Two-sample tests

The two-sample z-test

- Recall that the standard z-test is used to compare the observed outcome (e.g. sample mean, sample percentage) to the expected, i.e. the hypothesized parameter:

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}}$$

- The z-test can also be used to compare two populations, based on two samples

The two-sample z-test

First, suppose we have independence and we want to assess whether:

$p_1 =$ proportion in sample 1

is equal to

$p_2 =$ proportion in sample 2

"Nothing unusual is going on" means $p_1 = p_2$. It's common to look at the difference $p_2 - p_1$ instead:

$H_0: \; p_2 - p_1 = 0 \qquad\qquad H_1: \; p_2 - p_1 \neq 0$

p_1 is estimated by $\hat{p}_1 = 55\%$, p_2 by $\hat{p}_2 = 58\%$. The central limit theorem applies to the difference $\hat{p}_2 - \hat{p}_1$ just as it does to $\hat{p}_1$ and $\hat{p}_2$. So we can use a z-test:

The two-sample z-test

We can use a z-test for the difference $\hat{p}_2 - \hat{p}_1$:

$$z = \frac{\text{observed difference} - \text{expected difference}}{\text{SE of difference}} = \frac{(\hat{p}_2 - \hat{p}_1) - (p_2 - p_1)}{\text{SE of difference}}$$

An important fact is that if $\hat{p}_1$ and $\hat{p}_2$ are independent, then

$$\text{SE}(\hat{p}_2 - \hat{p}_1) = \sqrt{(\text{SE}(\hat{p}_1))^2 + (\text{SE}(\hat{p}_2))^2}.$$

Therefore we can estimate $\text{SE}(\hat{p}_2 - \hat{p}_1)$ with the plug-in principle or with the bootstrap.

Note that $p_2 - p_1 = 0$ in the z-statistic, as we compute it assuming H_0 is true.

The two-sample z-test

The two-sample z-test is applicable in the same way to the difference of two sample means in order to test for equality of two population means.

If the two samples are independent, then again

$$\mathrm{SE}(\bar{x}_2 - \bar{x}_1) \ = \ \sqrt{(\mathrm{SE}(\bar{x}_1))^2 + (\mathrm{SE}(\bar{x}_2))^2}$$

and $\mathrm{SE}(\bar{x}_1) = \frac{\sigma_1}{\sqrt{n_1}}$ is estimated by $\frac{s_1}{\sqrt{n_1}}$. Or estimate $\mathrm{SE}(\bar{x}_2 - \bar{x}_1)$ using the bootstrap.

All of the above two-sample tests require that the two samples are independent.

The two-sample z-test

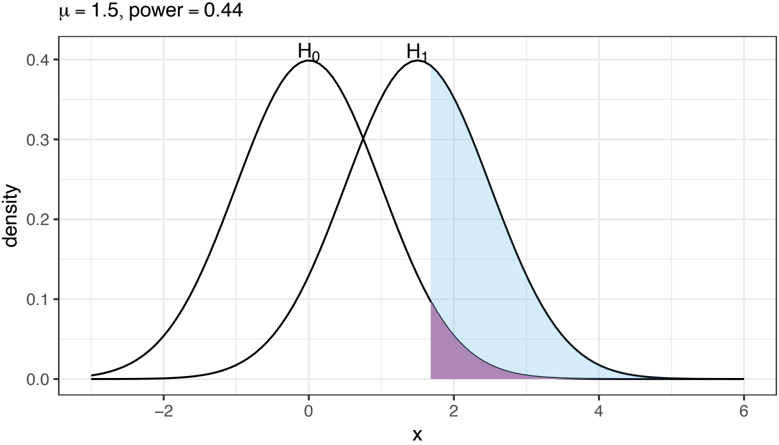
$$\mathrm{SE}(\bar{x}_2 - \bar{x}_1) \;=\; \sqrt{(\mathrm{SE}(\bar{x}_1))^2 + (\mathrm{SE}(\bar{x}_2))^2}$$

Recall that the square-root law applies to $\mathrm{SE}(\bar{x}_1)$ and $\mathrm{SE}(\bar{x}_2)$, hence also to $\mathrm{SE}(\bar{x}_2 - \bar{x}_1)$:

If the two samples are large, then the SE will be small.

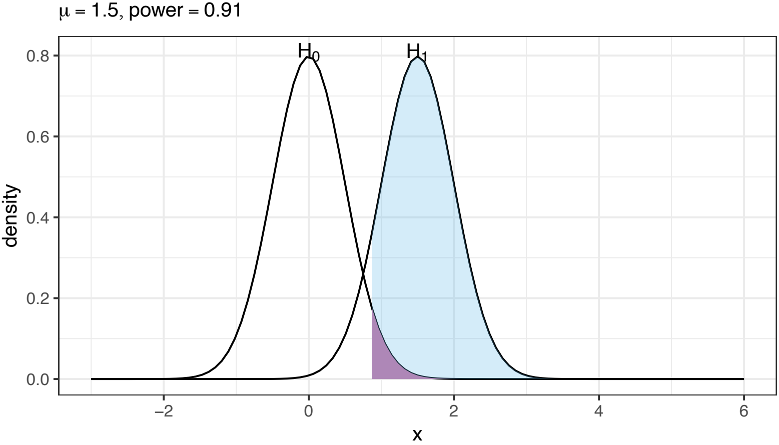
As a consequence, we will be able to detect smaller differences in the population means if the sample sizes are larger:

Statistic has SE=1 under null and alternative, means differ by 1.5



Power is 44%

Statistic has SE=0.5 under null and alternative, means differ by 1.5



Power is 91%

A real-world example: flight delays

Let's now consider a real-world example. We have the flight delay times for Delta and United for their first 30000 flights of 2015. Those data are from a Kaggle competition. The data are on canvas as "short-flights".

```r
mean(delta.delays)
```

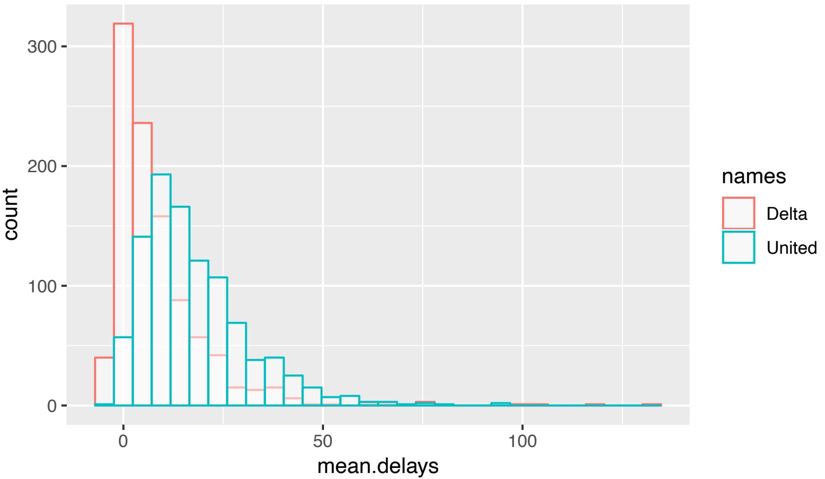
```
[1] 8.039098
```

```r
mean(united.delays)
```

```
[1] 17.88113
```

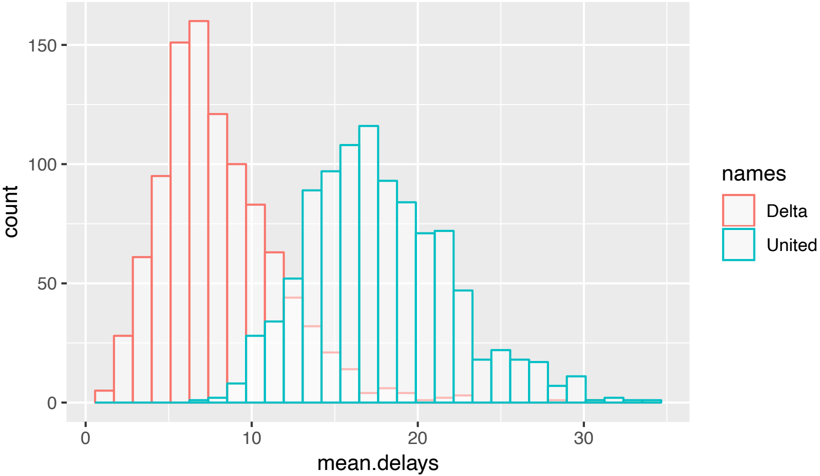
Suppose we had only small samples available to compare the means. We will see how larger sample sizes provide more power to declare that United has longer delays.

Histogram of 1000 means of samples of size 10



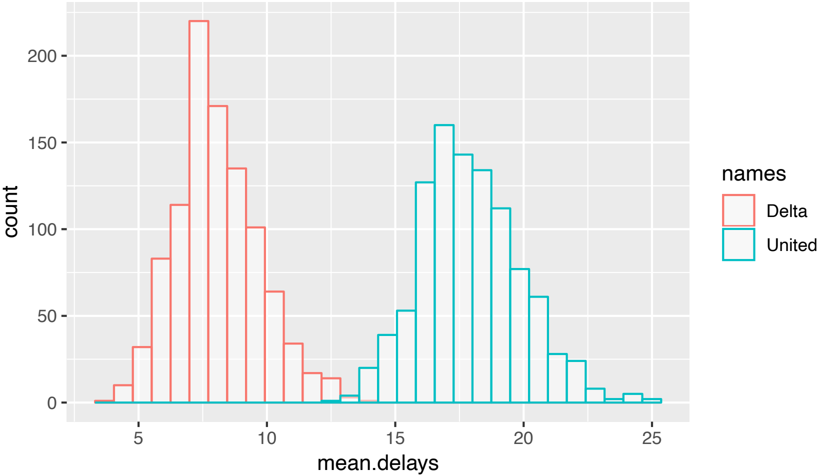
Mean delays using samples of size 10

Histogram of 1000 means of samples of size 100



Mean delays using samples of size 100

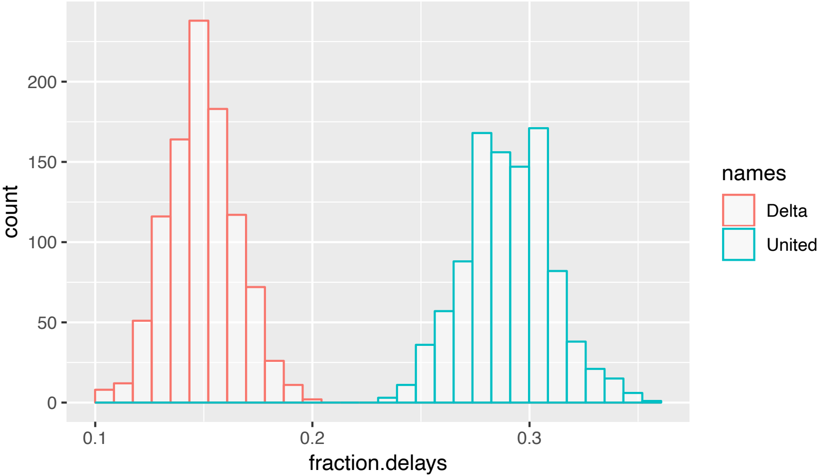
Histogram of 1000 means of samples of size 500



Mean delays using samples of size 500

A more complicated statistic

Let's look at the number of times each airline has delays more than 15 minutes. This is more robust against outliers (very long delays). The same conclusions apply.



The paired-difference test

Do husbands tend to be older than their wives?

The ages of five couples:

Husband's age	Wife's age	age difference
43	41	2
71	70	1
32	31	1
68	66	2
27	26	1

The two-sample z-test is not applicable since the two samples are not independent. Even if they were independent, the small differences in ages would not be significant since the standard deviations are large for husbands and also for the wives.

The paired-difference test

Since we have paired data, we can simply analyze the differences obtained from each pair with a regular z-test, which in this context of **matched pairs** is called **paired z-test**:

H$_0$: population difference has mean zero

$t \ = \ \frac{\bar{d}-0}{\mathrm{SE}(\bar{d})}$, where d_i is the age difference of the ith couple.

$\mathrm{SE}(\bar{d}) = \frac{\sigma_d}{\sqrt{n}}$. Estimate σ_d by $s_d = 0.55$. Then
$z = \frac{1.4-0}{0.55/\sqrt{5}} \ = \ 5.69$

The independence assumption is in the sampling of the couples.

The sign test

What if didn't know the age difference d_i but only if the husband was older or not?

We can test

H_0: half the husbands in the population are older than their wives

using 0/1 labels and a z-test, just as testing whether a coin is fair:

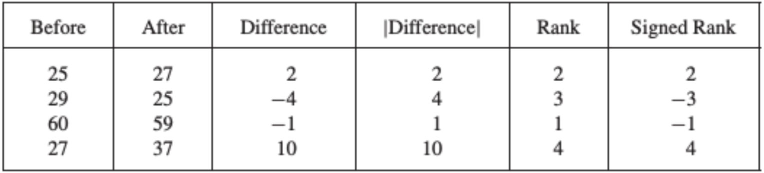
$$z \;=\; \frac{\text{sum of 1s}_{-\frac{n}{2}}}{\text{SE of sum}} \;=\; \frac{5-\frac{5}{2}}{\sqrt{5}\frac{1}{2}} \;=\; 2.24 \quad \text{since } \sigma = \frac{1}{2} \text{ on } \mathrm{H}_0.$$

The p-value of this **sign-test** is less significant than that of the paired z-test. This is because the latter uses more information, namely the size of the differences.

On the other hand, the sign test has the virtue of easy interpretation due to the analogy to coin tossing.

A Nonparametric Method - The Wilcoxon Signed Rank Test

A nonparametric test based on ranks can be constructed for **paired** samples. Example:



Before	After	Difference	|Difference|	Rank	Signed Rank
25	27	2	2	2	2
29	25	−4	4	3	−3
60	59	−1	1	1	−1
27	37	10	10	4	4

See chapter 11.3.2, p. 448 in: Rice, John A. Mathematical statistics and data analysis. Cengage Learning, 2006.

The Signed Rank Test - Test Statistic

1. Calculate the differences D_i and the absolute values of the differences and rank them.

2. Restore the signs of the differences to the ranks, obtaining signed ranks.

3. Calculate W_+, the sum of those ranks that have positive signs. In the example, this is $W_+ = 2 + 4 = 6$.

Why?

The idea is as follows:

- If there truly is no difference between the two paired conditions, we expect about half of the D_i to be positive and half negative. This means that W_+ will be neither too small or too large (since the ranks should neither all be very small or very large if D_i is negative and positive about half of the time each).
- However, if one condition generates larges values than the other, then W_+ will be more extreme.
- Therefore, we can use W_+ as a test statistic and reject for extreme values (very small or very large).

What is the Null Hypothesis?

- H_0: The distribution of D_i is symmetric about 0. For example, this would be true if each of the two groups are randomly assigned to e.g. treatment and control and the treatment has no effect at all.
- We need to know the distribution of W_+ under the null hypothesis. Then we need to find the "rejection region", i.e. we need to look at the tails of the null distribution in a way such that our test has level α.
- Under the null, the D_i are symmetric about 0. So e.g. the k-th value of D is equally likely to be positive or negative - and in particular any assignment of the signs to the integers $1, ..., n$ (ranks) is equally likely.
- There are 2^n such assignments. Computer packages can calculate this null distribution for you.

Notes

- If we have a large sample size, we can also use the CLT and calculate $z = \frac{W_{+,obs} - W_{+,exp}}{SE(W_{+})}$.
- Caution: Be careful about differences of 0 in the two groups and ties in the ranks!

Another non-parametric method: The Mann-Whitney Test

- Sometimes the Mann-Whitney test is also called the Wilcoxon rank sum test. Beware: The literature is not unanimous about the definitions of the Wilcoxon rank sum and Mann-Whitney tests.

- More information: Chapter 11.2.3 in Rice, John A. Mathematical statistics and data analysis. Cengage Learning, 2006.

- Let's consider it in a specific context: We have $m + n$ experimental units to assign to a **treatment and a control randomly**. Suppose n units are randomly chosen to be in the control and m to be in the treatment. We want to test the null that there is no effect of the treatment.

- Idea: If there is no difference between the two conditions, then any difference in the outcomes is due to the randomization.

Mann-Whitney Test

- Group all $m + n$ observations together and rank them in the order of increasing size (we will assume that there are no ties).
- Then, calculate the sum of the ranks of those observations that come from the control. If this sum is “too small” or “too large”, we reject the null.
- Key idea: It turns out that it is possible to calculate the distribution of the sum of the ranks under the null exactly, because under the null every assignment of ranks is equally likely.
- Computer programs such as R can calculate this for you, see e.g. `wilcox.test()` in R.

Duality of Confidence Intervals and Hypothesis Tests

- There is a duality between confidence intervals and hypothesis tests.
- Key message: The confidence interval consists of all values for which we would NOT reject the null hypothesis.
- Specifically, suppose we are testing $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$. Then the confidence interval contains all values of μ_0 for which we would accept the null hypothesis $H_0 : \mu = \mu_0$.