Lecture 8: Hypothesis Testing and Contingency Analysis

t-test

- The assumption that the sampling distribution will be **normally distributed** holds for large samples but not for small samples
- Sample size is **large**, use **z-test**
- When sample size is **small**, **t-test** is used
  - Statistical concept of t-distribution
  - Comparing means for **2 independent groups**
    - **unpaired t-test**
  - Comparing means for **2 matched groups**
    - **paired t-test**

t-test for 2 independent samples

- $X_1' - X_2' = 0.08157 - 0.03943 = 0.04$

- **Question**: What is the probability that the difference of 0.04 units between the two sample means has occurred purely by chance, i.e. due to sampling error alone?

Blood Pb concentrations

	Battery workers (occupationally exposed)	Control (not occupationally exposed)
	0.082	0.040
	0.080	0.035
	0.079	0.036
	0.069	0.039
	0.085	0.040
	0.09	0.046
	0.086	0.040
mean	**0.08157**	**0.03943**
std dev	**0.0067047**	**0.0035523**

t-test for 2 independent samples

- In general, we can denote the means of the two groups as μ_1 and μ_2.
- The null hypothesis indicates that the population means are equal, $H_0 : \mu_1 = \mu_2$.
- In contrast, the alternative hypothesis is one the following:
  - $\mathbf{H_A : \mu_1 > \mu_2}$ if we believe the mean for group 1 is greater than the mean for group 2.
  - $\mathbf{H_A : \mu_1 < \mu_2}$ if we believe the mean for group 1 is less than the mean for group 2.
  - $\mathbf{H_A : \mu_1 \neq \mu_2}$ if we believe the means are different but we do not specify which one is greater.

- We can also express these hypotheses in terms of the difference in the means:
  - $H_A : \mu_1 - \mu_2 > 0$,
  - $H_A : \mu_1 - \mu_2 < 0$, or
  - $H_A : \mu_1 - \mu_2 \neq 0$

Unpaired t-test

- We are testing the **hypothesis** that **battery workers** could have **higher blood** Pb levels than the **control group** of workers as they are occupationally exposed

- Note: conventionally, a p-value of 0.05 is generally recognized as low enough to reject the Null Hypothesis of "no difference"

Blood Pb concentrations

	Battery workers (occupationally exposed)	Control (not occupationally exposed)
	0.082	0.040
	0.080	0.035
	0.079	0.036
	0.069	0.039
	0.085	0.040
	0.09	0.046
	0.086	0.040
mean	**0.08157**	**0.03943**
std dev	**0.0067047**	**0.0035523**

Unpaired t-test

- **Null Hypothesis**: No difference in mean blood Pb level between battery workers and control group, i.e.
  - H0: $\mu_{battery} = \mu_{control}$

- t-score is given by

$$t = \frac{\overline{X}_1 - \overline{X}_2}{SE_{(\overline{X}_1 - \overline{X}_2)}} = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}}$$

with (n_1+n_2-2) degrees of freedom

Unpaired t-test

- For the given example

$$t = \frac{0.08157 - 0.03943}{0.002868}$$

= 14.7 with 12 d.f.

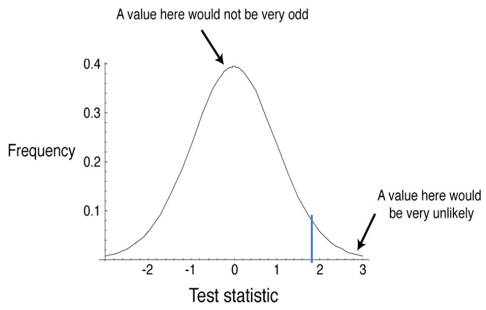
- P-value <0.001, **reject Null hypothesis**

- Some evidence, from the data, that battery workers in our study have higher blood Pb level than the control group on average

Blood Pb concentrations

	Battery workers (occupationally exposed)	Control (not occupationally exposed)
	0.082	0.040
	0.080	0.035
	0.079	0.036
	0.069	0.039
	0.085	0.040
	0.09	0.046
	0.086	0.040
mean	**0.08157**	**0.03943**
std dev	**0.0067047**	**0.0035523**

t-table



From our example:
t=14.7 with 12 d.f.

Value far exceeds 4.318, the critical value for statistical significance at the Pr=0.001 (0.1%) level when df=12 i.e. Pr < 0.001

	Probability			
df	.05	.02	.01	.001
1	12.706	31.821	63.657	636.619
2	4.303	6.965	9.925	31.598
3	3.182	4.541	5.841	12.924
4	2.776	3.747	4.604	8.610
5	2.571	3.365	4.032	6.869
6	2.447	3.143	3.707	5.959
7	2.365	2.998	3.499	5.408
8	2.306	2.896	3.355	5.041
9	2.262	2.821	3.250	4.781
10	2.228	2.764	3.169	4.587
11	2.201	2.718	3.106	4.437
12	2.179	2.681	3.055	4.318
13	2.160	2.650	3.012	4.221
14	2.145	2.624	2.977	4.140
15	2.131	2.602	2.947	4.073
16	2.120	2.583	2.921	4.015
17	2.110	2.567	2.898	3.965
18	2.101	2.552	2.878	3.922
19	2.093	2.539	2.861	3.883
. . .	. . .	. . .	. . .	. . .
25	2.060	2.485	2.787	3.725
26	2.056	2.479	2.779	3.707
27	2.052	2.473	2.771	3.690
28	2.048	2.467	2.763	3.674
29	2.045	2.462	2.756	3.659
30	2.042	2.457	2.750	3.646
40	2.021	2.423	2.704	3.551
60	2.000	2.390	2.660	3.460
120	1.980	2.358	2.617	3.373
α	1.960	2.326	2.576	3.291

Unpaired t-test assumptions

- Data are **normally distributed** in the population from which the **two independent samples** have been drawn
- The **two samples** are **random** and **independent**, i.e. observations in one group are not related to observations in the other group
- The two independent samples have been drawn from populations with the **same** (homogeneous) **variance**, i.e. $\sigma_1=\sigma_2$

Paired t-test

- Previous problem uses un-paired t-test as the two samples were matched
  - i.e. the two samples were independently derived

- Sometimes, we may need to deal with matched study designs

Patient	Fasting cholesterol	Postprandial cholesterol
1	198	202
2	192	188
3	241	238
4	229	226
5	185	174
6	303	315

Study involves 6 subjects acting as their own control (best match)

Paired t-test

- **Null hypothesis**: No difference in mean cholesterol levels between fasting and after eating states
  - H0: $\mu_{fasting} = \mu_{postprandial}$

Patient	Fasting cholesterol	Postprandial cholesterol	Difference (d)
1	198	202	-4
2	192	188	+4
3	241	238	+3
4	229	226	+3
5	185	174	+11
6	303	315	-12

$\bar{d} = 0.833$
$s_d = 7.885$
$n = 6$

Paired t-test

- t-score given by

$$t = \frac{\bar{d}}{SE_{\bar{d}}} = \frac{\bar{d}}{s_d / \sqrt{n}}$$

$$\frac{0.833}{3.219} = 0.259$$

with (n-1) degrees of freedom, where n is the # of pairs

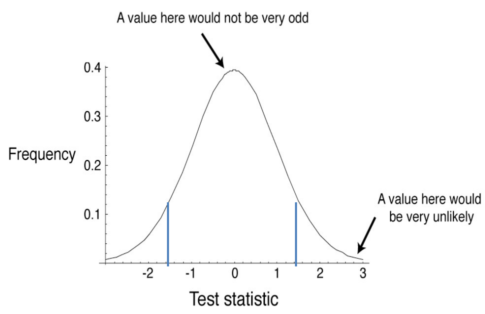
Patient	Difference (d)
1	-4
2	+4
3	+3
4	+3
5	+11
6	-12

$\bar{d}$ = 0.833
s_d = 7.885
n= 6

t-table

From our example:
t=0.259 with 5 d.f.

Value is very much lower than 2.571, the critical value for statistical significance at the Pr=0.05 (5%) level when df=5
i.e. Pr > 0.05



	Probability			
df	.05	.02	.01	.001
1	12.706	31.821	63.657	636.619
2	4.303	6.965	9.925	31.598
3	3.182	4.541	5.841	12.924
4	2.776	3.747	4.604	8.610
5	2.571	3.365	4.032	6.869
6	2.447	3.143	3.707	5.959
7	2.365	2.998	3.499	5.408
8	2.306	2.896	3.355	5.041
9	2.262	2.821	3.250	4.781
10	2.228	2.764	3.169	4.587
11	2.201	2.718	3.106	4.437
12	2.179	2.681	3.055	4.318
13	2.160	2.650	3.012	4.221
14	2.145	2.624	2.977	4.140
15	2.131	2.602	2.947	4.073
16	2.120	2.583	2.921	4.015
17	2.110	2.567	2.898	3.965
18	2.101	2.552	2.878	3.922
19	2.093	2.539	2.861	3.883
...	...	...	...	...
25	2.060	2.485	2.787	3.725
26	2.056	2.479	2.779	3.707
27	2.052	2.473	2.771	3.690
28	2.048	2.467	2.763	3.674
29	2.045	2.462	2.756	3.659
30	2.042	2.457	2.750	3.646
40	2.021	2.423	2.704	3.551
60	2.000	2.390	2.660	3.460
120	1.980	2.358	2.617	3.373
α	1.960	2.326	2.576	3.291

Paired t-test

- **Conclusion**: Insufficient evidence, from the data, to suggest that postprandial cholesterol levels are, on average, higher than fasting cholesterol levels
- **Action**: Should not reject the Null Hypothesis

Patient	Fasting cholesterol	Postprandial cholesterol
1	198	202
2	192	188
3	241	238
4	229	226
5	185	174
6	303	315

Common errors relating to t-test

- **Failure to recognize assumptions**
  - If assumption does not hold, explore data transformation or use of non-parametric methods

- **Failure to distinguish** between **paired** and **unpaired** designs

Contingency Analysis: Associations between Categorical variables

Association

- Examining relationship between 2 categorical variables
- Some examples of association:
  - Smoking and lung cancer
  - Number of defected sensors and season of the year
  - Ethic group and choice of Movie Genre
- Questions of interest when testing for association between two categorical variables
  - Does the **presence/absence** of **one factor** (variable) **influence** the presence/absence of the **other factor** (variable)?
- **Caution**
  - presence of an **association** does **not necessarily** imply **causation**

Hypothesis testing involving categorical data

Test the independence of **two or more categorical variables**

Chi-square is a test for statistical **association** between **two variables** and involving 2x2 tables or **contingency tables**

- Testing for associations involving **small, unmatched samples and small, matched samples**

Assumptions of the Chi-Square Test

1. The $\chi 2$ assumes that the data for the study is obtained through **random selection**

2. The **categories** are **mutually exclusive** i.e. each subject fits in only one category.

3. The **data** should be in the form of **frequencies** or **counts** of a particular category and **not in percentages**

4. The data should **not consist of paired samples**.
   - **Observations** should be **independent** of each other

5. More than **80% of the expected frequencies** must have a **value of more than 5**.
   - To tackle this problem: Either one should combine the categories only if it is relevant or obtain more data or use Fisher's exact test

Comparison between proportions

Treatment	Improvement	No improvement	Total
Arthritic drug	18	6	24
placebo	9	11	20
Total	27	17	44

- Proportion improved in drug group = 18/24 = 75%
- Proportion improved in control group = 9/20 = 45.0%
- **Question**: What is the probability that the observed difference of 30% is purely due to sampling error, i.e. chance in sampling?
- Use **chi-squared –test**

Chi-square test for statistical association

treatment	Improvement	No improvement	Total
Arthritic drug	18 (a)	6 (b)	24
placebo	9 (c)	11 (d)	20
Total	27	17	44

- Prob of selecting a person in drug group = 24/44
- Prob of selecting a person with improvement = 27/44
- Prob of selecting a person from drug group who had shown improvement= (24/44)*(27/44) = 0.3347 (assuming two independent events)
- Expected value for cell (a) =0.3347*44 = 14.73

Chi-square test for statistical association

treatment	Improvement	No improvement	Total
Arthritic drug	18 (14.73)	6 (9.27)	24
placebo	9 (12.27)	11 (7.73)	20
Total	27	17	44

- General formula for Chi-squared:

$$\chi^2 = \sum \frac{(obs - \exp)^2}{\exp}$$

- Chi-squared –test is always performed on categorical variables using absolute frequencies, never percentage or proportion

Chi-square test for statistical association

- For the given problem:

$$\sum \frac{(obs - \exp)^2}{\exp} = \frac{(18 - 14.73)^2}{14.73} + \frac{(6 - 9.27)^2}{9.27} + \frac{(9 - 12.27)^2}{12.27} + \frac{(11 - 7.73)^2}{7.73}$$

$$= 4.14 \text{ with 1 degree of freedom}$$

- **Chi-square** degree of freedom is given by: (no. of rows-1)*(no. of cols-1) = (2-1)*(2-1) = 1

How many of these 4 cells are free to vary if we keep the row and column totals constant?

18	6	24
9	11	20
27	17	44

χ^2 table

Critical values in the distributions of chi-squared
for different degrees of freedom

df	Probability .05	.02	.01	.001
1	3.841	5.412	6.635	10.827
2	5.991	7.824	9.210	13.815
3	7.815	9.837	11.345	16.266
4	9.488	11.668	13.277	18.467
5	11.070	13.388	15.086	20.515
6	12.592	15.033	16.812	22.457
7	14.067	16.622	18.475	24.322
8	15.507	18.168	20.090	26.125
9	16.919	19.679	21.666	27.877
10	18.307	21.161	23.209	29.588
11	19.675	22.618	24.725	31.264
12	21.026	24.054	26.217	32.909
13	22.362	25.372	27.688	34.528
14	23.585	26.873	29.141	36.123
15	24.996	28.259	30.578	37.697
16	26.296	29.633	32.000	39.252
17	27.587	30.995	33.409	40.790
18	28.869	32.346	34.805	42.312
19	30.144	33.687	36.191	43.820
20	31.410	35.020	37.566	35.315
21	32.671	36.343	38.932	46.797
22	33.924	37.659	40.289	48.268
23	35.172	38.968	41.638	49.728
24	36.415	40.270	42.980	51.179
25	37.652	41.566	44.314	52.620
26	38.885	42.856	45.642	54.052
27	40.113	44.140	46.963	55.476
28	41.337	45.419	48.278	56.893
29	42.557	46.693	49.588	58.302
30	43.773	47.962	50.892	59.703

observed χ^2 value of 4.14 exceeds critical value of 3.841 for P=0.05 but not critical value of 5.412 for P=0.02 at 1 d.f.

i.e. 0.05 > P > 0.02

Chi-square test for statistical association

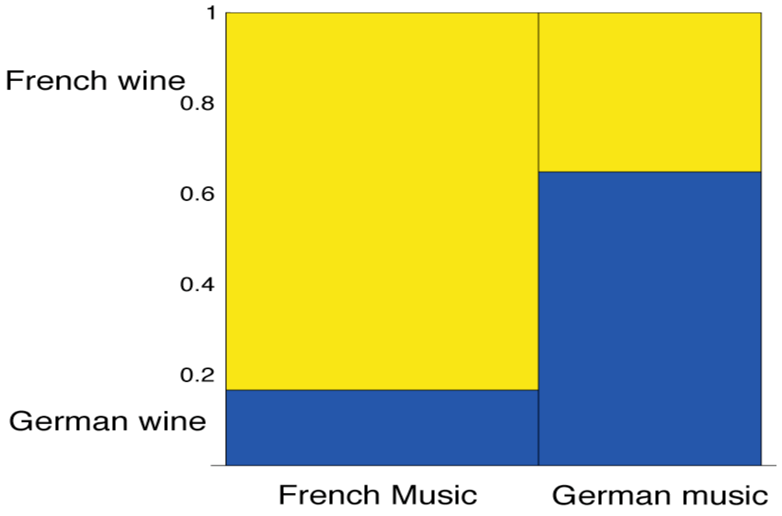
- Probability of getting an observed difference of 30% in improvement rates if the Null hypothesis of no association is correct is between 2% and 5%

- Hence, there is **some statistical evidence** from this study to suggest that treatment of arthritic patient with the drug can significantly improve grip strength

Another Example

Music and wine buying

OBSERVED	French music playing	German music playing	Totals
Bottles of French wine sold	40	12	52
Bottles of German wine sold	8	22	30
Totals	48	34	82



Hypotheses

- **H_0:** The nationality of the bottle of wine *is* ***independent*** of the nationality of the music played when it is sold.

- **H_A:** The nationality of the bottle of wine sold ***depends*** **on the nationality** of the music being played when it is sold.

Calculating the expectations

With independence,

Pr [French wine AND French music] = Pr [French wine] * Pr [French music]

Calculating the expectations

EXP.	French music	German music	Totals
French wine sold			52
German wine sold			30
Totals	48	34	82

Pr[French wine] = 52/82=0.634

Pr[French music] = 48/82= 0.585

If H_0 is true,
Pr[French wine AND French music] = (0.634)(0.585) = 0.37112

EXP.	French music	German music	Totals
French wine sold	0.37 (82) = 30.4	21.6	52
German wine sold	17.6	12.4	30
Totals	48	34	82

By H_0,
Pr[French wine AND French music] = (0.634)(0.585)=0.37112

$$X^2 = \frac{(Observed_i - Expected_i)^2}{Expected_i}$$

$$= \frac{(40 - 30.4)^2}{30.4} + \frac{(12 - 21.6)^2}{21.6} + \frac{(8 - 17.6)^2}{17.6} + \frac{(22 - 12.4)^2}{12.4}$$

$$= 20.0$$

Degrees of freedom

df= (# columns -1)(#rows -1)

For music/wine example,

df = (2-1)(2-1) = 1

Conclusion

So we can **reject the null hypothesis of independence**, and say that the nationality of the wine sold did depend on what music was played.

$\chi^2 = 20.0 >> \chi^2 = 10.83$, so we can say $P < 0.001$.

Assumptions

- This χ^2 test is just a special case of the χ^2 **goodness-of-fit test**, so the same rules apply.
- You can't have any expectation less than 1, and no more than 20% < 5.

Fisher's exact test

- For 2 x 2 contingency analysis.
- Does not make assumptions about the size of expectations
- When N<20 or N>20 but expected cell count is >=5 is less than 80% of cells.
- Programs will do it, but cumbersome to do by hand

	Men	**Women**	*Row Total*
Studying	***a***	***b***	*a* + *b*
Non-studying	***c***	***d***	*c* + *d*
Column Total	*a* + *c*	*b* + *d*	*a* + *b* + *c* + *d* (=*n*)

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} = \frac{\binom{a+b}{b}\binom{c+d}{d}}{\binom{n}{b+d}} = \frac{(a+b)!\;(c+d)!\;(a+c)!\;(b+d)!}{a!\;\;b!\;\;c!\;\;d!\;\;n!}$$

Fisher's exact test

Calculating the expectations

EXP.	Men	Women	Totals
Studying			12
Non-studying			4
Totals	12	4	16

A shortcut for calculating expectations (assuming H_0 is true):

$$Exp[row\ i,\ column\ j] = \frac{(row\ i\ total)(column\ j\ total)}{grand\ total}$$

Exp[Studying, Men] = 12*12/16 = 9

Comparing observed and expected

OBS.	Men	Women	Totals
Studying	12	0	12
Non-studying	0	4	4
Totals	12	4	16

EXP.	Men	Women	Totals
Studying	9	3	12
Non-studying	3	1	4
Totals	12	4	16

Too many of the expected are below 5, so we cannot use the ^{2} contingency test. Instead, we use a computer to do Fisher's exact test:

P = 0.00055, so we reject the H_0 of no association.