Some commonly used tests

One-sample t test

Given a normal random sample $y_1, \ldots, y_n$, with $Y_i \sim \mathcal{N}(\mu, \sigma^2)$, a classical testing problem on μ is of the form (for two-sided alternative, say)

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

The test statistic is given by

$$T = \frac{\overline{Y} - \mu_0}{\sqrt{\frac{S^2}{n}}} \sim t_{n-1}\,, \qquad \text{when } H_0 \text{ is true}$$

with the p-value given by

$$p = \Pr_{H_0}(|T| \geq |t_{obs}|)$$

which can be computed as $p = 2\Pr_{H_0}(T \geq |t_{obs}|) = 2\{1 - F_{t_{n-1}}(|t_{obs}|)\}$, since the t distribution is symmetric around 0.

Ex

$$T = \frac{\bar{Y} - \mu_0}{S/\sqrt{n}} \sim t_{n-1}$$

$n = 130$
$\bar{y} = 36.8$
$s^2 = 0.166$

WITH HYPOTHESIS $\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{cases}$

A PREVIOUS AVERAGE FOR BODY TEMPERATUS WAS 37, NOW WE KNOW IT'S 36.8. WE COMPUTE THE t_{OBS} AND COMPUTE THE P-VALUE:

$$t_{OBS} = \frac{36.8 - 37}{\sqrt{0.166}/\sqrt{130}} \simeq -5.59$$

THIS IS A VERY LARGE NUMBER FOR A t DISTRIBUTION, USUALY IT MOVES ON THE SAME VALUE AS STD NORMAL DISTRIBUTION. NOW, THE P-VALUE:

$$p = P_{H_0}(|T| \geq |t_{OBS}|) = 2 P_{H_0}(T \geq |t_{OBS}|) = 2\{1 - F_{t_{n-1}}(|t_{OBS}|)\}$$

$$p = 2(1 - F_{t_{n-1}}(|-5.59|)) = 1.23 \cdot 10^{-7}$$

→ VERY STRONG EVIDENCE THAT BODY TEMPERATURE IS DIFFERENT FROM 37

AND WHAT IF THE ASSUMPTION IS THAT IS SMALLER?

$$P(T < t_{obs}) = F_{t_{n-1}}(t_{obs})$$

$$F_{t_{n-1}}(-5.6) = 6.47 \cdot 10^{-8} \rightarrow \text{EVEN MORE STRONG}$$

Example

The DAAG book introduces the simple dataset `pair65`, about an experiment on the effect of heat on the stretchiness of elastic bands: a small sample of differences between two different conditions for 9 bands.

heated	ambient	difference
244	225	19
255	247	8
253	249	4
254	253	1
251	245	6
269	259	10
248	242	6
252	255	-3
292	286	6

Example (cont'd)

Focusing on the 9 differences on the amount of stretch, we test

$$\begin{cases} H_0 : \mu = 0 \\ H_1 : \mu \neq 0 \end{cases}$$

by means of the `t.test` function, resulting in significance at 5% level

```
##
##  One Sample t-test
##
## data:  difference
## t = 3.1131, df = 8, p-value = 0.01438
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   1.641939 11.024728
## sample estimates:
## mean of x
##  6.333333
```

Approximate tests

For large random samples, the Central Limit Theorem ensures that $\overline{Y} \dot{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$, being $\mu = E(Yi)$ and $\sigma^2 = \mathrm{var}(Y_i)$.

A test statistic for $H_0 : \mu = \mu_0$ is therefore

$$Z = \frac{\overline{Y} - \mu_0}{\sqrt{\frac{S^2}{n}}} \dot{\sim} \mathcal{N}(0,1)\,, \qquad \text{when } H_0 \text{ is true}$$

CAN USE BOTH
THE SUGGEST IS
THE SECOND

$\widehat{\pi}$:= VALUE OBSERVED

π_0 := VALUE ASSUMED UNDER NULL HYPOTESIS (TRUE)

The estimator of the variance S^2 can be replaced by a more suitable one. For example, for binary data, $Y_i \sim \mathcal{B}_i(1, \pi)$, commonly used test statistics are $Z = \dfrac{\widehat{\pi} - \pi_0}{\sqrt{\frac{\widehat{\pi}\,(1-\widehat{\pi})}{n}}}$ or $Z = \dfrac{\widehat{\pi} - \pi_0}{\sqrt{\frac{\pi_0\,(1-\pi_0)}{n}}}$, the latter being preferable.

Tests based on the CLT are instances of **approximate tests**, for which the property concerning the Type I error level holds only approximately.

Two sample t-test

Given two **independent normal samples**, represented by $X_i \sim \mathcal{N}(\mu_X, \sigma_X^2)\,, i = 1, \ldots, n_X$ and $Y_i \sim \mathcal{N}(\mu_Y, \sigma_Y^2)\,, i = 1, \ldots, n_Y$, the test statistic for testing the equality between the two means is

$$T = \frac{\overline{X} - \overline{Y}}{\mathrm{SE}(\overline{X} - \overline{Y})}$$

ASSUMING $\sigma_x^2 \neq \sigma_x^2$

with $\mathrm{SE}(\overline{X} - \overline{Y})$ estimated by $\sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}$.

A different formula is instead adopted is if it is possible to assume that $\sigma_X^2 = \sigma_Y^2$.

The distribution of T when H_0 is true is given by a suitable t distribution.

Like for the one-sample case, there are general formulas for large samples, employing the normal distribution.

EXAMPLE ON COIN TOSS

* CHECK R FILE FOR RESULTS

$X_i \sim N(\mu_x, \sigma_x^2)$ $Y_i \sim N(\mu_y, \sigma_y^2)$ BOTH ASSUMED INDIPENDENT

$\hat{x} = 26.05$
$\hat{y} = 17.46$ } SAMPLE MEANS

$S_x^2 = 414.05$
$S_y^2 = 196.72$ } SAMPLE VARIANCES

$n_x = 81$

$n_y = 82$

LET'S MAKE SOME HYPOTHESIS:

$$\begin{cases} H_0: \mu_x = \mu_y \Rightarrow \mu_x - \mu_y = 0 \\ H_1: \mu_x > \mu_y \Rightarrow \mu_x - \mu_y > 0 \end{cases}$$

WE COULD MAKE ANOTHER ASSUMPTION THAT STATE : $\sigma_x^2 = \sigma_y^2$

IF WE CAN ALSO MAKE A TEST AND ASSUME WE CAN USE

$$T = \frac{\bar{X} - \bar{Y} - (\mu_0 - \mu_1)}{SE(\bar{x} - \bar{y})} \sim t_{n_x + n_y - 2}$$

AND HAVE NOW 2 CASES.

WE CAN COMPUTE THE STD ERROR :

$$SE(\bar{X} - \bar{Y}) = \sqrt{\frac{(n_x - 1) S_x^2 + (n_y - 1) S_y^2}{n_x + n_y - 1} \cdot \left(\frac{1}{n_x} + \frac{1}{n_y}\right)}$$

IF WE DON'T MAKE THAT ASSUMPTION :

$$SE(\bar{X} - \bar{Y}) = \sqrt{\frac{S_x^2}{n_x} + \frac{S_y^2}{n_y}}$$

THIS SE FORMULA EVEN IF NOT IN STD BUT HAVE VERY LARGE SAMPLE

Paired t-test

Paired observations arise whenever each unit of a random sample of size n is observed twice, under different conditions, so that we end up again with two set of variables $X_i \sim \mathcal{N}(\mu_X, \sigma_X^2)\,, i = 1, \ldots, n$ and $Y_i \sim \mathcal{N}(\mu_Y, \sigma_Y^2)\,, i = 1, \ldots, n$.

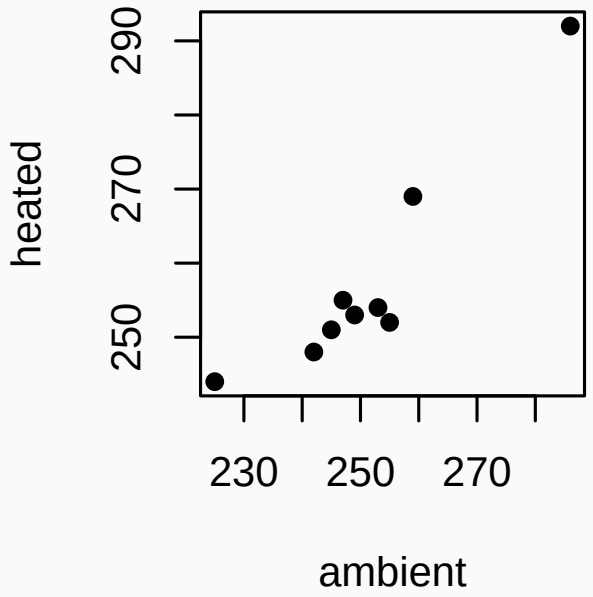
However, now the pair (X_i, Y_i) refers to the same unit, so that the two samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ **are no longer independent**.

The `pair65` data set is exactly of this nature. Like in that example, the resolution is to focus on the random sample of the n differences $D_i = X_i - Y_i$, for which $E(D_i) = \mu_X - \mu_Y$: for testing the equality of the two means μ_X and μ_Y we just apply the theory for the one-sample t-test, with $\mu_0 = 0$.

For the `pair65` data set, the p-value of about 0.014 suggests that heat may indeed have an effect on stretchiness.

Example

Even though the `pair65` data is very small, the fact that the two groups of observations are not independent is readily suggested by a scatterplot



By (blindly) applying the test for independent data we would get a p-value of about 0.40, hinting at a quite different conclusion.

UNTIL NOW WE FOUND TWO TIPES OF α:

- α FOR THE TEST
- α FOR THE CONFIDENCE INTERVALL

Relation between tests and confidence intervals

Main result

As displayed for the `pair65` data testing, the `t.test` R function returns also the confidence interval for the parameter under testing, in that case the true mean of the differences in stretchiness.

This is not by chance, since there is a close connection between hypothesis testing on the value of a certain parameter and confidence intervals for that parameter.

For the case of a mean, for example, the basic idea is that

> *If the confidence interval for μ does not contain zero, this is equivalent to rejection of the hypothesis that the true mean is zero.*

Important: the connection is between two-sided confidence intervals and two-sided alternative hypotheses. For one-sided alternative hypotheses, the connection is with one-sided confidence intervals.

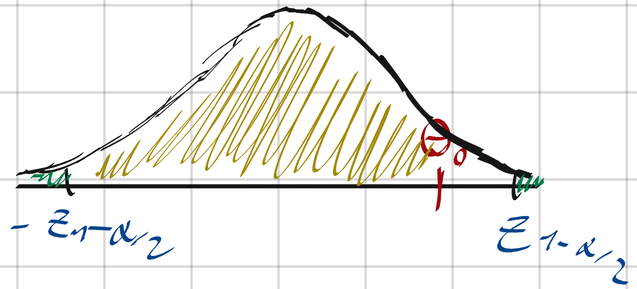
More precisely

The general result is as follows, and states a perfect equivalence between the two methods:

1. Given a method to find a confidence interval of level $(1-\alpha)\%$ for a certain scalar parameter θ, we can establish whether the p-value for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ is smaller than the significance level α by checking if θ_0 is included in the interval

2. Given a method to find a p-value for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, we can obtain a confidence interval of level $1-\alpha$ by selecting all the θ_0 values that will lead to a p-value larger than α

LET'S TRY TO SCATCH THE 1° METHOD



CONSTRUCT A CONFIDENCE INTERVAL
THEN, CHECK IF θ_0 IS IN THE INTEVAL
IF SO, WE EXPECT THAT THE P-VALUE IS LARGER THAN α, WHERE α IS THE SIGNIFICANT LEVEL OF THE TEST THAT IS EQUAL TO THE α SPECIFIED IN THE CONFIDENCE INTERVAL

ALL THE VALUE INSIDE THE $1-\alpha$ INTERVALL, WE ACCEPT THE NULL HYPOTHESIS
WE REJECT THE NULL HYPOTHESIS

Example: `pair65` data

The 95% and 99% confidence intervals for the mean of the differences are, respectively

IF WE SET $H_0: \mu = 0$, COMPUTE P-VALUE AND SET INTERVALS WITH (1-P-VALUE)

REJECT →	95%	1.6419	11.0247
REJECT →	99%	-0.4930	13.1596
ACCEPT →	98.56217%	0.0000	12.6667

⚠ JUST TO SHOW THE RELATIONS, IN PRACTICE IT DOESN'T HAPPEN

The 95% confidence interval does not contain zero, while the wider 99% does, implying that the hypothesis $\mu = 0$ is rejected for $\alpha = 0.05$, but not for $\alpha = 0.01$.

Note that for a confidence interval of level $1 - p = 0.9856217$, we obtain a lower limit exactly equal to 0: the p-value, in fact, corresponds to a significance level which is borderline between rejection and non-rejection of H_0.

Nonparametric tests

USED WHEN YOU DON'T KNOW FROM WHAT DISTRIBUTION THE DATA CAME FROM AND YOU ARE NOT CONFIDENT ON DOING ASSUMPTION ON THE DISTRIBUTION.

WHEN MAKING TEST, YOU SPECIFY ASSOCIATION BETWEEN VARIABLE OR IF A SAMPLE COME FROM A DISTRIBUTION. AND SO, ON.

YOU CAN TEST ALMOST WETHEVER YOU WANT BUT CONSTRUCTION MIGHT BE A LITTLE COMPLEX

Main idea behind nonparametric tests

Nonparametric tests specify only partially a statistical model for the data, so that they may provide more robust inferences than parametric tests with contaminated data, outliers or, more generally, in settings where model specification is hard.

This is sometimes useful, especially when only certain aspects of the data are of interest, or for checking the results obtained with a full model specification.

The details of such tests, and more generally the theory supporting their validity, would require a substantial amount of space. Here we just mention such solutions in passing, as a tool in the statistician's reservoir that at times may be a useful complement to parametric tests.

Wilcoxon rank sum and signed rank tests

The main idea of nonparametric tests is illustrated by the Wilcoxon rank sum test, which can be used to replace the t test when normality is doubtful, due to outliers or excessive rounding, for example.

The test uses the **ranks**, which are the index of each observation in the sample sorted in ascending order. For instance, for the `pair65` set of differences

difference	rank
19	9
8	7
4	3
1	2
6	5
10	8
6	5
-3	1
6	5

IN THE EXAMPLE ARE WRONG

WE CAN USE IT FOR EVALUATE THE SIMMETRY OR THE DIFFERENCE FROM TWO SAMPLES, WITHOUT BEING INFLUENCED BY THE OUTLIERS

WE ORDER OUR DIFFERENCE IN ABSOLUTE VALUE, IT DOESN'T CARE ABOUT SIGN, FROM SMALLEST TO LARGEST

Wilcoxon rank sum and signed rank tests

In the example, the R function `wilcox.test` returns a p-value of 0.017, which is very similar to what returned by the parametric test, thus reinforcing the conclusion.

There are also two-sample extensions, for both independent data or paired data (though the latter can be performed by considering the differences, as done here). The two-sample version (for independent samples) is known as *signed rank test* or *Mann-Whitney test*.