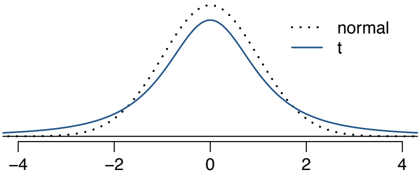
Comparing samples using *t*-tests

Paul M. Magwene

Fall 2016

The *t*-distribution

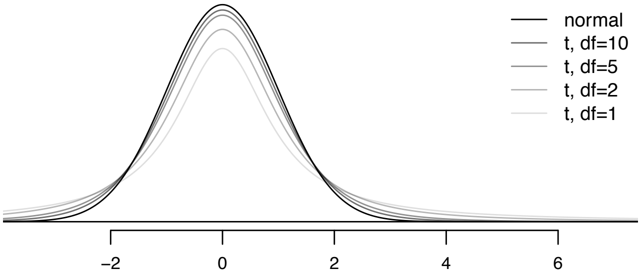
- When working with small samples, and the population standard deviation is unknown (usually the case), the uncertainty of the standard error estimate is addressed by using the t-distribution
- This distribution also has a bell shape, but its tails are thicker than the normal model's.
- Therefore observations are more likely to fall beyond two SDs from the mean than under the normal distribution.
- These extra thick tails are helpful for resolving our problem with a less reliable estimate the standard error (since *n* is small)



Credit: OpenIntro Stats, 3rd ed.

The *t*-distribution, cont.

- Centered at zero, like the standard normal distribution
- Has a single parameter, degrees of freedom (*df*)



What happens to the shape of the distribution as the degrees of freedom increases?

Credit: OpenIntro Stats, 3rd ed.

One sample inference using the *t*-distribution

- If $n < 30$, sample means follow a t-distribution with $SE = \frac{s}{\sqrt{n}}$ and $df = n - 1$
- Conditions:
  - independence of observations (random sample, n < 10% of population)
  - No extreme skew
- Confidence interval for mean:

$$\text{point estimate} \pm t^\star_{df} \times SE$$

- Hypothesis testing, test statistic:

$$T = \frac{\text{point estimate} - \text{null value}}{SE}$$

Credit: OpenIntro Stats, 3rd ed.

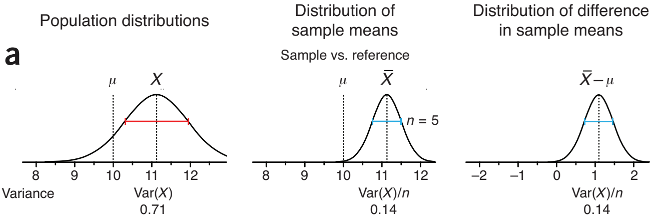
One-sample t-test, example

You treat a sample of mice with a drug, X, and measure the expression of the gene YFG1 following treatment. For a sample of five mice you observe the following expression values:

- X = {11.25, 10.5, 12, 11.75, 10}

This data set can be downloaded from the class wiki as the file `gene-expression-1sample.csv`

One-sample t-test, relevant distributions and comparisons



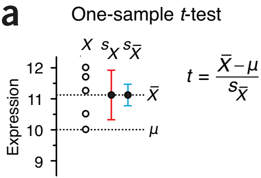


Image credit: Krzywinski & Altman 2014

Hands-on, one-sample t-test

You have established from previous studies that the expression level of YFG1 in control mice is 10. You wish to determine whether the average expression of YFG1 in mice treated with drug X differs from control mice.

1. Write down an appropriate null and alternative hypothesis.
2. Calculate 95% confidence intervals for the mean expression of YFG1 in the sample data set
3. Calculate the test statistic T
4. Calculate a p-value giving the probability that you'd find an average expression value at least as extreme as that observed, if the null hypothesis was true

Hint: Recall the `dt`, `pt`, `qt` and `rt` functions for working with the t-distribution

t-tests in R

The R function `t.test` makes it easy to calculate one- and two-sample t-tests

1. Read the help documentation on `t.test`
2. Carry out a one-sample t-test for the gene expression data set, using the null hypothesis that average gene expression for YFG1 is 10
3. Do the results given by `t.test` correspond to the your "by hand" calculations from the previous exercise? If not, go back and figure out where your previous calculations went awry.

Two-sample t-test

The two-sample t-test is used to determine if two populations, represented by observed samples, have equal means.

Test statistic:

$$T = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

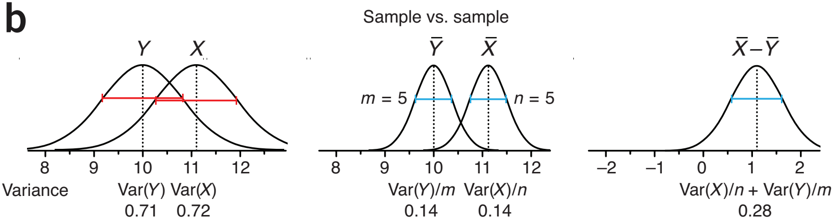
Two-sample t-tests, example

You treat samples of mice with two drugs, X and Y. We want to know if the two drugs have the same average effect on expression of the gene YFG1. The measurements of YFG1 in samples treated with X and Y are as follows:

- $X = \{11.25, 10.5, 12, 11.75, 10\}$
- $Y = \{8.75, 10, 11, 9.75, 10.5\}$

This data set is available on the course wiki as `gene-expression-2sample.csv`

Two-sample t-tests, relevant distributions and comparisons



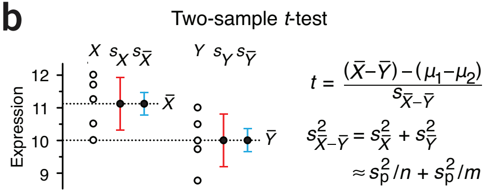


Image credit: Krzywinski & Altman 2014

Hands-on, two-sample t-tests in R

The `t.test` function we introduced earlier can be used to carry out two-sample t-tests.

1. Re-read the documentation for `t.test`
2. Use ggplot and `geom_point` to create a figure showing the expression values (y-axis) as a function of drug treatment (x-axis)
3. Carry out a two sample t-test comparing mean expression values for drugs X and Y.

Paired t-test

Paired t-tests are used in the case where sample observations are correlated, such as “before-after” studies where the same individual or object at different time points or paired comparisons where some aspect of the same object is measured in two different conditions. The repeated measurement of the same individual/object means that we can’t treat the two sets of observations as independent.

Paired t-test, test statistic

- Let the variable of interest for object i in the paired conditions be designated x_i and y_i
- Let $D_i = y_i - x_i$ be the paired difference for object i
- Let $\overline{D}$ be the mean difference and s_D be the standard deviation of the differences
- The standard error of mean difference is $SE(\overline{D}) = \frac{s_D}{\sqrt{n}}$

The test statistic is

$$T = \frac{\overline{D}}{SE(\overline{D})}$$

Under the null hypothesis, this statistic follows a t-distribution with $n - 1$ degrees of freedom.

Paired t-test, visual representation

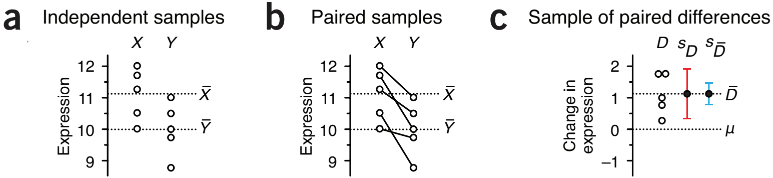


Image credit: Krzywinski & Altman 2014

Paired t-test, example

You measure the expression of the gene YFG1 in five mice. You then treat those five mice with drug X and measure gene expression again.

- YFG1 expression before treatment = {11, 10, 10.5, 8.87, 9.75}
- YFG1 expression after treatment = {12, 11.75, 11.25, 10.5, 10}

This data set is available as `gene-expression-paired.csv` from the course wiki.

Hands-on, paired t-test in R

1. Use the `t.test` function to carry out a paired t-test for paired gene expression data set.
2. Compare your results to the two-sample t-test for the same data if you treat it as unpaired. How do the results differ? Can you provide an intuitive explanation for why you would get different results when treating the data as paired vs unpaired?