Statistical Analysis: Crash Course

December 1, 2022

Slides available at: `tynanochse.com`

Why use statistical tests?

- Short answer: we are scientists and we like precision
- Longer answer: when we perform an experiment, sometimes we think there is a difference between the data we collect and the data we expect. But when we perform a statistical test, it turns out the difference we see isn't really "significant." This can happen the other way too: we think our results fall in line with what is expected, but in fact they are significantly different.

Chi-square test

Use when doing a comparison between **categorical** samples.

	Attribute 1	Attribute 2	Totals
Category 1	observed	observed	Total
Category 2	observed	observed	Total
	Total	Total	

Every χ^2-test will have

- Null hypothesis: There are no relationships between the categorical variables
- Significane threshold ($p \leq 0.05$ is standard)

When doing the test, "expected" data is the data you should see if the null hypothesis is true. We then compute

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

This means "calculate the fraction $\frac{(O-E)^2}{E}$ for every pair of observed-versus-expected data, and then add them all up". Also compute the "degrees of freedom:"

$$DF = (\text{rows} - 1)(\text{columns} - 1)$$

So if we have two "categories" and each category has two data-points, then the degrees of freedom is equal to $(2 - 1) \cdot (2 - 1) = 1$.

These two numbers can be used to produce a p-value.

If $p \leq 0.05$, then we reject the null hypothesis.

If $p > 0.05$, we fail to reject the null hypothesis.

	Cats	Dogs	Totals
Undergraduates	207	282	489
Graduates	231	242	473
	438	524	962

Expected value for undergraduates who prefer cats is $\frac{438 \cdot 489}{962}$

$$\frac{(O-E)^2}{E} = \frac{(207 - 222.64)^2}{222.64} = 1.099$$

Expected value for undergraduates who prefer dogs is $\frac{524 \cdot 489}{962}$

$$\frac{(O-E)^2}{E} = \frac{(282 - 266.36)^2}{266.36} = 0.918$$

$$\chi^2 = 1.099 + 0.918 + 1.136 + 0.949 = 4.102$$

and

$$DF = (2 - 1)(2 - 1) = 1$$

(Two categories = “freshmen” and “seniors”, and two choices for each category = “dogs” and “cats”)

Now we ask what is the probability of getting a χ^2 value of 4.102 when $DF = 1$:

$$p = 0.043$$

We see that $p < 0.05$, so we reject the null hypothesis with 95% certainty.

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.953	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69

t-test

Use when doing a comparison between a sample and a population, or when doing a comparison between samples.

Sample	1	2	3	4	5	$\cdots$	n
Data	d_1	d_2	d_3	d_4	d_5	$\cdots$	d_n

t-tests only work when the population is approximately normally distributed and n is not too small.

Every t-test will have

- Null hypthesis: there is no difference between sample mean and proposed population mean.
- Confidence interval (usually 95%)

We calculate

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}$$

and

$$DF = n - 1$$

$\bar{x}$ is the mean of the sample.
μ_0 is the null hypothesis mean.
s is the sample standard deviation.
n is the sample size.

Use DF and the condifence level to look up a “critical value”

$$t^*$$

If $t < t^*$, then we fail to reject the null hypothesis (results are not statistifcally significant).

If $t > t^*$, then we reject the null hypothesis.

A biologist wants to test a special fertilizer on her crops to see if it has an affect on crop yield. She tests the fertilizer on 25 random plants and measures yield. On average, she is able to get a yield of 75 per plant without the fertilizer.

Sample	1	2	3	4	5	$\cdots$	25
Data	76	81	80	79	82	$\cdots$	79

Null hypothesis: the fertilizer has no significant affect on crop yield. From the data we compute $\bar{x} = 79$ and $s = 10$. So

$$t = \frac{79 - 75}{\frac{10}{\sqrt{25}}} = 2$$

and

$$DF = 25 - 1 = 24$$

We find $t^* = 2.064$. Since $t < t^*$, we fail to reject the null hypothesis (i.e there is not a significant affect on crop yield by using the fertilizer).

Degrees of freedom	Significance level					
	20% (0.20)	10% (0.10)	5% (0.05)	2% (0.02)	1% (0.01)	0.1% (0.001)
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.598
3	1.638	2.353	3.182	4.541	5.841	12.941
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.859
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.405
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.767
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.043	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
40	1.303	1.684	2.021	2.423	2.704	3.551
60	1.296	1.671	2.000	2.390	2.660	3.460
120	1.289	1.658	1.980	2.158	2.617	3.373
∞	1.282	1.645	1.960	2.326	2.576	3.291

We can also do comparisons between two sample groups.

Suppose the researcher in the previous example wants to compare the affect of two different fertilizers on crop yield. She randomly picks 15 plants to give fertilizer A and 12 plants to get fertilizer B. She collects the crop yield data from each group and finds

$$\text{average yield of crops using A } = \bar{x}_1 = 79$$

$$\text{average yield of crops using B } = \bar{x}_2 = 87$$

and

$$s_1 = 10$$

$$s_2 = 9$$

For comparisons between groups, we compute

$$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$DF = (n_1 - 1) + (n_2 - 2)$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

for our example we have

$$s_p^2 = \frac{(15 - 1)10^2 + (12 - 1)9^2}{25} = 91.64$$

so

$$t = \frac{87 - 79}{\sqrt{91.64}\sqrt{\frac{1}{15} + \frac{1}{12}}} = 2.1578$$

and

$$DF = (15 - 1) + (12 - 1) = 25$$

With 25 degrees of freedom and a confidence level of 95% we get

$$t^* = 2.06$$

Since

$$t = 2.1578 > 2.06 = t^*$$

we reject the null hypothesis (there is a significant difference between average crop yeilds)

ANOVA

ANOVA = analysis of variance

Comparison between several groups

	Group 1	Group 2	Group 3	Group 4
Sample size	n_1	n_2	n_3	n_4
Sample mean	$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$	$\bar{x}_4$
Sample SD	s_1	s_2	s_3	s_4

k is the number of groups (so $k = 4$ in the example above).
$N = n_1 + n_2 + n_3 + n_4$ is the number of samples in the data. Data point $x_{i,j}$ is the j^{th} sample from group i, e.g $x_{2,1}$ is the value of sample 1 from group 2.

$$\bar{x}_2 = \frac{\sum_j x_{2,j}}{n_2}$$

Set $\bar{x}$ to be the average of the entire sample

Any ANOVA test has

- Null hypothesis: the population means are all equal
- Confidence interval/significance threshold (95%/$p \leq 0.05$)

For a simple formula for the F-statistic, we assume sample standard deviations are all equal: $s_1 = s_2 = s_3 = s_4$. Then

$$F = \frac{V_E}{V_U}$$

where

$$V_E = \frac{\sum_i n_i(\bar{x}_i - \bar{x})^2}{k-1}$$

and

$$V_U = \frac{\sum_i \sum_j (x_{i,j} - \bar{x}_i)^2}{N-k}$$

$$DF_1 = k-1,\ DF_2 = N-k$$

Use DF_1, DF_2, and threshold 0.05 to look up critical F-value F^*.

If $F > F^*$, we reject the null hypothesis.

If $F < F^*$, we fail to reject the null hypothesis.

A nutritionist is studying the affect of different diets on weight-loss. Three diets are considered, and from the population of several hundred participants in the study, 5 people from each diet program are randomly sampled. The sample participants' weight loss is tabulated below (positive number means weight loss):

Low cal	Low fat	Low carb	control
8	2	3	2
9	4	5	2
6	3	4	-1
7	5	2	0
3	1	3	3

So we have 4 groups and 5 observations from each group. This means $k = 4$, $n_1 = n_2 = n_3 = n_4 = 5$ and $N = 20$.

- Null hypothesis: the population means in weight-loss are all equal (i.e diet does not have a significant affect on weight-loss).
- Threshold 0.05.

	Low cal	Low fat	Low carb	control
Sample size	5	5	5	5
Sample mean	6.6	3.0	3.4	1.2

We compute $\bar{x} = 3.6$ and

$$V_E = \frac{5(6.6-3.6)^2 + 5(3-3.6)^2 + 5(3.4-3.6)^2 + 5(1.2-3.6)^2}{4-1}$$

so $V_E = 25.27$.

To compute V_U, we look at the data for each group:

$$\sum_{j=1}^{5}(x_{1,j}-\bar{x}_1)^2 = (8-6.6)^2+(9-6.6)^2+(6-6.6)^2+(7-6.6)^2+(3-6.6)^2$$

$$\sum_{j=1}^{5}(x_{2,j}-\bar{x}_2)^2 = (2-3)^2+(4-3)^2+(3-3)^2+(5-3)^2+(1-3)^2$$

etc... to get

$$V_U = \frac{21.4+10.0+5.4+10.6}{20-4} = 2.9625$$

So

$$F = \frac{V_E}{V_U} = 8.523$$

$$DF_1 = 4 - 1 = 3 \ , DF_2 = 20 - 4 = 16$$

With significance threshold 0.05, we get

$$F^* = 3.239$$

Since $F > F^*$ we reject the null hypothesis (i.e there is a difference in population means between the four diets)

Appendix 4a

5 per cent Points of the *F*-distribution

Column represents degrees of freedom (ν_1) for numerator of *F*-test
Row represents degrees of freedom (ν_2) for denominator of *F*-test

	1	2	3	4	5	6	7	8	9	10	12	24	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	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	19.38	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	8.812	8.785	8.745	8.638	8.526
4	7.709	6.944	6.591	6.388	6.256	6.163	6.094	6.041	5.999	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	4.772	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	4.099	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	3.677	3.637	3.575	3.410	3.230
8	5.318	4.459	4.066	3.838	3.688	3.581	3.500	3.438	3.388	3.347	3.284	3.115	2.928
9	5.117	4.256	3.863	3.633	3.482	3.374	3.293	3.230	3.179	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	3.020	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	2.896	2.854	2.788	2.609	2.405
12	4.747	3.885	3.490	3.259	3.106	2.996	2.913	2.849	2.796	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	2.714	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	2.646	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	2.588	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	2.538	2.494	2.425	2.235	2.010

Summary

Three different types of significance tests... when to use each one?

1. χ^2 tests if there is a relationship between two variables
2. t-test tests if two groups are statistically different
3. ANOVA tests if many groups are statistically different

Examples

Scenario 1: You have a sample of a bunch of different acidic compounds with ph values between 1 and 6. You know that compound A reacts with acidic compounds, but you want to know if there is a relationship between the ph of the acid and the rate of the reaction with compound A.

Scenario 2: You have two vials of blue liquid. You know the liquids consist of some metal ion in solution: one solution contains Cu^{+2} and one containes Cr^{+2}. You also are given samples of the elemental metal extracted from each solution (you know which metal comes from which vial). You want to decide which vial contains copper ion, and so you take several samples of the elemental metals and test if it reacts with water and acid.

Scenario 3: You have 5 samples of a strong base, a weak base, a strong acid, and a week acid. You know all compounds react with solution B. For each sample you calculate the reaction rate/velocity with solution B. You want to know if there is a signicant difference in average reaction times.