Stat 411/511

Some other Two Sample Procedures

Oct 28 2015

Charlotte Wickham

stat511.cwick.co.nz

Your turn

Fill in the column for Wilcoxon Rank Sum on the test summary worksheet (posted on web).

Wilcoxon Rank Sum

	Wilcoxon Rank Sum test
Null hypothesis*	The difference in population means (or medians) is zero. OR The treatment effect is zero.
Assumptions	•Two populations have the same shape •Equal population standard deviations •Independence of subjects within and between groups.
Robust to assumptions?	(for 1 and 2 above) Still valid but tests a different null hypothesis (null: two populations are identical)
Resistant to outliers?	Resistant
Test statistic	Sum of the ranks in the smaller group

Equality of standard deviations

The **Wilcoxon Rank Sum** test (if you are using it to talk about means or medians), and the **two-sample *t*-test** both assume the population standard deviations are the same.

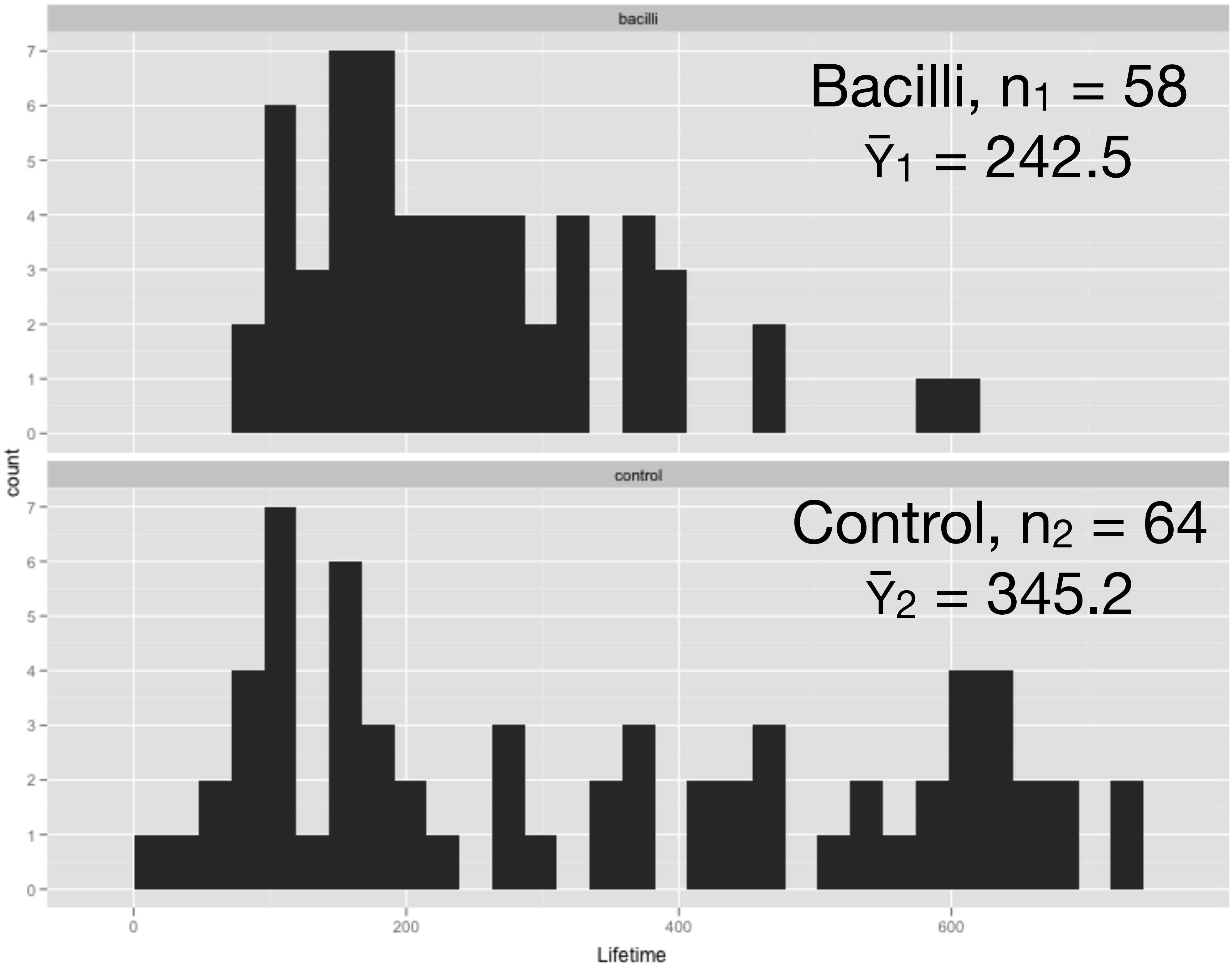
Welch's *t*-test is an alternative two-sample *t*-test for that does not assume the population standard deviations are the same, but it still assumes the populations are **Normally** distributed (although it’s robust with large sample sizes).

Sometimes our question of interest is about spread not center. **Levene's test** is a test for equal population standard deviations. You **shouldn't** use Levene's test to choose between the two-sample t-test and Welch's t-test.

Welch's *t*-test

```
qplot(Lifetime, data = ex0211) + facet_wrap(~ Group, ncol = 1)
```

The data are survival times (in days) of guinea pigs that were randomly assigned either to a control group or to a treatment group that received a dose of tubercle bacilli.



	Lifetime	Group
1	18	control
2	36	control
3	50	control
	.	
	.	
	.	
65	76	bacilli
66	93	bacilli
67	97	bacilli
68	107	bacilli

Welch's *t*-test

Allows different population standard deviations.

Instead of a **pooled** estimate of **one** standard deviation, we use **each sample** to estimate **its own population standard deviation**.

Leads to a **different standard error** on the difference in averages, and a **different degrees of freedom.**

Same t-statistic

The two-sample *t*-ratio:

$$\frac{(\overline{Y_2} - \overline{Y_1}) - (\mu_2 - \mu_1)}{\boxed{\mathrm{SE}_{\overline{Y_2} - \overline{Y_1}}}}$$

Different estimate of standard error

$$SE_{\overline{Y_2} - \overline{Y_1}} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Exact theory is hard, a *t*-distribution with a special degrees of freedom

(see pg 97 in Sleuth) is a good approximation.

No `var.equal = TRUE`

```
> t.test(Lifetime ~ Group, data = ex0211)

  Welch Two Sample t-test

data:  Lifetime by Group
t = -3.2296, df = 97.807, p-value = 0.001689
alternative hypothesis: true difference in means
is not equal to 0
95 percent confidence interval:
 -165.80689  -39.59289
sample estimates:
mean in group bacilli mean in group control
             242.5345              345.2344
```

Small p-value

We have convincing evidence that TB changes the mean lifetime of guinea pigs (Welch's t-test, two-sided p-value = 0.002).

It is estimated that the TB reduced the mean lifetime by 102.7 days.

With 95% confidence, it is estimated that the TB reduced the mean lifetime by between 39.6 and 165.8 days.

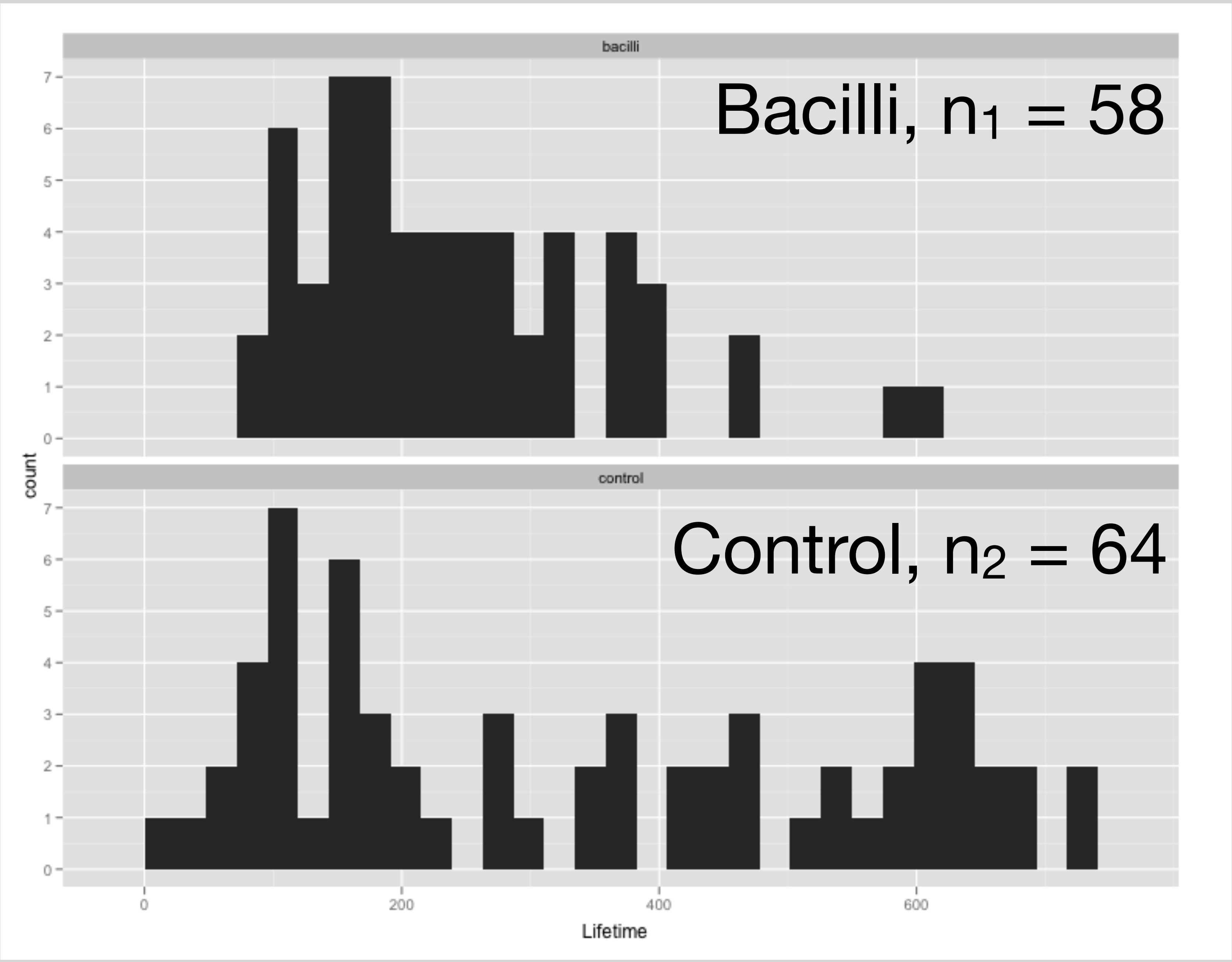
Welch's t-test is almost as good as the usual two sample t-test when the equal variance assumption is met, and much better when the assumption is violated.

Some recommend,

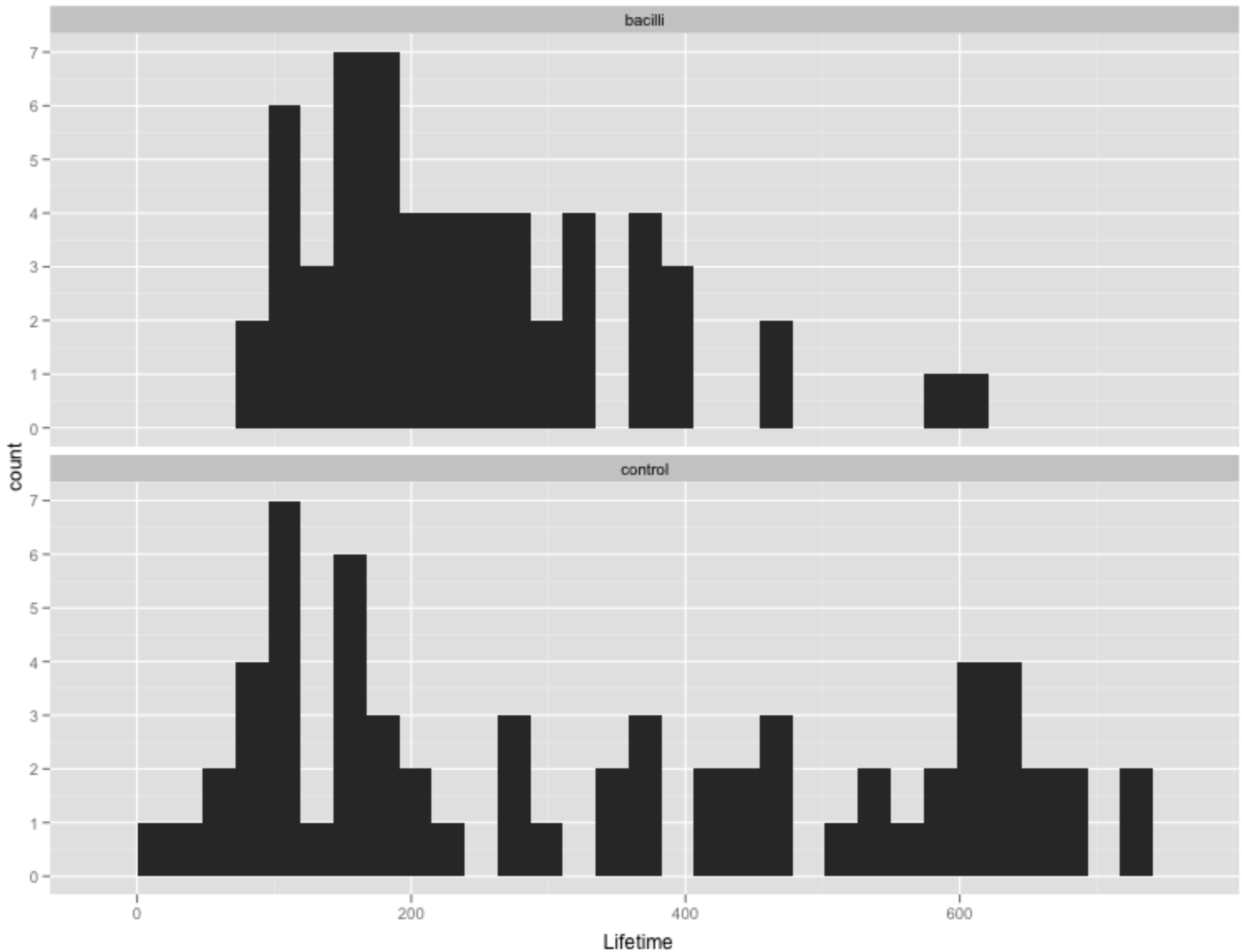
“always use Welch’s t-test”

More complicated models tend to make the equal variance assumption.

Your turn



What are we failing to communicate by saying "It is estimated that the TB reduced the lifetime by 102.7 days. "?



Caveat: Additive treatment effect isn't completely adequate.

The reduction in lifetime might depend on the Guinea pigs lifetime to begin with.

Levene's Test

Levene's test provides a formal way to test the assumption of equal population standard deviations.

Null: The standard deviation of population 1 is the same as the standard deviation of population 2, $\sigma_1 = \sigma_2$

Alternative: The standard deviation of population 1 is **not** the same as the standard deviation of population 2, $\sigma_1 \neq \sigma_2$

Brown Forsythe

Levene's test procedure

Do a **two sample t-test** (two-sided) on

$Z_1 = |Y_1 - \text{median}_1|$, and $Z_2 = |Y_2 - \text{median}_2|$

A **small** p-value is **evidence** that the mean absolute deviation from the median in population 1 is different to the mean absolute deviation from the median in population 2.

A **small** p-value is **evidence** the populations have **different** standard deviations.

Levene's test	Welch's t-test
The difference in population standard deviations is zero.	The difference in population means is zero. OR The treatment effect is zero.
•Normal populations of absolute deviations from median •Equal population standard deviations of absolute deviations from median •Independence of subjects within and between groups.	•Normal populations •Independence of subjects within and between groups.
Sleuth says it is robust.	•Robust to non-Normal populations with large samples.
Not resistant	Not resistant
two sample t-statistic on $Z_1 = \|Y_1 - median_1\|$ & $Z_2 = \|Y_2 - median_2\|$	$((\bar{Y}_2 - \bar{Y}_1) - (\mu_2 - \mu_1))$ / $SE_{\bar{Y}2 - \bar{Y}1}$ with different SE to two-sample t-test

Power

All statistical tests have a **test-statistic**.

We know what the test-statistic should look like if the null hypothesis is true, called the **null distribution**.

Two sample, paired t-test, Welch's t-test: the test statistics look like a **mathematical curve**, the **Student's t-distribution**

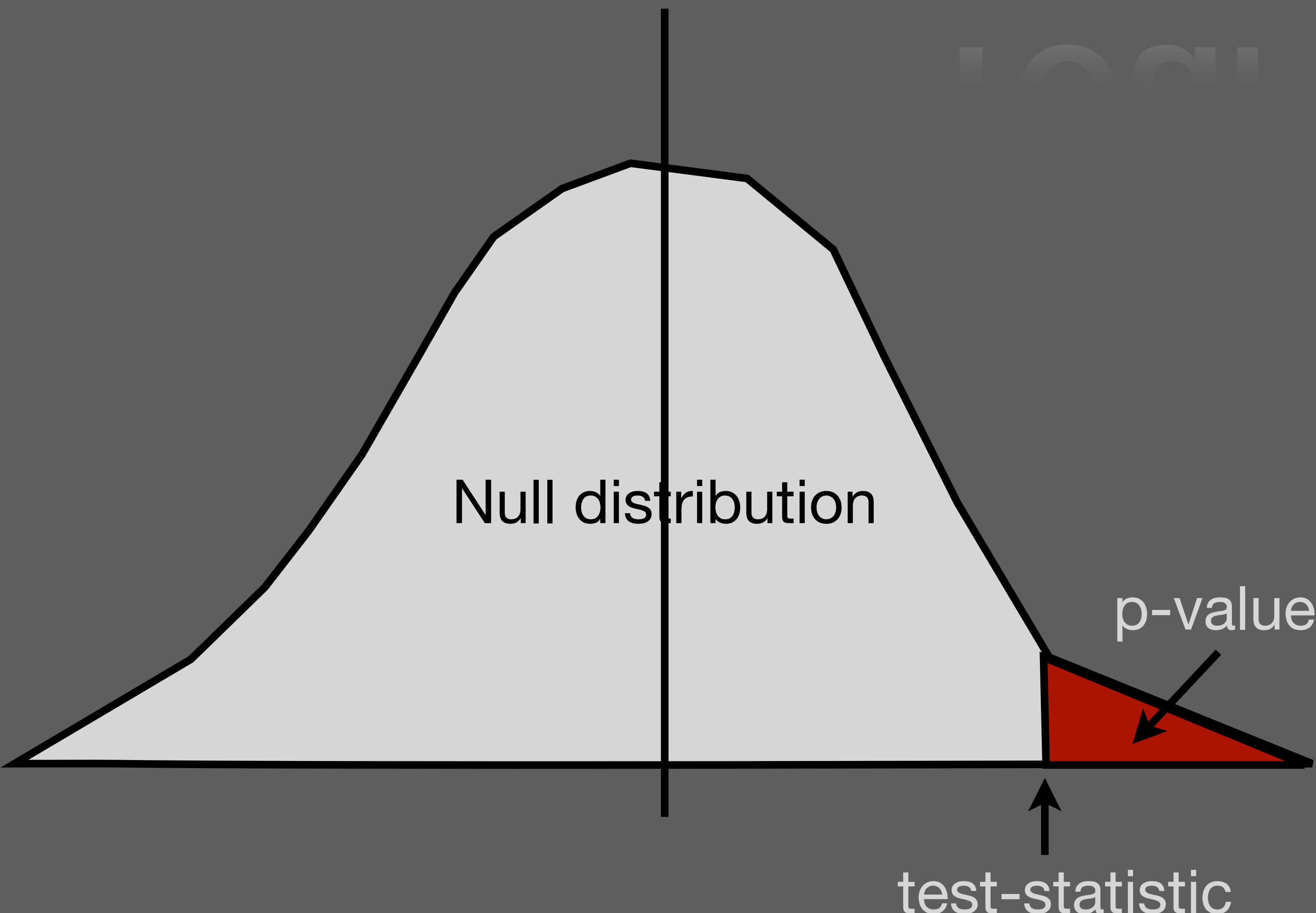
Wilcoxon Rank Sum, Wilcoxon Signed Rank, Sign test: the test statistics look like a **histogram** of possible test statistics under **permutations/randomizations of group assignment.**

The distance of the test-statistic from the center of the null distribution tells us how likely our data would be if the null hypothesis was true.

We measure this with the **p-value**, the probability of seeing as or more extreme test-statistic if the null hypothesis was true.

A **small p-value**, says "this data is unlikely if the null was true", we take that as evidence against the null.

Your turn



1. What are the possible states of truth when we get a small p-value?

2. What are the possible states of truth when we get a large p-value?

Hint #1: Paul the octopus correctly predicted all 7 of Germany's World Cup Soccer matches, can he predict soccer games?

Hint #2: OJ Simpson was declared not guilty, does that mean he was innocent?

If the p-value is small

Either:

The null hypothesis is false

Or:

The null hypothesis is true but we got an unlikely dataset by chance.

If the p-value is large

Either:

The null hypothesis is true

Or:

The null hypothesis is false but we didn't have enough evidence to reject it.

Types of errors

		Decision based on test	
		Fail to reject null	Reject null
Truth	Null is true	No error	Type I (false positive)
	Null is false	Type II (false negative)	No error

Types of errors

		Decision based on test	
		Fail to reject null	Reject null
Truth	Null is true	No error	With probability, **α** **significance level**
	Null is false	Type II (false negative)	With probability, β **power**

The **significance level** of a test is our cutoff on the **p-value** for declaring "the null hypothesis is rejected". It is the probability we will reject the null, when in fact the null is true.

The **power** of a test, is the probability we will correctly reject the null, when the null is false.

A power of 0.5 means, even when the null is false, in half of all possible samples we won't reject the null.

Generally we fix the **significance level** (at 0.05, say), and do what we can to **maximize power**.

For the **t-tests**, larger sample size, bigger true difference in means (treatment effect) and smaller standard deviation, all lead to larger power.

Using fewer assumptions often comes with a decrease in power.