

Stat 411/511

STATISTICAL POWER & CHOOSING A TEST

Feb 13 2012

Charlotte Wickham

stat511.cwick.co.nz

Correction

The randomization test assumes equal population standard deviations.

It is robust to the assumption when the sample sizes are close to equal.

Test summary updated online

What do the tests have in
common?

Commonalities

Two-sample t-test, paired t-test, Welch's t-test, Wilcoxon Rank Sum test, Wilcoxon Sign Rank Test, Sign test.

All test the null hypothesis that the two groups have the same center (mean or median depending on the test).

You can use the hypothesis and summary flowchart for all of them, but be careful to use "median" instead of "mean" when appropriate.

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

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

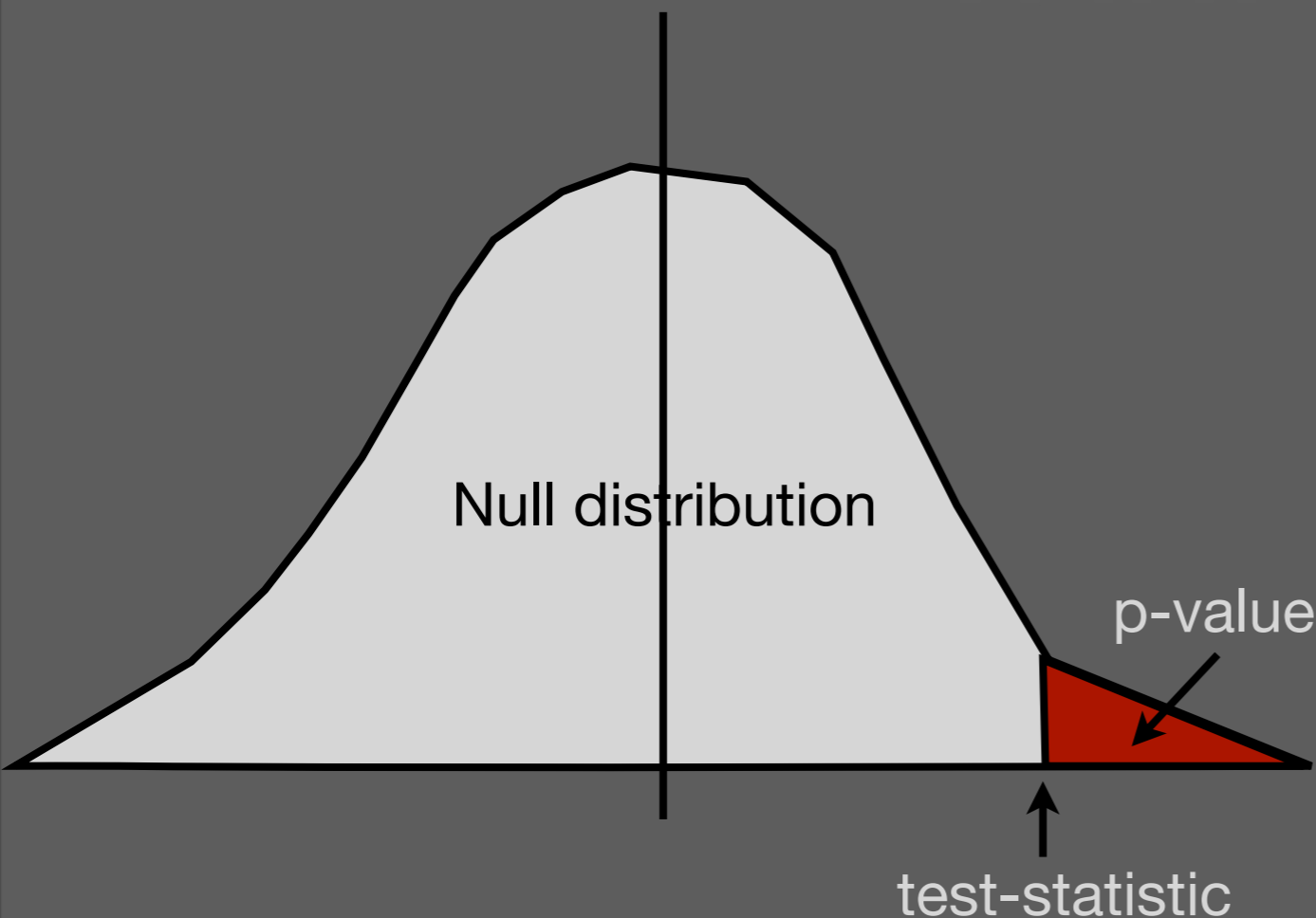
Wilcoxon Rank Sum, Wilcoxon Signed Rank, Sign test: the test statistics look like a **histogram** of possible test statistics under **permutations/randomizations of group assignment**.

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 make a decision.

Types of errors

Decision based on test

Truth

	Fail to reject null	Reject null
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

Truth

	Fail to reject null	Reject null
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 there is a 50/50 chance we won't declare a difference even if there is one.

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.

```
> power.t.test(sd = 0.25, sig.level = 0.05, delta = 0.2, power = 0.8, type = "paired")
```

Paired t test power calculation

```
      n = 14.3028
  delta = 0.2
     sd = 0.25
sig.level = 0.05
  power = 0.8
alternative = two.sided
```

Schizophrenia Study

NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs

```
> power.t.test(sd = 0.25, sig.level = 0.05, delta = 0.1, power = 0.8, type = "paired")
```

Paired t test power calculation

```
      n = 51.00957
  delta = 0.1
     sd = 0.25
sig.level = 0.05
  power = 0.8
alternative = two.sided
```

NOTE: n is number of *pairs*, sd is std.dev. of *differences* within pairs

```
>
```

How do the tests differ?

What are the differences?

- 1.** There are two classes of tests: those for two independent samples, and those for two paired samples (one sample of differences).
- 2.** Within a class the tests have different assumptions
- 3.** Within a class the tests have different resistance to outliers

How to choose a test

Step 1: Do you have two independent groups, or two paired groups?

Step 2: What assumptions seem reasonable?

Step 3: Are outliers an issue?

Flowchart posted on web.

One way to proceed through checking assumptions.

Checking of assumptions can be done with data at hand, and sometimes you have external information to help.

Treating an assumption as **not violated** when **it is violated**, will result in an **invalid test**.

Treating an assumption **as violated** when **it is not violated**, will result in a **valid test**, but it may be less efficient.



You might need a bigger sample to get the same power

Your turn

Which analysis would be appropriate?

Ex 29, pg 108

Ex 32 pg 111