Inspire…Educate…Transform.

Stat Skills

Chi-Square Distribution, F Distribution, ANOVA, Correlation and Covariance

Dr. Anand Jayaraman

ajayaraman@gmail.com

Apr 9, 2017

Thanks to Dr.Sridhar Pappu for the material

What do I remember from yesterday? Do you mean other than my way back here?

Review

- Student t-distribution
  - Small samples or σ unknown
  - Depends on degrees of freedom (sample size)
- Confidence Intervals
  - Margin of Error
- Hypothesis Testing
  - Null Hypothesis and Alternate Hypothesis
  - Compute relevant statistic and check if it lies in the critical region
  - Types of Errors in Hypothesis testing

Review

- Two sample t-test for the mean
  - Dependent samples – paired t-test
  - Unrelated samples – unpaired t-test
- Connection between Hypothesis testing formalism and confidence intervals

Two sample t-Test : unpaired data

The Central Limit Theorem states that the difference in two sample means, $\overline{x_1} - \overline{x_2}$, is normally distributed for large sample sizes (both n_1 and $n_2 \geq 30$) whatever the population distribution.

Also, $\mu_{\overline{x_1}-\overline{x_2}} = \mu_1 - \mu_2$ *[Recall E(X-Y)=E(X)-E(Y)]*

and $\sigma_{\overline{x_1}-\overline{x_2}} = \sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}$ *[Recall Var(X-Y)=Var(X)+Var(Y)]*

$$z = \frac{observed\ difference - expected\ difference}{SE\ of\ the\ difference} = \frac{(\overline{x_1}-\overline{x_2})-(\mu_1-\mu_2)}{\sqrt{\frac{{\sigma_1}^2}{n_1}+\frac{{\sigma_2}^2}{n_2}}}$$

This is the test statistic for a 2-sample z-test.

Two-Sample t-Test for Unpaired Data

$H_0: \mu_1 = \mu_2; H_1: \mu_1 \neq \mu_2$

Test statistic, $t = \frac{(\overline{x_1} - \overline{x_2})}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}}$

Assuming the two samples come from populations with the **same standard deviation** (Rule of thumb: The ratio between the higher *s* and the lower *s* is less than 2), pooled variance can be used to calculate SE.

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

$t = \frac{(\overline{x_1} - \overline{x_2})}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ with $(n_1 + n_2 - 2)$ degrees of freedom.

Insomnia Treatment

A statistics professor claims that his lectures can cure insomnia. You want to test the claim. You collect 30 patients with sleeping trouble and divide them into 2 groups of 15 each.

The control group were asked to follow their usual routine while the other group was exposed to 1-hour of his lecture on t-distribution shortly after dinner. The time taken to sleep was measured for each group.

Two sample t-test

Time taken to sleep (hr)					
Treated Subjects			Control Subjects		
0.81	0.56	0.46	1.15	1.15	0.92
1.06	0.45	0.43	1.28	0.72	0.67
0.43	0.88	0.37	1.00	0.79	0.76
0.54	0.73	0.73	0.95	0.67	0.82
0.68	0.43	0.93	1.06	1.21	0.82

$n_2 = 15$

$\overline{x_2} = 0.633$

$s_2 = 0.216$

$s_2{}^2 = 0.0467$

$n_1 = 15$

$\overline{x_1} = 0.931$

$s_1 = 0.202$

$s_1{}^2 = 0.0408$

Hypothesis Testing

What is the null hypothesis?

$H_0: \mu_1 - \mu_2 = 0$ (The lecture has no impact)

What is the alternative hypothesis?

$H_1: \mu_1 - \mu_2 \neq 0$

Is it a one-tailed test or a two-tailed test?

Two-tailed

What could be a possible hypothesis for a one-tailed test?

The lecture helps people sleep better.

Two sample t-test

At α = 0.05, determine if there is a significant difference between the two groups.

$$s_p{}^2 = \frac{(n_1-1)s_1{}^2+(n_2-1)s_2{}^2}{(n_1-1)+(n_2-1)}; t = \frac{(\overline{x_1}-\overline{x_2})}{s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \text{ with } (n_1+n_2-2) \text{ df.}$$

$$s_p{}^2 = \frac{(15-1)*0.0408+(15-1)*0.0467}{(15-1)+(15-1)} = 0.04375; s_p = 0.209$$

$$t = \frac{0.931-0.633}{0.209*\sqrt{\frac{1}{15}+\frac{1}{15}}} = 3.91$$

You can find the p-value for this t-score or knowing that the t-score is way more than the critical value for 28 df (~ 2) at this significance level, you see that it is in the critical region in the right tail.

Hypothesis Testing

Will you reject the null hypothesis or fail to do so?

Reject. That means lecture does affect the time-to-sleep.

Does it increase or decrease the time to sleep and by how much?

As the treated patients slept in shorter time (0.633 hr) compared to the control group (0.931 hr), the lecture reduces the time to sleep by 0.298 hr.

R code: t.test(data1, data2, alternative="two.sided")

Confidence Intervals

Margin of Error ME= $t_{n-1,\frac{\alpha}{2}}$ *S.E = $t_{n-1,\frac{\alpha}{2}} * s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$

=2.048*0.0763 = 0.156

$$(\overline{x_1}-\overline{x_2})-ME \leq \mu_1-\mu_2 \leq (\overline{x_1}-\overline{x_2})+ME$$
$$0.298-0.156 \leq \mu_1-\mu_2 \leq 0.298+0.156$$

95% CI: (0.142, 0.454)

Note zero difference is unlikely as at 95% Confidence Level, the difference ranges between 0.142 hr and 0.454 hr, with a point estimate for the difference in sleep time being 0.298 hr.

Two-Sample t-Test for Unpaired Data

Welch's t-test using Welch-Satterthwaite equation for df

$H_0: \mu_1 = \mu_2$; $H_1: \mu_1 \neq \mu_2$; Test statistic, $t = \frac{(\overline{x_1} - \overline{x_2})}{\sqrt{\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}}}$

for **unequal standard deviations** for the two populations.

The degrees of freedom in this case are calculated as:

$$\nu = \frac{\left(\frac{{s_1}^2}{n_1} + \frac{{s_2}^2}{n_2}\right)^2}{\left[\frac{\left(\frac{{s_1}^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{{s_2}^2}{n_2}\right)^2}{n_2 - 1}\right]}$$, rounded off to the nearest integer.

R code: t.test(data1, data2, alternative="two.sided", var.equal=FALSE)

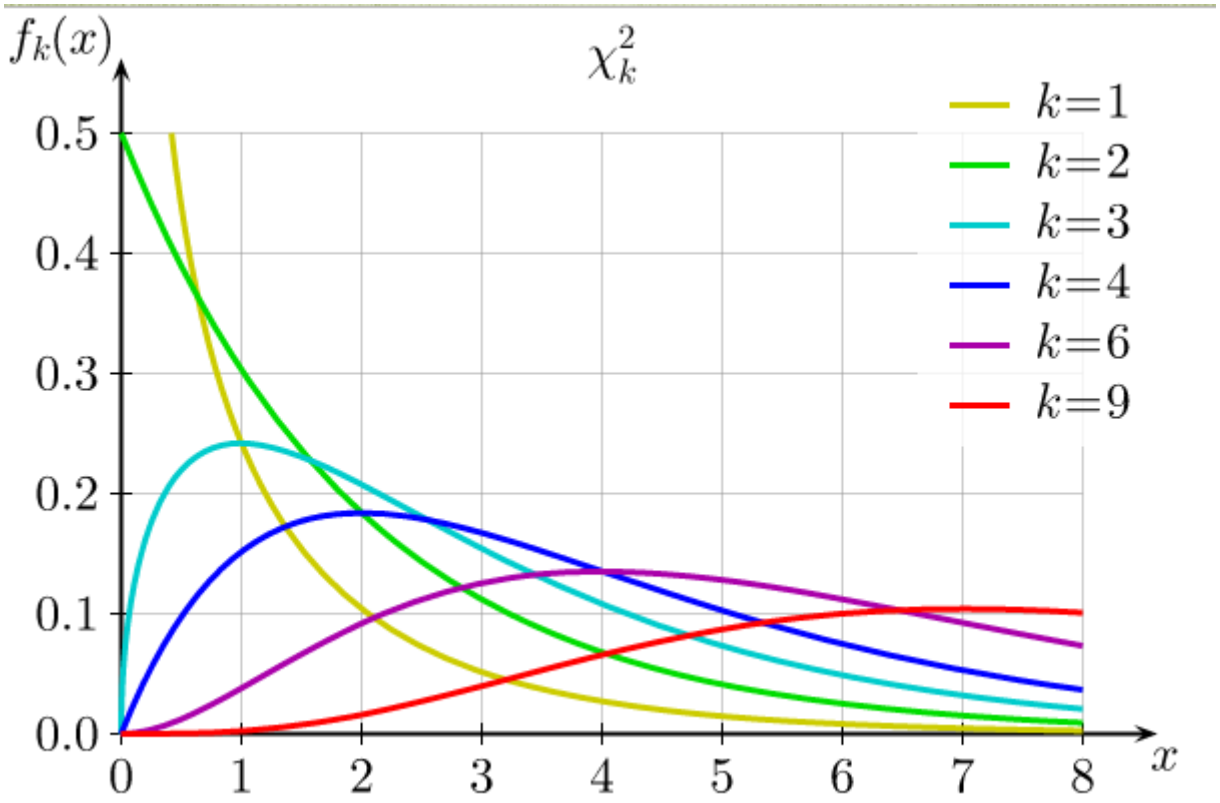
χ^2 DISTRIBUTION

Suppose you modeled a situation using a probability distribution and have an expectation of how things will shape up in the long run.

But what if, the observed data and expected data are not the same? How would you know if the difference is due to normal fluctuations or if your model was incorrect?

Chi Square Distribution

- Let X ~ N(0,1)
- How does a distribution of X^2 look like?
- How does the distribution of sum of squares of two random picks looks like? i.e $X_1^2 + X_2^2$



χ^2 distribution

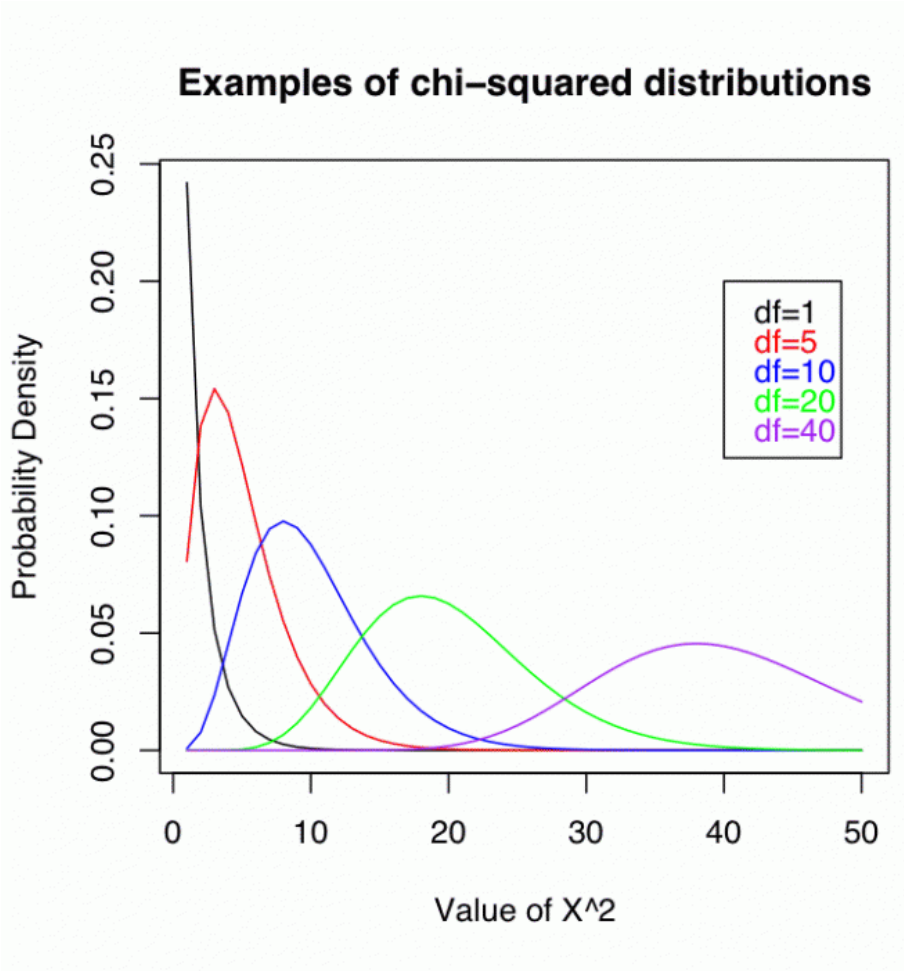
$$Recall\ z = \frac{X - \mu}{\sigma}$$

$$z^2 = \frac{(X - \mu)^2}{\sigma^2}$$

$$z^2 = \chi^2_{(1)}$$

χ^2 distribution is a distribution of the squared deviates.

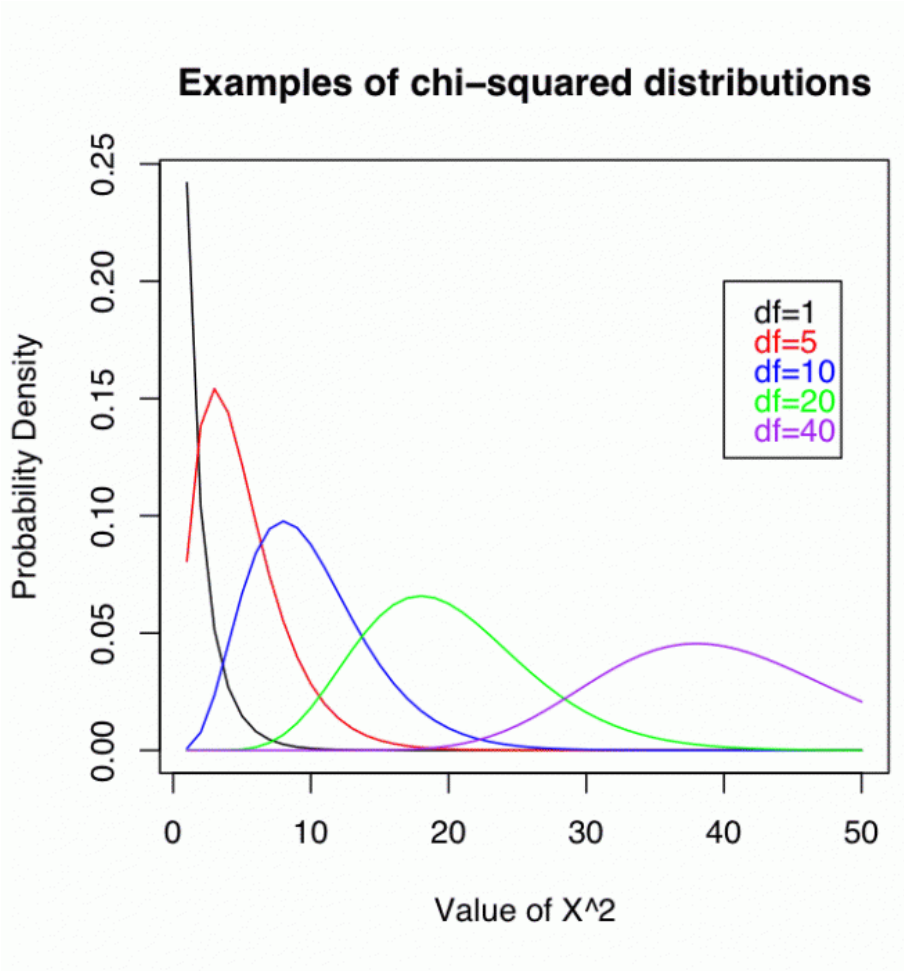
The shape depends on number of squared deviates added together.



χ^2 distribution

$X^2 \sim \chi^2_{(\nu)}$, where ν represents the degrees of freedom.

When ν is greater than 2, the shape of the distribution is skewed positively gradually becoming approximately normal for large ν.



Properties of X^2 random variable

- A X^2 random variable takes values between 0 and ∞.
- Mean of a χ^2 distribution is ν.
- Variance of a χ^2 distribution is 2ν.
- The shape of the distribution is skewed to the right.
- As ν increases, Mean gets larger and the distribution spreads wider.
- As ν increases, distribution tends to normal.

Let us say you are running a casino and the slot machines are causing you headaches. You had designed them with the following expected probability distribution, with X being the net gain from each game played.

x	-2	23	48	73	98
P(X=x)	0.977	0.008	0.008	0.006	0.001

You collected some statistics and found the following frequency of peoples' winnings.

x	-2	23	48	73	98
Frequency	965	10	9	9	7

You want to compare the actual frequency with the expected frequency.

x	-2	23	48	73	98
P(X=x)	0.977	0.008	0.008	0.006	0.001

x	Observed Frequency	Expected Frequency
-2	965	977
23	10	8
48	9	8
73	9	6
98	7	1

Are these differences significant and if they are, is it just pure chance?

χ^2 test to the rescue

χ^2 distribution uses a test statistic to look at the difference between the expected and the actual, and then returns a probability of getting observed frequencies as extreme.

$X^2 = \sum \frac{(O-E)^2}{E}$, where O is the observed frequency and E the expected frequency.

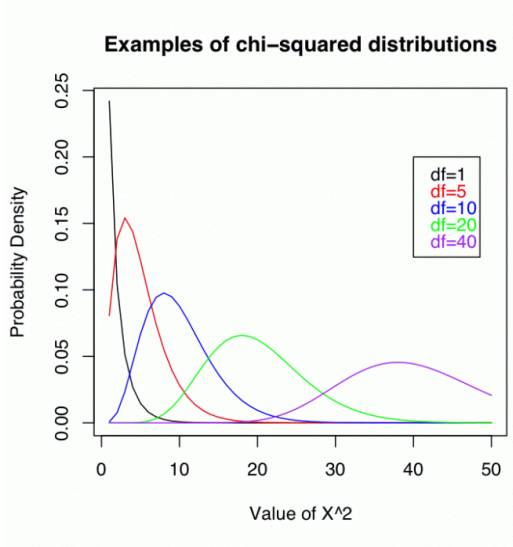
x	Observed Frequency	Expected Frequency
-2	965	977
23	10	8
48	9	8
73	9	6
98	7	1

$X^2 = 38.272$

Is this high?

To find this, we need to look at the χ^2 distribution.

x	Observed Frequency	Expected Frequency
-2	965	977
23	10	8
48	9	8
73	9	6
98	7	1



In the above case, we had 5 frequencies to calculate. However, since the TOTAL expected frequency has to be equal to the TOTAL observed frequency (**RESTRICTION**), calculating 4 would give the 5^{th}. Therefore, there are 5-1=4 degrees of freedom.

ν = (number of classes) – (number of restrictions), or

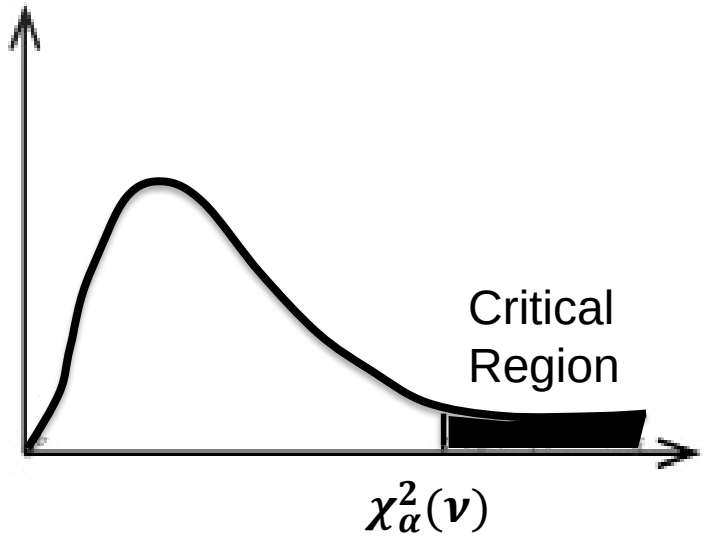
ν = (number of classes) – 1 – (number of parameters being estimated from sample data)

How do we know the Significance of the difference?

One-tailed test using the upper tail of the distribution as the critical region.

A test at significance level α is written as $\chi^2_\alpha(\nu)$. The critical region is to its right.

Higher the value of the test statistic, the bigger the difference between observed and expected frequencies.



Using χ^2 probability tables

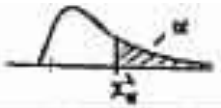
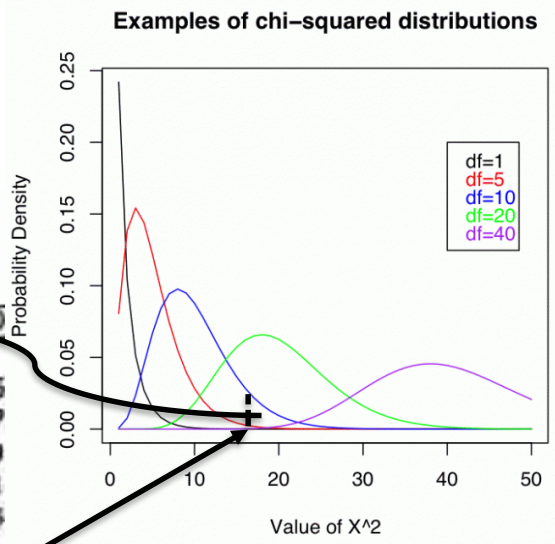


TABLE OF CHI-SQUARE DISTRIBUTION

ν \ α	0.995	0.99	0.98	0.975	0.95	0.90	0.80	0.20	0.10	0.05	0.025	0
1	0.0^4393	0.0^3157	0.0^3628	0.0^3982	0.00393	0.0158	0.0642	1.642	2.706	3.841	5.024	5
2	0.0100	0.0201	0.0404	0.0506	0.103	0.211	0.446	3.219	4.605	5.991	7.378	7
3	0.0717	0.115	0.185	0.216	0.352	0.584	1.005	4.642	6.251	7.815	9.348	9
4	0.207	0.297	0.429	0.484	0.711	1.064	1.649	5.989	7.779	9.488	11.143	11
5	0.412	0.554	0.752	0.831	1.145	1.610	2.343	7.289	9.236	11.070	12.832	13
6	0.676	0.872	1.134	1.237	1.635	2.204	3.070	8.558	10.645	12.592	14.449	15
7	0.989	1.239	1.564	1.690	2.167	2.833	3.822	9.803	12.017	14.067	16.013	16.622	18.475	20.278	24.322
8	1.344	1.646	2.032	2.180	2.733	3.490	4.594	11.030	13.362	15.507	17.535	18.168	20.090	21.955	26.125
9	1.735	2.088	2.532	2.700	3.325	4.168	5.380	12.242	14.684	16.919	19.023	19.679	21.666	23.589	27.877
10	2.156	2.558	3.059	3.247	3.940	4.865	6.179	13.442	15.987	18.307	20.483	21.161	23.209	25.188	29.588
11	2.603	3.053	3.609	3.816	4.575	5.578	6.989	14.631	17.275	19.675	21.920	22.618	24.725	26.757	31.264
12	3.074	3.571	4.178	4.404	5.226	6.304	7.807	15.812	18.549	21.026	23.337	24.054	26.217	28.300	32.909
13	3.565	4.107	4.765	5.009	5.892	7.042	8.634	16.985	19.812	22.362	24.736	25.472	27.688	29.819	34.528
14	4.075	4.660	5.368	5.629	6.571	7.790	9.467	18.151	21.064	23.685	26.119	26.873	29.141	31.319	36.123
15	4.601	5.229	5.985	6.262	7.261	8.547	10.307	19.311	22.307	24.996	27.488	28.259	30.578	32.801	37.697
16	5.142	5.812	6.614	6.908	7.962	9.312	11.152	20.465	23.542	26.296	28.845	29.633	32.000	34.267	39.252
17	5.697	6.408	7.255	7.564	8.672	10.085	12.002	21.615	24.769	27.587	30.191	30.995	33.409	35.718	40.790
18	6.265	7.015	7.906	8.231	9.390	10.865	12.857	22.760	25.989	28.869	31.526	32.346	34.805	37.156	42.312
19	6.844	7.633	8.567	8.907	10.117	11.651	13.716	23.900	27.204	30.144	32.852	33.687	36.191	38.582	43.820
20	7.434	8.260	9.237	9.591	10.851	12.443	14.578	25.038	28.412	31.410	34.170	35.020	37.566	39.997	45.315
21	8.034	8.897	9.915	10.283	11.591	13.240	15.445	26.171	29.615	32.671	35.479	36.343	38.932	41.401	46.797
22	8.643	9.542	10.600	10.982	12.338	14.041	16.314	27.301	30.813	33.924	36.781	37.659	40.289	42.796	48.268
23	9.260	10.196	11.293	11.688	13.091	14.848	17.187	28.429	32.007	35.172	38.076	38.968	41.638	44.181	49.728
24	9.886	10.856	11.992	12.401	13.848	15.659	18.062	29.553	33.196	36.415	39.364	40.270	42.980	45.558	51.179
25	10.520	11.524	12.697	13.120	14.611	16.473	18.940	30.675	34.382	37.652	40.646	41.566	44.314	46.928	52.620



CSE 7315c

Uses of χ^2 distribution

- To test **goodness of fit**.
- To test **independence** of two variables.

Steps to test Goodness-of-fit

You want to see if there is sufficient evidence at the 5% significance level to say the slot machines have been rigged.

What are the null and alternate hypotheses?

H_0: The slot machine winnings per game follow the described probability distribution, i.e., they are not rigged.

H_1: The slot machine winnings per game do not follow this distribution.

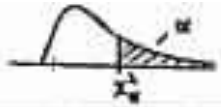
What are the expected frequencies and degrees of freedom?

x	Observed Frequency	Expected Frequency
-2	965	977
23	10	8
48	9	8
73	9	6
98	7	1

$\nu = 4$

What is the critical region?

TABLE OF CHI-SQUARE DISTRIBUTION



ν \ α	0.995	0.99	0.98	0.975	0.95	0.90	0.80	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.001
1	0.0^4393	0.0^3157	0.0^3628	0.0^3982	0.00393	0.0158	0.0642	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.827
2	0.0100	0.0201	0.0404	0.0506	0.103	0.211	0.446	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.815
3	0.0717	0.115	0.185	0.216	0.352	0.584	1.005	4.642	6.251	7.815	9.348	9.837	11.345	12.838	16.268
4	0.207	0.297	0.429	0.484	0.711	1.064	1.649	5.989	7.779	9.488	11.143	11.668	13.277	14.860	18.465
5	0.412	0.554	0.752	0.831	1.145	1.610	2.343	7.289	9.236	11.070	12.832	13.388	15.086	16.750	20.517
6	0.676	0.872	1.134	1.237	1.635	2.204	3.070	8.558	10.645	12.592	14.449	15.033	16.812	18.548	22.457
7	0.989	1.239	1.564	1.690	2.167	2.833	3.822	9.803	12.017	14.067	16.013	16.622	18.475	20.278	24.322
8	1.344	1.646	2.032	2.180	2.733	3.490	4.594	11.030	13.362	15.507	17.535	18.168	20.090	21.955	26.125
9	1.735	2.088	2.532	2.700	3.325	4.168	5.380	12.242	14.684	16.919	19.023	19.679	21.666	23.589	27.877
10	2.156	2.558	3.059	3.247	3.940	4.865	6.179	13.442	15.987	18.307	20.483	21.161	23.209	25.188	29.588
11	2.603	3.053	3.609	3.816	4.575	5.578	6.989	14.631	17.275	19.675	21.920	22.618	24.725	26.757	31.264
12	3.074	3.571	4.178	4.404	5.226	6.304	7.807	15.812	18.549						

$\chi^2_{5\%}(4) = 9.488$. This means the critical region is $X^2 > 9.488$.

Is the test statistic inside or outside the critical region?

Since $X^2 = 38.27$ and the critical region is $X^2 > 9.488$, this means X^2 is inside the critical region.

Will you accept or reject the null hypothesis?

Reject. There is sufficient evidence to reject the hypothesis that the slot machine winnings follow the described probability distribution.

This sort of hypothesis test is called a **goodness of fit** test. This test is used whenever you have a set of values that should fit a distribution, and you want to test whether the data actually does.

χ^2 goodness of fit works for any probability distribution

Distribution	Condition	ν
Binomial	You know *p* (probability of success or the proportion of successes in a population)	$\nu = n - 1$
	You don't know *p* and have to estimate it from observed frequencies	$\nu = n - 2$
Poisson	You know λ	$\nu = n - 1$
	You don't know λ, and have to estimate it from observed frequencies	$\nu = n - 2$
Normal	You know μ and σ^2	$\nu = n - 1$
	You don't know μ and σ^2, and have to estimate them from observed frequencies	$\nu = n - 3$

The 108 Medical Emergency Service received calls during 150 5-minute intervals as follows. Is the distribution Poisson at $\alpha=0.01$?

# of calls per 5-min interval	Frequency
0	18
1	28
2	47
3	21
4	16
5	11
6 or more	9

Step 1: Decide H_0 and H_1

H_0: The frequency distribution is Poisson.

H_1: The frequency distribution is not Poisson.

Step 2: Find expected frequencies and degrees of freedom

# of calls per 5-min interval	Observed Frequency	Total Calls = # of calls/interval X Frequency
0	18	0
1	28	28
2	47	94
3	21	63
4	16	64
5	11	55
6 or more	9	54
TOTAL	**150**	**358**

Expected frequency = Probability * Total

$$P(X = r) = \frac{e^{-\lambda}\lambda^{r}}{r!}$$

$$\lambda = Avg\ no.\ of\ calls\ per\ 5\ \text{min}\ interval = \frac{Total\ \#\ of\ calls}{Total\ \#\ of\ intervals}$$

$$\lambda = \frac{358}{150} = 2.39$$

Step 2: Find expected frequencies and degrees of freedom

Expected frequencies are obtained by multiplying expected probabilities by the total frequency. $P(X = r) = \frac{e^{-\lambda}\lambda^{r}}{r!}$, where $\lambda = 2.39$

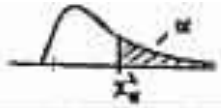
ν = 7-2 = 5

# of calls per 5-min interval	Expected Probability	Expected Frequency	Observed Frequency
0	0.0916	13.74	18
1	0.2190	32.85	28
2	0.2617	39.25	47
3	0.2085	31.27	21
4	0.1246	18.69	16
5	0.0595	8.93	11
6 or more	0.0526	3.56	9
TOTAL		**150.00**	**150**

How do you find this value?

Step 3: Determine the critical region

TABLE OF CHI-SQUARE DISTRIBUTION



ν \ α	0.995	0.99	0.98	0.975	0.95	0.90	0.80	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.001
1	0.0^4393	0.0^3157	0.0^3628	0.0^3982	0.00393	0.0158	0.0642	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.827
2	0.0100	0.0201	0.0404	0.0506	0.103	0.211	0.446	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.815
3	0.0717	0.115	0.185	0.216	0.352	0.584	1.005	4.642	6.251	7.815	9.348	9.837	11.345	12.838	16.268
4	0.207	0.297	0.429	0.484	0.711	1.064	1.649	5.989	7.779	9.488	11.143	11.668	13.277	14.860	18.465
5	0.412	0.554	0.752	0.831	1.145	1.610	2.343	7.289	9.236	11.070	12.832	13.388	15.086	16.750	20.517
6	0.676	0.872	1.134	1.237	1.635	2.204	3.070	8.558	10.645	12.592	14.449	15.033	16.812	18.548	22.457
7	0.989	1.239	1.564	1.690	2.167	2.833	3.822	9.803	12.017	14.067	16.013	16.622	18.475	20.278	24.322
8	1.344	1.646	2.032	2.180	2.733	3.490	4.594	11.030	13.362	15.507	17.535	18.168	20.090	21.955	26.125
9	1.735	2.088	2.532	2.700	3.325	4.168	5.380	12.242	14.684	16.919	19.023	19.679	21.666	23.589	27.877
10	2.156	2.558	3.059	3.247	3.940	4.865	6.179	13.442	15.987	18.307	20.483	21.161	23.209	25.188	29.588
11	2.603	3.053	3.609	3.816	4.575	5.578	6.989	14.631	17.275	19.675	21.920	22.618	24.725	26.757	31.264
12	3.074	3.571	4.178	4.404	5.226	6.304	7.807	15.812	18.549	21.026	23.337	24.054	26.217	28.300	32.909

$\chi^2_{1\%}(5) = 15.086$. This means the critical region is $X^2 > 15.086$.

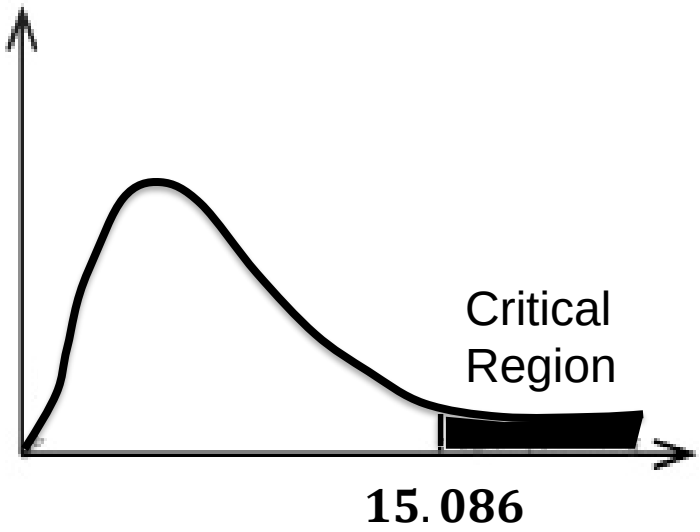
Step 4: Calculate the test statistic X^2

# of calls per 5-min interval	Observed Frequency	Expected Frequency	$\frac{(O-E)^2}{E}$
0	18	13.74	1.32
1	28	32.85	0.72
2	47	39.25	1.53
3	21	31.27	3.37
4	16	18.69	0.39
5	11	8.93	0.48
6 or more	9	3.56	2.66
TOTAL	**150**	**150**	**10.46**

$X^2 = 10.46$

Step 5: See whether the test statistic is in the critical region

X^2 = 10.46, which is less than the critical value of 15.086. It is NOT in the critical region.



Step 6: Make your decision

There is not enough evidence to reject the null hypothesis that the distribution is Poisson.

Business Implication: Now that 108 services management knows that the distribution is Poisson, it can plan the staffing of the call centre more efficiently.

Another Example: χ^2 Goodness-of-fit Test

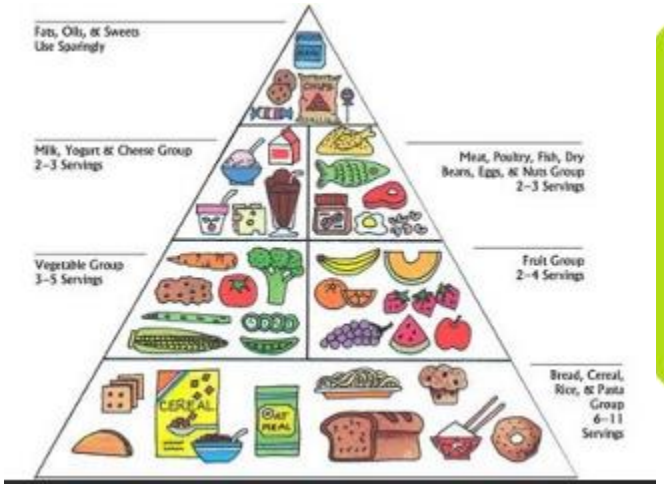
USDA intends to update the nutritional guideline for ideal diet. 90 adults are shown 3 different diagrams (D1, D2, D3) explaining the proper proportion of food groups, and asked to identify the easiest to follow. The group's picks are given below.

Are there any significant differences in the acceptability of the diagrams at 5% significance level?

D1: 23

D2: 39

D3: 28



Step 1: Decide H_0 and H_1

H_0: All Diagrams are likely to be equally selected.

H_1: There is significant difference in the likelihood of different diagrams being selected.

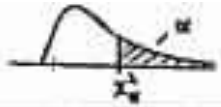
Step 2: Find Expected Frequencies and Degrees of Freedom

Leaflet	Observed Frequency	Expected Frequency	$\frac{(O-E)^2}{E}$
D1	23	30	1.63
D2	39	30	2.70
D3	28	30	0.13
TOTAL	**90**	**90**	**4.46**

$$\nu = 2$$

Step 3: Determine the critical region

TABLE OF CHI-SQUARE DISTRIBUTION



ν \ α	0.995	0.99	0.98	0.975	0.95	0.90	0.80	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.001
1	0.0^4393	0.0^3157	0.0^3628	0.0^3982	0.00393	0.0158	0.0642	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.827
2	0.0100	0.0201	0.0404	0.0506	0.103	0.211	0.446	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.815
3	0.0717	0.115	0.185	0.216	0.352	0.584	1.005	4.642	6.251	7.815	9.348	9.837	11.345	12.838	16.268
4	0.207	0.297	0.429	0.484	0.711	1.064	1.649	5.989	7.779	9.488	11.143	11.668	13.277	14.860	18.465
5	0.412	0.554	0.752	0.831	1.145	1.610	2.343	7.289	9.236	11.070	12.832	13.388	15.086	16.750	20.517
6	0.676	0.872	1.134	1.237	1.635	2.204	3.070	8.558	10.645	12.592	14.449	15.033	16.812	18.548	22.457
7	0.989	1.239	1.564	1.690	2.167	2.833	3.822	9.803	12.017	14.067	16.013	16.622	18.475	20.278	24.322
8	1.344	1.646	2.032	2.180	2.733	3.490	4.594	11.030	13.362	15.507	17.535	18.168	20.090	21.955	26.125
9	1.735	2.088	2.532	2.700	3.325	4.168	5.380	12.242	14.684	16.919	19.023	19.679	21.666	23.589	27.877
10	2.156	2.558	3.059	3.247	3.940	4.865	6.179	13.442	15.987	18.307	20.483	21.161	23.209	25.188	29.588
11	2.603	3.053	3.609	3.816	4.575	5.578	6.989	14.631	17.275	19.675	21.920	22.618	24.725	26.757	31.264
12	3.074	3.571	4.178	4.404	5.226	6.304	7.807	15.812	18.549						

$\chi^2_{5\%}(2) = 5.991$. This means the critical region is $X^2 > 5.991$.

The observed value is 4.46. This is outside the critical region and so null hypothesis cannot be rejected.

What is the business implication/decision?

Leaflet	Observed Frequency	Expected Frequency
D1	23	30
D2	39	30
D3	28	30

While the results are not statistically significant, with a large enough sample of 90 Adults, it can be clearly seen that D2 shows a lot more preference over the other two.

Statistical insignificance doesn't mean there is no difference. The experiment simply may not have the power to detect this. Further thought is warranted when such frustrating results are obtained.

χ^2 independence test

Your casino is facing another issue. You think you are losing more money from one of the croupiers on the blackjack tables. You want to test if the outcome of the game is dependent on which croupier is leading the game.



Possible Outcomes

	Croupier A	Croupier B	Croupier C
Win	43	49	22
Draw	8	2	5
Lose	47	44	30

Observed Results

χ^2 independence test

The process is the same as before. The null hypothesis assumes that choice of croupier is independent of the outcome, and is rejected if there is sufficient evidence against it.

However, a **contingency table** has to be drawn to find the expected frequencies using probability.

χ^2 independence test

	Croupier A	Croupier B	Croupier C	Total
Win	43	49	22	114
Draw	8	2	5	15
Lose	47	44	30	121
Total	98	95	57	250

$$P(Win) = \frac{Total\ Wins}{Grand\ Total} = \frac{114}{250}$$

$$P(A) = \frac{Total\ A}{Grand\ Total} = \frac{98}{250}$$

If croupier and the outcome are independent,

$$P(Win\ and\ A) = \frac{Total\ Wins}{Grand\ Total} X \frac{Total\ A}{Grand\ Total}$$

χ^2 independence test

	Croupier A	Croupier B	Croupier C	Total
Win	43	49	22	114
Draw	8	2	5	15
Lose	47	44	30	121
Total	98	95	57	250

$$Expected\ Frequency\ of\ Win\ and\ A$$

$$= Grand\ Total\ X\ \frac{Total\ Wins}{Grand\ Total}\ X\ \frac{Total\ A}{Grand\ Total} = \frac{Total\ Wins\ X\ Total\ A}{Grand\ Total}$$

$$= \frac{Row\ Total\ X\ Column\ Total}{Grand\ Total}$$

χ^2 independence test – Finding expected frequencies

	Croupier A	Croupier B	Croupier C	Total
Win	43	49	22	114
Draw	8	2	5	15
Lose	47	44	30	121
Total	98	95	57	250

	Croupier A	Croupier B	Croupier C
Win	(114*98)/250	(114*95)/250	(114*57)/250
Draw	(15*98)/250	(15*95)/250	(15*57)/250
Lose	(121*98)/250	(121*95)/250	(121*57)/250

χ^2 independence test – Calculating X^2

	Observed	Expected	$\frac{(O-E)^2}{E}$
A	43	44.688	0.0638
	8	5.88	0.7644
	47	47.432	0.0039
B	49	43.32	0.7447
	2	5.7	2.4018
	44	45.98	0.0853
C	22	25.992	0.6131
	5	3.42	0.7299
	30	27.588	0.2109
	$\sum O = 250$	$\sum E = 250$	$\sum \frac{(O-E)^2}{E} = 5.618$

χ^2 independence test – Calculating ν

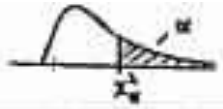
	Croupier A	Croupier B	Croupier C
Win			
Draw			
Lose			

We calculated 9 but really need to calculate 4 and figure out the rest using the total frequency of each row and column. In general, the degrees of freedom will be *(m-1)(n-1)* where *m* is the number of columns and *n* the number of rows.

χ^2 independence test – Determine critical region

Let us say we need 1% significance level to see if the outcome is independent of the croupier.

TABLE OF CHI-SQUARE DISTRIBUTION

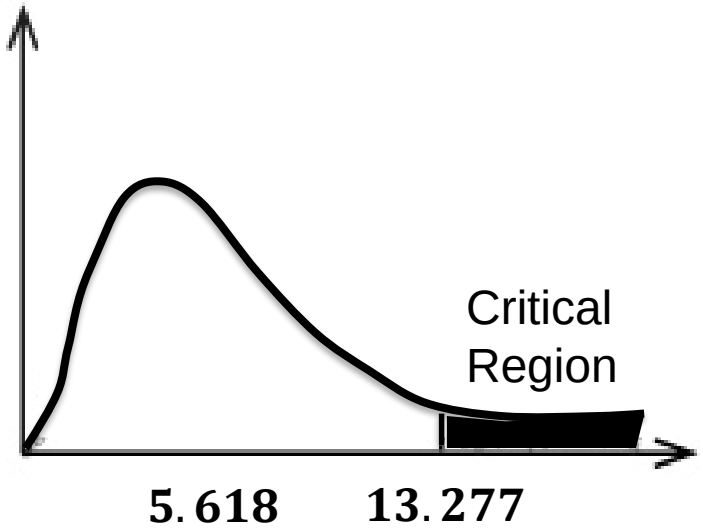


ν \ α	0.995	0.99	0.98	0.975	0.95	0.90	0.80	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.001
1	0.0^4393	0.0^3157	0.0^3628	0.0^3982	0.00393	0.0158	0.0642	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.827
2	0.0100	0.0201	0.0404	0.0506	0.103	0.211	0.446	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.815
3	0.0717	0.115	0.185	0.216	0.352	0.584	1.005	4.642	6.251	7.815	9.348	9.837	11.345	12.838	16.268
4	0.207	0.297	0.429	0.484	0.711	1.064	1.649	5.989	7.779	9.488	11.143	11.668	13.277	14.860	18.465
5	0.412	0.554	0.752	0.831	1.145	1.610	2.343	7.289	9.236	11.070	12.832	13.388	15.086	16.750	20.517
6	0.676	0.872	1.134	1.237	1.635	2.204	3.070	8.558	10.645	12.592	14.449	15.033	16.812	18.548	22.457
7	0.989	1.239	1.564	1.690	2.167	2.833	3.822	9.803	12.017	14.067	16.013	16.622	18.475	20.278	24.322
8	1.344	1.646	2.032	2.180	2.733	3.490	4.594	11.030	13.362	15.507	17.535	18.168	20.090	21.955	26.125
9	1.735	2.088	2.532	2.700	3.325	4.168	5.380	12.242	14.684	16.919	19.023	19.679	21.666	23.589	27.877
10	2.156	2.558	3.059	3.247	3.940	4.865	6.179	13.442	15.987	18.307	20.483	21.161	23.209	25.188	29.588
11	2.603	3.053	3.609	3.816	4.575	5.578	6.989	14.631	17.275	19.675	21.920	22.618	24.725	26.757	31.264
12	3.024	3.571	4.178	4.404	5.226	6.304	7.807	15.812	18.549	21.026	23.337	24.054	26.217	28.300	32.909

$\chi^2_{1\%}(4) = 13.277$. This means the critical region is $X^2 > 13.277$.

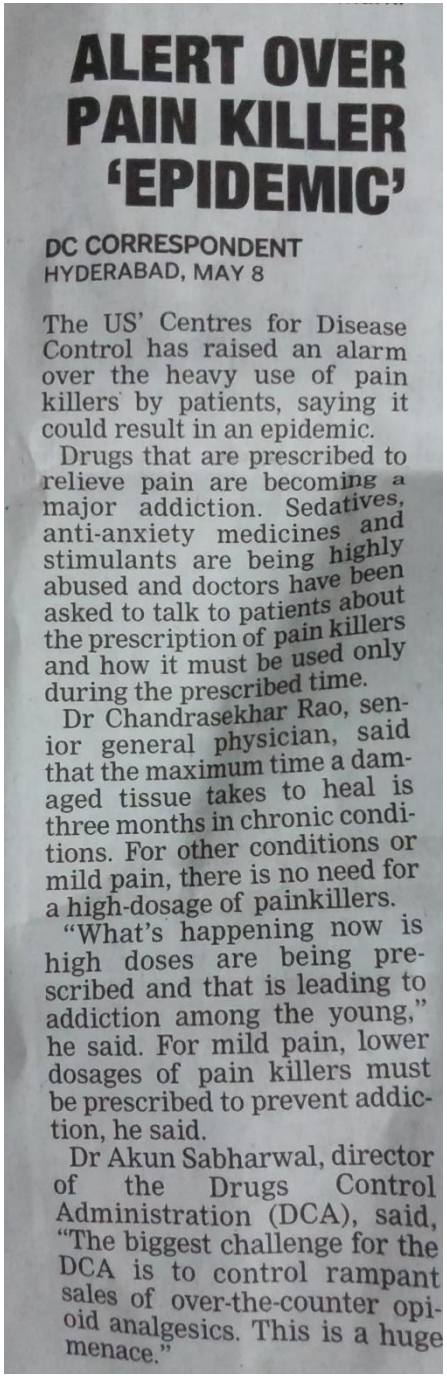
χ^2 independence test – Decision

Since calculated $X^2 = 5.618$, it is outside the critical region, and hence we accept the null hypothesis.



χ^2 independence test

There is widespread abuse of prescription drugs for a number of reasons like getting high, reducing appetite, relieving tension, feeding an addiction, etc.



ALERT OVER PAIN KILLER 'EPIDEMIC'

DC CORRESPONDENT
HYDERABAD, MAY 8

The US' Centres for Disease Control has raised an alarm over the heavy use of pain killers by patients, saying it could result in an epidemic.

Drugs that are prescribed to relieve pain are becoming a major addiction. Sedatives, anti-anxiety medicines and stimulants are being highly abused and doctors have been asked to talk to patients about the prescription of pain killers and how it must be used only during the prescribed time.

Dr Chandrasekhar Rao, senior general physician, said that the maximum time a damaged tissue takes to heal is three months in chronic conditions. For other conditions or mild pain, there is no need for a high-dosage of painkillers.

"What's happening now is high doses are being prescribed and that is leading to addiction among the young," he said. For mild pain, lower dosages of pain killers must be prescribed to prevent addiction, he said.

Dr Akun Sabharwal, director of the Drugs Control Administration (DCA), said, "The biggest challenge for the DCA is to control rampant sales of over-the-counter opioid analgesics. This is a huge menace."

χ^2 independence test

The National Council on Alcoholism and Drug Dependence wants to understand if there is dependence of the type of prescription drug abuse on the age of the patient. A random poll of 309 patients is taken as shown below. At α = 0.01, are the two variables independent?

	Pain relievers	Tranquilizers /Sedatives	Stimulants	TOTAL
21-34	26	95	18	139
35-55	41	40	20	101
>55	24	13	32	69
TOTAL	91	148	70	309

Step 1: Decide H_0 and H_1

H_0: Type of prescription drug abuse is independent of age.

H_1: Type of prescription drug abuse is not independent of age.

Step 2: Find expected frequencies and degrees of freedom

OBSERVED

	Pain relievers	Tranquilizers/ Sedatives	Stimulants	TOTAL
21-34	26	95	18	139
35-55	41	40	20	101
>55	24	13	32	69
TOTAL	91	148	70	309

EXPECTED

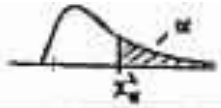
	Pain relievers	Tranquilizers/ Sedatives	Stimulants	TOTAL
21-34	40.94	66.58	31.49	139
35-55	29.74	48.38	22.88	101
>55	20.32	33.05	15.63	69
TOTAL	91	148	70	309

$\nu = 4$

CSE 7315c

Step 3: Determine the critical region

TABLE OF CHI-SQUARE DISTRIBUTION



ν \ α	0.995	0.99	0.98	0.975	0.95	0.90	0.80	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.001
1	0.0^4393	0.0^3157	0.0^3628	0.0^3982	0.00393	0.0158	0.0642	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.827
2	0.0100	0.0201	0.0404	0.0506	0.103	0.211	0.446	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.815
3	0.0717	0.115	0.185	0.216	0.352	0.584	1.005	4.642	6.251	7.815	9.348	9.837	11.345	12.838	16.268
4	0.207	0.297	0.429	0.484	0.711	1.064	1.649	5.989	7.779	9.488	11.143	11.668	13.277	14.860	18.465
5	0.412	0.554	0.752	0.831	1.145	1.610	2.343	7.289	9.236	11.070	12.832	13.388	15.086	16.750	20.517
6	0.676	0.872	1.134	1.237	1.635	2.204	3.070	8.558	10.645	12.592	14.449	15.033	16.812	18.548	22.457
7	0.989	1.239	1.564	1.690	2.167	2.833	3.822	9.803	12.017	14.067	16.013	16.622	18.475	20.278	24.322
8	1.344	1.646	2.032	2.180	2.733	3.490	4.594	11.030	13.362	15.507	17.535	18.168	20.090	21.955	26.125
9	1.735	2.088	2.532	2.700	3.325	4.168	5.380	12.242	14.684	16.919	19.023	19.679	21.666	23.589	27.877
10	2.156	2.558	3.059	3.247	3.940	4.865	6.179	13.442	15.987	18.307	20.483	21.161	23.209	25.188	29.588
11	2.603	3.053	3.609	3.816	4.575	5.578	6.989	14.631	17.275	19.675	21.920	22.618	24.725	26.757	31.264
12	3.074	3.571	4.178	4.404	5.226	6.304	7.807	15.812	18.549	21.026	23.337	24.054	26.217	28.300	32.909

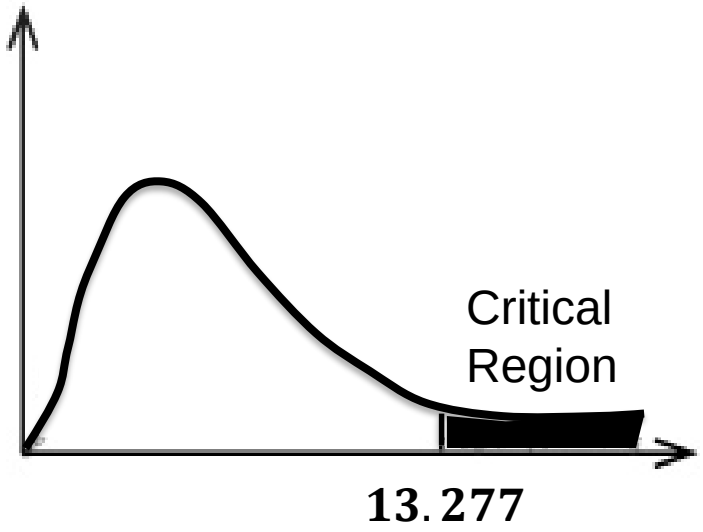
$\chi^2_{1\%}(4) = 13.277$. This means the critical region is $X^2 > 13.277$.

Step 4: Calculate the test statistic X^2

	Observed	Expected	$\frac{(O-E)^2}{E}$
Pain relievers	26	40.94	
	41	29.74	
	24	20.32	
Tranquilizers/ Sedatives	95	66.58	
	40	48.38	
	13	33.05	
Stimulants	18	31.49	
	20	22.88	
	32	15.63	
	$\sum O = 309$	$\sum E = 309$	$\sum \frac{(O-E)^2}{E} = 59.41$

Step 5: See whether the test statistic is in the critical region

$X^2 = 59.41$, which is greater than the critical value of 13.277. It is in the critical region.



Step 6: Make your decision

There is enough evidence to reject the null hypothesis that the type of prescription drug abuse and age are independent.

Testing Hypotheses about a Variance

Sample estimate of population variance is given by

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

Multiplying the variance estimate by *n-1* gives the sum of squares. Dividing by population variance gives a random variable distributed as chi-squared with *n-1* degrees of freedom.

$$\therefore \chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

Recall $\chi^2 = \frac{\sum(x_i - \bar{x})^2}{\sigma^2}$

Testing Hypotheses about a Variance

A manufacturing company produces bearings of 2.65 cm in diameter. A major customer requires that the variance in diameter be no more than 0.001 cm^2. The manufacturer tests 20 bearings using a precise instrument and gets the below values. Assuming the diameters are normally distributed, can the population of these bearings be rejected due to high variance at 1% significance level?

Data: 2.69, 2.66, 2.64, 2.59, 2.62, 2.63, 2.69, 2.66, 2.63, 2.65, 2.57, 2.63, 2.70, 2.71, 2.64, 2.65, 2.59, 2.66, 2.62, 2.57

Testing Hypotheses about a Variance

What are null and alternate hypotheses?

H_0: $\sigma^2 \leq 0.001$; H_1: $\sigma^2 > 0.001$

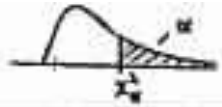
How many degrees of freedom?

Since n=20, df=19.

Testing Hypotheses about a Variance

What is the critical region?

TABLE OF CHI-SQUARE DISTRIBUTION



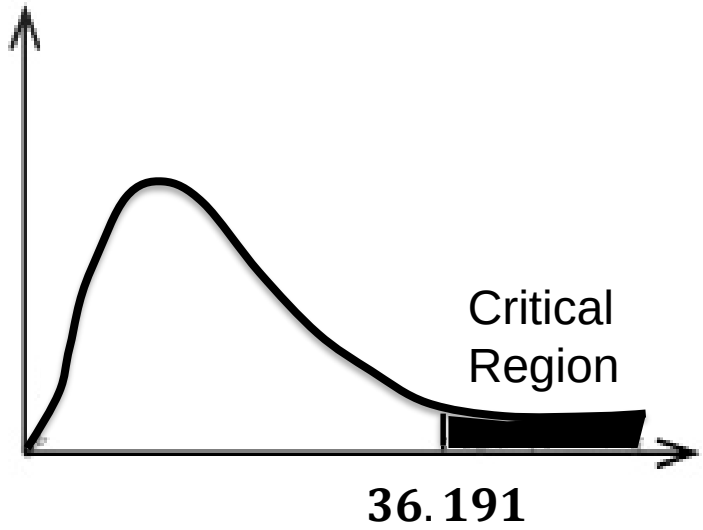
ν \ α	0.995	0.99	0.98	0.975	0.95	0.90	0.80	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.001
1	0.0^4393	0.0^3157	0.0^3628	0.0^3982	0.00393	0.0158	0.0642	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.827
2	0.0100	0.0201	0.0404	0.0506	0.103	0.211	0.446	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.815
3	0.0717	0.115	0.185	0.216	0.352	0.584	1.005	4.642	6.251	7.815	9.348	9.837	11.345	12.838	16.268
4	0.207	0.297	0.429	0.484	0.711	1.064	1.649	5.989	7.779	9.488	11.143	11.668	13.277	14.860	18.465
5	0.412	0.554	0.752	0.831	1.145	1.610	2.343	7.289	9.236	11.070	12.832	13.388	15.086	16.750	20.517
6	0.676	0.872	1.134	1.237	1.635	2.204	3.070	8.558	10.645	12.592	14.449	15.033	16.812	18.548	22.457
7	0.989	1.239	1.564	1.690	2.167	2.833	3.822	9.803	12.017	14.067	16.013	16.622	18.475	20.278	24.322
8	1.344	1.646	2.032	2.180	2.733	3.490	4.594	11.030	13.362	15.507	17.535	18.168	20.090	21.955	26.125
9	1.735	2.088	2.532	2.700	3.325	4.168	5.380	12.242	14.684	16.919	19.023	19.679	21.666	23.589	27.877
10	2.156	2.558	3.059	3.247	3.940	4.865	6.179	13.442	15.987	18.307	20.483	21.161	23.209	25.188	29.588
11	2.603	3.053	3.609	3.816	4.575	5.578	6.989	14.631	17.275	19.675	21.920	22.618	24.725	26.757	31.264
12	3.074	3.571	4.178	4.404	5.226	6.304	7.807	15.812	18.549	21.026	23.337	24.054	26.217	28.300	32.909
13	3.565	4.107	4.765	5.009	5.892	7.042	8.634	16.985	19.812	22.362	24.736	25.472	27.688	29.819	34.528
14	4.075	4.660	5.368	5.629	6.571	7.790	9.467	18.151	21.064	23.685	26.119	26.873	29.141	31.319	36.123
15	4.601	5.229	5.985	6.262	7.261	8.547	10.307	19.311	22.307	24.996	27.488	28.259	30.578	32.801	37.697
16	5.142	5.812	6.614	6.908	7.962	9.312	11.152	20.465	23.542	26.296	28.845	29.633	32.000	34.267	39.252
17	5.697	6.408	7.255	7.564	8.672	10.085	12.002	21.615	24.769	27.587	30.191	30.995	33.409	35.718	40.790
18	6.265	7.015	7.906	8.231	9.390	10.865	12.857	22.760	25.989	28.869	31.526	32.346	34.805	37.156	42.312
19	6.844	7.633	8.567	8.907	10.117	11.651	13.716	23.900	27.204	30.144	32.852	33.687	36.191	38.582	43.820
20	7.434	8.260	9.237	9.591	10.851	12.443	14.578	25.038	28.412	31.410	34.170	35.020	37.566	39.997	45.315

$$\chi^2_{0.01,19} = 36.191$$

Testing Hypotheses about a Variance

What is the observed χ^2 value?

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{19 * 0.001621}{0.001} = 30.8$$



Is it in critical region? $\chi^2_{0.01,19} = 36.191$

No.

Will you reject or fail to reject the null hypothesis?

Fail to reject.

What is the business decision?

The population variance is within specification limits required by the customer and hence the bearings can be shipped.

F DISTRIBUTION

F distribution

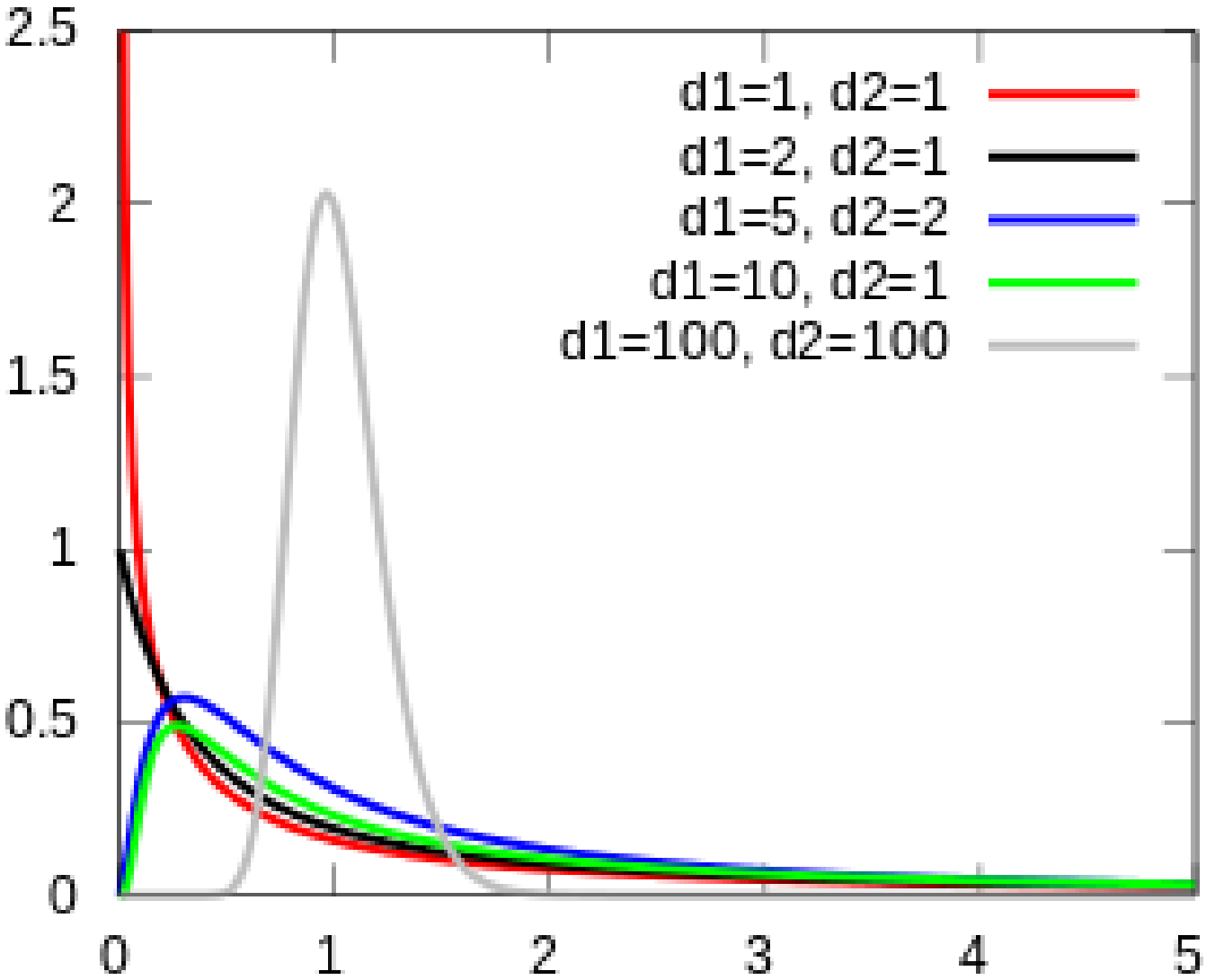
- χ^2 was useful in testing hypotheses about a single population variance.
- Sometimes we want to test hypotheses about difference in variances of two populations:
  - Is the variance of 2 stocks the same?
  - Do parts manufactured in 2 shifts or on 2 different machines or in 2 batches have the same variance or not?
  - Is the powder mix for tablet granulations homogeneous?
  - Is there variability in assayed drug blood levels in a bioavailability study?
  - Is there variability in the clinical response to drug therapy of two samples?

F distribution

- Ratio of 2 variance estimates: $F = \frac{s_1^2}{s_2^2} = \frac{est.\sigma_1^2}{est.\sigma_2^2}$
- Ideally, this ratio should be about 1 if 2 samples come from the same population or from 2 populations with same variance, but sampling errors cause variation.
- *Recall* $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$. So, F is also a ratio of 2 chi-squares, each divided by its degrees of freedom, i.e.,

$$F = \frac{\frac{\chi^2_{\nu_1}}{\nu_1}}{\frac{\chi^2_{\nu_2}}{\nu_2}}$$

F distribution



Hypothesis test for 2 sample variances

A machine produces metal sheets with 22mm thickness. There is variability in thickness due to machines, operators, manufacturing environment, raw material, etc. The company wants to know the consistency of two machines and randomly samples 10 sheets from machine 1 and 12 sheets from machine 2. Thickness measurements are taken. Assume sheet thickness is normally distributed in the population.

The company wants to know if the variance from each sample comes from the same population variance (population variances are equal) or from different population variances (population variances are unequal).

How do you test this?

Hypothesis test for 2 sample variances

Data

Machine 1		Machine 2	
22.3	21.9	22.0	21.7
21.8	22.4	22.1	21.9
22.3	22.5	21.8	22.0
21.6	22.2	21.9	22.1
21.8	21.6	22.2	21.9
		22.0	22.1
$s_1^2 = 0.11378$	$n = 10$	$s_2^2 = 0.02023$	$n = 12$

Ratio of sample variances, $F = \frac{s_1^2}{s_2^2} = \frac{0.11378}{0.02023} = 5.62$

Hypothesis test for 2 sample variances

What are null and alternate hypotheses?

$H_0: \sigma_1^2 = \sigma_2^2;\ H_1: \sigma_1^2 \neq \sigma_2^2$

Is it a one-tailed test or a two-tailed test?

Two-tailed.

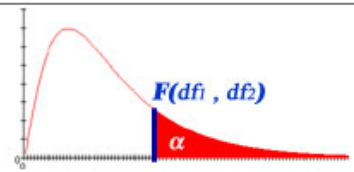
What are numerator and denominator degrees of freedom?

$\nu_1 = 10 - 1 = 9; \nu_2 = 12 - 1 = 11$

Hypothesis test for 2 sample variances

Reading an F-table.

F Table for $\alpha = 0.025$



/	df_1=1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
df_2=1	647.7890	799.5000	864.1630	899.5833	921.8479	937.1111	948.2169	956.6562	963.2846	968.6274	976.7079	984.8668	993.1028	997.2492	1001.414	1005.598	1009.800	1014.020	1018.258
2	38.5063	39.0000	39.1655	39.2484	39.2982	39.3315	39.3552	39.3730	39.3869	39.3980	39.4146	39.4313	39.4479	39.4562	39.465	39.473	39.481	39.490	39.498
3	17.4434	16.0441	15.4392	15.1010	14.8848	14.7347	14.6244	14.5399	14.4731	14.4189	14.3366	14.2527	14.1674	14.1241	14.081	14.037	13.992	13.947	13.902
4	12.2179	10.6491	9.9792	9.6045	9.3645	9.1973	9.0741	8.9796	8.9047	8.8439	8.7512	8.6565	8.5599	8.5109	8.461	8.411	8.360	8.309	8.257
5	10.0070	8.4336	7.7636	7.3879	7.1464	6.9777	6.8531	6.7572	6.6811	6.6192	6.5245	6.4277	6.3286	6.2780	6.227	6.175	6.123	6.069	6.015
6	8.8131	7.2599	6.5988	6.2272	5.9876	5.8198	5.6955	5.5996	5.5234	5.4613	5.3662	5.2687	5.1684	5.1172	5.065	5.012	4.959	4.904	4.849
7	8.0727	6.5415	5.8898	5.5226	5.2852	5.1186	4.9949	4.8993	4.8232	4.7611	4.6658	4.5678	4.4667	4.4150	4.362	4.309	4.254	4.199	4.142
8	7.5709	6.0595	5.4160	5.0526	4.8173	4.6517	4.5286	4.4333	4.3572	4.2951	4.1997	4.1012	3.9995	3.9472	3.894	3.840	3.784	3.728	3.670
9	7.2093	5.7147	5.0781	4.7181	4.4844	4.3197	4.1970	4.1020	4.0260	3.9639	3.8682	3.7694	3.6669	3.6142	3.560	3.505	3.449	3.392	3.333
10	6.9367	5.4564	4.8256	4.4683	4.2361	4.0721	3.9498	3.8549	3.7790	3.7168	3.6209	3.5217	3.4185	3.3654	3.311	3.255	3.198	3.140	3.080
11	6.7241	5.2559	4.6300	4.2751	4.0440	3.8807	3.7586	3.6638	3.5879	3.5257	3.4296	3.3299	3.2261	3.1725	3.118	3.061	3.004	2.944	2.883
12	6.5538	5.0959	4.4742	4.1212	3.8911	3.7283	3.6065	3.5118	3.4358	3.3736	3.2773	3.1772	3.0728	3.0187	2.963	2.906	2.848	2.787	2.725

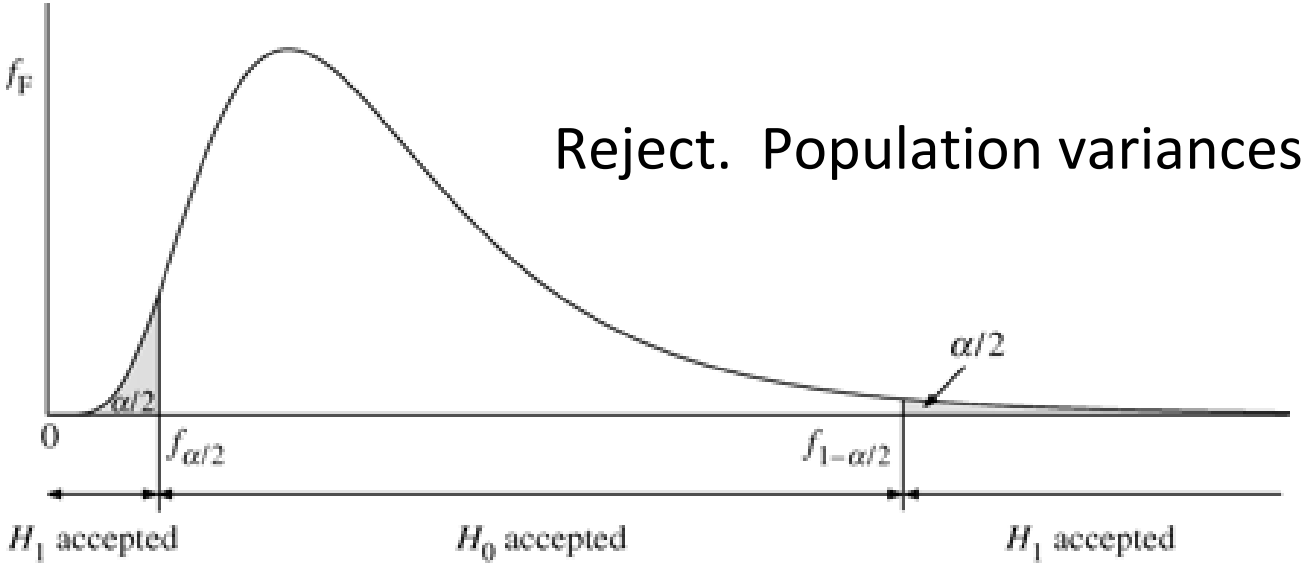
$$F_{0.025,9,11} = 3.5879; F_{0.975,9,11} = \frac{1}{F_{0.025,9,11}} = 0.2787$$

CSE 7315c

Hypothesis test for 2 sample variances

$$F_{0.025,9,11} = 3.5879; F_{0.975,9,11} = \frac{1}{F_{0.025,9,11}} = 0.2787; F_{observed} = 5.62$$

Will you reject the null hypothesis or not?



Reject. Population variances are not equal.

Hypothesis test for 2 sample variances

What are the business implications?

Variance in machine 1 is higher than in machine 2. Machine 1 needs to be inspected for any issues.

Applications of F Distribution

- Test for equality of variances.
- Test for differences of means in ANOVA.
- Test for regression models (slopes relating one continuous variable to another, e.g., Entrance exam scores and GPA)

Relations among Distributions – Children of the Normal

- χ^2 is drawn from the normal – N(0,1) deviates squared and summed.
- *F* is the ratio of 2 chi-squares, each divided by its *df*.
- A χ^2 divided by its *df* is a variance estimate, i.e., a sum of squares divided by the degrees of freedom.
- $F=t^2$. If you square *t*, you get an *F* with 1 *df* in the numerator, i.e., $t^2_{(v)} = F_{(1,v)}$
  (see http://www-ist.massey.ac.nz/dstirlin/CAST/CAST/HsimpleAnova/simpleAnova3.html for an example using this relationship)

ANOVA

The purpose of ANOVA (Analysis of Variance) is to test for significant differences between means of different groups.

Let us say 3 groups of students were given 3 different memory pills and their scores in an exam recorded. We want to understand if the differences are due to within group differences or between group differences.

Group 1	Group 2	Group 3
3	5	5
2	3	6
1	4	7

$\bar{X}_1 = 2$ $\bar{X}_2 = 4$ $\bar{X}_3 = 6$

$$\bar{\bar{X}} = \frac{3+2+1+5+3+4+5+6+7}{9} = 4$$

$Total\ Sum\ of\ Squares, SST$
$= (3-4)^2+(2-4)^2+(1-4)^2+(5-4)^2+(3-4)^2+(4-4)^2+(5-4)^2+(6-4)^2+(7-4)^2 = \mathbf{30}$
(Do you see any similarities with the formula for variance?)

When there are *m* groups and *n* members in each group, the degrees of freedom are $\boldsymbol{mn-1}$, since we can calculate one member knowing the overall mean.

How much of this variation is coming from within the groups and how much from between the groups?

Group 1	Group 2	Group 3
3	5	5
2	3	6
1	4	7
$\bar{X}_1 = 2$	$\bar{X}_2 = 4$	$\bar{X}_3 = 6$

$$\bar{\bar{X}} = \frac{3+2+1+5+3+4+5+6+7}{9} = 4$$

$Total\ Sum\ of\ Squares\ Within, SSW$
$= (3-2)^2+(2-2)^2+(1-2)^2+(5-4)^2+(3-4)^2+(4-4)^2+(5-6)^2+(6-6)^2+(7-6)^2 = \mathbf{6}$
When there are *m* groups and *n* members in each group, the degrees of freedom are $\boldsymbol{m(n-1)}$, since we can calculate one member knowing the group mean.

$Total\ Sum\ of\ Squares\ Between, SSB = 3(2-4)^2 + 3(4-4)^2+3(6-4)^2 = \mathbf{24}$
When there are *m* groups, the degrees of freedom are $\boldsymbol{m-1}$.

SST = SSW + SSB
Also, for degrees of freedom, $\boldsymbol{mn-1 = m(n-1)+(m-1)}$

Group 1	Group 2	Group 3
3	5	5
2	3	6
1	4	7
$\bar{X}_1 = 2$	$\bar{X}_2 = 4$	$\bar{X}_3 = 6$

$$\bar{\bar{X}} = \frac{3 + 2 + 1 + 5 + 3 + 4 + 5 + 6 + 7}{9} = 4$$

Given that mean of group 3 is highest and that of group 1 lowest, can we conclude that the pills given to group 3 had a larger impact or is it just variation within the group?

Let us have a null hypothesis that the population means of the 3 groups from which the samples were taken have the same mean, i.e., the pills do not have an impact on the performance in the exam. $\mu_1 = \mu_2 = \mu_3$. Let us also have a significance level, $\alpha = 0.10$.

What is the alternate hypothesis?

The pills have an impact on performance.

Group 1	Group 2	Group 3
3	5	5
2	3	6
1	4	7

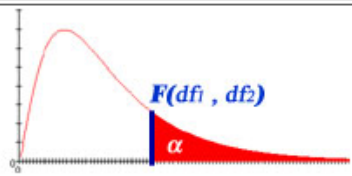
$\bar{X}_1 = 2$ $\bar{X}_2 = 4$ $\bar{X}_3 = 6$

$$\bar{\bar{X}} = \frac{3+2+1+5+3+4+5+6+7}{9} = 4$$

The test statistic used is F-statistic.

$$F - statistic = \frac{\frac{SSB}{df_{SSB}}}{\frac{SSW}{df_{SSW}}} = \frac{\frac{24}{2}}{\frac{6}{6}} = 12$$

If numerator is much bigger than the denominator, it means variation **between** means has bigger impact than variation **within**, thus rejecting the null hypothesis.

F Table for α = 0.10	

\	df_1=1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
df_2=1	39.86346	49.50000	53.59324	55.83296	57.24008	58.20442	58.90595	59.43898	59.85759	60.19498	60.70521	61.22034	61.74029	62.00205	62.26497	62.52905	62.79428	63.06064	63.32812
2	8.52632	9.00000	9.16179	9.24342	9.29263	9.32553	9.34908	9.36677	9.38054	9.39157	9.40813	9.42471	9.44131	9.44962	9.45793	9.46624	9.47456	9.48289	9.49122
3	5.53832	5.46238	5.39077	5.34264	5.30916	5.28473	5.26619	5.25167	5.24000	5.23041	5.21562	5.20031	5.18448	5.17636	5.16811	5.15972	5.15119	5.14251	5.13370
4	4.54477	4.32456	4.19086	4.10725	4.05058	4.00975	3.97897	3.95494	3.93567	3.91988	3.89553	3.87036	3.84434	3.83099	3.81742	3.80361	3.78957	3.77527	3.76073
5	4.06042	3.77972	3.61948	3.52020	3.45298	3.40451	3.36790	3.33928	3.31628	3.29740	3.26824	3.23801	3.20665	3.19052	3.17408	3.15732	3.14023	3.12279	3.10500
6	3.77595	3.46330	3.28876	3.18076	3.10751	3.05455	3.01446	2.98304	2.95774	2.93693	2.90472	2.87122	2.83634	2.81834	2.79996	2.78117	2.76195	2.74229	2.72216
7	3.58943	3.25744	3.07407	2.96053	2.88334	2.82739	2.78493	2.75158	2.72468	2.70251	2.66811	2.63223	2.59473	2.57533	2.55546	2.53510	2.51422	2.49279	2.47079

The df are 2 for numerator and 6 for denominator.

F_c, the critical F-statistic, therefore, is 3.46330. 12 is way higher than this and hence we reject the null hypothesis. That means the pills do have an impact on the performance.

ANOTHER EXAMPLE

A wind turbine manufacturer is testing 3 different designs of the turbines. It picks 3 different sites in the same district to install each model of the turbine. The mean power output (MW) over the day is measured for 9 consecutive days in each of the sites.

We want to understand if the differences are due to within-group differences or between-group differences.

Model 1			Model 2			Model 3		
3	4	3	3	5	7	5	5	5
2	5	5	6	7	6	6	5	7
4	3	3	4	4	8	7	6	6

$\bar{X}_1 = 3.56\ MW$ $\bar{X}_2 = 5.56\ MW$ $\bar{X}_3 = 5.78\ MW$

$$\bar{\bar{X}} = \frac{134}{27} = 4.96\ MW$$

$Total\ Sum\ of\ Squares, SST$
$= (2-4.96)^2 + 5 * (3-4.96)^2 + 4 * (4-4.96)^2 + 7 * (5-4.96)^2 + 5 * (6-4.96)^2$
$+ 4 * (7-4.96)^2 + (8-4.96)^2 = \mathbf{62.96}$

$Total\ Sum\ of\ Squares\ Within, SSW$
$= (2-3.56)^2 + 4 * (3-3.56)^2 + 2 * (4-3.56)^2 + 2 * (5-3.56)^2 + (3-5.56)^2 + 2 * (4-5.56)^2$
$+ (5-5.56)^2 + 2 * (6-5.56)^2 + 2 * (7-5.56)^2 + (8-5.56)^2 + 4 * (5-5.78)^2 + 3 * (6-5.78)^2$
$+ 2 * (7-5.78)^2 = \mathbf{36.00}$

$Total\ Sum\ of\ Squares\ Between, SSB$
$= 9 * (3.56-4.96)^2 + 9 * (5.56-4.96)^2 + 9 * (5.78-4.96)^2 = \mathbf{26.96}$

What is the null hypothesis?

All 3 sites from which the samples were taken have the same population-mean, i.e., the turbine design does not have an impact on the power production. That is $\mu_1 = \mu_2 = \mu_3$.

Let us also specify a significance level, $\alpha = 0.10$.

What is the alternate hypothesis?

The turbine design does impact the power output.

Compute the statistic

$$F - statistic = \frac{\frac{SSB}{df_{SSB}}}{\frac{SSW}{df_{SSW}}} = \frac{\frac{26.96}{2}}{\frac{36}{24}} = 8.99$$

If numerator is much bigger than the denominator, it means variation **between** means has bigger impact than variation **within**, thus rejecting the null hypothesis.

F Table for α = 0.10

\	df_1=1	2	3	4	5	6	7	8	9	10	12	
df_2=1	39.86346	49.50000	53.59324	55.83296	57.24008	58.20442	58.90595	59.43898	59.85759	60.19498	60.70521	61
2	8.52632	9.00000	9.16179	9.24342	9.29263	9.32553	9.34908	9.36677	9.38054	9.39157	9.40813	9
3	5.53832	5.46238	5.39077	5.34264	5.30916	5.28473	5.26619	5.25167	5.24000	5.23041	5.21562	5
4	4.54477	4.32456	4.19086	4.10725	4.05058	4.00975	3.97897	3.95494	3.93567	3.91988	3.89553	3
5	4.06042	3.77972	3.61948	3.52020	3.45298	3.40451	3.36790	3.33928	3.31628	3.29740	3.26824	3
6	3.77595	3.46330	3.28876	3.18076	3.10751	3.05455	3.01446	2.98304	2.95774	2.93693	2.90472	2
7	3.58943	3.25744	3.07407	2.96053	2.88334	2.82739	2.78493	2.75158	2.72468	2.70251	2.66811	2
8	3.45792	3.11312	2.92380	2.80643	2.72645	2.66833	2.62413	2.58935	2.56124	2.53804	2.50196	2
9	3.36030	3.00645	2.81286	2.69268	2.61061	2.55086	2.50531	2.46941	2.44034	2.41632	2.37888	2
10	3.28502	2.92447	2.72767	2.60534	2.52164	2.46058	2.41397	2.37715	2.34731	2.32260	2.28405	2
11	3.22520	2.85951	2.66023	2.53619	2.45118	2.38907	2.34157	2.30400	2.27350	2.24823	2.20873	2
12	3.17655	2.80680	2.60552	2.48010	2.39402	2.33102	2.28278	2.24457	2.21352	2.18776	2.14744	2
13	3.13621	2.76317	2.56027	2.43371	2.34672	2.28298	2.23410	2.19535	2.16382	2.13763	2.09659	2
14	3.10221	2.72647	2.52222	2.39469	2.30694	2.24256	2.19313	2.15390	2.12195	2.09540	2.05371	2
15	3.07319	2.69517	2.48979	2.36143	2.27302	2.20808	2.15818	2.11853	2.08621	2.05932	2.01707	1
16	3.04811	2.66817	2.46181	2.33274	2.24376	2.17833	2.12800	2.08798	2.05533	2.02815	1.98539	1
17	3.02623	2.64464	2.43743	2.30775	2.21825	2.15239	2.10169	2.06134	2.02839	2.00094	1.95772	1
18	3.00698	2.62395	2.41601	2.28577	2.19583	2.12958	2.07854	2.03789	2.00467	1.97698	1.93334	1
19	2.98990	2.60561	2.39702	2.26630	2.17596	2.10936	2.05802	2.01710	1.98364	1.95573	1.91170	1
20	2.97465	2.58925	2.38009	2.24893	2.15823	2.09132	2.03970	1.99853	1.96485	1.93674	1.89236	1
21	2.96096	2.57457	2.36489	2.23334	2.14231	2.07512	2.02325	1.98186	1.94797	1.91967	1.87497	1
22	2.94858	2.56131	2.35117	2.21927	2.12794	2.06050	2.00840	1.96680	1.93273	1.90425	1.85925	1
23	2.93736	2.54929	2.33873	2.20651	2.11491	2.04723	1.99492	1.95312	1.91888	1.89025	1.84497	1
24	2.92712	2.53833	2.32739	2.19488	2.10303	2.03513	1.98263	1.94066	1.90625	1.87748	1.83194	1
25	2.91774	2.52831	2.31702	2.18424	2.09216	2.02406	1.97138	1.92925	1.89469	1.86578	1.82000	1

The *df* are 2 for numerator and 24 for denominator.

F_c, the critical F-statistic, therefore, is 2.53833.

Our F=8.99 is way higher than this and hence we reject the null hypothesis. That means the turbine design does have an impact on the power production.

Anova: Single Factor

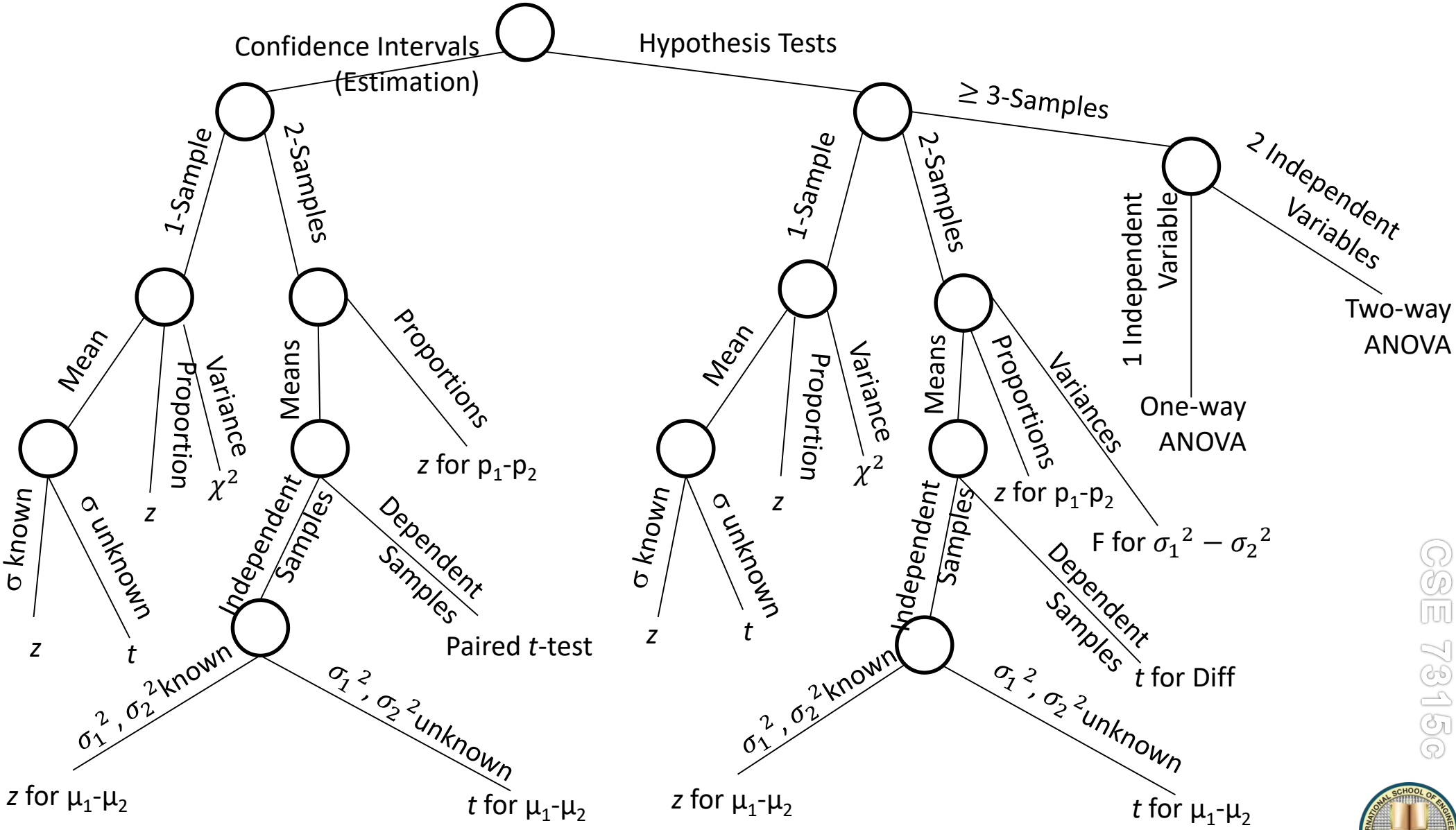
SUMMARY

Groups	*Count*	*Sum*	*Average*	*Variance*
Group1	9	32	3.55556	1.027778
Group2	9	50	5.55556	2.777778
Group3	9	52	5.77778	0.694444

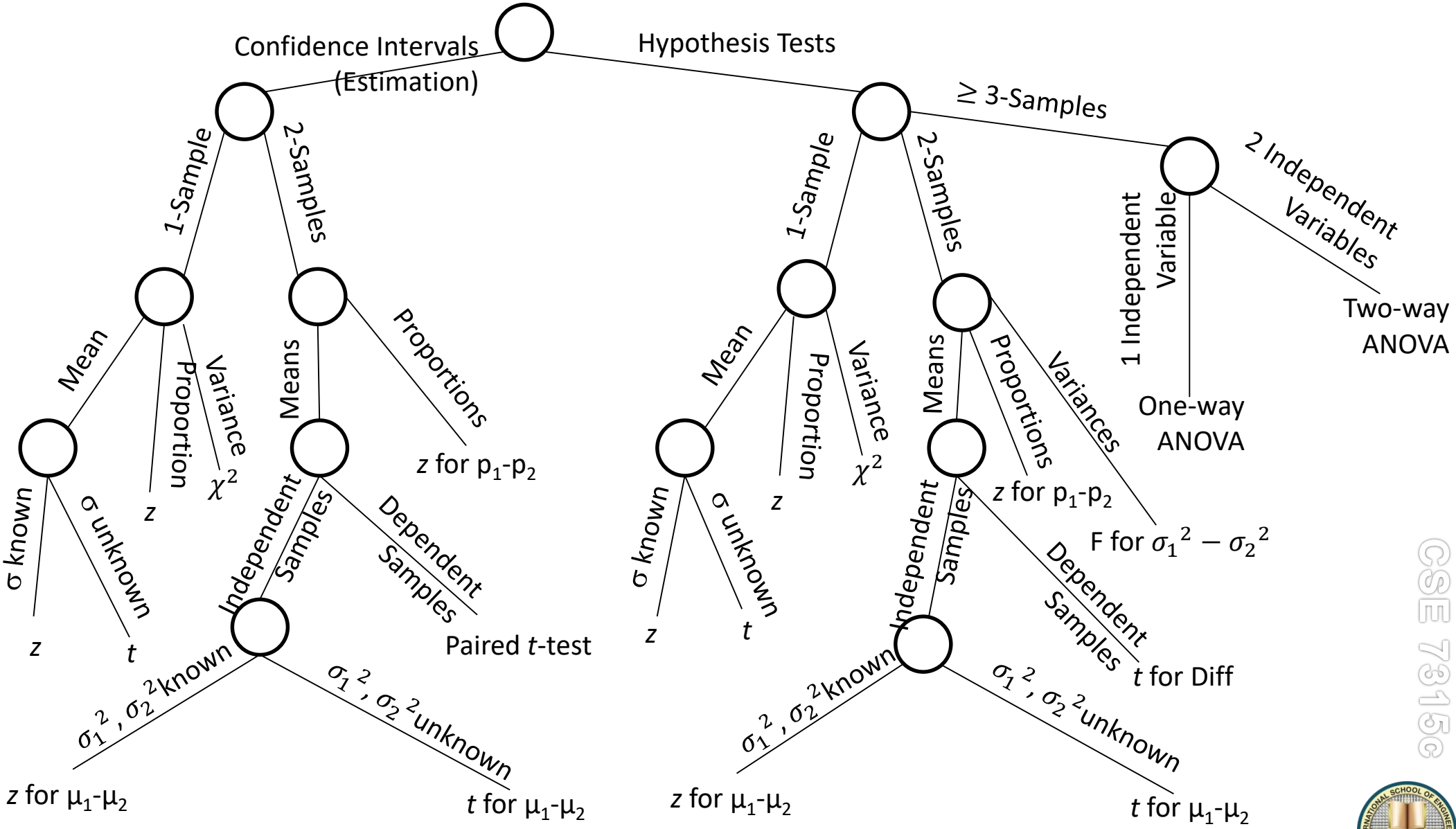
ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Between Groups	26.963	2	13.4815	8.987654	0.0012	2.5383
Within Groups	36	24	1.5			
Total	62.963	26				

Tree Diagram Taxonomy of Inferential Techniques



Tree Diagram Taxonomy of Inferential Techniques



CORRELATION, COVARIANCE AND REGRESSION

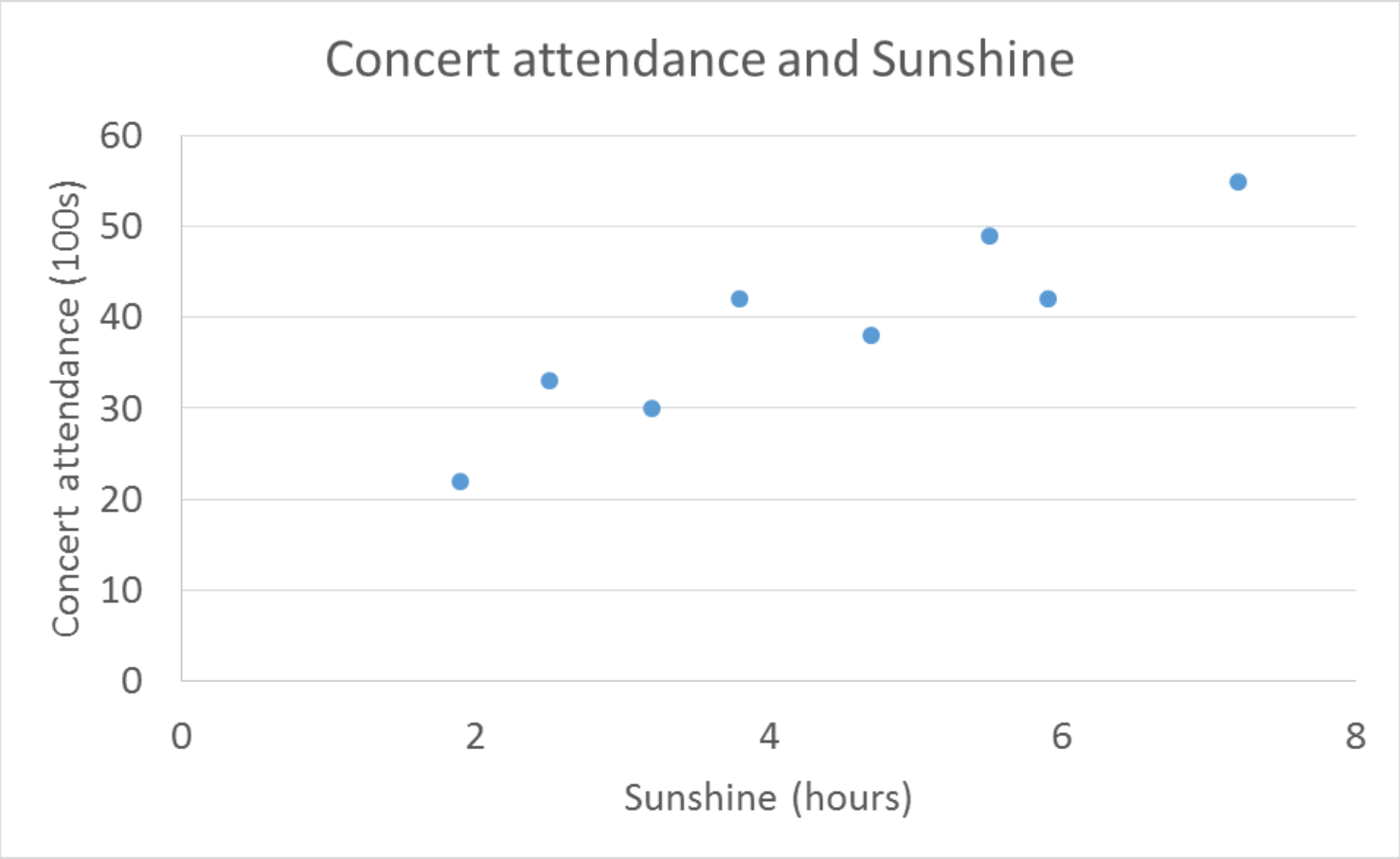
Image Source: http://blurtonline.com/wp-content/uploads/2013/06/Shaky-Knees-1514.jpeg;
Last accessed: May 1, 2014

Sunshine (hours)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
Concert attendance (100s)	22	33	30	42	38	49	42	55

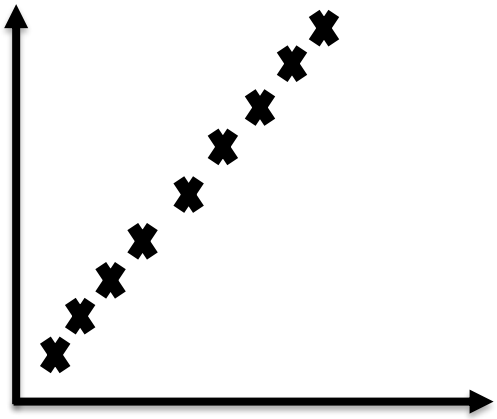
- The band makes a loss if less than 3500 people attend.
- Based on predicted hours of sunshine, can we predict ticket sales?
- Are sunshine and concert attendance correlated?

Sunshine (hours)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
Concert attendance (100s)	22	33	30	42	38	49	42	55

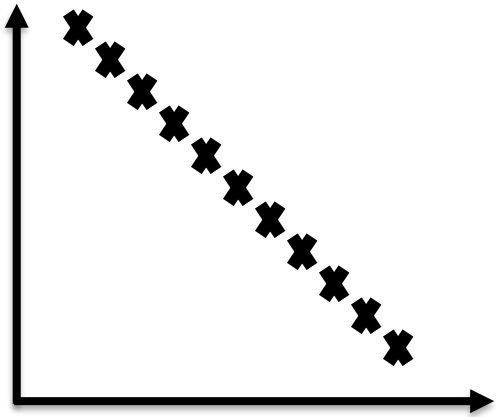
- Independent variable (explanatory) – Sunshine – Plotted on X-axis
- Dependent variable (response) – Concert attendance – Plotted on Y-axis



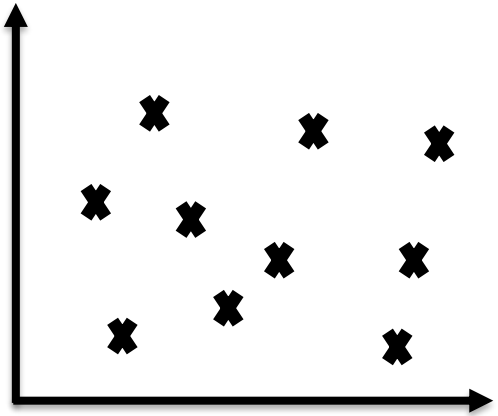
- Hours of sunshine and concert attendance are correlated, i.e., in general, longer sunshine hours indicate higher attendance.



Positive Linear Correlation

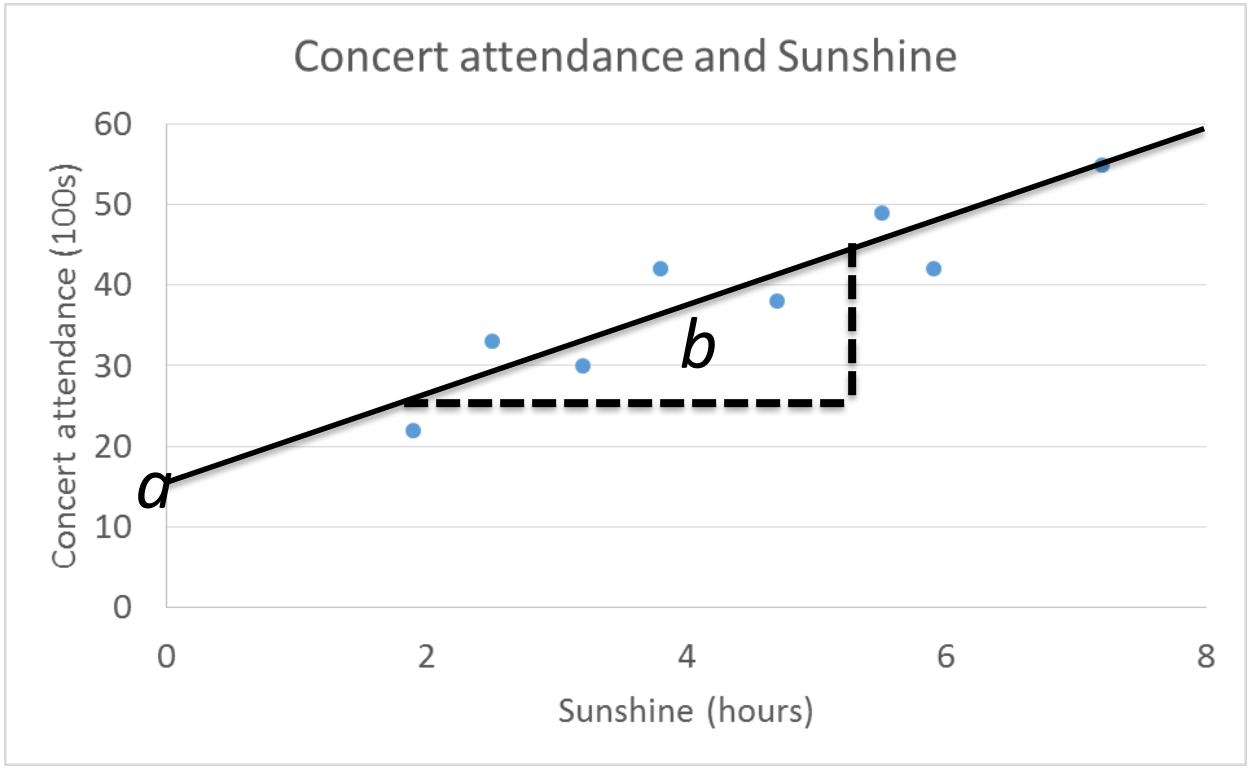


Negative Linear Correlation



No Correlation

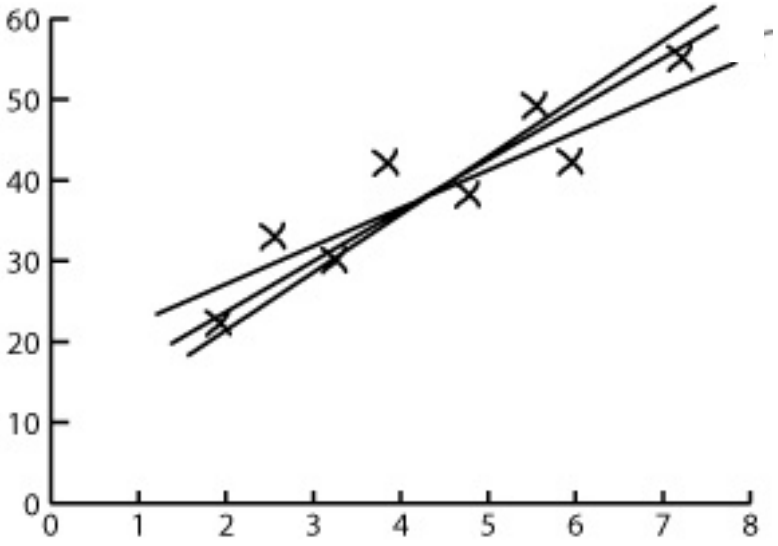
We need to find the equation of the line.

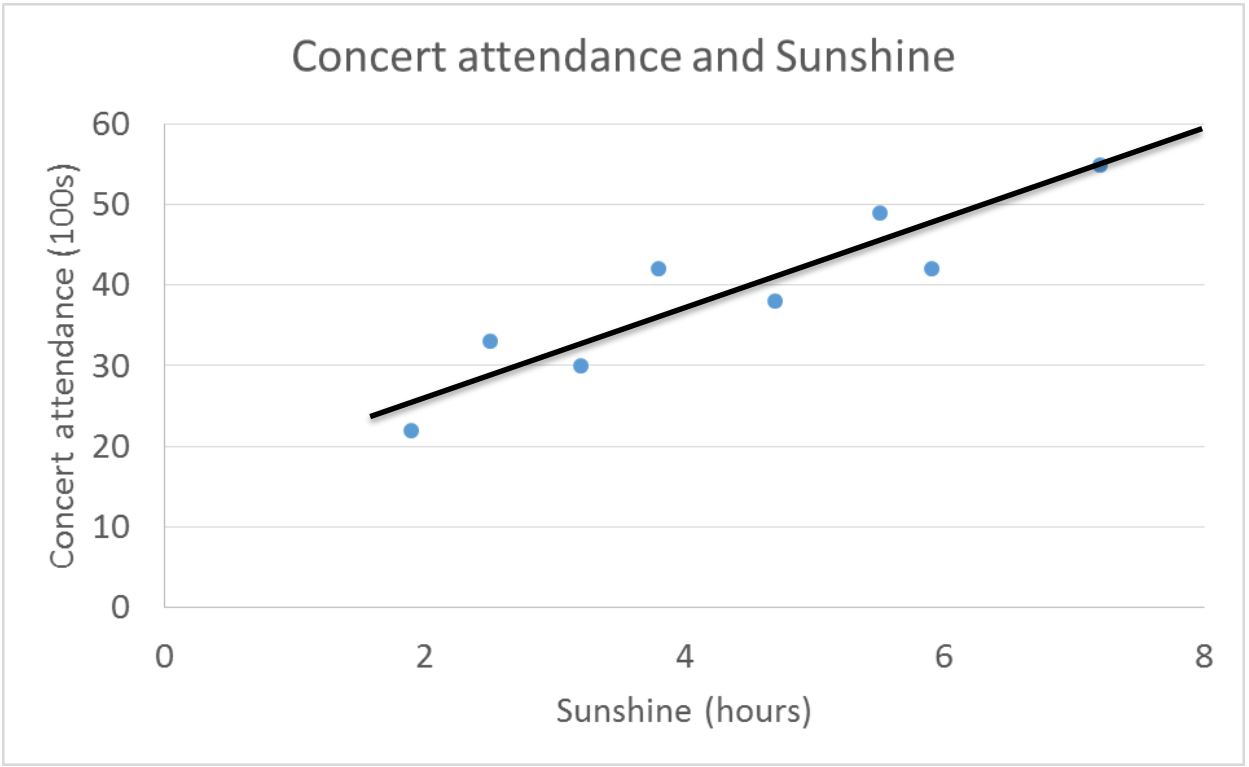


$y = a + bx$

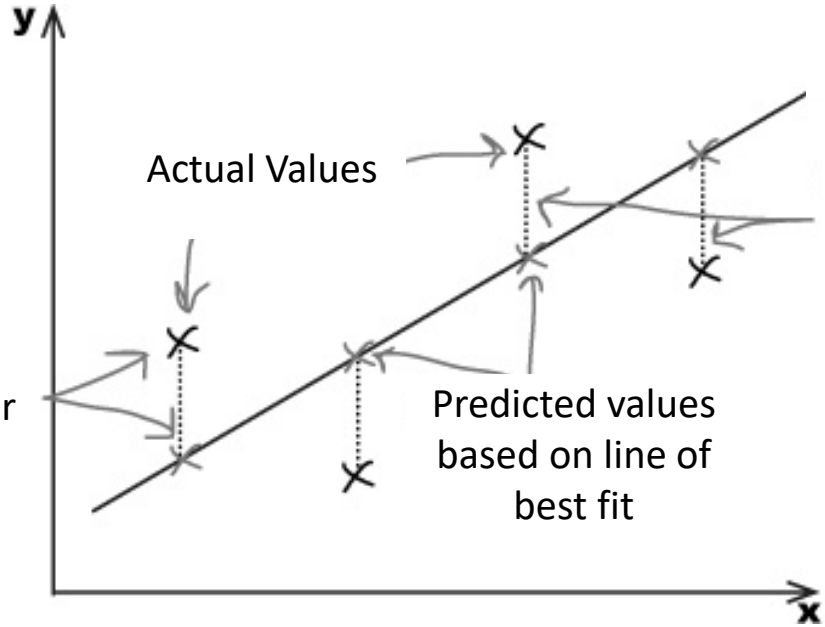
Sunshine (hours)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
Concert attendance (100s)	22	33	30	42	38	49	42	55

- Line of best fit





We need to minimize errors.



Line of best fit is the line that minimizes all the distances (residuals) between the Actual and Estimated values

We could do that by minimizing $\sum(y_i - \hat{y}_i)$, where y_i is the actual value and $\hat{y}_i$ its estimate. $(y_i - \hat{y}_i)$ is also known as the **residual**.

We need to minimize errors.

Just as we did when finding variance, we find the **sum of squared errors** or SSE. *Note in variance calculations, we subtract mean, $\bar{y}$, not $\hat{y}_i$.*

$$SSE = \sum (y_i - \hat{y}_i)^2$$

The value of *b*, the slope, that minimizes the SSE is given by

$$b = \frac{\sum((x - \bar{x})(y - \bar{y}))}{\sum(x - \bar{x})^2}$$

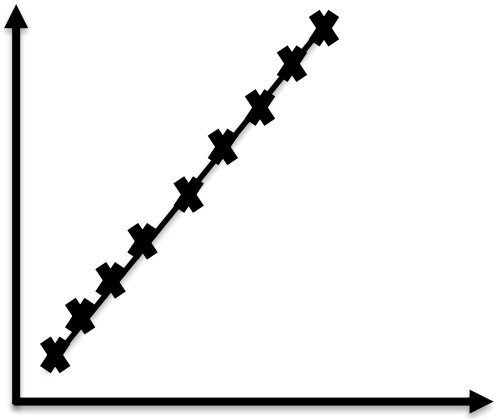
The value of *b*, the slope, that minimizes the SSE is given by

$$b = \frac{\sum((x - \bar{x})(y - \bar{y}))}{\sum(x - \bar{x})^2}$$

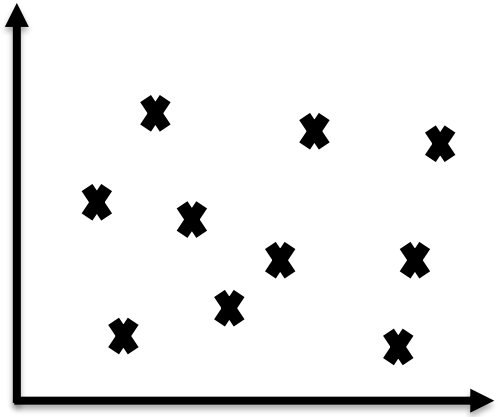
How do you calculate *a*? The line of best fit must pass through $(\bar{x}, \bar{y})$. Substituting in the equation y = *a* + *b*x, we can find *a*.

This method of fitting the line of best fit is called **least squares regression**.

But how do you know how good this fitted line is?



Perfect Linear Correlation



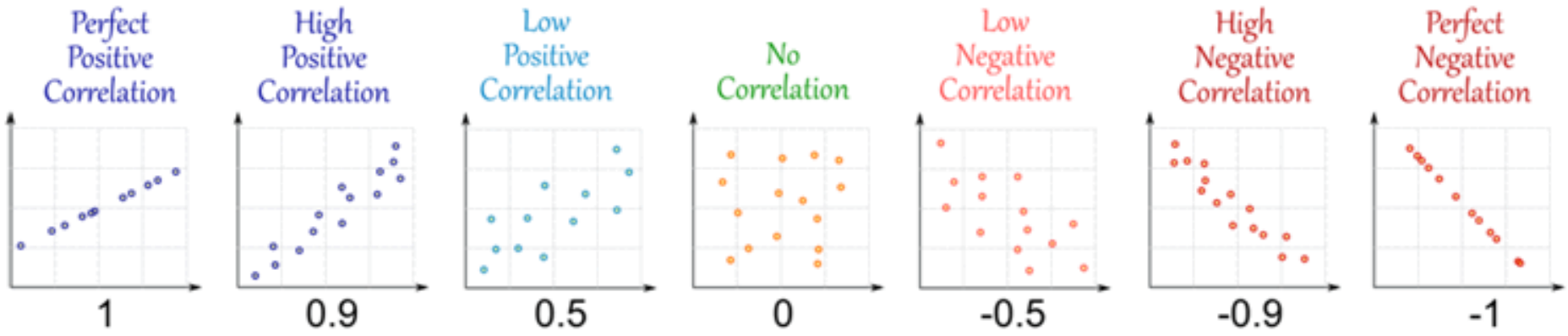
No Linear Correlation

The fit of the line is given by **correlation coefficient *r***.

$$r = \frac{b\ s_x}{s_y}$$

Correlation Coefficient

Correlation coefficient, *r*, is a number between -1 and 1 and tells us how well a regression line fits the data.



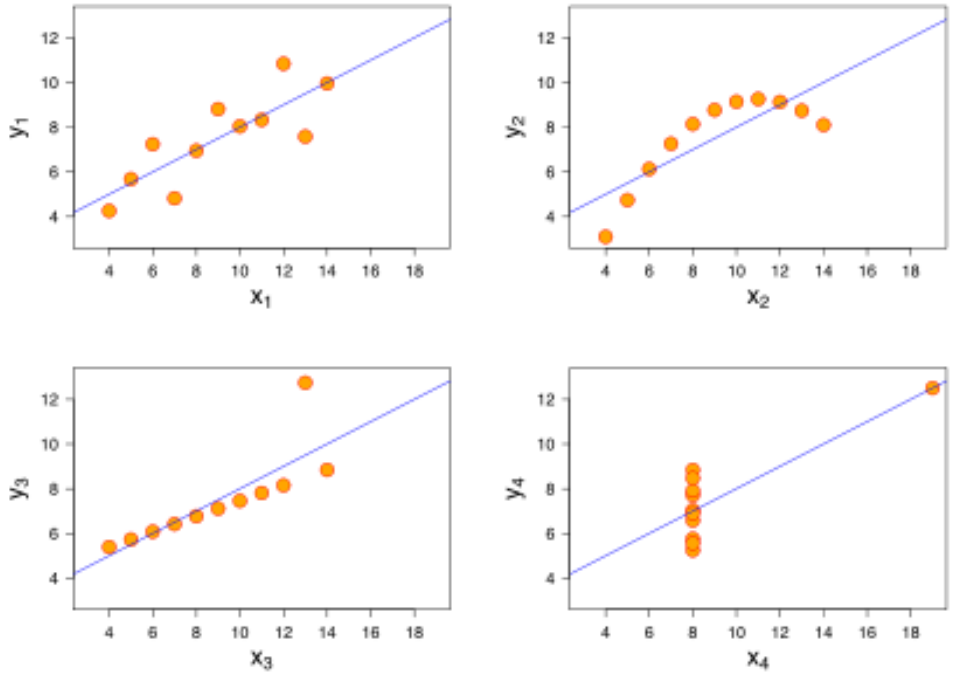
It gives the strength and direction of the relationship between two variables.

Img source: https://www.mathsisfun.com/data/correlation.html
Access date: 8/1/2017

Anscombe's Quartet

Anscombe's quartet							
I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.04	10	9.1	10	7.46	8	6.6
8	6.95	8	8.1	8	6.77	8	5.8
13	7.58	13	8.7	13	12.7	8	7.7
9	8.81	9	8.8	9	7.11	8	8.8
11	8.33	11	9.3	11	7.81	8	8.5
14	9.96	14	8.1	14	8.84	8	7
6	7.24	6	6.1	6	6.08	8	5.3
4	4.26	4	3.1	4	5.39	19	13
12	10.8	12	9.1	12	8.15	8	5.6
7	4.82	7	7.3	7	6.42	8	7.9
5	5.68	5	4.7	5	5.73	8	6.9

Property	Value
Mean of x in each case	9 (exact)
Sample variance of x in each case	11 (exact)
Mean of y in each case	7.50 (to 2 decimal places)
Sample variance of y in each case	4.122 or 4.127 (to 3 decimal places)
Correlation between x and y in each case	0.816 (to 3 decimal places)
Linear regression line in each case	$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)



THE STORY

We started by looking at various types of data (Categorical/Numerical) and we sought methods to describe the central tendency of the data. We discussed Mean, Median and Mode.

We realized to understand the data better, along with the central-tendency, we need to understand the spread. We looked at Range, Interquartile Range, Variance and Standard Deviation.

We discussed Box and Whisker plots as a way to represent both central-tendency and the spread in the data.

We moved on to basic Probability theory and described rules of probabilities using Venn Diagrams. We talked about conditional probabilities and that led to Bayes Theorem.

We then looked at some of the commonly occurring Probability Distributions and their properties, and looked at the expected values, their variance and the probabilities of various possible outcomes.

Then we saw how the *Sampling Distributions of Means* tend to normal distribution irrespective of how the population is distributed and learned how to describe populations based on available sample data.

We then looked at Confidence Intervals to properly describe the conclusions about populations based on samples.

Then we studied Hypothesis Tests as an alternative inferential technique to prove our claims. Of course, there are errors in these tests too.

We then studied various statistical tests to test hypotheses.

We looked at how to analyze results and find differences between what we expect and what we get, through χ^2 Distributions (goodness-of-fit).

We studied ANOVA, 2-sample t-tests and F test as a means of understanding significant differences between means and variances.

We also studied Independence, Correlation and Covariance between variables and learned about Regression basics.

International School of Engineering

Plot 63/A, Floors 1&2, Road # 13, Film Nagar, Jubilee Hills, Hyderabad - 500 033

For Individuals: +91-9502334561/63 or 040-65743991

For Corporates: +91-9618483483

Web: http://www.insofe.edu.in

Facebook: https://www.facebook.com/insofe

Twitter: https://twitter.com/Insofeedu

YouTube: http://www.youtube.com/InsofeVideos

SlideShare: http://www.slideshare.net/INSOFE

LinkedIn: http://www.linkedin.com/company/international-school-of-engineering