Statistical Significance and Analysis of Data

Daniel Swenson

University of California, Merced

April 19, 2017

Statistical Significance

What is **statistical significance**, and why is it important?

Let's examine this with a specific example. Suppose we have two different groups of students — say, representing different majors. In a particular course, a sample of five students from one group has the scores

$$45, 89, 85, 68, 84$$

and a sample of five students from the other group has the scores

$$89, 79, 82, 76, 88$$

Statistical Significance

The mean score of the first group is 74.2 and the mean score of the second group is 82.8. This is a mean difference of 8.2, which is a fairly practically significant difference (assuming that the scores are percentages).

Statistical Significance

But, could this mean difference be "due to chance"? It would be embarrassing to make the conclusion that students with these different majors perform differently on the exam, if this mean difference is, in fact, just "due to chance" (i.e., just a statistical artifact).

Statistical Significance

Statistical significance is the concept that pertains to addressing the question of whether or not the observed results could be "due to change." In this case, the **statistical test** that should be applied is called the *t*-test. It can be applied in Excel or other statistical software.

Statistical Significance

The t-test returns a value called a p-value. Basically, the p-value is the probability of seeing a result as extreme as the one you actually did see, under the assumption that the results obtained actually are due to chance alone.

In this case, the p-value of the t-test turns out to be 0.361. It is customary to claim a statistically significant difference only if the p-value is less than 0.05; so, this result would not be considered statistically significant.

Statistical Significance

In this example, the **significance level** of the test would be 0.05. A test where we consider the result to be significant at the 0.05 level has a **false positive** rate of 0.05 (5%). That is, 5% of such tests will be statistically significant, even if there is no true underlying relationship in the population being studied.

Statistical Significance

You might wonder — what would we do if we really did suspect that students with these majors generally tend to score differently on this exam?

It turns out that the p-value of the t-test is affected by three things:

- The difference in mean scores between the two groups;
- The variation within each of the two groups; and
- The sample size of each of the two groups.

The last one is what we have control over; so, in practice, we would sample larger groups of students and run another t-test.

Statistical Significance

If we had three or more groups represented in our data (e.g., three or more majors), we could use a generalization of the t-test called **ANOVA** (for **Analysis of Variance**).

Similar to the t-test, the p-value of the ANOVA is affected by three things:

- The variation between the groups;
- The variation within each group; and
- The sample size of each group.

Statistical Significance

Sometimes, it's more convenient to use special statistical software to do statistical tests. For example, not all spreadsheet programs can easily compute a p-value for a correlation. However, special statistical software can do this easily.

One of these programs is called `R`. `R` is used widely and can be downloaded for free. In the following examples, we will make use of `R` (sometimes also making use of the `xlsx` package to import Excel spreadsheets into `R`).

Linear Regression

Linear regression is a type of analysis that is done when trying to "describe" an output variable (e.g., an exam score) by some input variables (e.g., how much the student studied for the exam, how much they went to office hours, etc.).

Linear Regression

Basically, linear regression assumes that each equal-sized change in an input variable (e.g., hours studied) has a corresponding constant-sized change in the output variable (e.g., the exam grade). The size of the change is called a *coefficient* in the linear regression model. Each coefficient has its own p-value.

Example - Math 5 Midterm 1 Grades

Example. Math 5 Data From Fall 2016. Let's do linear regression on the Midterm 1 data.

To load the data, we use `data = read.xlsx("/home/daniel/Desktop/Math 5 Midterm 1 Data.xlsx",1)`. (The path will depend on where you've put the file.)

Example - Math 5 Midterm 1 Grades

After using `attach(data)`, we use `summary(lm(Exam.One~STEM+PALS+Clinic.Hours+Hours.of.Sleep.a.Night+Study))` to get the following regression table.

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)              53.8203    11.7160   4.594 4.68e-05 ***
STEM                     -0.2313     9.1270  -0.025  0.97991
PALS                    -15.3126     8.4063  -1.822  0.07640 .
Clinic.Hours            -35.1360    12.8002  -2.745  0.00919 **
Hours.of.Sleep.a.Night    2.6146     1.5867   1.648  0.10762
Study                    -0.4287     1.1067  -0.387  0.70062
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Example - Math 5 Midterm 1 Grades

Since only `Clinic.Hours` (and the intercept term) is statistically significant, we look at the coefficient: -35. However, if we enter `Clinic.Hours`, we see that only two students in this sample (of $n = 44$ students) actually used the clinic hours; so, the statistical significance is spurious ("fake") in this case.

Example - Math 5 Midterm 1 Grades

In conclusion, this data doesn't present any evidence that any of the factors considered here are associated with the students' exam one score. (To make sure that the lack of statistical significance isn't just due to the sample size, we could consider a similar set of data with a larger sample size.)

Example - Math 11 Exam Grades and Hours of Sleep

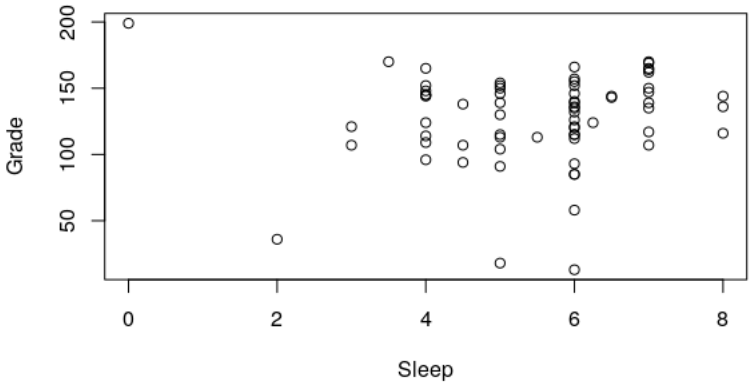
Example. Math 11 Exam Grades and Hours of Sleep. Let's now look at the summer 2016 data regarding Math 11 exam grades and hours of sleep, which can be loaded in a similar way to the previous dataset.

Example - Math 11 Exam Grades and Hours of Sleep

The command `summary(lm(Grade~Sleep))` gives a p-value of 0.608, which is not statistically significant. So, no statistically significant association is observed in this sample between the student's exam grade and the number of hours of sleep they get each night.

Example - Math 11 Exam Grades and Hours of Sleep

This is also evident in the plot formed by `plot(Sleep,Grade)`.



Correlation

A **correlation** is a quantity that can be computed to see if these is an association between two variables. It has its own p-value. The correlation of two quantities is always a number between -1 and 1.

Example - Math 32 Grade Forecasting Data

Example. Math 32 Grade Forecasting Data. Suppose that we just want to examine two aspects of the Math 32 grade forecasting data (in the relevant `RData` file) — whether the students' forecasted grade was "higher than expected" and whether they intend to "study more."

The `RData` file can be loaded into your `R` environment with the command `load("~/Desktop/Student Responses Regarding Math 32 Grade Predictor.RData")`. (The path will depend on where you've placed the `RData` file on your hard drive.)

Example - Math 32 Grade Forecasting Data

The command `cor.test(StudentResponses$HigherThanExpected, StudentResponses$StudyingMore)` gives the following output.

```
        Pearson's product-moment correlation

data:  StudentResponses$HigherThanExpected and StudentResponses$StudyingMore
t = -0.017848, df = 43, p-value = 0.9858
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.2960192  0.2910446
sample estimates:
        cor
-0.00272184
```

Example - Math 32 Grade Forecasting Data

The correlation is nearly zero, and the p-value is 0.9858. It seems likely that these two quantities are not related to each other.

Miscellaneous Considerations

It is impossible to mention all of the possible considerations that could come into play when analyzing data in the scope of just one workshop. Here are some considerations that are generally good to keep in mind:

- ANOVA can be done in a **hierarchical** way; that is, it can assess whether one quantity "describes" another well, *controlling for another quantity*. (For example, we might want to know if a student's study time is associated with a higher exam grade, *controlling for* the student's percentage grade in their previous mathematics course.)
- The statistical methods mentioned in this workshop are what are technically called **parametric methods**. This means that they make certain assumptions about the data. If these assumptions are not true, the resulting p-values may not be reliable. There are methods called **non-parametric methods** that do not make these assumptions; one of their shortcomings is that their analysis has a higher **false negative** rate in contexts where parametric methods *could have been used*.

Miscellaneous Considerations

- Using continuous data (e.g., percentage grades) is better than discrete data (e.g., letter grades) if possible, because discretizing data leads to an increase in the false negative rate of statistical tests.
- Without randomization, association (e.g., correlation) doesn't imply causation. In academic settings, randomization generally isn't practical. So, our conclusions about causation based on survey data will typically be of a somewhat speculative nature.

Miscellaneous Considerations

- If several inputs are fed into a linear regression model, then some of the p-values will probably be statistically significant "by chance". This is the **multiple testing problem** — when many statistical tests are run at once, some of them may be significant "by chance." When comparing regression models, it is possible to use quantities that take this into account, such as **AIC** (for **Akaike information criterion**). AIC basically includes the number of input variables as a measure of the complexity of the model and tries to balance **goodness-of-fit** of the model against the **complexity** of the model.

Miscellaneous Considerations

- The goodness-of-fit of a linear regression model is measured using the **R-squared** value. A high R-squared value (close to 1, i.e., 100%) implies a good fit; a low R-squared value (close to 0) implies a bad fit. However, the R-squared value can be made artificially high simply by including many input variables in the model. This phenomenon is called **overfitting**. It can be addressed by **pruning the model** by removing the input variables with the high p-values and examining the R-squared value of the reduced model. Alternatively, the model with highest AIC (see the previous bullet point) can be selected as the "best" model.

Miscellaneous Considerations

- To access an R help file for a specific command, we use the ? command. (For example, to access the help file for the `plot` command, enter `? plot` at the R prompt.) However, the help files can be somewhat cryptic, so Google can be helpful as well. (For example, you could Google "How do I make good-looking plots in R?")
- Sometimes when using R, you'll realize that you need a specific package in order to implement a particular functionality. (Often you'll end up realizing this as a result of a Google search, such as "How do I import Excel spreadsheets into R?") Say we want to use the `xlsx` package — we use the command `install.packages("xlsx")` followed by `library(xlsx)`. We are now ready to use the functions from the `xlsx` package.
- If you want to use RStudio, you'll need to install R first, and *then* install RStudio.
- All of the commands that work in R will also work in RStudio.

Recommended Reading

- *Elementary Statistics* by Mario F. Triola
- *Introductory Statistics with R* by Peter Dalgaard
- The Wikipedia pages on
  - Hypothesis testing
  - Statistical significance
  - Correlation
  - Linear regression
  - t-test
  - p-value
  - ANOVA
  - Parametric statistics
  - Non-parametric statistics
  - Multiple testing problem
  - Akaike information criterion