Lecture Notes Week 9

1 Recap: χ^2 test

When do we use the χ^2 test?

We have a set of discrete (or discretized, by binning) data and want to compare this to a proposed discrete (or discretized) probability distribution. We then set up a null Hypothesis "The data could have resulted from the proposed probability distribution".

For example we can have k possible outcomes (like the 6 results of throwing a standard die) with equal probability $1/k$ (i.e. 1/6 in this case). Or we have a Gaussian distribution, and we choose bins like

$$(-\infty, -2,), [-2, -1), [-1, 0), [0, 1), [1, 2), [2, \infty)$$

each of which has their own probability (found via the cumulative distribution function. Note that we have to cover *all possible outcomes*, and hence the probabilities must sum up to 1.

We can then compute the χ^2 metric (introduced in the last lecture) and the degrees of freedom ν of the problem, set a level of significance and see if the null Hypothesis passes the test.

Note that:

- $\chi^2 \geq 0$ always, and the test is always one-sided, i.e. if the χ^2 value is too large, the null hypothesis fails.
- The most likely outcome is generally *not* $\chi^2 = 0$, which would mean that the result corresponds *exactly* to the expectation values, but rather a finite value which depends of ν
- The χ^2 distribution describes the distribution of the sum of squares of k independent standard normal random variables, i.e. the "uncertainty of the uncertainty"

2 1-factor ANOVA

2.1 What is ANOVA?

With the z and Student t test we have compared either the mean of a sample to a number, or the mean of one sample to the mean of another sample. Now what if we have a large data set that we can split into many groups. For example, we have measured the height of all students at EPFL and group them by their sections. Our starting point would be a null hypothesis like

"the mean of the probability distributions from which the heights of the students in each section originate is the same for all sections"

Note that this is *not* the same as saying "the (measured) mean height of the students is the same in each section". Because there will always be some degree of variation! The question we are trying to ask is: Is the variation between groups larger than what you would just expect from random fluctuations, such that we should conclude that there is a correlation between the section someone is part of and their height.

In this example we could say that the "factor" is "the section you belong to" and this factor has k "levels" which would be the total number of sections. A factor can be numerical (it could for example be the year you attend, or your weight in kg, binned in steps of 1kg) but it does not have to be (as in our example above).

2.2 How to perform ANOVA

We can then perform an "analysis of variance" (ANOVA) to test the null hypothesis:

For each group/level, we find how many elements (e.g. students) it contains, we call this N_{Si} so for example N_{S1} is the number of students in the first section. We call N_{ST} the total number of elements. For that group we will have a mean and an unbiased estimator for the variance given by:

$$\overline{X}_i = \frac{1}{N_{Si}} \sum_{j=1}^{N_{Si}} \underset{\uparrow \atop \text{one element}}{X_{i,j}} \qquad S_i^2 = \frac{1}{N_{Si}-1} \sum_{j=1}^{N_{Si}} \left(X_{i,j} - \overline{X}_i\right)^2$$

And we have a total (also known as global) number of elements and mean given by:

$$\overline{X}_T = \frac{\sum_{i=1}^{k} \sum_{j=1}^{N_{Si}} X_{i,j}}{\sum_{i=1}^{k} N_{Si}} \underset{\uparrow \atop \text{iff all } N_{Si} \text{ same}}{=} \frac{1}{k} \sum_{i=1}^{k} \overline{X}_i$$

$$= \frac{\sum_{i=1}^{k} N_{Si} \overline{X}_i}{\sum_{i=1}^{k} N_{Si}} \quad \text{in general (weighted sum...)}$$

In ANOVA, it is common not to work with the variance itself, but with something called the "square sum". The total square sum SS_T is the sum of the difference of each element from the total mean, suquared. When can then divide this total into the parts that come from variations *between* the groups, SS_B and variations *inside* the groups, SS_E where the E stands for "error". We call it error because this is then considered the part of the variation which is taken to be just random or coming from unknown factors. To make the notation less cluttered, we use a simplified sum notation, where

$$\sum_i = \sum_{i=1}^{k}$$

that is we sum over all elements of the noted index.

$$SS_T = \sum_i \sum_j (X_{i,j} - \bar{X}_T)$$

$$= \sum_i \sum_j \left(x_{i,j} - \bar{x}_i + \bar{x}_i - \bar{X}_T \right)^2$$

$$\ldots$$

$$= \underbrace{SS_B}_{\sum_{i=1}^{k} N_{Si} (\bar{X}_i - \bar{X}_T)^2} + \underbrace{SS_E}_{\sum_{i=1}^{k} \sum_{j=1}^{N_{Si}} (X_{i,j} - \bar{X}_i)^2}$$

$$\sum_{i=1}^{k} \sum_{j=1}^{N_{Si}} (X_{i,j} - \bar{X}_i)^2 = \sum_{i=1}^{k} (N_{Si} - 1) \cdot S_i^2$$

To compare the two contributions, we look at the "mean squared sum" *between* MS_B, and the "mean squared error" *inside* the group, MS_E. To compute them, we devide by the degrees of freedom of each. Between the groups

$$\nu_B = k - 1$$

(similar to the χ^2 test), because we lose one degree of freedom from the global mean. Inside the groups

$$\nu_E = \sum_{i=1} (N_{Si} - 1)$$

because we lose one degree of freedom for each group's mean.

$$MS_B = \frac{SS_B}{\underbrace{k-1}_{\nu_B}} \qquad MS_E = \frac{SS_E}{\sum_{i}^{k} (N_{Si} - 1)} = \frac{SS_E}{\underbrace{N_{ST} - k}_{\nu_E}}$$

It is insightful to consider the case where each group has the same number of elements, i.e. $N_{Si} = N_S$ for all i.

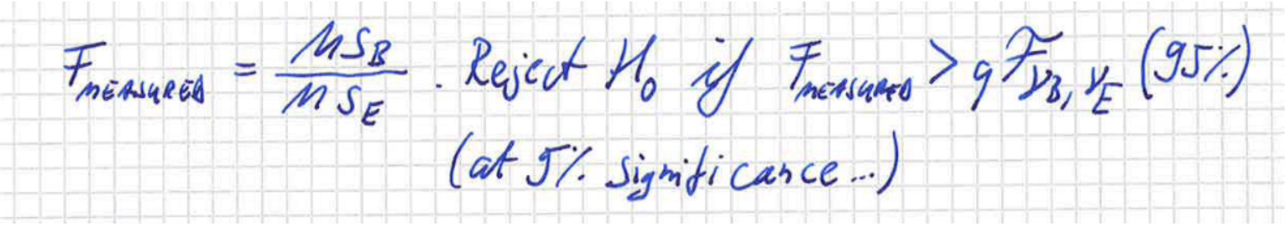
for N_{Si} all the same (all $= N_S$)

$$MS_B = \frac{N_{Si} \sum_{i=1}^{k} (\bar{x}_i - \bar{X}_T)^2}{k-1} = N_S \cdot S_B^2 \qquad MS_E = \frac{\cancel{(N_S - 1)} \sum_{i=1}^{k} S_i^2}{k \cancel{(N_S - 1)}} = \overline{S_i^2}$$

if all groups equivalent $= N_S \cdot \frac{\sigma^2}{N_S} = \sigma^2$

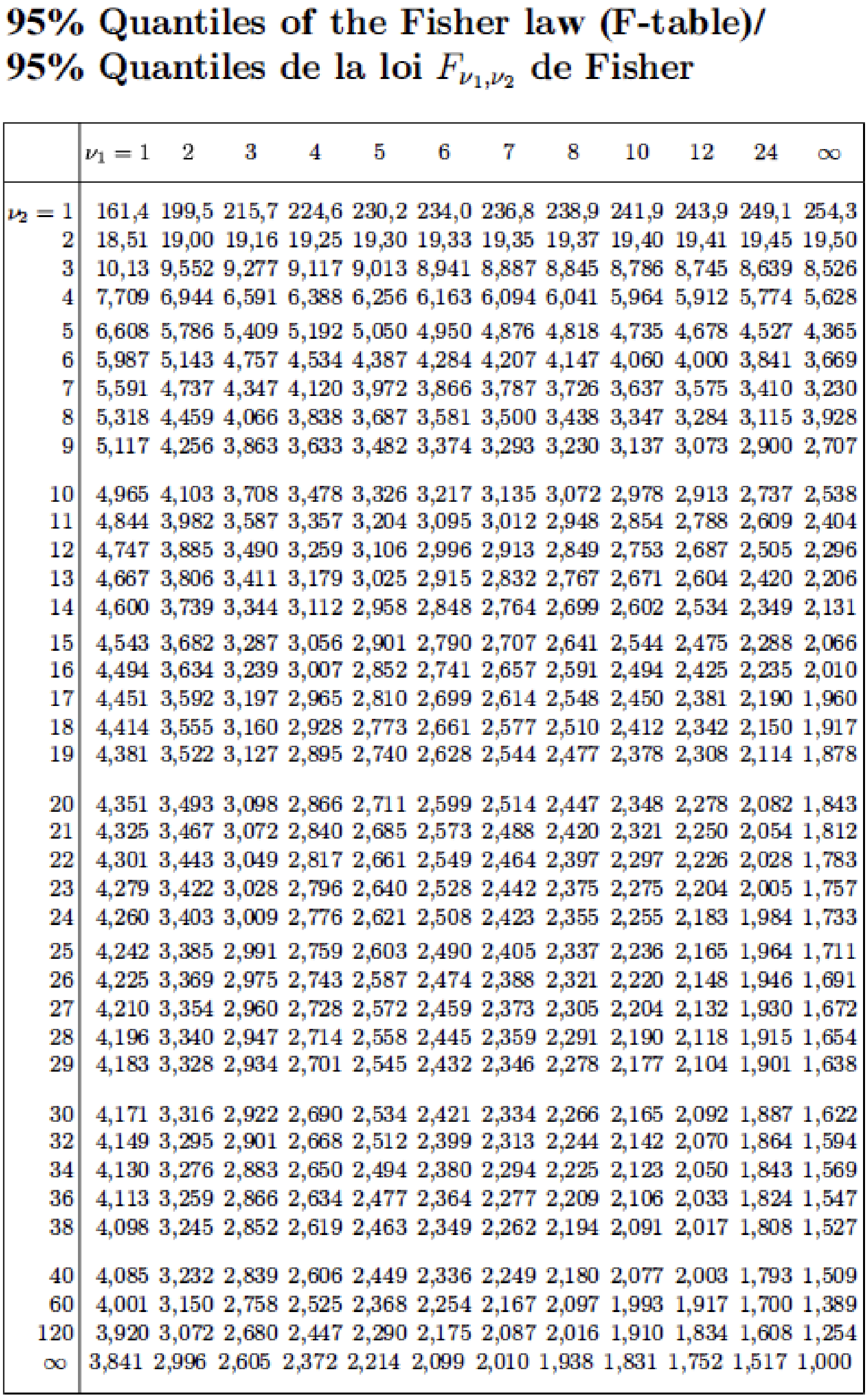
(mean of all variances)

Now we take the ratio of the MS_B and MS_E to compute the Fisher statistic, which we then compare to a critical value from a table.



$$F_{MEASURED} = \frac{MS_B}{MS_E} .$$ Reject H_0 if $F_{MEASURED} > 9 \mathcal{F}_{\nu_B, \nu_E}$ (95%) (at 5% significance ...)

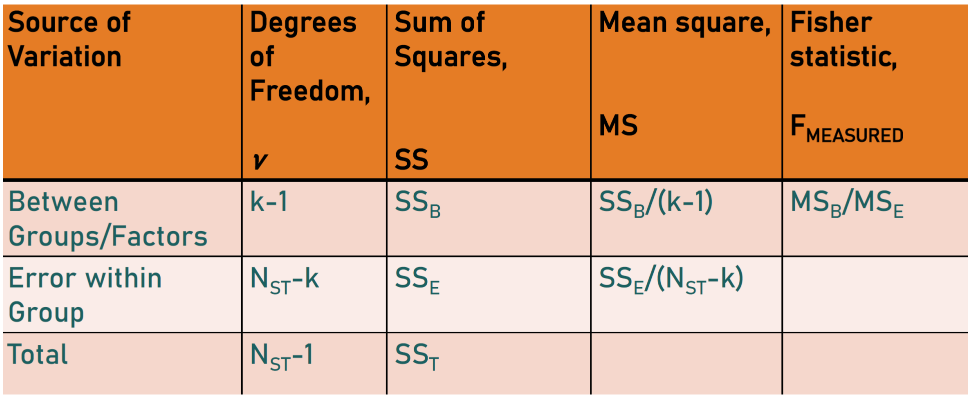
Note that the critical value depends on the chosen significance, and on the degrees of freedom:



95% Quantiles of the Fisher law (F-table)/ 95% Quantiles de la loi F_{ν_1,ν_2} de Fisher

	$\nu_1 = 1$	2	3	4	5	6	7	8	10	12	24	∞
$\nu_2 = 1$	161,4	199,5	215,7	224,6	230,2	234,0	236,8	238,9	241,9	243,9	249,1	254,3
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,40	19,41	19,45	19,50
3	10,13	9,552	9,277	9,117	9,013	8,941	8,887	8,845	8,786	8,745	8,639	8,526
4	7,709	6,944	6,591	6,388	6,256	6,163	6,094	6,041	5,964	5,912	5,774	5,628
5	6,608	5,786	5,409	5,192	5,050	4,950	4,876	4,818	4,735	4,678	4,527	4,365
6	5,987	5,143	4,757	4,534	4,387	4,284	4,207	4,147	4,060	4,000	3,841	3,669
7	5,591	4,737	4,347	4,120	3,972	3,866	3,787	3,726	3,637	3,575	3,410	3,230
8	5,318	4,459	4,066	3,838	3,687	3,581	3,500	3,438	3,347	3,284	3,115	3,928
9	5,117	4,256	3,863	3,633	3,482	3,374	3,293	3,230	3,137	3,073	2,900	2,707
10	4,965	4,103	3,708	3,478	3,326	3,217	3,135	3,072	2,978	2,913	2,737	2,538
11	4,844	3,982	3,587	3,357	3,204	3,095	3,012	2,948	2,854	2,788	2,609	2,404
12	4,747	3,885	3,490	3,259	3,106	2,996	2,913	2,849	2,753	2,687	2,505	2,296
13	4,667	3,806	3,411	3,179	3,025	2,915	2,832	2,767	2,671	2,604	2,420	2,206
14	4,600	3,739	3,344	3,112	2,958	2,848	2,764	2,699	2,602	2,534	2,349	2,131
15	4,543	3,682	3,287	3,056	2,901	2,790	2,707	2,641	2,544	2,475	2,288	2,066
16	4,494	3,634	3,239	3,007	2,852	2,741	2,657	2,591	2,494	2,425	2,235	2,010
17	4,451	3,592	3,197	2,965	2,810	2,699	2,614	2,548	2,450	2,381	2,190	1,960
18	4,414	3,555	3,160	2,928	2,773	2,661	2,577	2,510	2,412	2,342	2,150	1,917
19	4,381	3,522	3,127	2,895	2,740	2,628	2,544	2,477	2,378	2,308	2,114	1,878
20	4,351	3,493	3,098	2,866	2,711	2,599	2,514	2,447	2,348	2,278	2,082	1,843
21	4,325	3,467	3,072	2,840	2,685	2,573	2,488	2,420	2,321	2,250	2,054	1,812
22	4,301	3,443	3,049	2,817	2,661	2,549	2,464	2,397	2,297	2,226	2,028	1,783
23	4,279	3,422	3,028	2,796	2,640	2,528	2,442	2,375	2,275	2,204	2,005	1,757
24	4,260	3,403	3,009	2,776	2,621	2,508	2,423	2,355	2,255	2,183	1,984	1,733
25	4,242	3,385	2,991	2,759	2,603	2,490	2,405	2,337	2,236	2,165	1,964	1,711
26	4,225	3,369	2,975	2,743	2,587	2,474	2,388	2,321	2,220	2,148	1,946	1,691
27	4,210	3,354	2,960	2,728	2,572	2,459	2,373	2,305	2,204	2,132	1,930	1,672
28	4,196	3,340	2,947	2,714	2,558	2,445	2,359	2,291	2,190	2,118	1,915	1,654
29	4,183	3,328	2,934	2,701	2,545	2,432	2,346	2,278	2,177	2,104	1,901	1,638
30	4,171	3,316	2,922	2,690	2,534	2,421	2,334	2,266	2,165	2,092	1,887	1,622
32	4,149	3,295	2,901	2,668	2,512	2,399	2,313	2,244	2,142	2,070	1,864	1,594
34	4,130	3,276	2,883	2,650	2,494	2,380	2,294	2,225	2,123	2,050	1,843	1,569
36	4,113	3,259	2,866	2,634	2,477	2,364	2,277	2,209	2,106	2,033	1,824	1,547
38	4,098	3,245	2,852	2,619	2,463	2,349	2,262	2,194	2,091	2,017	1,808	1,527
40	4,085	3,232	2,839	2,606	2,449	2,336	2,249	2,180	2,077	2,003	1,793	1,509
60	4,001	3,150	2,758	2,525	2,368	2,254	2,167	2,097	1,993	1,917	1,700	1,389
120	3,920	3,072	2,680	2,447	2,290	2,175	2,087	2,016	1,910	1,834	1,608	1,254
∞	3,841	2,996	2,605	2,372	2,214	2,099	2,010	1,938	1,831	1,752	1,517	1,000

To get a nice overview, we produce an "ANOVA Table"



Source of Variation	Degrees of Freedom, ν	Sum of Squares, SS	Mean square, MS	Fisher statistic, $F_{MEASURED}$
Between Groups/Factors	k-1	SS_B	SS_B/(k-1)	MS_B/MS_E
Error within Group	N_{ST}-k	SS_E	SS_E/(N_{ST}-k)	
Total	N_{ST}-1	SS_T		

Note that for only two groups, $k = 2$ ($\nu = 1$), the ANOVA Fisher test is equivalent to the Student t-test: the squareroot of $\mathcal{F}_{\infty,\nu_\mathcal{E}}$ is the same as the corresponding two-sided value from the t-table. So we can see ANOVA as an extension of the Student t-test to a number of groups larger than two.