Differences between Two Groups

Marco Baroni

May 19, 2005

1 Typical scenarios

- Do Brits use longer words than Americans?
- Are translated texts more collocational than original texts (according to a certain "collocativity score" we can compute for each document)?
- How likely is it that the mean difference in performance of tagger A and B in the 10 folds of our experiment was due to chance?
- Is the unstressed vowel derived from mid-high /o/ higher than the unstressed vowel derived from mid-low /O/?
- In general, any scenario in which we have samples of (independent) values coming from two groups, and we wonder whether the differences we observe between the groups (e.g., in their means or medians) are significant or due to chance.

2 The classic *t-test*

- The statistical reasoning and (some of the) mathematics involved is not too complicated, but they would require a few hours by themselves, so I assume you read the books and I will just discuss intuitions behind the test, the necessary assumptions and how to do it in R.

2.1 The statistical setting

- You are interested in knowing whether two (hopefully random) samples $x_1, x_2, ..., x_n$ and $y_1, y_2, ...y_m$ come from underlying *populations* having the same or different distributions.
- E.g., the samples could be samples of word lengths from the Brown and LOB, and the population distributions could be the "true" distributions of word lengths in American and British English (whatever that means).

- A population distribution is characterized by its class (normal, Poisson, etc.) and a certain number of *parameters* – e.g., a *mean* μ and a *variance* σ.
- However, we almost never know the mean and the other parameters of the "true" populations – what we have are statistics we can compute from our random samples – e.g., the standardized difference of the means of the two samples.
- In turn, the difference of sample means will have its own distribution, whose shape will depend on that of the underlying population.
- Then, we can make an hypothesis about the underlying population, and, given what we know about the relation between the underlying population distribution and the distribution of the statistic we can estimate from our data, we can tell how likely our sample statistic would be if our hypothesis about the underlying population was right.
- In the case in which we compare two groups, typically, the interesting hypothesis is that the populations are different, but the tested *null hypothesis* is that $\mu_x - \mu_y = 0$ (and thus we hope to find out that the sample statistic value we got is very unlikely under the null hypothesis).
- In practice, the hypothesis testing steps (not only with t-test) are as follow:
  - From our sample data, we compute a *test statistic* (a score), by plugging various values we can extract from our data into a fixed formula.
  - This test statistic has the characteristic that, under various assumptions, we know how likely the various values it takes are *if* the null hypothesis is true, i.e., we can assign a *p-value* to the score obtained in the previous step.
  - The lower the *p-value*, the less likely it is that the data did indeed come from the population hypothesized under the null hypothesis.
  - In our case, the lower the *p-value*, the less likely it is that the two population means are identical (i.e., that the μ of the population of mean differences is 0), and thus the more likely it is that the two samples come from different populations.
  - It is up to the experimenter to decide how happy she/he is about the *p-value* obtained – in statistical parlance, whether to accept or reject the null hypothesis.
  - Popular rejection thresholds are 0.05 and 0.01.[1]
- Important points:

[1] One should decide the rejection threshold *before* running the test. Also, notice that the *p-values* to the gazillionth decimal point reported by statistical packages are not so meaningful, since typically our measurements and the assumptions we made do not justify reporting a *p-value* at such level of granularity. It makes more sense to report $p < 0.01$ than $p = 0.0078342$.

- Were the assumptions necessary to interpret the test statistic truly met? (Randomness and normality are serious issues in many corpus linguistics settings.)
- Is the null hypothesis reasonable enough that rejecting it is an interesting thing to do?
- What are the underlying "populations" we want to extend our results to? (A HUGE issue in corpus linguistics!)

2.2 The *t* statistic for differences between means

- The classic t statistic:

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{N_x} + \frac{s_y^2}{N_y}}} \tag{1}$$

  where $\bar{x}$ and $\bar{y}$ are the sample means, s_x and s_y are the sample standard deviations and N_x and N_y are the sample sizes.

- If the assumption of normality of the distribution of the difference of sample means holds (see 2.3 below), it can be shown through the magic of mathematical statistics that, if the null hypothesis is true (there is no difference between the population means), then t has a t-distribution with $\mu = 0$ and shape determined by the parameter ν (the *degrees of freedom*).

- The more the t we found is far from 0, the less likely it is that the true difference between populations is 0.

- In the case in which we can assume that the two underlying populations have the same variance, the parameter ν equals $N_x + N_y - 2$; in the case in which we do not assume that the two underlying populations have the same variance, the parameter ν must be derived from a more complicated formula, but it still depends on the sample sizes (and on the variances).

- The t-distributions corresponding to lower ν's are flatter, more spread-out cousins of the standard normal distribution (i.e., extreme values of t will have larger *p-values* than equivalent values of the standard normal distributions, so that stronger evidence will be needed to reject the null hypothesis).

- Indeed, you probably noticed that the formula to compute t is similar to the one for calculating a z-score for the difference of sample means under the hypothesis that the population difference is 0. We do not refer to a standard normal distribution table because we do not know the true standard deviation of the differences between the means and we are estimating it from the data. The flatness of the t-distribution is our "punishment" for not using the true value of the standard deviation.

- When the sample sizes – on which ν depends – are low, our sample-based estimates of population parameters are less reliable, thus we are more uncertain about the inferences, and thus we must refer to a flatter t-distribution.
- For larger ν's, the t-distribution becomes identical to the standard normal.
- How do $\bar{x} - \bar{y}$, the s's and the N's affect t?
- Notice that, because of the N's in the denominators of the denominator, for very large samples even uninterestingly small differences between the means will produce high t values, and thus low p's (a very concrete problem with corpus data!)

2.3 Assumptions of the *t-test*

- The two samples must be random and independent (but see below on *paired* t-test).
- Data for which mean computation is reasonable operation.
- The default R implementation does not require that the underlying populations have equal variance; older implementations do.
- The variable $\bar{X} - \bar{Y}$ (distribution of the sample mean differences) must be normally distributed, which will happen either when X and Y come from normally distributed populations or when the sample sizes are large.

2.3.1 The central limit theorem

- This second way of meeting the normality condition is justified by the mind-blowing *central limit theorem*, which, simplifying a lot, says that, as N increases, the distribution of the means of samples of size N becomes approximately normal (with same mean as underlying population, and variance σ^2/N.).
- Notice difference between distribution of underlying population and distribution of means of samples from the population: for large samples, the latter becomes normal, no matter what the shape of the former is.
- It is instructive to see the central limit theorem "happening" in R.

```
> hist(runif(10000))
> qqnorm(runif(10000))
> u <- replicate(10000,mean(runif(5)))
> hist(u)
> u <- replicate(10000,mean(runif(10)))
> hist(u)
> u <- replicate(10000,mean(runif(15)))
```

```
> hist(u)
> u <- replicate(10000,mean(runif(20)))
> hist(u)
> u <- replicate(10000,mean(runif(25)))
> hist(u)
> qqnorm(u)
...
```

- The more the underlying distribution is skewed, the higher N must be for the central limit theorem to be applicable (does any very skewed distribution come to mind?)
- The central limit theorem happening more slowly with a skewed underlying distribution:

```
> hist(rpois(10000,1))
> qqnorm(rpois(10000,1))
> u <- replicate(10000,mean(rpois(5,1)))
> hist(u)
> qqnorm(u)
> u <- replicate(10000,mean(rpois(10,1)))
> hist(u)
> u <- replicate(10000,mean(rpois(15,1)))
> hist(u)
> u <- replicate(10000,mean(rpois(20,1)))
> hist(u)
> u <- replicate(10000,mean(rpois(100,1)))
> hist(u)
> u <- replicate(10000,mean(rpois(500,1)))
> hist(u)
> qqnorm(u)
```

2.4 The *t-test* in R

```
> brown <- read.table("brown.stats",header=TRUE)
> lob <- read.table("lob.stats",header=TRUE)

> t.test(brown$se,lob$se)

        Welch Two Sample t-test

data:  brown$se and lob$se
t = -0.1416, df = 987.538, p-value = 0.8874
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -4.93277  4.26877
```

```
sample estimates:
mean of x mean of y
   99.152    99.484

# Notice:
# Confidence intervals
# Non-integer degrees of freedom: R's default t-test does not
# assume equality of variances, and formula for nu in such
# case can lead to non-integer values

# Following is one-tailed test where alternative hypothesis
# is that Brown's mean is smaller than LOB's mean
# FOR DIDACTIC PURPOSES ONLY, NOT APPROPRIATE FOR
# SERIOUS BROWN/LOB COMPARISONS!

> t.test(brown$se,lob$se,alternative="l")

        Welch Two Sample t-test

data:  brown$se and lob$se
t = -0.1416, df = 987.538, p-value = 0.4437
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
     -Inf 3.527977
sample estimates:
mean of x mean of y
   99.152    99.484

> t.test(brown$towl,lob$towl)

        Welch Two Sample t-test

data:  brown$towl and lob$towl
t = 3.4416, df = 981.116, p-value = 0.0006026
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.03488570 0.12744478
sample estimates:
mean of x mean of y
 4.271106  4.189941
```

2.5 Other themes in t-testing

- Special form of t-test formula for *paired samples* (pairing will reduce variance)[2] – Sara P's tagging performance comparison data.
- Tests of the assumptions of normality and (if necessary) equal variance.

3 A non-parametric alternative: The Mann-Whitney test

- No assumptions about population distributions and parameters.
- Works with data at least on ordinal scale.
- As with Spearman's correlation, we transform the values of the two variables into ranks.
- Transforming scores into ranks reduces the space of possible outcomes we have to take into account, and thus makes exact tests viable.
- In the case of Mann-Whitney, procedure is as follows:[3]
  - Put x and y values in same table, order them, and assign rank.
  - Separate x and y values again.
  - Sum all ranks of x (but you could pick y and final result would be the same).
  - Subtract this sum from the maximum possible value that it could take in theory (e.g., if both X and Y are samples of two values, the maximum theoretical value for the sum of x's ranks is $4 + 3 = 7$; if the actual sum is $2 + 3 = 5$, the test statistic is $7 - 5 = 2$).
  - Count how many possible outcomes would give a result for this statistic (known as U) that is as extreme or more extreme than the one obtained empirically, with respect to the value of U that would be obtained under the null hypothesis that ranks are equally spread between the two samples (the null hypothesis' U can be calculated with the formula $(N_x N_y)/2$).
  - It can also be shown that, as the sample sizes increase, the sampling distribution of the sum of ranks approximates a normal distribution, and one can avoid the computational cost associated with the exact testing approach by computing the appropriate z-scores.

[2] This has to do, in magic ways, with the fact that correlated variables have non-0 covariance, and a term derived from covariance is subtracted from the sum of single population variances when computing the variance and standard deviation of the differences between sample means.

[3] Different procedures, leading to essentially equivalent statistics, are also found in the literature. The one I am reporting here is from `http://faculty.vassar.edu/lowry/webtext.html`. R uses a different method to compute U.

– Like with t-test, a paired sample version is available.
– Notice that Mann-Whitney test (or t-test, for that!) does not tell us if the two populations have similar of different *shapes* – (to compare overall shapes of distribution, consider test based on empirical distribution function of the Kolmogorov-Smirnov type).

3.1 The Mann-Whitney test in R

```
# again, default is two-tailed alternative -- this time,
# I'm not even going to try one-tailed test for illustrative
# purposes, as medians/avg ranks are identical

> wilcox.test(brown$se,lob$se)

        Wilcoxon rank sum test with continuity correction

data:  brown$se and lob$se
W = 128498, p-value = 0.4437
alternative hypothesis: true mu is not equal to 0

> wilcox.test(brown$se,lob$se,exact=TRUE)

        Wilcoxon rank sum test with continuity correction

data:  brown$se and lob$se
W = 128498, p-value = 0.4437
alternative hypothesis: true mu is not equal to 0

Warning message:
Cannot compute exact p-value with ties in:
wilcox.test.default(brown$se, lob$se, exact = TRUE)

> wilcox.test(brown$towl,lob$towl)

        Wilcoxon rank sum test with continuity correction

data:  brown$towl and lob$towl
W = 138845, p-value = 0.002432
alternative hypothesis: true mu is not equal to 0

> wilcox.test(brown$towl,lob$towl,exact=TRUE)

# it gets stuck, although I'm not sure of whether
# this is because of computational load of exact
# test, or because Eros is wasting gollum's precious
```

```
# RAM with a script called kalimba_de_luna.pl
```

3.2 More than two samples

- If you need to test for differences among n groups (where $n > 2$), do not run two-sample tests for all possible variable pairs.
- By running many tests of the same null hypothesis (that the n groups do not differ), we increase chances of rejecting it although it was correct.
- In such cases, use One Way ANOVA instead of t-test and Kruskal-Wallis test instead of Mann-Whitney.