Crash Course in Statistics for Neuroscience Center Zurich University of Zurich

Dr. C.J. Luchsinger

7 Test theory (incl 1-way-ANOVA)

Further readings: Chapter 8, 10, 11 & 12 in Stahel or Chapter 8 in Cartoon Guide

7.1 Where do I find which test & some remarks

7.5.1, page 67: $p = 0.5$ and related in Binomial case
7.5.2, page 69: "z"-Test
7.5.3, page 70: 1-sample-t-Test
7.5.4, page 71: 2-sample-t-Test
7.6, page 73: χ^2-Test for independence in contingency tables
7.7.2, page 75: 1-Way-ANOVA

Some remarks about the above choice: There are very many tests in various situations and a Crash Course in Statistics can not cover all of them - especially not in depth. I suggest further readings especially in Stahel where many more tests are presented and the typical question "when should I use which test" is treated on several occasions more deeply. I further strongly suggest that you use the consulting services offered by University of Zurich and ETH (free of charge) with large expertise. Having followed this course you can follow their suggestions much better.

7.2 Two introductory examples

7.2.1 Binomial: the p

Researcher competing with you says (s)he is able to grow nerves with a method (s)he developed such that it works in 60 % of all cases (600 out of 1'000 mice). Up till now it was only 50 % of all cases. You can not believe this result and repeat the experiment with 1'000 mice. When will you say (s)he was cheating? Your result then is 566.

7.2.2 1-Way-ANOVA

We test 4 methods in growing nerves. Results are as below. Is there any significant difference between the 4 methods (or is $\mu_1 = \mu_2 = \mu_3 = \mu_4$)?

7.3 Notation

5 Minutes "Bradford-Hill" and "Sir Karl Raimund Popper"; see links on website for more.

What we saw in above two examples is that there are 2 *possible* hypothesis. We are going to call them Null Hypothesis $\mathcal{H}_0$ and Alternative Hypothesis $\mathcal{H}_1$. *Either* $\mathcal{H}_0$ *or* $\mathcal{H}_1$ is true, but we don't know which is true. Then we get data (566 out of 1'000 mice). Depending on which hypothesis is true, the random variable that generated this data will have a different distribution (for example a $\text{Bin}(1'000, 0.5)$ vs $\text{Bin}(1'000, 0.6)$). We then have to make a decision: say whether we **provisionally** accept $\mathcal{H}_0$ or $\mathcal{H}_1$. Before we get that data, we have some time to discuss a strategy. All we have is the following situation:

We have to either say: "we **provisionally** accept $\mathcal{H}_0$" or "we **provisionally** accept $\mathcal{H}_1$".
4 things can happen:

1. $\mathcal{H}_0$ is true and we **provisionally** accept $\mathcal{H}_0$ [NO ERROR]
2. $\mathcal{H}_0$ is true and we **provisionally** accept $\mathcal{H}_1$ [TYPE 1 ERROR]
3. $\mathcal{H}_1$ is true and we **provisionally** accept $\mathcal{H}_1$ [NO ERROR]
4. $\mathcal{H}_1$ is true and we **provisionally** accept $\mathcal{H}_0$ [TYPE 2 ERROR]

In German: Fehler 1. bzw. 2. Art

Examples:

* $\mathcal{H}_0$: fair dice (1/6) vs $\mathcal{H}_1$: at least two (!) probabilities are not 1/6.
* $\mathcal{H}_0$: fair coin (1/2) vs $\mathcal{H}_1$: probability not equal to 1/2.
* $\mathcal{H}_0$: drug does not change blood pressure vs $\mathcal{H}_1$: drug does lead to lower blood pressure

What are the Hypothesis in 7.2.2 (first make general assumptions)?

Critical Region (see 7.2.1): where we reject the Null Hypothesis $\mathcal{H}_0$; **critical values** are at the edge. [In German: **Ablehnungsbereich**, wir lehnen dort die $\mathcal{H}_0$-Hypothese ab, Grenzen des Ablehnungsbereichs sind die **kritischen Werte**].

Probability for type 1 error is $\alpha \in [0, 1]$; usually 10, 5, 2.5, 1, 0.1, 0.01, 0.001 %, also called size of test [in German: Grösse des Tests, Risiko 1. Art]

Probability for type 2 error is $\beta \in [0, 1]$. It is usually not pre-specified (in German: vorgegeben) as the α above. We want it to be small (see power just below too).

Power is $(1 - \beta)$ (in German: Macht). We want the power to be large: "The hypothesis testing design problem is: Choose a test to maximize power subject to a pre-specified size." (in German: Wir suchen zu einem vorgegebenen α einen Test, welcher die Macht maximiert - und damit das Risiko 2. Art minimiert).

We *can* choose good or bad tests. To be more precise and using our newly learnt definitions: Given an α, there are different tests with different β's. We want a small β, that is a large power:

Time to solve exercise 7.1.

Test Statistic: function of the data $x_1, \ldots, x_n$ which we use to solve our test problem: for example $\overline{x}$ or $\sum_i x_i$.

P-Value: Under $\mathcal{H}_0$ (if $\mathcal{H}_0$ is true) and we have data: what is the probability of observing a test statistic **at least as extreme (or even more extreme)** as with our data: Example with $n = 1$:

Why $\alpha = 0.1$ or $\alpha = 0.05$? Why not $\alpha = 0.5$ or better: $\alpha = \beta$ (would be fair)? $\mathcal{H}_0$ is usually the common, current knowledge in research or a careful assumption:

* ”In dubio pro reo”
* ”Old drug is better” (we know side effects of the old drug)
* ”Drug does not alter blood pressure” (pharmaceutical company must prove they are better and not the FDA)
* ”There is no God” (scientific point of view - does not mean we are atheists)

We are therefore very conservative such that we do not want results which are simply a result of pure chance and not of a systematic effect to become common opinion in science!

7.4 General Recipe

We illustrate it with following situation: We know (or assume) data $x_1, x_2, \ldots, x_9$ comes from a $\mathcal{N}(0,1)$ or a $\mathcal{N}(2,1)$ RV. We don't know whether the mean is 0 or 2.

Steps in Hypothesis Testing	**Example**
Set up both hypothesis	$\mathcal{H}_0 : \mu = 0$ vs $\mathcal{H}_1 : \mu = 2$
Set α and (if possible) n	$\alpha = 0.05, n = 9$
Choose a good test statistic	$\frac{1}{9}\sum_{i=1}^{9} X_i \ (= \overline{X})$
Find distribution of test statistic under $\mathcal{H}_0$	$\mathcal{N}(0, 1/9)$
Critical value	1.64/3≐0.547
Get data	$x_1, \ldots, x_9$
Reject or accept $\mathcal{H}_0$	$\overline{x} < 0.547$: accept $\mathcal{H}_0$, otherwise reject $\mathcal{H}_0$

Alternatively to above recipe (from Step 5 onwards): Compute P-Value and compare with α.

7.5 Some classical (parametrical) tests

7.5.1: Binomial RV: is p equal to some p_0 (for example 0.5)? [exact and approximate solution]

Remember 7.2.1: We have (we take 10 times less for Maths sake) 100 Be(p), success/failure in some experiment with mice. One of your competitors claims in a paper that he has success probability $p = 0.6$ (so far only $p = 0.5$ was known). You now try to reproduce these results. Be aware that it is not so clear how to choose the hypothesis...

You will need $P[\text{Bin}(100, 0.6) < 52] \doteq 0.05$ for the exact solution (I have that number from the computer):

Approximate solution using the CLT (see page 41):

We could now change the hypothesis; you then need: $P[\mathrm{Bin}(100, 0.5) > 58] \doteq 0.05$

7.5.2: $x_1, \ldots, x_n$ from $\mathcal{N}(\mu, \sigma^2)$ with σ^2 known: is μ equal to some μ_0 (for example equal to 0)? [”z”-TEST]

Example: $x_1, \ldots, x_n$ being length of nerves that have grown under some particular circumstances. Generally known that these can be modelled with a $\mathcal{N}(10, 4)$-RV. We slightly changed setting and are now interested, whether mean is larger or not:

7.5.3: $x_1, \dots, x_n$ from $\mathcal{N}(\mu, \sigma^2)$ with σ^2 unknown: is μ equal to some μ_0 (for example equal to 0)? [1 SAMPLE T-TEST]

Motivation/Example same as in 7.5.2, but realistically you don't know the variance. You might want to (re)visit 4.2.5 and especially look at equation (6.11).

Solve Exercise 7.2 now.

7.5.4: $x_1, \ldots, x_m$ from $\mathcal{N}(\mu_1, \sigma^2)$, $y_1, \ldots, y_n$ from a $\mathcal{N}(\mu_2, \sigma^2)$, independent and variance is equal but unknown. Is $\mu_1 = \mu_2$? [2 SAMPLE T-TEST]

This expression has a t_{m+n-2}-distribution (we omit the proof). Solve exercise 7.4 now.

7.5.5: Paired t-Test

Say you have medical data from 50 patients such that for example blood pressure is taken before and after medical treatment. We could now take these $2 \times 50 = 100$ data points and treat them with the 2-Sample-T-Test (this test needs independece of **all** data!) to check whether means are equal (medical treatment has no effect). But although different patients can been seen as behaving independently of one another, there might be dependence such that a person having high level before might have high level after treatment too. Besides - more important - you are giving away important information. What could you do instead? Problems?

7.6 An example of a non-parametrical test: χ^2-Test for independence in contingency tables

We only treat the 2×2 case here [in German: Vierfeldertafel]. Generalization in Stahel, Chapter 10.

$n = 100$ people tested in Maths; 46 male, 54 female; 12 male fail, 22 female fail. Analyze this situation properly.

Mathematical model is as follows (not quite stringent but common notation): $X = 0$ for male, $X = 1$ for female, $Y = 0$ for failed, $Y = 1$ for passed. With

$$P[X = i, Y = j] =: \pi_{ij},$$

we have $\pi_{00} = 0.12, \pi_{01} = 0.34, \pi_{10} = 0.22, \pi_{11} = 0.32$. Null-Hypothesis is then

$$P[X = i, Y = j] = P[X = i]P[Y = j] =: \pi_{i.}\pi_{.j}$$

for all i, j. We have $\pi_{0.} = 0.46, \pi_{1.} = 0.54, \pi_{.0} = 0.34, \pi_{.1} = 0.66$. We must discuss this Null-Hypothesis (in practice) together.

We now define: $N_{ij} := n\pi_{ij}$ and the $N_{.i}$ accordingly. Mathematicians have shown, that the expression

$$u := \sum_{i,j} \frac{(N_{ij} - N_{i.}N_{.j}/n)^2}{N_{i.}N_{.j}/n}. \qquad (\chi^2 - \text{Test for Independence})$$

has a χ^2-Distribution with 1 df if n goes to infinity (for us: if n is large). Compute the example with Maths-Test:

Famous example with death penalty in the U.S.: 326 criminals; 166 African American, 160 white; 17 of 166 African American were sentenced to death, 19 of 160 white where sentenced to death. Gives P-Value of 76.89 %. Therefore no racial discrimination when only looking at these data. But in fact, there *was* racial discrimination. How is this possible?

Solve exercise 7.6.

7.7 ANOVA (1 way)

7.7.1 Intro

This section is a prototype of Statistics at it's best. It's the most simple case of ANOVA (=**An**alysis **of Va**riance), including the 2-sample-t-test as a special (and even simpler) case. We are first going to treat one-way ANOVA, in German: Ein-Weg-Varianzanalyse.

Suppose a company want's to develop a good fertilizer (in German: Dünger). We examine growth of crop under 4 conditions (you can also think of the growth of nerves under different circumstances): Call them I-IV. Results from experiment are as follows:

I	II	III	IV
33.3	35.5	29.6	38.5
47.8	35.4	33.4	42.4
44.4	47.6	32.8	45.5
42.9	38.8	38.8	38.9
40.9		42.8	38.9
35.5			44.5

Table 7.1: for example, 4th plant with treatment I grew 42.9 cm high. 2 plants with treatment II were destroyed through fire; one plant was accidently of a wrong type in treatment III - such is life. We will be able to treat this problem never the less:

We are interested whether there is some significant difference between the different treatments. One idea would be to compare the means of pairs of 2 with tests we have already treated in this course. That would then give us

$$\binom{4}{2} = 6$$

tests. If we had 7 treatments, we would even have

$$\binom{7}{2} = 21$$

tests. If we test them all at the 5 % level (1 out of 20), we can expect that we make a type I error (given $\mathcal{H}_0$ is true) with high probability. We should therefore be more precise in what we are testing and find some more sophisticated method (see 7.7.2). We will come back to these not only philosophical questions in 7.7.3.

7.7.2 1-Way-ANOVA

Model with k groups ($k = 4$ in the above example):

$$Y_{ij} = \mu_j + E_{ij} \quad (1 \leq j \leq k; 1 \leq i \leq n_j). \tag{7.1}$$

Let n_j be the number of plants in group j ((6,4,5,6) in our example). Total number is

$$n = \sum_{j=1}^{k} n_j \; .$$

The number 42.9 is then data that came out of RV Y_{41}. E_{ij} will be modelled as iid $\mathcal{N}(0, \sigma^2)$-RV's. **DANGER: same variance in all groups!** Therefore:

$$Y_{ij} \sim \mathcal{N}(\mu_j, \sigma^2)$$

and the n_j data in group j have been generated independently of each other. Want to know, whether the k means are equal or not. Null-Hypothesis is

$$\mu_1 = \mu_2 = \ldots = \mu_k. \tag{$\mathcal{H}_0$ – Hypothesis}$$

We are surely going to estimate the mean of each group in all k groups: for example in group 1, we will estimate μ_1 through

$$\hat{\mu}_1 := \bar{Y}_{.1} := \frac{\sum_{i=1}^{n_1} Y_{i1}}{n_1};$$

in general

$$\hat{\mu}_j := \bar{Y}_{.j} := \frac{\sum_{i=1}^{n_j} Y_{ij}}{n_j},$$

where $1 \leq j \leq k$. OK, now we have k *estimated* means! They will all be different and one of them is going to be the largest. And now?

Statistical reasoning and experience says: even under $\mathcal{H}_0$, one of the estimated means is going to be the largest and anyway, there is going to be some variation amongst those estimated means in any case. For us to reject the Null-Hypothesis, we expect the largest to be *significantly* larger than the others. But how much larger is that? Depending on how large σ is, there is going to be a large variation in the data anyway. And we don't even know this σ. We need a good test-statistic!

Well, under the Null Hypothesis we simply have one iid-sample of length n and can use all data points equally to estimate a **G**rand **M**ean GM

$$GM := \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} Y_{ij}}{n}.$$

In a second step, we can show that (home work for people who like maths)

$$\sum_{j=1}^{k} \sum_{i=1}^{n_j} (Y_{ij} - GM)^2 = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\bar{Y}_{.j} - GM)^2 + \sum_{j=1}^{k} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{.j})^2. \qquad \text{(Fundi ANOVA)}$$

This is the fundamental equation of analysis of variance. It states: the total sum of squares of deviations from the grand mean (left side) is equal to the sum of squares of deviations **between treatment** means and the grand mean (first summand) plus the sum of squares of deviations **within treatments** (second summand).

Don't panic: we are not going to solve exercises related to this expression - we will look at it with the computer.

What do you think: given we are not in $\mathcal{H}_0$, which of the two sums is going to be much larger than the other (suppose σ is small)? Look at extreme cases to solve such problems.

Mathematical Statisticians have shown, that

$$V := \frac{\sum_{j=1}^{k} n_j (\bar{Y}_{.j} - GM)^2/(k-1)}{\sum_{j=1}^{k} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_{.j})^2/(n-k)}$$

has a $F_{k-1,n-k}$-distribution under the $\mathcal{H}_0$-Hypothesis (we omit this proof). We are going to reject the Null Hypothesis if V has large values.

This last test setting was also useful to show that you don't have to know all the mathematics behind a test statistic. If the data has been analyzed properly with the right tests etc., you can interpret the statistical results with what you know from simpler cases as treated in other sections.

7.7.3 Multiple Tests - Bonferroni

We are now going to look at multiple tests again. You may find out in 7.7.2 that the different treatments lead to different results. Then you might want to test pairs of treatments.

Bonferroni: Say you want to make m different tests with the data at hand. You must then take care of the α: If you want the final α to be at most 5 %, then you must perform each of the m tests at the level of

$$\alpha/m.$$

If you then reject the $\mathcal{H}_0$ hypothesis as soon as one of those m tests has a significant result, your OK. The reason is as follows:

$$\begin{aligned} P[\text{at least one test is significant}] &= P[\cup_{l=1}^{m}\{\text{l} - \text{th test is significant}\}] \\ &\leq \sum_{l=1}^{m} P[\text{l} - \text{th test is significant}] = \sum_{l=1}^{m} \alpha/m = \alpha. \end{aligned}$$

7.8 Interpretation of Results - famous mistakes

Warning: Often heard, stays wrong never the less: "$\mathcal{H}_0$ is true with 95 % probability!"
True view of problem: *Either* $\mathcal{H}_0$ *or* $\mathcal{H}_1$ is true, 0 or 100 %. We never know (exceptions are trivial cases) which hypothesis is true. But we allow us to make a type 1 error in 5 % of the cases where $\mathcal{H}_0$ is true. In other words: given $\mathcal{H}_0$ is true, the test statistic will lie out of the critical region with 95 % probability. If not we reject $\mathcal{H}_0$ - although it is true.

See introductory example 7.2.1: **Wrong:** "Probability for 566 in a Bin(1000, 0.6) is so small that we reject $\mathcal{H}_0$." **True view of problem:** Probability for 600 (the mean!) in a Bin(1000, 0.6) is not much larger... . You must take the **whole tail of the distribution** (one or two-sided). You look at the **probability for such an extreme value or a value which is even more extreme.**

If you reject the $\mathcal{H}_0$, it is possible that $\mathcal{H}_1$ is true. But it is also possible that you reject $\mathcal{H}_0$ by chance ("α-Pech"), although it is true. You can make α smaller to make this more unlikely.

It is also possible that assumptions you made are wrong. For example data may not be normally distributed and you use a t-test with small sample (mathematical reason: therefore so called CLT not relevant).

To end this part, some correct statements:

The test statistic gave a value which is in the critical region. We therefore reject $\mathcal{H}_0$.

The test statistic gave a value which is outside of the critical region. We therefore accept $\mathcal{H}_0$.

P-Value (Probability for such an event or an even more extreme event under $\mathcal{H}_0$) is smaller than α. We therefore reject $\mathcal{H}_0$.

P-Value is larger than α. We therefore accept $\mathcal{H}_0$.

P-Value is 3.4 %. This means that if $\mathcal{H}_0$ is true, such values as we observe - or even more extreme values - only happen with probability 0.034. If we have an α larger than 0.034, we reject $\mathcal{H}_0$, if we have an α smaller than 0.034, we accept $\mathcal{H}_0$.

7.9 Exercises

7.1 Let $\mathcal{H}_0$ be $\mathcal{N}(0,1)$. We have $n = 1$ to keep the maths easy. Choose $\alpha = 0.05$ and test against

a) $\mathcal{H}_1$ being $\mathcal{N}(1,1)$ and compute the β too.

b) $\mathcal{H}_1$ being $\mathcal{N}(2,1)$ and compute the β too.

c) $\mathcal{H}_1$ being $\mathcal{N}(3,1)$ and compute the β too.

d) $\mathcal{H}_1$ being $\mathcal{N}(4,1)$ and compute the β too.

e) Summarize results from a)-d). Does it make sense?

Obviously: 1.64 and 1.96 are very important numbers for statisticians!

7.2 Medical treatment: You have 51 patients, measure blood pressure before treatment and after treatment. Data at hand is difference: $x_1, \ldots, x_{51}$. Pharmaceutical company claims, blood pressure is lower with treatment than without. σ has been estimated to be 8.4; $\overline{x} = -2.3$. Make a statistical test.

7.3 Reproduce R-Results for CI and Test in Situation for 7.5.3 by hand (use Data.txt).

7.4 Test if means are equal in Data.txt "2nd"-Dataset.

7.5 Tabletts are weighted. We got the following weight in grams:

$$1.19, 1.23, 1.18, 1.21, 1.27, 1.17, 1.15, 11.4.$$

a) Test, whether the average weight is 1.2 g (two-sided) at 5 %

b) Test, whether the average weight is less than 1.2 g (one-sided) at 5 %

Give precisely the two hypothesis.

7.6 Data from 1915: we are examining the number of people that get typhoid fever (Typhus). Among 6815 people who got a vaccine only 56 people got ill. Among 11668 people who were not vaccinated, there were 272 cases of typhoid fever. Is this a significant difference?