Hypothesis testing and using t-tests

Practical Fundamentals of Probability and Statistics (NPFL 081)

by Martin Holub
Charles University in Prague, 2017–18

Terms

Statistical hypothesis
A statement about the parameters describing a population (not a sample).

Significance level of a test (α)
The test's probability of incorrectly rejecting the null hypothesis.

Statistic
A value calculated from a sample, often to summarize the sample for comparison purposes (e.g. the *t-statistic* or *t-value*). The **critical region** of a hypothesis test is the set of all statistic outcomes which cause the *null hypothesis* to be rejected in favor of the *alternative hypothesis*.

Critical value
Corresponds to a given significance level. Determines the boundary between those samples resulting in a test statistic that leads to *rejecting the null hypothesis* and those that lead to *a decision not to reject the null hypothesis*. If the calculated value from the statistical test is greater than the critical value, then the null hypothesis is rejected in favour of the alternative hypothesis, and vice versa.

Definition of $\mathbf{t_k(\alpha)}$:

$$P(\,|T|\, \geq t_k(\alpha)) = \alpha$$

p-value
The probability, assuming the null hypothesis is true, of observing a result at least as extreme as the test statistic.

The null hypothesis rejection using the p-value
If the p-value is less than the required significance level, then we say the null hypothesis is rejected at the given level of significance.

What is the t-test useful for?

1) One Sample t-test
To test if the mean of a (normally distributed) population is equal to a given value.

2) Paired t-test
To test if the difference of the means of two populations is equal to a given value, assuming that the given sample contains paired individuals.

Remarks on t-test and p-value

Meaning of p-value
– p-value is the least significance level (alpha) for which we would reject the null hypothesis.
– Or, p-value is the probability of observing a sample that is
"at least as extreme as the observed one" *if the null hypothesis were true.*

Explanation: The principle of t-test is based on the fact that when you randomly choose samples of size n from a normally distributed population, then

$$T = \frac{\bar{X} - \mu}{S} \sqrt{n} \sim t_{n-1}$$

where $\bar{X}$ is the sample mean, S is the sample standard deviation, and μ is the population mean.

Exercise 1
The sample generated from a normal population is 8, 8, 9, 10, 12, 16.
The null hypothesis is H_0: $\mu = 9$.
Calculate the t-statistic and the p-value.
What do the values imply? Will the null hypothesis be rejected?

Exercise 2
The results of 6-fold cross-validation experiment – accuracy of two different classifiers:

sample	1	2	3	4	5	6
classifier **A**	81.80%	81.00%	82.20%	80.90%	81.50%	81.60%
classifier **B**	81.50%	81.10%	82.00%	81.10%	81.40%	81.40%

Which of the two classifiers is better?

Exercise 3 – t-test

In city A, men's height was measured on a sample of 15 people:
171.2 163.5 172.9 208.4 178.4 164.3 180.5 173.6 185.7 183.4 177.8 158.1 182.0 181.1 182.8

In city B, men's height was measured on a sample of 7 people:
180.6 189.0 166.9 174.1 167.2 170.3 181.8

Is the expected value of men's height in the two cities different?

Test it for confidence levels 90%, 95%, and 99%.

Compute t-values, critical values. p-values, and confidence intervals.
Do it by definition.
Then evaluate the tests and make conclusions.

Exercise 4 – paired t-test

When doing 10-fold cross validation, the following accuracies were found
(the same data were used for both classifiers A and B):

* classifier A:
  0.853 0.859 0.863 0.871 0.832 0.848 0.863 0.860 0.850 0.849

* classifier B:
  0.851 0.848 0.862 0.871 0.835 0.836 0.860 0.859 0.841 0.843

a) Compare confidence intervals of classifiers accuracy
   for confidence levels 90%, 95%, and 99%.

b) Then do paired t-test, and test if the expected accuracies are equal
   for confidence levels 90%, 95%, and 99%.

Demo code – heights of men, with mean 180.3 cm and variance 100

```
experiments = 10^5
M = 180.3
population = rnorm(10^6, M, sqrt(100))
count = 0
for(i in 1:experiments) {
  s = population[sample(1:10^6, 10)]
  test = t.test(s, conf.level=0.99)
  if( test$conf.int[1] < M  &  M < test$conf.int[2] ) {count=count+1}
}
return( count/experiments )
```

```
# another way to do the same
```

```
experiments = 10^5
M = 180.3
count = 0
for(i in 1:experiments) {
  s = rnorm(10, 180.3, sqrt(100))
  test = t.test(s, conf.level=0.99)
  if( test$conf.int[1] < M  &  M < test$conf.int[2] ) {count=count+1}
}
return( count/experiments )
```

EXAMPLE 1
One Sample t-test

```
> x <- c(8, 8, 9, 10, 12, 16)
> t.test(x, mu = 9, conf.level = 0.95 )

      One Sample t-test

data:  x
t = 1.1921, df = 5, p-value = 0.2867
alternative hypothesis: true mean is not equal to 9
95 percent confidence interval:
  7.265422 13.734578
sample estimates:
mean of x
     10.5
```

The result and its interpretation:

- p-value > alpha (= 0.05)
- therefore we cannot reject the null hypothesis (that mu = 9), because the risk of doing wrong would be more than 5%
- the reason is that the data **does *not*** show that the mean is different from 9 (at the given significance level)

EXAMPLE 2
One Sample t-test – the same data at a smaller confidence level

```
> t.test(x, mu = 9, conf.level = 0.70 )

      One Sample t-test

data:  x
t = 1.1921, df = 5, p-value = 0.2867
alternative hypothesis: true mean is not equal to 9
70 percent confidence interval:
  9.045691 11.954309
sample estimates:
mean of x
     10.5
```

The result and its interpretation:

- p-value < alpha (= 0.30)
- therefore we reject the null hypothesis (that $\mu = 9$), and we do not do wrong with confidence level (only) 70%
- the reason is that the data **does** show that the mean is different from 9 (at the given significance level)

Exercise 1 – Solution

To compute the p-value given the data from the example we can calculate:

```
> x <- c(8, 8, 9, 10, 12, 16)
> T <- (mean(x) - 9)*sqrt(length(x))/sd(x)
> pvalue <- pt( -abs(T), df=length(x)-1 ) * 2
```

Exercise 2 – Solution

sample mean = 0.0833, variance = 0.0377

$T = 1.0518 < t_5(0.05) = 2.571$

The null hypothesis H_0: $\mu = 0$ cannot be rejected.