groups.
10 rats in the Control group received a standard diet.
10 rats in the Deficient group received a calcium deficient diet.
For both groups, a 0.15% lead-acetate solution was available to drink.
The amount of solution (Y) consumed by each rat was measured.

The goal of the study is to compare mean lead consumption for the two treatments. It seems obvious, but notice that use of a Control treatment is critical! This provides a benchmark to which the Deficient treatment is compared.

Summary Statistics:	
Control	**Deficient**
$\bar{y}_C = 5.06$	$\bar{y}_D = 8.56$
$s_C = 1.189$	$s_D = 1.471$
$n_C = 10$	$n_D = 10$

Let: μ_1= population mean for group 1
μ_2= population mean for group 2
σ_1= population standard deviation for group 1
σ_2= population standard deviation for group 2

We want to make inference about the **difference** between population means using sample means.

Hypothesis test: Are the means the same?

$$H_0 : \mu_1 = \mu_2 \qquad (\text{or } \mu_1 - \mu_2 = 0)$$
$$H_A : \mu_1 \neq \mu_2 \qquad (\text{or } \mu_1 - \mu_2 \neq 0)$$

Strategy:

Estimate the difference; standardize it; then evaluate whether the result is far enough from zero to reject H_0.

$$t = \frac{(\bar{y}_1 - \bar{y}_2) - 0}{\text{std. error of } (\bar{y}_1 - \bar{y}_2)}$$

General form of *t*:

$$t = \left[\frac{est - (hyp.val.)}{\text{std. error of } (est)} \right]$$

Standard Error and Assumptions

If the data are **normally distributed** and/or we have large sample sizes, then $\bar{y}_1$ and $\bar{y}_2$ are normally distributed and are **independent**, their sum or difference is also normally distributed.

Since $\bar{y}_1$ and $\bar{y}_2$ are **independent** then

$$Var(\bar{y}_1 - \bar{y}_2) = Var(\bar{y}_1) + Var(\bar{y}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

If we assume that $\sigma_1 = \sigma_2 = \sigma$ (**equal variance**) then

$$Var(\bar{y}_1 - \bar{y}_2) = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

Hence the std. error of $(\bar{y}_1 - \bar{y}_2) = s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$

We estimate s^2 by a "pooled estimate"

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

(Think of s_p^2 as a weighted average of s_1^2 and s_2^2)

df=n_1+n_2-2

Pooled Two-Sample t-test (Equal Variances)

Assumptions: Independent random samples, equal variances, normally distributed data and/or large sample sizes.

Test Statistic: $$t = \frac{\bar{y}_1 - \bar{y}_2 - \Delta_0}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

Alternative Hypothesis:	**Rejection Region:**
(1) $H_A: \mu_1 - \mu_2 > \Delta_0$	$t \geq t_{\alpha,\, n1+n2-2}$
(2) $H_A: \mu_1 - \mu_2 < \Delta_0$	$t \leq -t_{\alpha,\, n1+n2-2}$
(3) $H_A: \mu_1 - \mu_2 \neq \Delta_0$	$\lvert t \rvert \geq t_{\alpha/2,\, n1+n2-2}$

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|.

Rat Lead Example: Two-sided alternative

1. **Hypotheses:**

   H_0: $\mu_C-\mu_D=0$ vs H_A: $\mu_C-\mu_D\neq 0$

2. **Test Statistic:**

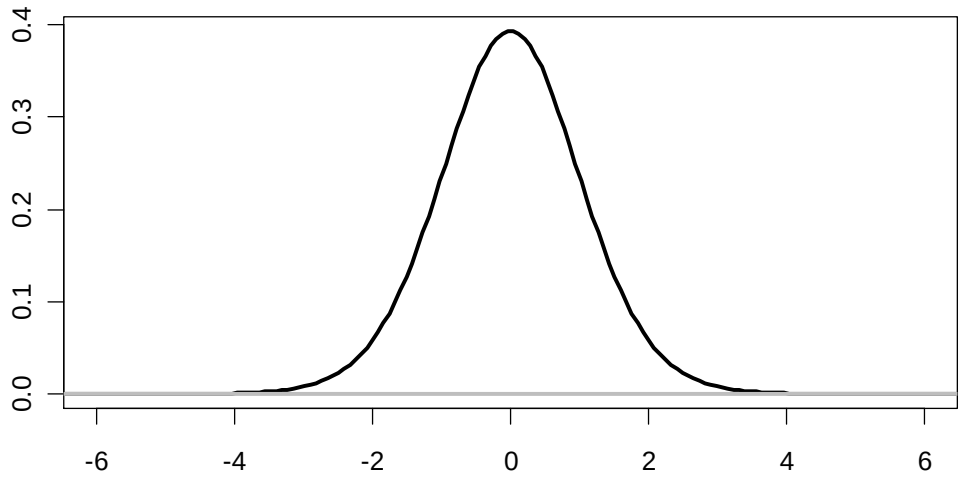
$$s_p = \sqrt{\frac{9(1.189)^2 + 9(1.471)^2}{18}} = 1.337$$

$$t = \frac{5.06 - 8.56}{1.337\sqrt{\frac{1}{10} + \frac{1}{10}}} = -5.85$$

3. **Rejection Region:**

   α=0.05, df=10+10-2=18

   Reject H0 if $|t| > t_{\alpha/2}$=2.101



p-value (in R):

```
p = 2*(1-pt(5.85,df=18))
```

= 0.000015

4. **Conclusion:**

   $|t| = 5.85 > t_{\alpha/2} = 2.101$

   We have evidence of a difference between the true population means. Rats on calcium deficient diet consume more lead solution.

See "**Two-sample t-test**" example.

Rat Lead Example: One-sided alternative

Suppose we want only to consider the research alternative that Deficient group consumes **more** of the lead acetate solution than the Control group. (In practice, this should be decided in advance!)

1. **Hypotheses:**

   H_0: $\mu_C-\mu_D \geq 0$ vs H_A: $\mu_C-\mu_D<0$

2. **Test Statistic (same as before):**

$$t = \frac{5.06-8.56}{1.337\sqrt{\frac{1}{10}+\frac{1}{10}}} = -5.85$$

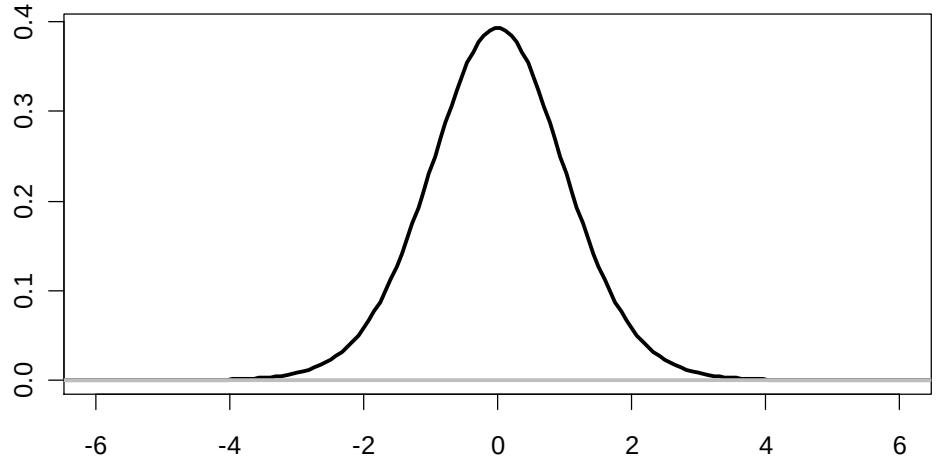
3. **Rejection Region:**

   α=0.05, df=10+10-2=18

   Reject H0 if $t < -t_\alpha = -1.734$.

4. **Conclusion:**

$t = -5.85 < -t_\alpha = -1.734$



p-value (in R):

```
p=pt(-5.85, df=18)
```

= 0.000007

We have evidence that $\mu_C<\mu_D$. Rats on calcium deficient diet consume more lead solution.

Two-Sample t Confidence Interval

The (1-α)100% confidence interval for μ_1-μ_2 is:

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2} \text{ std error of } (\bar{y}_1 - \bar{y}_2)$$

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2}\ s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where the table value $t_{\alpha/2}$ is determined from the Student's t-distribution with df $=n_1+n_2-2$.

Assumptions: Independent random samples, equal variances, normally distributed data and/or large sample sizes.

Rat Lead Example (95% CI, df=18):

$$(5.06 - 8.56) \pm (2.101)(1.337)\sqrt{\frac{1}{10} + \frac{1}{10}}$$

$$-3.50 \pm 1.26 \quad \rightarrow \quad (-4.76, -2.24)$$

Since the CI does not include 0, we have evidence that there is a difference between the population means. The CI will give same conclusion as two-sided test.

Two-Sample t-test and CI in R/Rcmdr

- In R, use the function t.test().
- In Rcmdr, choose Statistics -> Means -> Independent samples t-test.
- NOTE: Can assume equal or unequal variances.

- For the Rat Lead Example: (See "**Two-Sample t-test**")

```
> t.test(y ~ trt, var.equal=TRUE, data=ratlead)
     Two Sample t-test
data:  y by trt
t = -5.8507, df = 18, p-value = 1.532e-05
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
 -4.756813 -2.243187
sample estimates:
mean in group control mean in group deficient
                 5.06                    8.56
```

2. Welch-Satterthwaite t-test (Unequal variances)

Assumptions: Independent random samples, normally distributed data and/or large sample sizes.

Test Statistic: $$t = \frac{\bar{y}_1 - \bar{y}_2 - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where $$\text{df}' = \frac{(n_1 - 1)(n_2 - 1)}{(n_2 - 1)c^2 + (1 - c)^2(n_1 - 1)}$$ where $$c = \frac{s_1^2 / n_1}{s_1^2 / n_1 + s_2^2 / n_2}$$

Alternative Hypothesis:	**Rejection Region:**
(1) $H_A: \mu_1 - \mu_2 > \Delta_0$	$t \geq t_{\alpha,df'}$
(2) $H_A: \mu_1 - \mu_2 < \Delta_0$	$t \leq -t_{\alpha,df'}$
(3) $H_A: \mu_1 - \mu_2 \neq \Delta_0$	$\lvert t \rvert \geq t_{\alpha/2,df'}$

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|.

Comments on the Welch-Satterthwaite t-test

1. In R, use the `t.test()` function with `var.equal=FALSE` (default).
2. The use of the t-distribution is approximate for the Sattherthwaite t test, but this approach is good and very common!
3. When n1=n2, then the t-test statistics will be the same whether or not we assume equal variances. The df will still be different.
4. Confidence intervals can be calculated allowing unequal variance. Use Sattertwaite df'.

$$\bar{y}_1 - \bar{y}_2 \pm t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

5. $\min(n_1\text{-}1, n_2\text{-}1) \leq$ df' (Satt df) $\leq n_1+n_2\text{-}2$
   df will be "small" if s_1^2 very different from s_2^2
   df will be "large" if s_1^2 close to s_2^2

How do we decide between Pooled or Satterthwaite tests?

Recall that the Pooled variance two-sample t-test assumes equal variances, while the Welch-Satterthwaite t-test allows unequal variances.

Pooled variance t-test is generally used unless the variances are believed to be unequal. See simulation results from O&L and later in these notes.

In Chapter 7, we will discuss a formal test of $H0$: $\sigma_1^2 = \sigma_2^2$

For now, you can use a rule of thumb:
If $s_{max}/s_{min} < 2$ then assume "equal" variances (used Pooled t-test).
If $s_{max}/s_{min} \geq 2$ then do not assume equal variances (use Welch-Satterthwaite t-test)

Rat Lead Example: Allowing Unequal Variances

Using the rule of thumb from the previous slide:
$s_D/s_C = 1.471/1.189 = 1.237 < 2$
So, in practice we could use the pooled variance t-test. But for illustration, we will rerun the analysis allowing unequal variances.
Results are very similar to original analysis!
Note: Satterthwaite df' = 17.24

Hypothesis Test:
Because the sample sizes are equal, the test statistic is unchanged.

$$t = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{5.06 - 8.56}{\sqrt{\frac{(1.189)^2}{10} + \frac{(1.471)^2}{10}}} = -5.85$$

We can calculate the two-sided p-value using R:

```
p=2*(1-pt(5.85, df = 17.24))
```

Result: p=0.000018

Rat Lead 95% Confidence Interval:

From R: `qt(0.975, df = 17.24)` = 2.108

$$5.06 - 8.56 \pm (2.108)\sqrt{\frac{(1.189)^2}{10} + \frac{(1.471)^2}{10}}$$

$$-3.50 \pm 1.262 \rightarrow (-4.76, -2.24)$$

How Sensitive is t-test to Unequal Variances?

In these simulations, 1000 samples are from normal populations with equal **means** (H_0: $\mu_1 - \mu_2 = 0$ is true) but different **variances.** Pooled variance and Welch-Satterthwaite t-tests were run using stated $\alpha = 0.05$. Empirically estimate the Type I error rate (proportion of times that the H_0 is falsely rejected).

n_1	n_2	σ_1	σ_2	Pooled t-test	WS t-test
15	15	1	2	0.060	0.055
10	20	1	2	0.017	0.044
20	10	1	2	0.114	0.059

Conclusions from the simulation:

1. If sample sizes are equal, or nearly equal, then the effect of unequal variances on the t-test is minimal. It doesn't matter "much" whether Pooled or Welch-Satterthwaite test is used; Pooled test is preferred unless s_1^2 and s_2^2 are very different (in which case, use Welch-Satterthwaite).

2. If sample sizes are unequal, then the effect of unequal variances on the t-test is more serious:
   A. When the group with the smaller sample size has the larger variance, there are too many false rejections. This is considered serious because there are too many false claims of significance.
   B. When the group with the larger sample size has the larger variance, there are too few false rejections. This is considered not as serious, because we would have too few false claims of significance (conservative), but it wastes power.

3. Practical Considerations

We have now gone through the details of running a two-sample t-test (or constructing the corresponding confidence interval).

But now we consider several practical considerations:

- Statistical vs Practical Significance
- ASA guidelines on p-values
- Writing up results
- Effect sizes
- Assumption of independence

Statistical vs Practical Significance

"Statistical significance" of a study is generally determined by a statistical test (or CI).

We have already seen that statistical significance depends on the magnitude of the difference ($\overline{y_1} - \overline{y_2}$) but also the sample size (n).

Practical significance generally has to do with the magnitude of the difference.

Practical significance is somewhat harder to define (at least for me) because every reader must judge for themselves what is practically significant.

Especially when the sample size is large, it is possible to obtain a small p-value even if the estimated difference is very small. In other words, it is possible to have statistical significance without any practical significance!

ASA Statement on Statistical Significance and p-values (2016)

1. P-values can indicate how incompatible the data are with a specified statistical model.
2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.
    - The p-value is NOT a statement about the truth of a null hypothesis. We CANNOT conclude that there is "no difference" based on a large p-value.
    - Absence of evidence is not evidence of absence.
3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.
    - Cannot justify "bright line" rule at $p = 0.05$ (or any other value).
    - Even rare cases that require yes/no decision, want to consider many contextual factors not just p-value!

ASA Statement on Statistical Significance and p-values (2016)

4. Proper inference requires full reporting and transparency.
   - Conducting multiple analyses of the data and reporting only those with certain *p*-values (typically those passing a significance threshold) renders the reported *p*-values essentially uninterpretable.
5. A *p*-value, or statistical significance, does not measure the size of an effect or the importance of a result.
   - Statistical significance is not equivalent to scientific, human, or economic significance.
6. By itself, a *p*-value does not provide a good measure of evidence regarding a model or hypothesis.
   - Researchers should recognize that a *p*-value without context or other evidence provides limited information.

Moving to a World Beyond "p < 0.05" (The American Statistician, 2019)

Don't say "Statistically Significant"

- We are NOT recommending that the calculation and use of continuous p-values be discontinued.
- Where p-values are used, they should be reported as continuous quantities (e.g. p = 0.08).
- For the integrity of scientific publishing and research dissemination, therefore, whether a p-value passes any arbitrary threshold should not be considered at all when deciding which results to present or highlight.

An Idea: Consider writing most of the text for the Abstract BEFORE running the statistical analysis. This forces the author to identify the most important research questions and variables. After analysis, simply add in the actual results.

Moving to a World Beyond "p < 0.05" (The American Statistician, 2019)

Acronym: ATOM

Accept uncertainty.

- Accompany every point estimate with a measure of uncertainty such as a standard error or interval estimate.

Be thoughtful, open and modest.

- Thoughtful researchers begin above all else with clearly expressed objectives.
- They invest in producing solid data.
- They consider not one but a multitude of data analysis techniques.
- Thoughtful research includes careful consideration of the definition of meaningful effect size. As a researcher you should communicate this up front, before data are collected and analyzed.

Writing up results

We want to give the reader enough information to recreate our results. At the same time, we want to keep things brief.

We should never present just a p-value without giving the reader more information! It is critical to provide information about the estimated values, variability and sample size. Do not "round off" the p-value to < or > 0.05.

CI's can also be used as an alternative to hypothesis testing.

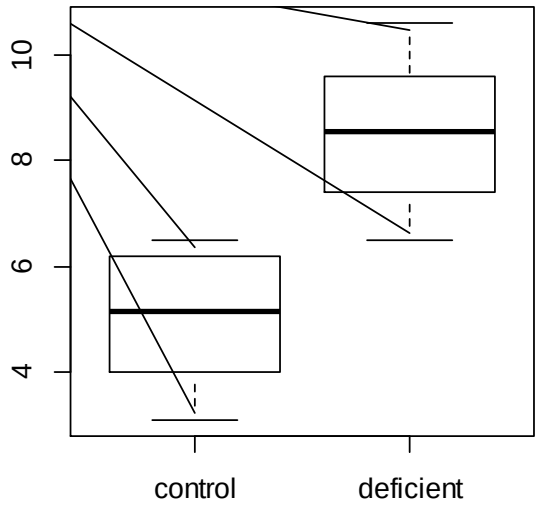
I find that the summary statistics (means, SE (or sd) and sample size) for each group are most helpful.

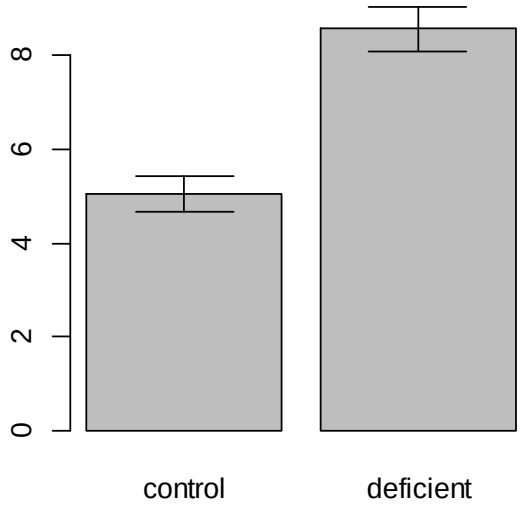
Only present a table or graph if there is something to say about it!

Rat Lead Example:

R software was used for statistical analysis. The two-sample t-test assuming equal variance was used to compare mean lead consumption for the two groups. The table below gives the mean (and SE) for each group and the p-value from the two-sample t-test.

Response	Control (n=10)	Deficient (n=10)	p-value
Lead Consumption	5.06 (0.38)	8.56 (0.47)	<0.001





Effect Sizes and Cohen's d

Some people advocate the use of "effect sizes". In the scenario where we are comparing two means, the unstandardized effect size is just the estimated difference between the two means.

Cohen's d is an example of a standardized effect size:

$$d = \frac{\overline{y_1} - \overline{y_2}}{pooled\ sd} = \frac{\overline{y_1} - \overline{y_2}}{s_p}$$

In R, you can use the **`effsize`** package.

```
> library(effsize)
> cohen.d(y ~ trt, data = ratlead)
Cohen's d
d estimate: -2.61651 (large)
95 percent confidence interval:
      inf       sup
-3.999808 -1.233212
```

What does it mean to have "independent" observations? How can this assumption be violated?

Suppose an investigator is investigating an anti-fungal treatment on plants. They have two treatments under consideration: Mock and Active. Treatments are applied to the leaves of plants.

We consider several possible designs.

Design1: A total of 12 plants are grown to a fixed age. 6 plants are randomly assigned to receive Mock trt and 6 are randomly assigned to receive Active trt. The treatment is applied to a single leaf from each plant. We record a total of 12 measurements.

Different plants -> Independent Obs -> 2 sample t (df = 10)

Design2: A total of 6 plants are grown to a fixed age. Each plant has one leaf treated with Active and another leaf treated with Mock. We record a total of 12 measurements.

Paired observations -> Paired t-test or CI.

Design3: 2 plants are grown to a fixed age. One plant has 6 leaves treated with Active trt. The other plant has 6 leaves treated with Mock trt. We record a total of 12 measurements.

Do not use this design! Unreplicated or “pseudo” replicated. NOT independent observations.

Design4: A total of 6 plants are grown to a fixed age. 3 plants have Active trt applied to 2 leaves. 3 plants have Mock trt applied to 2 leaves. We record a total of 12 measurements.

Two levels of replication:
"Bio" reps = plants
"Tech" reps = leaves within plants
Not independent observations, do NOT use 2 sample t-test with df = 10.

Option1: Average over 2 leaves per plant and use total of 6 observations for analysis. Use 2 sample t-test with df = 4.

Option2: Mixed model (nested) analysis discussed in STAT512.

Pseudo replication was defined by Hurlbert in 1984 as “the use of inferential statistics to test for treatment effects with data from experiments where either treatments are not replicated, or replicates are not statistically independent.”

From Manly:
“When dependent data are analyzed as if they are independent, the sample size used is larger than the effective number of independent observations….To avoid this, a good rule to follow is that statistical inferences should be based on only one value from each independently sampled unit, unless the dependence in the data is properly handled in the analysis.”

4. Sample Size and Power for Two Sample t-tests

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.

We will consider several cases:

1. Find the n (per group) 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 two-sample one-sided t-test.
3. Use R to compute power for two-sample two-sided t-test.
4. Use Lenth's on-line power calculator to compute power for two-sample one- or two-sided t-tests.

NOTE: When planning, people typically use:

- $n_1=n_2$ (equal sample sizes for the two groups)
- $\sigma_1=\sigma_2$ (equal standard deviations for the two groups)

"Power" Case 1: Find the n required so that the expected width of a 100(1-α)% CI for $\mu_1 - \mu_2$ is approximately 2E (or the ME is E).

A 100(1-α)% CI is of the form:

$$\bar{y}_1 - \bar{y}_2 \pm ME \text{ where } ME = t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\text{and if } n_1 = n_2 = n, \text{ then } ME = t_{\alpha/2} s_p \sqrt{\frac{2}{n}}$$

We will use a conjectured value for s_p, then try different values of n.

Two Approaches:

A. Use R 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.

$$n = \frac{2\left(t_{\alpha/2}\right)^2 {s_p}^2}{E^2}$$

"Power" Case 1A Example

You want a 95% confidence interval for the <u>difference</u> between μ_1 and μ_2 to have total width about 10mg (or ME=5mg), and you conjecture that σ=4mg.

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

Based on these results, a value of n1=n2=7 will result in a 95% ME < 5 (or a total CI width < 10)

n1	n2	ME
5	5	5.83
6	6	5.15
7	7	4.66
8	8	4.29
9	9	4.00
10	10	3.76

"Power" Case 1B Example

You want a 95% confidence interval for the <u>difference</u> between μ_1 and μ_2 to have total width about 10mg (E=5mg), and you conjecture that σ=4mg.

1. Take **initially** $t_{\alpha/2} = 2$. $$n = \frac{2(2)^2 4^2}{5^2} = 5.1 \cong 5$$

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

$t_{\alpha/2}$ =2.306 $$n = \frac{2(2.306)^2 4^2}{5^2} = 6.8 \cong 7$$

3. Check the resulting value based on n1=n2=7. ME= 4.66 < 5 -> OK!

Note: This is n <u>per group.</u>

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 **difference μ1 - μ2** (under H_A) and **σ.** (Coming up with reasonable conjectures is often the hardest part!)

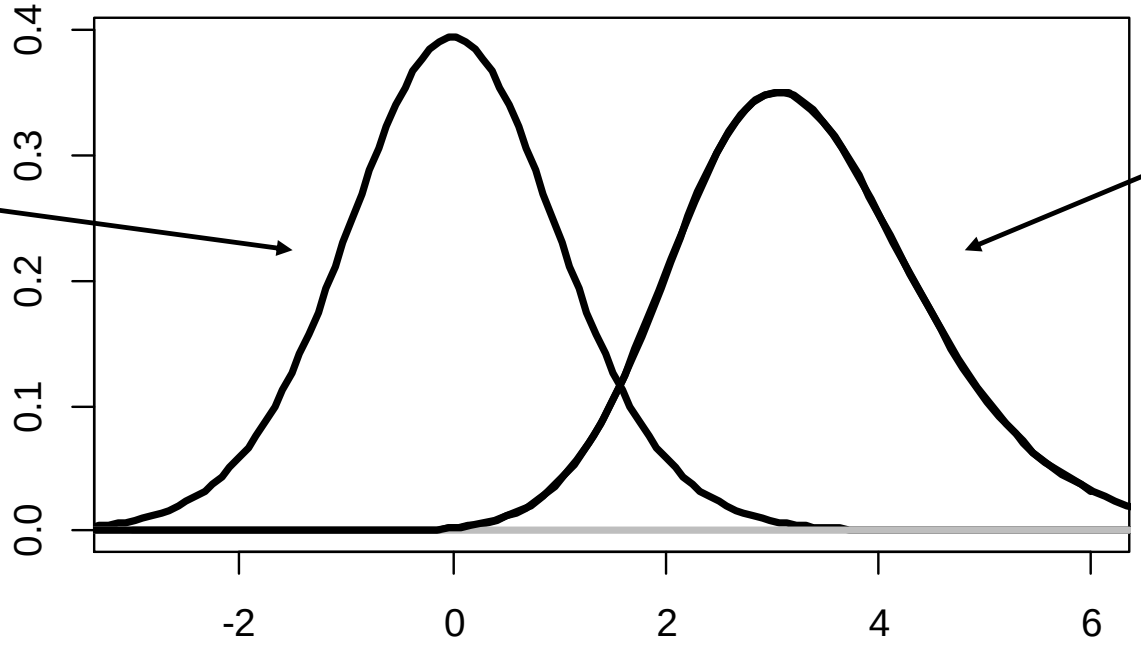
Example: $\alpha=0.05$, $n_1=n_2=n=9$, $df=9+9-2=16$

Step 1: Set up the hypothesis test:

$H_0: \mu_1 - \mu_2 \leq 0$ $H_A: \mu_1 - \mu_2 > 0$ $(\mu_1 > \mu_2)$

We will reject H_0 if $t > t_\alpha = 1.746$ (the Rejection Region).

Distbn of t when H_0 is true ($\mu_1 - \mu_2 = 0$)



Distbn of t when H_A is true ($\mu_1 - \mu_2 = 6$)

Step 2: Identify your conjectures about σ and about the true means.
We conjecture: $\sigma = 4$, $\mu_1 = 18$, $\mu_2 = 12$. So, $\mu_1 - \mu_2 = 6$.

When the alternative is true, the distribution of *t* is not centered at zero; it is "non-central", centered at its "noncentrality" parameter:

$$\lambda = \frac{\mu_1 - \mu_2}{\sigma\sqrt{\frac{2}{n}}} = \frac{18-12}{4.0\sqrt{\frac{2}{9}}} = 3.18$$

The noncentrality parameter describes the difference between the true means in terms of $\text{std. deviation of } (\bar{y}_1 - \bar{y}_2)$.

Step 3: Power can then be computed in R:

`power = 1-pt(1.746, df = 16, ncp = 3.18)=` 0.919

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 Two-Sample t-test**".)

Power for Two-Sample t-test in R

```
>power.t.test(n=9, delta=6, sd=4,
sig.level=0.05, type="two.sample",
alternative="one.sided")

     Two-sample t test power calculation
              n = 9
          delta = 6
             sd = 4
      sig.level = 0.05
          power = 0.9189915
    alternative = one.sided

NOTE: n is number in *each* group
```

n = 9 → Sample size (per group!)

delta = 6 → Conjectured diff $|\mu_1 - \mu_2|$

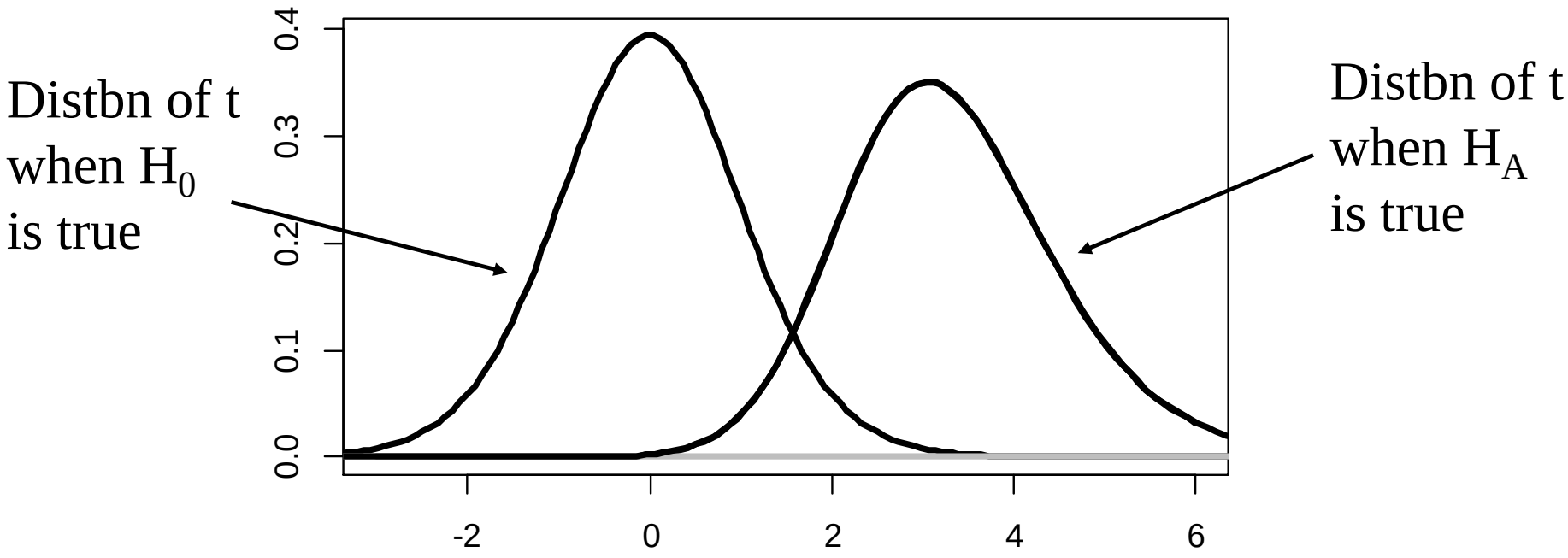
sd = 4 → Conjectured std deviation (σ)

sig.level = 0.05 → Significance level (α)

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

$H_0: \mu_1 - \mu_2 = 0$ $H_A: \mu_1 - \mu_2 \neq 0$
A modification/continuation of previous example.

Reject H_0 if $|t| > t_{\alpha/2}=2.120$ (df=2*9-2=16) and sum power from both tails.



For the example with n = 9 per group, the critical value with df=16 is $t_{\alpha/2}$ = 2.120. Then power can be computed in R using:

```
power = pt(-2.120, df=16, ncp=3.18)
      +(1-pt(2.120, df=16, ncp=3.18))
```

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

Power for n=9 (per group) is 0.847.

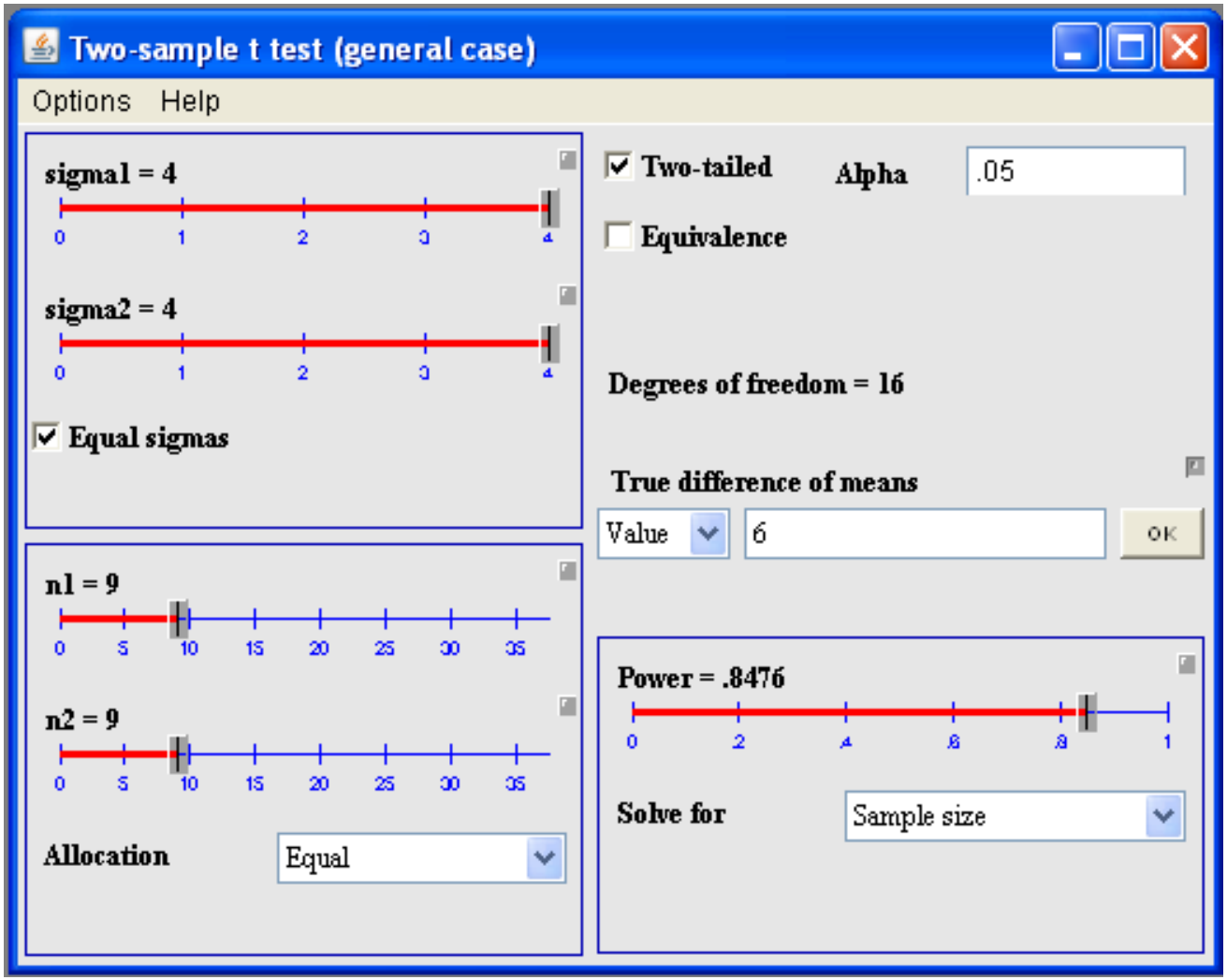
Notes:

- Power for the two-sided test is **lower** than power for the one-sided test. (This is because the one-sided test was able to "concentrate" its power in only the one direction.)
- The R package `pwr` contains some additional power calculations, including the option to allow for unequal sample sizes.
- When using other power calculators (besides R or Lenth), watch out for whether the n per group (n1, n2) or total n (n1+n2) is given!!!!!!

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

* Choose Two-sample t-test

Conjectured Standard Deviation (σ)



Sample Size

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

Conjectured Difference $|\mu_1 - \mu_2|$

Power

Writing up a sample size justification

We want to give the reader enough information to recreate our results. At the same time, we want to keep things brief.

With sample size justifications, I try to keep things simple, but realistic.

In practice, I often try several "what if" calculations. But when I write things up, I tend to report the power for a single set of conditions (ex: sample size and conjectures).

If your "conjectured" values come from a published article, consider providing the reference. If your "conjectured" values come from pilot data, say so.

If you have multiple response variables, you can focus on the most important or run multiple power calculations.

Example1: Lenth's online power calculator was used to calculate power for a two-sample t-test with $\alpha = 0.05$. Based on a difference between means of 6 and standard deviation of 4 and n = 9 subjects per group, the power was found to be 0.85.

Example2: Lenth's online power calculator was used to calculate power for a two-sample t-test with $\alpha = 0.05$. Based on a difference between means of 6 and standard deviation of 4, to achieve 90% power a sample size of n=11 subjects per group is required.

5. Wilcoxon Rank Sum (Two-Sample) Test

Nonparametric alternative to the two-sample t-test

Oxygen Example: Is there evidence of a difference in dissolved oxygen (ppm) at upstream sites and downstream sites?

Upstream:	**4.8**	**5.2**	**5.0**	**4.9**	**5.1**
Downstream:	**5.0**	**4.7**	**4.9**	**4.8**	**4.9**

Sort the data:

data	group	rank	rank(with ties)
5.2	1		
5.1	1		
5.0	1		
5.0	2		
4.9	1		
4.9	2		
4.9	2		
4.8	1		
4.8	2		
4.7	2		

T_1=sum of ranks in group 1 = **34**

Wilcoxon Rank Sum Test

One-sided test H_0: populations are identical

H_A: population 1 is shifted to the right

(Larger values of T_1 support the alternative)

Reject H0 if $T_1 > T_U$ (T_U obtained from Table 5b)

For the example, $T_U = 36$ from Table 5b, with $\alpha=0.05$

Since $T_1=34$ is not greater than $T_U=36$, DO NOT reject at $\alpha=0.05$.

Two-sided test H_0: populations are identical

H_A: population 1 is shifted either direction

(Larger or smaller values of T_1 support H_A)

Reject H0 if $T_1 > T_U$ or $T_1 < T_L$ (T_L , T_U from Table 5a)

For the example, $T_U= 37$ and $T_L=18$ from Table 5a, with $\alpha=0.05$

Since $T_1= 34$ is not in the rejection region, DO NOT reject H0.

Wilcoxon (Two-Sample) Rank Sum Test in R/Rcmdr

- In R, use the function wilcox.test().
- In Rcmdr, choose Statistics -> Nonparametric Tests -> Two Sample Wilcoxon Test.
- Group ordering is determined by alphabetical order. This is important when using a one-sided alternative!
- The test statistic is $W=T_1-n_1(n_1+1)/2$.
- See also wilcox_test() in the coin package for exact tests!
- For the Oxygen Example: (See "**Wilcoxon Two-Sample**")

```
> wilcox.test(Oxygen~Loc,data=oxygen)
Warning: cannot compute exact p-value with ties

Wilcoxon rank sum test with continuity correction
data:  Oxygen by Loc
W = 6, p-value = 0.2017
alternative hypothesis: true location shift is
not equal to 0
```

What null hypothesis really being tested by the Wilcoxon Rank Sum Test?

Wilcoxon's Rank Sum test assumes that the two distributions are identical under H0, and identical, except for a shift under HA. The H0 that the two distributions are identical includes both the assumptions of equal means and equal medians. If H0 or HA really hold, then the test can be interpreted as a test of equal means and as a test of equal medians.

In practice, researchers are usually not willing to believe such restrictive assumptions (for the same reasons they are not willing to believe normality, equality of variance, etc.). In practice, the Wilcoxon Rank Sum test is often applied to groups that have very different shapes. In that case the hypotheses can be stated:

$$H0: P(Y_1 > Y_2) = 0.5 \text{ versus } HA: P(Y_1 > Y_2) \neq 0.5$$

where Y_1 is a randomly sampled observation from group 1,

and Y_2 is a randomly sampled observation from group 2.

This is not quite the same thing as a test of equal medians, but in practice the Wilcoxon Rank sum test is usually interpreted as an approximate test of medians. It is not usually thought of as a test of means.

Normal approximation for Wilcoxon Rank Sum Test

Note: These formulas are included so that you will know what R is doing (with exact=FALSE, correct=FALSE). We won't be calculating these by hand.

The test statistic is $z = \dfrac{T_1 - \mu_T}{\sigma_T}$ where

$$\mu_T = \frac{n_1(n_1 + n_2 + 1)}{2}, \quad \sigma_T = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}}$$

z is approximately standard normal, so use Z-table to get P-values.

For the dissolved oxygen example, we have $\mu_T = 27.5, \ \sigma_T = 4.79$

and $z = \dfrac{T_1 - \mu_T}{\sigma_T} = \dfrac{34 - 27.5}{4.79} = 1.36$ So 2-sided P-value = 0.174

One-sided P values are computed in the usual manner.

(Ties were ignored in the above formulas. See book for more details.)

Simulation Study: Type I Error Rates

O&L discuss a simulation (where H_0: $\mu_1=\mu_2$ is true) and empirically estimate the Type I error rate (proportion of times that the H_0 is falsely rejected). We consider the performance of the **Wilcoxon and t-tests**. **From O&L Table 6.13, stated α=0.05.**

n1, n2	Test	Normal	Heavy Tailed	Skewed
5, 5	t-test	0.044	0.024	0.049
5, 5	Wilcoxon	0.046	0.051	0.049
5, 15	t-test	0.047	0.056	0.041
5, 15	Wilcoxon	0.048	0.046	0.049
15, 15	t-test	0.052	0.030	0.046
15, 15	Wilcoxon	0.054	0.046	0.046

Wilcoxon test performs well (observed α close to stated 0.05).

t-test appears robust to skew (observed α close to stated 0.05), but performance is not as good for (extremely) heavy tailed distribution.

Simulation Study: Power

We continue to compare the performance of the **Wilcoxon and t-tests**. This time we look at the results when there is a difference between means (HA is true) and look at the observed power.

From O&L Table 6.13, stated α=0.05.

n1, n2	Test	Normal	Heavy Tailed	Skewed
5, 5	t-test	**0.523**	0.288	**0.545**
5, 5	Wilcoxon	0.503	**0.408**	0.537
5, 15	t-test	**0.724**	0.282	**0.723**
5, 15	Wilcoxon	0.694	**0.576**	0.688
15, 15	t-test	**0.947**	0.333	**0.935**
15, 15	Wilcoxon	0.933	**0.839**	0.927

When data is from a normal or skewed population, t-test offers (slightly) higher power. When data is from an (extremely) heavy-tailed distribution, Wilcoxon offers noticeably higher power.

6. Paired Sampling (Paired t-test and CI)

Example (Problem 6.36 in O&L):

Study the effect of Benzedrine on heart rate in dogs.

Study period 1: shot of Benzedrine, H.R. measured after 2hrs

"Wash-out period" = recovery time

Study period 2: shot of placebo (saline) H.R. measured after 2hrs.

14 dogs, each got **both** (half had treatments and periods reversed)

P	B	D=diff
250	258	-8
271	285	-14
243	245	-2
252	250	2
266	268	-2
272	278	-6
293	280	13
296	305	-9
301	319	-18
298	308	-10
310	320	-10
286	293	-7
306	305	1
309	313	-4

Summary Statistics:

Placebo	**Benzedrene**	**Differences**
$\bar{y}_1$ = 282.36	$\bar{y}_2$ = 287.64	$\bar{d}$ = -5.29
s_1 = 23.14	s_2 = 25.39	s_d = 7.63
n_1=14	n_2=14	n_d = 14

Analysis of Paired Data

When we have paired data we consider the differences (d).

We denote the mean of the d's as $\bar{d}$

Note that $\bar{d} = (\bar{y}_1 - \bar{y}_2)$.

The mean of the difference is equal to the difference of the means.

Since $\bar{d}$ is like the mean of a single sample, we can use a modification of the one-sample t-test.

Caution: Not a two-sample t-test, because the samples are **not** independent (**same dogs!**). Each dog serves as its own control.

Paired t-test and CI

Assumptions: random sample, paired data, normally distributed differences and/or large sample size.

Confidence Interval: $\bar{d} \pm t_{\alpha/2} \frac{s_d}{\sqrt{n_d}}$

Test Statistic: $t = \frac{\bar{d} - \Delta_0}{s_d / \sqrt{n_d}}$

Note: df=n_d-1

Alternative Hypothesis:	**Rejection Region:**
(1) $H_A: \mu_D > \Delta_0$	$t \geq t_\alpha$
(2) $H_A: \mu_D < \Delta_0$	$t \leq -t_\alpha$
(3) $H_A: \mu_D \neq \Delta_0$	$\|t\| \geq t_{\alpha/2}$

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|.

Paired t-test and CI in R/Rcmdr

- In R, use the function t.test(,paired=T).
- In Rcmdr, choose Statistics -> Means -> Paired samples t-test.

- For the Dog Example: (See "**Paired t-test**")

```
> t.test(dog$P, dog$B, paired = T)

        Paired t-test
data:  dog$P and dog$B
t = -2.592, df = 13, p-value = 0.02234
alternative hypothesis: true difference in
means is not equal to 0
95 percent confidence interval:
 -9.6912541 -0.8801745
sample estimates:
mean of the differences
              -5.285714
```

Paired t-test and CI (Example)

Test with α=0.05:

1.H0: μ_D=0 vs HA:$\mu_D \neq 0$

2.t = -5.286/(7.630/$\sqrt{14}$) = -2.59

3.Reject H0 if $|t|>t_{0.025}$= 2.160 (df=13)

4.We find evidence that there is a difference between the means.

p-value (using R):

```
p=2*(1-pt(2.59,13))
```

p-value = 0.0223

95% Confidence Interval

$$-5.286 \pm 2.160 \frac{7.630}{\sqrt{14}}$$

(–9.691,–0.881)

“Bracketing” the p-value

- If I am running a test “by hand” using only a distribution table, what can I say about the p-value?
- For the dog example, after comparing our test statistic to the critical value (using α=0.05), we Reject H0. Hence p-value < 0.05.
- To be more specific, look at Table 2.

We see that $2.160 < 2.59 < 2.650$.

Then $0.025 > \alpha/2 > 0.01$

Hence 0.05 > 2-sided p-value > 0.02

Notes about Paired t-test

1. The most obvious form of pairing occurs when we observe both treatments on each subject. However, other types of pairing can exist. Classic example: identical twins.
2. The primary advantage of pairing is that differences often have smaller variability because subject to subject variability is eliminated. In other words, by accounting for subject to subject variability we are better able to detect the treatment difference.
3. Possible disadvantages include possible negative relationships within pairs (fatigue) or insufficient washout.
4. Paired t-test requires the differences to be normally distributed, not the individual observations!

7. Wilcoxon (Paired) Signed Rank test

Nonparametric Alternative to the Paired t-test

Illustration using Benzedrine dogs data.

1) Order differences by **absolute** value
2) Assign ranks to differences
3) Sum the ranks by sign of differences

T_+= sum of positive ranks = 16.

T_-= sum of negative ranks = 89.

n = number of nonzero differences

(**Caution**: n may be smaller than the sample size).

diff	rank
-18	
-14	
13	
-10	
-10	
-9	
-8	
-7	
-6	
-4	
-2	
-2	
2	
1	

Wilcoxon Signed Rank test

H_0: the distribution of differences is symmetric about zero
(or Δ_0, in which case you subtract off Δ_0)

(1) H_A: differences tend to be larger than zero.

Reject H0 if $T = T_- \le T_\alpha$ (from Table 6 in O&L.)

(2) H_A: differences tend to be smaller than zero.

Reject H0 if $T = T_+ \le T_\alpha$

(3) H_A: either (1) or (2) (two-sided)

Reject H0 if T= **the smaller of T_- and T_+ is $\le T_\alpha$**

In Benzedrine example, we do the two-sided test:

n=14 (all the differences are nonzero), α=0.05,

and from Table 6: $T_{0.05} = 21$.

The smaller of T_- and T_+ is 16.

Reject H_0 because $16 \le 21$.

Wilcoxon (Paired) Signed-Rank Test in R/Rcmdr

- In R, use the function wilcox.test(,paired=T)
- In Rcmdr, choose Statistics -> Nonparametric tests -> Paired Samples Wilcoxon Test.
- When ties, use wilcoxsign_test from the coin package!
- For the Dog Example: (See "**Paired t-test**")

```
> wilcox.test(dog$P,dog$B,paired=T)
Warning in wilcox.test.default: cannot compute
exact p-value with ties

Wilcoxon signed rank test with continuity
correction
data:  dog$P and dog$B
V = 16, p-value = 0.02365
alternative hypothesis: true location shift is
not equal to 0
```

Wilcoxon Signed Rank test notes

1. By default (if exact is not specified), an exact p-value is computed if the samples contain less than 50 finite values and there are no ties. Otherwise, a normal approximation is used.
2. For larger sample sizes (still no ties), you can use the option exact=TRUE to get the exact p-values. (Could take awhile to run.)
3. If the data contains ties, you can use the wilcoxsign_test() from the coin package, but need to specify distribution="exact".
4. The test assumes symmetry of the differences: If distribution of differences is skewed, then positive differences will tend to have higher ranks than negative differences. Usually this is not much of a problem, because differences between y's are usually close to symmetric, even when y's are not.

8. Bootstrap two-sample t CI

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.

This works just like the single-sample case, except we use the two-sample t-statistic (unequal variances), and resample (with replacement) separately within each group:

Compute $t^* = \dfrac{\bar{y}_1^{\,*} - \bar{y}_2^{\,*} - (\bar{y}_1 - \bar{y}_2)}{\sqrt{\dfrac{\left(s_1^{\,*}\right)^2}{n_1} + \dfrac{\left(s_2^{\,*}\right)^2}{n_2}}}$ from the new sample

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

The bootstrap 95% t interval is then:

$$(\bar{y}_1 - \bar{y}_2) - t^*_{0.025}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, (\bar{y}_1 - \bar{y}_2) - t^*_{0.975}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Example: Arsenic data. Arsenic samples from shallow tube wells in Bangla Desh. Compare wells constructed in 1994 with those constructed in 1998.

Y = Arsenic concentration (ug/L)

EPA limit 5 ug/L !!

http://www.bgs.ac.uk/arsenic/bangladesh/datadownload.htm

See: "**Bootstrap two-sample t CI"** for R computations.

Results: Boxplots suggest extreme skewness.

10,000 bootstrap replications gives: $t_{0.025} = 1.612$ $t_{0.975} = -3.011$

$$(\bar{y}_1 - \bar{y}_2) = 95.0 - 84.44 = 10.56 \qquad \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{92.219^2}{24} + \frac{42.0395^2}{18}} = 21.27$$

Put numbers into the formula: $10.56 - 1.612(21.27),\ 10.56 - (-3.011)(21.27)$

-23.73, 74.62

Compare to the usual t-interval (unequal variances): -32.67, 53.79

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 two-sample t CI**" for example.