LECTURE 2 – **Significance Tests**

In order to evaluate the statistical significance of an experiment, we need a p-value. However, there are many different problems that require different tests, each yielding a p-value.

The sign test

This is the oldest significance test invented. It has few assumptions and it is based on counting. It was first used by John Arbuthnot in 1710 in order to know whether male and female births are equally likely or not (he was wondering whether God accounted for the fact that many men died in war by producing more male births).

Null hypothesis: Male and female births are equally likely.

John checked around 0.5 Million births over 82 years in London. He found that in every single year male births were more common than female births. Assuming a probability of 0.5 (as per null hypothesis) the probability of observing this outcome due to chance is 0.582, which is 2.1e-25. As this value is less than 0.05 (the significance level), it means that it is extremely unlikely for it to be due to chance therefore there must be a cause. This allowed John to reject the Null Hypothesis. Then, he wrongly concluded that God did this. Hypothesis testing allows you to reject the hypothesis, not to assign its cause to something else.

This is called sign test because for every year John deducted male births to female births and got a positive/negative sign and then just counted the positive signs.

The Y-chromosome is smaller so it weighs less so it swims faster and the probability of impregnating is higher (52%-48%)

The z-test

This test is performed when the parameters of the population are known (and It relies on the central limit theorem – it does not work for small samples).

Let’s assess if the drug NZT works (a drug that supposedly increase IQ)

Null Hypothesis: NZT does not improve IQ

Here we know that IQ distributed normally with a mean of 100 and standard deviation of 15. So we give the drug to 25 people and calculate the mean of those people’s IQ, being 109 (by getting the sample mean we are betting on the Law of Large Numbers because we hope that it represents the sample’s population). Now we can calculate z.

Text

Description automatically generated

Check the z table and get a p-value: 0.001

Therefore the fact that the sample of people got a higher IQ is not due to chance, we reject the null hypothesis and conclude that the drug is effective.

Table

Description automatically generated

The t-test

The t-test avoids both the need of the population parameter and it is not based on the CLT. Thus, it is the most commonly used test.

It does however rely on several assumptions: (1) The sample parameters (like the mean) can be interpreted meaningfully, (2) Data is distributed normally, (3) Homogeneity of variance (the variability within each sample is similar). These assumptions are normally violated but t-test is very robust.

Degrees of freedom are the number of independent pieces of information in a dataset (the more degrees of freedom, the more stable the parameter estimate is).

A picture containing schematic

Description automatically generatedIn t-test we use the available data to get the sample mean and sample standard deviation:

A picture containing text, clock

Description automatically generated When calculating the sample mean, we use all the data and therefore it is a linear combination of the data and we lose one independent variable (from n to n-1). That is why when calculating the sample std dev we use n-1 (the denominator is not sample size, it’s really independent variables).

For example, if we calculate 9 sampe means for a sample size of 90 in an ANOVA, we end up with 81 (90-9) degrees of freedom left.

With many degrees of freedom (around 30), the t-distribution approaches the normal distribution (“the Central Limit Theorem kicks in at 30”). At low degrees of freedom, the t-distribution has fat tails therefore an extreme value is more likely.

Nowadays, our datasets are very large therefore there are many degrees of freedom and no need for the population parameters.

There are several versions of the t-test:

* 2 independent groups, usually comparing a treatment and a control group (or two treatment groups). df = n1+n2-2
* 1 independent group, usually the same group is tested before and after a treatment. df = n-1

For example, let’s say you are a data scientist that wants to check if a drug boosts mental speed but it is very expensive to make so you only make a batch of 3.

Table

Description automatically generatedIf we do the t-test for independent groups we get this:

Icon

Description automatically generatedA picture containing text, clipart

Description automatically generatedA picture containing text

Description automatically generatedA picture containing text, clock

Description automatically generatedText

Description automatically generated

A picture containing icon

Description automatically generatedA picture containing text, scissors, brass knucks, tool

Description automatically generatedText

Description automatically generatedText

Description automatically generatedTable

Description automatically generatedIf we do the t-test for paired groups we get this:

A picture containing text, watch, gauge

Description automatically generated

The Welch’s t-test

Text

Description automatically generated with medium confidenceIn most applications variances are not homogeneous so this test is used (this is the test we should be using most of the time). It is exactly the same as the t-test but without that assumption.

ANOVA test

ANOVA is used so compare more than two groups

*Is reducing a dataset to its sample means always appropriate?*

No. For example, when categorical data uses numbers as categories, taking the average of those numbers is meaningless. Cat: 1, dog: 2, horse: 3. Python: “the mean between a cat and a horse is a dog”

We manage this data counting its frequencies and using a Chi squared test.

The Chi-Squared test

This is a significance test for categorical data and it is non-parametric. In general non-parametric tests have fewer assumptions but they are less powerful (more data is needed to achieve equivalently significant results).

A picture containing text

Description automatically generatedThe way Chi-Squared works is by comparing expected category counts with observed category counts, square the difference and sum them up.

For example, is Zodiac sign predictive of being a serial killer?

Chart, histogram

Description automatically generatedIcon

Description automatically generatedTable

Description automatically generateddf = # of categories – marginal = 12-1 = 11

We conclude that p-value is 0.34 (the area from 12,27 onwards covers 34% of the distribution), which is more than 5% therefore it is not statistically significant.

*Other occasions where reducing to sample means is not appropiate*

When it comes to user ratings for example. If two movies are rated 3 and 4 by one user and 2 an 5 by another user, is it fair to say that both the user’s mean rating is 3.5? No, the psychological distance between 2 and 3 is not the same as 3 and 4 or 4 and 5.

Moreover, if the population distribution contains outliers the mean is also not a good representation of the data:

Chart, histogram

Description automatically generated

The Mann-Whitney U test

It is an analogy to the t-test (which tests whether two samples have the same mean) because the Mann-Whitney test checks whether two samples come from populations with the same median. Because the median is robust to outliers!

This is how it works: Having ratings data from two samples, we arrange them all in rank order from the smallest to the largest value, regardless of which sample they came from. If both samples come from the same underlyng distribution/population, the sum of the ranks should be similar for both samlples (assuming equal n).

Chart, histogram

Description automatically generatedBUT, even if two samples have the same median, they can have a different distribution:

If you care about the distribution itself we should use a different test: Kolmogorov-Smirnov test.

Chart

Description automatically generatedThe Kolmogorov-Smirnov test – KS test

No comparisons of means or medians, it tests whether the underlying distributions are the same by comparing the cumulative distribution function of samples.

This is how it works: Plots the cdf of the two empirical samples, finds the point of largest separation and assigns it to the test statistic.

Diagram

Description automatically generatedSUMMARY

When to use each test

z-test: If you know the population parameters.

t-test: If you don’t know the population parameters and/or your sample size is small.

F-test: If you want to compare more than two groups.

Χ2-test: If you have categorical data.

U-test: If you are interested in medians.

Etc.

Graphical user interface, application, Word

Description automatically generated