



Inspire...Educate...Transform.

## Stat Skills

**Chi-Square Testing, ANOVA, 2-Sample z-Test, Correlation, Covariance, Regression**

**Dr. Sridhar Pappu**

**Executive VP – Academics, INSOF**

June 21, 2015

# $\chi^2$ DISTRIBUTION

So you modeled a situation using a probability distribution and got a good idea of how things will shape up in the long run. But what if what you expected and what you observed are not the same? How would you know if the difference is due to normal fluctuations or if your model was incorrect?

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.

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

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

<b>x</b>	-2	23	48	73	98
<b>Frequency</b>	965	10	9	9	7

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

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

<b>x</b>	<b>Observed Frequency</b>	<b>Expected Frequency</b>
-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?

# $\chi^2$ test to the rescue

$\chi^2$  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.

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

$$\chi^2 = 38.272$$

Is this high?

To find this, we need to look at the  $\chi^2$  distribution.

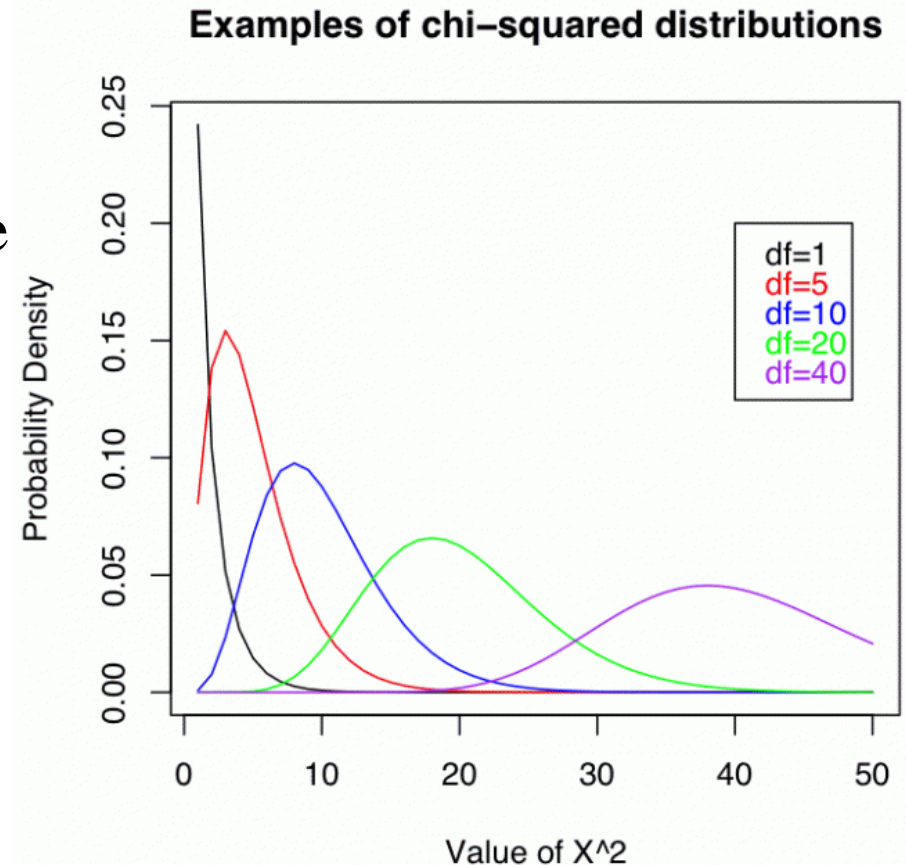
# $\chi^2$ distribution

$$X^2 \sim \chi^2(\nu)$$

$\nu$  represents the degrees of freedom. It is the # of independent variables used to calculate the test statistic  $X^2$ .

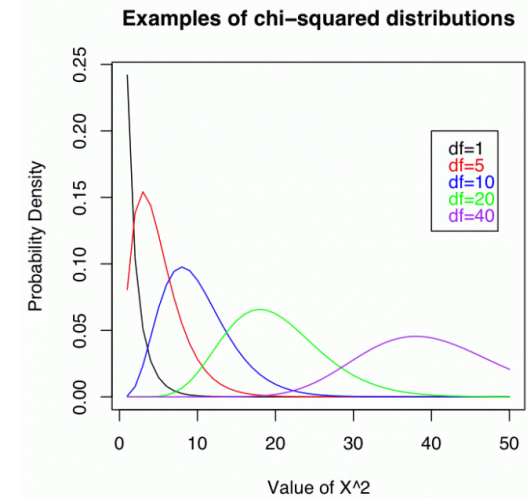
When  $\nu$  is 1 or 2, the probability of getting low values of the test statistic,  $X^2$ , is higher than getting high values, i.e., observed frequencies are expected to be close to the expected values.

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





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<sup>th</sup>. Therefore, there are  $5-1=4$  degrees of freedom.

**$\nu$  = (number of classes) – (number of restrictions), or**

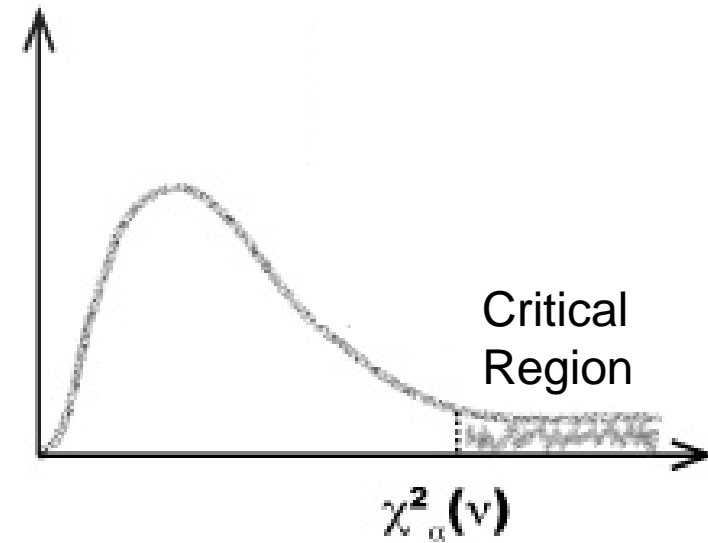
**$\nu$  = (number of classes) – 1 – (number of parameters being estimated from sample data)**

# How do we know the Significance of the difference?

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

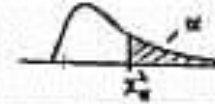
A test at significance level  $\alpha$  is written as  $\chi^2_{\alpha}(\nu)$ . The critical region is to its right.

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



# Using $\chi^2$ probability tables

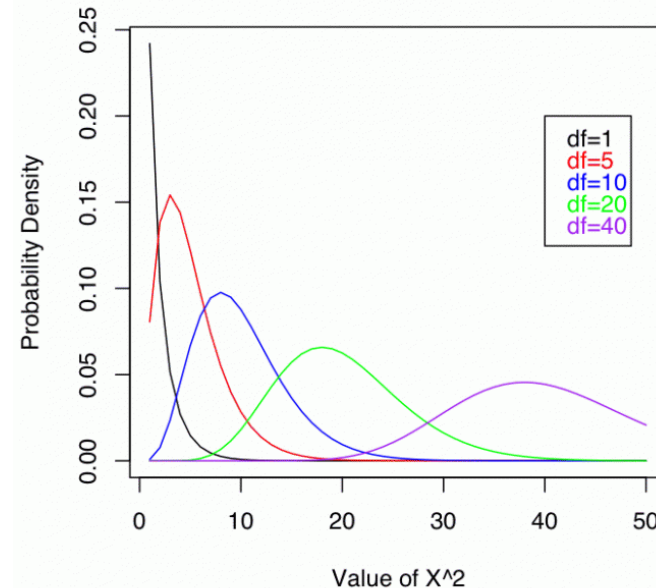
TABLE OF CHI-SQUARE DISTRIBUTION



$\alpha$	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
$\nu$															
1	0.00393	0.00446	0.00493	0.00541	0.00591	0.00643	0.00700	0.00764	0.00833	0.00910	0.00990	0.01075	0.01175	0.01287	0.01426
2	0.0100	0.0137	0.0175	0.0214	0.0256	0.0300	0.0347	0.0400	0.0455	0.0514	0.0577	0.0643	0.0713	0.0788	0.0878
3	0.0717	0.0878	0.1038	0.1206	0.1381	0.1561	0.1746	0.1935	0.2127	0.2321	0.2517	0.2715	0.2915	0.3117	0.3321
4	0.207	0.236	0.266	0.297	0.328	0.359	0.390	0.421	0.452	0.483	0.514	0.545	0.576	0.606	0.637
5	0.412	0.475	0.538	0.601	0.664	0.727	0.790	0.853	0.916	0.979	1.042	1.105	1.168	1.231	1.294
6	0.676	0.738	0.801	0.864	0.927	0.990	1.053	1.116	1.179	1.242	1.305	1.368	1.431	1.494	1.557

$$P(\chi_{\alpha}^2(\nu) \geq x) = \alpha.$$

Examples of chi-squared distributions



# Properties of $\chi^2$ random variable

- A  $\chi^2$  random variable takes values between 0 and  $\infty$ .
- Mean of a  $\chi^2$  distribution is  $\nu$ .
- Variance of a  $\chi^2$  distribution is  $2\nu$ .
- The shape of the distribution is skewed to the right.
- As  $\nu$  increases, Mean gets larger and the distribution spreads wider.
- As  $\nu$  increases, distribution tends to normal.

# Uses of $\chi^2$ 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.

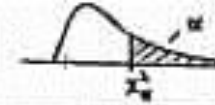
# What are the expected frequencies and degrees of freedom?

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

$$v = 4$$

# What is the critical region?

TABLE OF CHI-SQUARE DISTRIBUTION



$\alpha$	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
$\nu$															
1	0.00393	0.00446	0.00493	0.00541	0.00591	0.00641	0.00691	0.00741	0.00791	0.00841	0.00891	0.00941	0.00991	0.01041	0.01091
2	0.0100	0.0136	0.0173	0.0211	0.0250	0.0289	0.0329	0.0368	0.0408	0.0448	0.0488	0.0528	0.0568	0.0608	0.0648
3	0.0717	0.0842	0.0968	0.1094	0.1219	0.1345	0.1471	0.1597	0.1723	0.1849	0.1975	0.2101	0.2227	0.2353	0.2479
4	0.207	0.230	0.253	0.276	0.299	0.322	0.345	0.368	0.391	0.414	0.437	0.460	0.483	0.506	0.529
5	0.412	0.455	0.498	0.541	0.584	0.627	0.670	0.713	0.756	0.799	0.842	0.885	0.928	0.971	1.014
6	0.676	0.739	0.802	0.865	0.928	0.991	1.054	1.117	1.180	1.243	1.306	1.369	1.432	1.495	1.558
7	0.989	1.070	1.151	1.232	1.313	1.394	1.475	1.556	1.637	1.718	1.799	1.880	1.961	2.042	2.123
8	1.344	1.433	1.522	1.611	1.699	1.788	1.877	1.966	2.055	2.144	2.233	2.322	2.411	2.499	2.588
9	1.735	1.833	1.930	2.027	2.124	2.221	2.318	2.415	2.512	2.609	2.706	2.803	2.899	2.996	3.093
10	2.156	2.253	2.350	2.447	2.544	2.641	2.738	2.835	2.932	3.029	3.126	3.223	3.319	3.416	3.513
11	2.603	2.700	2.797	2.894	2.991	3.088	3.185	3.282	3.379	3.476	3.573	3.669	3.766	3.863	3.960
12	3.074	3.171	3.268	3.365	3.462	3.559	3.656	3.753	3.850	3.947	4.044	4.141	4.238	4.335	4.432

$\chi^2_{5\%}(4) = 9.488$ . 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.

# $\chi^2$ goodness of fit works for any probability distribution

Distribution	Condition	$\nu$
<b>Binomial</b>	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	$\nu = n - 2$
<b>Poisson</b>	You know $\lambda$	$\nu = n - 1$
	You don't know $\lambda$ , and have to estimate it from observed frequencies	$\nu = n - 2$
<b>Normal</b>	You know $\mu$ and $\sigma^2$	$\nu = n - 1$
	You don't know $\mu$ and $\sigma^2$ , 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

To calculate expected frequencies, we need to know the probabilities for each value. And to know the probabilities, we need  $\lambda$ .

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

# 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
<b>TOTAL</b>	<b>150</b>	<b>358</b>

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

## 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  $\lambda = 2.39$

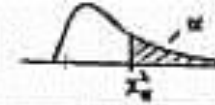
# of calls per 5-min interval	Expected Probability	Expected Frequency
0	0.0916	13.74
1	0.2190	32.85
2	0.2617	39.25
3	0.2085	31.27
4	0.1246	18.69
5	0.0595	8.93
6 or more	0.0526	3.56
<b>TOTAL</b>		<b>150.00</b>

$$v = 7 - 2 = 5$$

How do you find this value?

## Step 3: Determine the critical region

TABLE OF CHI-SQUARE DISTRIBUTION



$\alpha$	0.995	0.99	0.98	0.975	0.95	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.001
$\nu$																				
1	0.00393	0.00446	0.00493	0.00541	0.00591	0.00643	0.00697	0.00754	0.00813	0.00875	0.00939	0.01006	0.01076	0.01150	0.01227	0.01308	0.01392	0.01480	0.01572	0.01668
2	0.01000	0.01098	0.01199	0.01303	0.01412	0.01526	0.01645	0.01769	0.01900	0.02037	0.02181	0.02332	0.02490	0.02656	0.02830	0.03012	0.03202	0.03400	0.03606	0.03830
3	0.0717	0.0784	0.0853	0.0923	0.0995	0.1069	0.1145	0.1223	0.1303	0.1385	0.1469	0.1555	0.1643	0.1733	0.1825	0.1919	0.2015	0.2113	0.2213	0.2315
4	0.207	0.223	0.240	0.258	0.276	0.295	0.314	0.334	0.354	0.375	0.396	0.417	0.438	0.460	0.482	0.504	0.527	0.550	0.573	0.597
5	0.412	0.438	0.465	0.492	0.520	0.548	0.576	0.605	0.634	0.663	0.692	0.721	0.750	0.780	0.810	0.840	0.870	0.900	0.930	0.960
6	0.676	0.703	0.731	0.759	0.787	0.816	0.845	0.874	0.903	0.933	0.962	0.991	1.021	1.051	1.081	1.111	1.141	1.171	1.201	1.231
7	0.989	1.017	1.046	1.075	1.104	1.134	1.163	1.193	1.223	1.253	1.283	1.313	1.343	1.373	1.403	1.433	1.463	1.493	1.523	1.553
8	1.344	1.373	1.403	1.433	1.463	1.493	1.523	1.553	1.583	1.613	1.643	1.673	1.703	1.733	1.763	1.793	1.823	1.853	1.883	1.913
9	1.735	1.765	1.795	1.825	1.855	1.885	1.915	1.945	1.975	2.005	2.035	2.065	2.095	2.125	2.155	2.185	2.215	2.245	2.275	2.305
10	2.156	2.186	2.216	2.246	2.276	2.306	2.336	2.366	2.396	2.426	2.456	2.486	2.516	2.546	2.576	2.606	2.636	2.666	2.696	2.726
11	2.603	2.633	2.663	2.693	2.723	2.753	2.783	2.813	2.843	2.873	2.903	2.933	2.963	2.993	3.023	3.053	3.083	3.113	3.143	3.173
12	3.074	3.104	3.134	3.164	3.194	3.224	3.254	3.284	3.314	3.344	3.374	3.404	3.434	3.464	3.494	3.524	3.554	3.584	3.614	3.644

$\chi^2_{1\%}(5) = 15.086$ . 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
<b>TOTAL</b>	<b>150</b>	<b>150</b>	<b>10.46</b>

$$X^2 = 10.46$$



## 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.

# $\chi^2$ 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.



A black and white photograph of three croupiers (two women and one man) standing behind a curved blackjack table. Arrows point from each croupier down to a table below.

Possible Outcomes	Croupier A	Croupier B	Croupier C	Observed Results
<b>Win</b>	43	49	22	
<b>Draw</b>	8	2	5	
<b>Lose</b>	47	44	30	

# $\chi^2$ 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.

# $\chi^2$ 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

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

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

If croupier and the outcome are independent,

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

# $\chi^2$ 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*

$$\begin{aligned}
 &= \text{Grand Total} \times \frac{\text{Total Wins}}{\text{Grand Total}} \times \frac{\text{Total A}}{\text{Grand Total}} \\
 &= \frac{\text{Total Wins} \times \text{Total A}}{\text{Grand Total}} = \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}
 \end{aligned}$$

# $\chi^2$ 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	$(114 \cdot 98) / 250$	$(114 \cdot 95) / 250$	$(114 \cdot 57) / 250$
Draw	$(15 \cdot 98) / 250$	$(15 \cdot 95) / 250$	$(15 \cdot 57) / 250$
Lose	$(121 \cdot 98) / 250$	$(44 \cdot 95) / 250$	$(121 \cdot 57) / 250$

# $\chi^2$ independence test – Calculating $X^2$

	Observed	Expected	$\frac{(O - E)^2}{E}$
A	43		
	8		
B	47		
	49		
C	2		
	44		
	22		
	5		
	30		
	$\sum O = 250$	$\sum E =$	$\sum \frac{(O - E)^2}{E} = 5.004$



# $\chi^2$ independence test – Calculating $\nu$

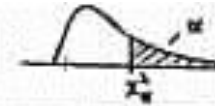
	Croupier A	Croupier B	Croupier C
Win			
Draw			
Lose			

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.

# $\chi^2$ 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



$\alpha$	0.995	0.99	0.98	0.975	0.95	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.05	0.025	0.01	0.005	0.001
$\nu$																			
1	0.00393	0.00446	0.00500	0.00554	0.00607	0.00675	0.00756	0.00854	0.00970	0.01107	0.01267	0.01453	0.01668	0.01928	0.02239	0.02615	0.03005	0.03438	0.03919
2	0.01000	0.01098	0.01200	0.01305	0.01406	0.01515	0.01632	0.01758	0.01892	0.02034	0.02184	0.02342	0.02508	0.02683	0.02866	0.03057	0.03256	0.03463	0.03678
3	0.0717	0.0742	0.0768	0.0794	0.0820	0.0847	0.0875	0.0903	0.0931	0.0959	0.0988	0.1017	0.1046	0.1075	0.1104	0.1133	0.1162	0.1191	0.1220
4	0.207	0.212	0.217	0.222	0.227	0.232	0.237	0.242	0.247	0.252	0.257	0.262	0.267	0.272	0.277	0.282	0.287	0.292	0.297
5	0.412	0.418	0.424	0.430	0.436	0.441	0.447	0.453	0.459	0.465	0.471	0.477	0.483	0.489	0.495	0.501	0.507	0.513	0.519
6	0.676	0.682	0.688	0.694	0.700	0.706	0.712	0.718	0.724	0.730	0.736	0.742	0.748	0.754	0.760	0.766	0.772	0.778	0.784
7	0.989	0.995	1.001	1.007	1.013	1.019	1.025	1.031	1.037	1.043	1.049	1.055	1.061	1.067	1.073	1.079	1.085	1.091	1.097
8	1.344	1.350	1.356	1.362	1.368	1.374	1.380	1.386	1.392	1.398	1.404	1.410	1.416	1.422	1.428	1.434	1.440	1.446	1.452
9	1.735	1.741	1.747	1.753	1.759	1.765	1.771	1.777	1.783	1.789	1.795	1.801	1.807	1.813	1.819	1.825	1.831	1.837	1.843
10	2.156	2.162	2.168	2.174	2.180	2.186	2.192	2.198	2.204	2.210	2.216	2.222	2.228	2.234	2.240	2.246	2.252	2.258	2.264
11	2.603	2.609	2.615	2.621	2.627	2.633	2.639	2.645	2.651	2.657	2.663	2.669	2.675	2.681	2.687	2.693	2.699	2.705	2.711
12	3.074	3.080	3.086	3.092	3.098	3.104	3.110	3.116	3.122	3.128	3.134	3.140	3.146	3.152	3.158	3.164	3.170	3.176	3.182

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

# $\chi^2$ independence test – Decision

Since calculated  $X^2 = 5.004$ , it is outside the critical region, and hence we accept the null hypothesis.

# $\chi^2$ independence test

A restaurant management wants to know if the type of beverage ordered with lunch is independent of the age of the consumer. A random poll of 309 lunch customers is taken as shown below. At  $\alpha = 0.01$ , are the two variables independent?

	Coffee/Tea	Soft Drink	Water	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 beverage preferred is independent of age.

$H_1$ : Type of beverage preferred is not independent of age.

## Step 2: Finding expected frequencies and degrees of freedom

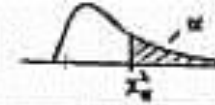
	Coffee/Tea	Soft Drink	Water	TOTAL
21-34	26	95	18	139
35-55	41	40	20	101
>55	24	13	32	69
TOTAL	91	148	70	309

	Coffee/Tea	Soft Drink	Water	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

$$v = 4$$

## Step 3: Determine the critical region

TABLE OF CHI-SQUARE DISTRIBUTION



$\alpha$	0.995	0.99	0.98	0.975	0.95	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.001
$\nu$																				
1	0.00393	0.00446	0.00493	0.00541	0.00591	0.00643	0.00700	0.00761	0.00827	0.00898	0.00975	0.01058	0.01147	0.01242	0.01344	0.01453	0.01569	0.01693	0.01826	0.01978
2	0.01000	0.01098	0.01197	0.01296	0.01396	0.01496	0.01597	0.01699	0.01802	0.01906	0.02011	0.02117	0.02224	0.02332	0.02441	0.02551	0.02662	0.02774	0.02887	0.03001
3	0.0717	0.0742	0.0768	0.0794	0.0820	0.0847	0.0874	0.0901	0.0929	0.0957	0.0985	0.1013	0.1041	0.1069	0.1097	0.1125	0.1153	0.1181	0.1210	0.1238
4	0.207	0.216	0.225	0.234	0.243	0.252	0.261	0.270	0.279	0.288	0.297	0.306	0.315	0.324	0.333	0.342	0.351	0.360	0.369	0.378
5	0.412	0.422	0.432	0.441	0.450	0.459	0.468	0.477	0.486	0.495	0.504	0.513	0.522	0.531	0.540	0.549	0.558	0.567	0.576	0.585
6	0.676	0.686	0.695	0.704	0.713	0.722	0.731	0.740	0.749	0.758	0.767	0.776	0.785	0.794	0.803	0.812	0.821	0.830	0.839	0.848
7	0.989	1.000	1.010	1.020	1.030	1.040	1.050	1.060	1.070	1.080	1.090	1.100	1.110	1.120	1.130	1.140	1.150	1.160	1.170	1.180
8	1.344	1.356	1.367	1.379	1.390	1.401	1.412	1.423	1.434	1.445	1.456	1.467	1.478	1.488	1.499	1.510	1.521	1.531	1.542	1.553
9	1.735	1.748	1.760	1.772	1.784	1.796	1.808	1.819	1.831	1.843	1.854	1.866	1.878	1.889	1.901	1.912	1.924	1.935	1.947	1.958
10	2.156	2.169	2.181	2.193	2.205	2.217	2.229	2.241	2.253	2.265	2.277	2.289	2.301	2.313	2.325	2.337	2.349	2.361	2.373	2.385
11	2.603	2.616	2.628	2.641	2.653	2.665	2.677	2.689	2.701	2.713	2.725	2.737	2.749	2.761	2.773	2.785	2.797	2.809	2.821	2.833
12	3.074	3.087	3.099	3.111	3.123	3.135	3.147	3.159	3.171	3.183	3.195	3.207	3.219	3.231	3.243	3.255	3.267	3.279	3.291	3.303

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

## Step 4: Calculate the test statistic $X^2$

		Observed	Expected	$\frac{(O - E)^2}{E}$
Coffee/Tea	{	26	40.94	
		41	29.74	
		24	20.32	
Soft Drink	{	95	66.58	
		40	48.38	
		13	33.05	
Water	{	18	31.49	
		20	22.88	
		32	15.63	
		$\sum O = 309$	$\sum E =$	$\sum \frac{(O - E)^2}{E} = 59.41$



## 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 preferred beverage and age are independent.

# 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 WUQ 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*

$$= (3 - 4)^2 + (2 - 4)^2 + (1 - 4)^2 + (5 - 4)^2 + (3 - 4)^2 + (4 - 4)^2 + (5 - 4)^2 + (6 - 4)^2 + (7 - 4)^2 = \mathbf{30}$$

When there are  $m$  groups and  $n$  members in each group, the degrees of freedom are  **$mn - 1$** , 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*

$$= (3 - 2)^2 + (2 - 2)^2 + (1 - 2)^2 + (5 - 4)^2 + (3 - 4)^2 + (4 - 4)^2 + (5 - 6)^2 + (6 - 6)^2 + (7 - 6)^2 = 6$$

When there are  $m$  groups and  $n$  members in each group, the degrees of freedom are  $m(n - 1)$ , since we can calculate one member knowing the group mean.

$$\text{Total Sum of Squares Between, SSB} = 3(2 - 4)^2 + 3(4 - 4)^2 + 3(6 - 4)^2 = 24$$

When there are  $m$  groups, the degrees of freedom are  $m - 1$ .

$$\text{SST} = \text{SSW} + \text{SSB}$$

$$\text{Also, for degrees of freedom, } mn - 1 = m(n - 1) + (m - 1)$$

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 WUQ.  $\mu_1 = \mu_2 = \mu_3$ . 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 is called 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 alpha=.10

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.

numerators →

df2/df1	1	2	3	4	5	6	7	8	9	10	12
1	39.86346	49.50000	53.59324	55.83296	57.24008	58.20442	58.90595	59.43898	59.85759	60.19498	60.705
2	8.52632	9.00000	9.16179	9.24342	9.29263	9.32553	9.34908	9.36677	9.38054	9.39157	9.408
3	5.53832	5.46238	5.39077	5.34264	5.30916	5.28473	5.26619	5.25167	5.24000	5.23041	5.215
4	4.54477	4.32456	4.19086	4.10725	4.05058	4.00975	3.97897	3.95494	3.93567	3.91988	3.895
5	4.06042	3.77972	3.61948	3.52020	3.45298	3.40451	3.36790	3.33928	3.31628	3.29740	3.268
6	3.77595	3.46330	3.28876	3.18076	3.10751	3.05455	3.01446	2.98304	2.95774	2.93693	2.904
7	3.58943	3.25744	3.07407	2.96053	2.88334	2.82739	2.78493	2.75158	2.72468	2.70251	2.668
8	3.45792	3.11312	2.92380	2.80643	2.72645						
9	3.36030	3.00645	2.81286	2.69268	2.61061						
10	3.28502	2.92447	2.72767	2.60534	2.52164						

denominators →

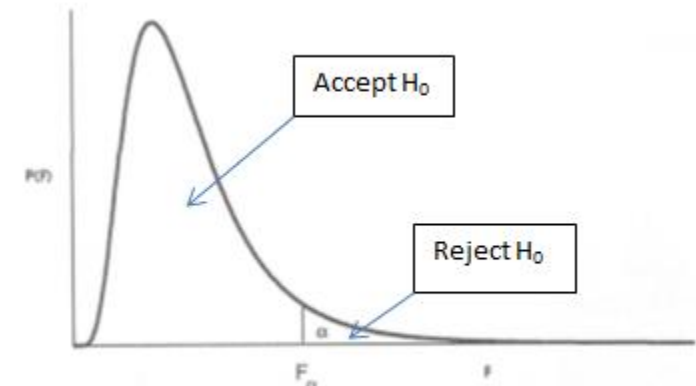


Figure K.1: The F distribution

# Problem

Dr. Dakshinamurthy is on a tour presenting Demystifying Data Science workshops in Columbus, London and Dubai. The workshop is the same each time but it is presented to High Level managers in Columbus, Midlevel managers in London and Low Level managers in Dubai. INSOFE believes that the feedback of the workshops will not vary with the audience.

On a scale of 1-10, with 10 being highest, Dr. Murthy got the following randomly selected scores. Use ANOVA to determine whether there is a significant difference in feedback according to manager level. Assume  $\alpha = 0.05$ . Discuss the business implications of your findings.

# Problem

Columbus (HL)	London (ML)	Dubai (LL)
7	8	5
7	9	6
8	8	5
7	10	7
9	9	4
	10	8
	8	

# Solution

## Hypothesis

$$H_0: \mu_1 = \mu_2 = \mu_3$$

$H_1$ : At least one of the means is different from the others

**Known:**  $\alpha = 0.05$

**Test:**  $\bar{X}_1 = 7.6$      $\bar{X}_2 = 8.86$      $\bar{X}_3 = 5.83$      $\bar{\bar{X}} = \frac{\sum n}{3} = 7.5$

*Total Sum of Squares, SST*

$$= 1 * (4 - 7.5)^2 + 2 * (5 - 7.5)^2 + 1 * (6 - 7.5)^2 + 4 * (7 - 7.5)^2 + 5 * (8 - 7.5)^2 + 3 * (9 - 7.5)^2 + 2 * (10 - 7.5)^2 = 1 * 3.5^2 + 4 * 2.5^2 + 4 * 1.5^2 + 9 * 0.5^2 = \mathbf{48.5}$$

*Total Sum of Squares Between, SSB*  $= 5 * (7.6 - 7.5)^2 + 7 * (8.86 - 7.5)^2 + 6 * (5.83 - 7.5)^2 = \mathbf{29.61}$

*Total Sum of Squares Within, SSW*

$$= (7 - 7.6)^2 + (7 - 7.6)^2 + \dots (8 - 8.86)^2 + (9 - 8.86)^2 + \dots (5 - 5.83)^2 + (6 - 5.83)^2 + \dots = \mathbf{18.89}$$

Columbus (HL)	London (ML)	Dubai (LL)
7	8	5
7	9	6
8	8	5
7	10	7
9	9	4
	10	8
	8	

# Solution

## Test (continued)

$$df_{SST} = 18 - 1 = \mathbf{17}$$

$$df_{SSB} = 3 - 1 = \mathbf{2}$$

$$df_{SSW} = (5 - 1) + (7 - 1) + (6 - 1) = \mathbf{15}$$

$$\frac{SSB}{df_{SSB}} = \frac{29.61}{2} = \mathbf{14.805}$$

$$\frac{SSW}{df_{SSW}} = \frac{18.89}{15} = \mathbf{1.259}$$

$$F \text{ value} = \frac{14.805}{1.259} = \mathbf{11.756}$$

# Solution

## Action

$F \text{ value} = 11.756$

Critical value is 3.68.

Percentage Points of the  $F$  Distribution (Continued)

		$\alpha = .05$							
$\nu_2$	$\nu_1$	Numerator Degrees of Freedom							
		1	2	3	4	5	6	7	8
1	1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88
2	1	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37
3	1	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85
4	1	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04
5	1	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82
6	1	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15
7	1	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73
8	1	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44
9	1	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23
10	1	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07
11	1	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95
12	1	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85
13	1	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77
14	1	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70
15	1	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64
16	1	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59
17	1	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55
18	1	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51
19	1	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48
20	1	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45
21	1	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42
22	1	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40
23	1	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37

Since the observed value is greater than the critical value, the decision will be to **REJECT** the null hypothesis.

# Solution

Columbus (HL)	London (ML)	Dubai (LL)
7	8	5
7	9	6
8	8	5
7	10	7
9	9	4
	10	8
	8	

## Business Implication

There is significant difference between the means of the feedback. This means the level of managers has an impact on feedback. INSOFE can determine that the content for these workshops is most suited to the midlevel and least to the low level managers and take necessary steps to meet the needs of all audiences.

Anova: Single Factor						
SUMMARY						
Groups	Count	Sum	Average	Variance		
Columbus (High Level)	5	38	7.6	0.8		
London (Midlevel)	7	62	8.857143	0.809524		
Dubai (Low Level)	6	35	5.833333	2.166667		
ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	29.60952	2	14.80476	11.75573	0.000849	3.68232
Within Groups	18.89048	15	1.259365			
Total	48.5	17				



# TWO-SAMPLE Z-TEST FOR MEANS



- Do two samples come from the same population?
- If they come from different populations, what is the difference in the means of the two populations?
  - Does the average cost of a two-bedroom flat differ between Bengaluru and Hyderabad? What is the difference?
  - What is the difference in the strength of steel produced under two different temperatures?
  - Does the effectiveness of Head & Shoulders anti-dandruff shampoo differ from Pantene anti-dandruff shampoo?
  - What is the difference in the productivity of men and women on an assembly line under certain conditions?

The Central Limit Theorem states that the difference in two sample means,  $\bar{x}_1 - \bar{x}_2$ , is normally distributed for large sample sizes (both  $n_1$  and  $n_2 \geq 30$ ) whatever the population distribution.

$$\text{Also, } \mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$$

$$\text{and } \sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \quad [\text{Recall } \text{Var}(X-Y) = \text{Var}(X) + \text{Var}(Y)]$$

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

# Hypothesis Testing

A random sample of 32 advertising managers and another sample of 34 auditing managers across India is taken for a salary survey.

You are testing whether there is a difference in the average wage of the two types of managers.

# Hypothesis Testing

Advertising Managers				Auditing Managers				
74256	80742	96767	71115	69962	60053	48036	43649	67814
96234	39672	77242	67574	55052	66359	67160	63369	71492
89807	45652	67056	59621	57828	61261	37386	59676	
93261	93083	64276	62483	63362	77136	59505	54449	
103030	63384	74194	69319	37194	66035	72790	46394	
74195	57791	65360	35394	99198	54335	71351	71804	
75932	65145	73904	86741	61254	42494	58653	72401	
54270	59045	68508	57351	73065	83849	63508	56470	

$$n_1 = 32$$

$$\bar{x}_1 = 70700$$

$$\sigma_1 = 16253$$

$$\sigma_1^2 = 264160$$

$$n_2 = 34$$

$$\bar{x}_2 = 62187$$

$$\sigma_2 = 12900$$

$$\sigma_2^2 = 166410$$

# Hypothesis Testing

What is the null hypothesis?

$$H_0: \mu_1 - \mu_2 = 0$$

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?

One is paid more than the other.

# Hypothesis Testing

At  $\alpha = 0.05$ , determine if there is a significant difference between the two types of managers.

$$z = \frac{(70700 - 62187) - (0)}{\sqrt{\frac{264160}{32} + \frac{166410}{34}}} = 2.35$$

You can find the p-value for this z-score (0.0094, which is less than 0.025) or knowing that the z-score for the desired significance level is  $\pm 1.96$ , 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.

Who earns more on an average and by how much?

Advertising managers earn 8513 more on an average.

# Confidence Intervals

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Rewriting,

$$(\bar{x}_1 - \bar{x}_2) - Z \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \leq \mu_1 - \mu_2 \leq (\bar{x}_1 - \bar{x}_2) + Z \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$



# Confidence Intervals

A study is conducted to estimate the difference between middle-income and low-income shoppers in terms of average amount saved on grocery bills per week by using coupons. 60 middle-income and 80 low-income shoppers are randomly sampled and their purchases monitored for a week.

Middle-income shoppers	Low-income shoppers
$n_1 = 60$	$n_2 = 80$
$\bar{x}_1 = \$5.84$	$\bar{x}_1 = \$2.67$
$\sigma_1 = \$1.41$	$\sigma_1 = \$0.54$

# Confidence Intervals

Construct a 98% CI to estimate the difference between the average amount saved with coupons by the two groups ( $z=2.33$ ).

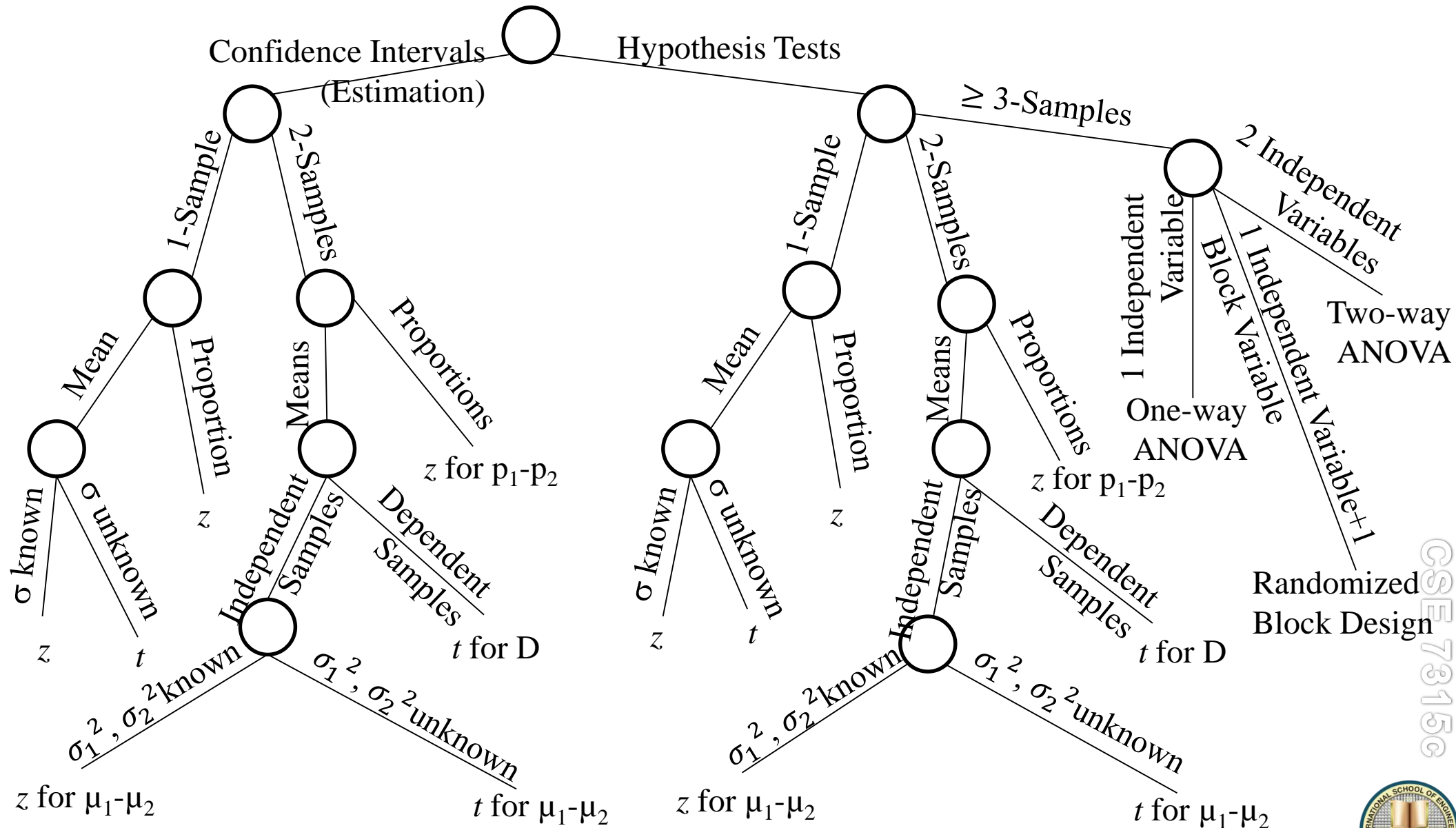
$$(5.84 - 2.67) - 2.33 \sqrt{\frac{1.41^2}{60} + \frac{0.54^2}{80}} \leq \mu_1 - \mu_2 \leq (5.84 - 2.67) + 2.33 \sqrt{\frac{1.41^2}{60} + \frac{0.54^2}{80}}$$

$$3.17 - 0.45 \leq \mu_1 - \mu_2 \leq 3.17 + 0.45$$

$$2.72 \leq \mu_1 - \mu_2 \leq 3.62$$

Note zero difference is unlikely as at 98% Confidence Level, the difference ranges between \$2.72 and \$3.62, with a point estimate for the difference in mean savings being \$3.17.

# Tree Diagram Taxonomy of Inferential Techniques



# CORRELATION, COVARIANCE AND REGRESSION



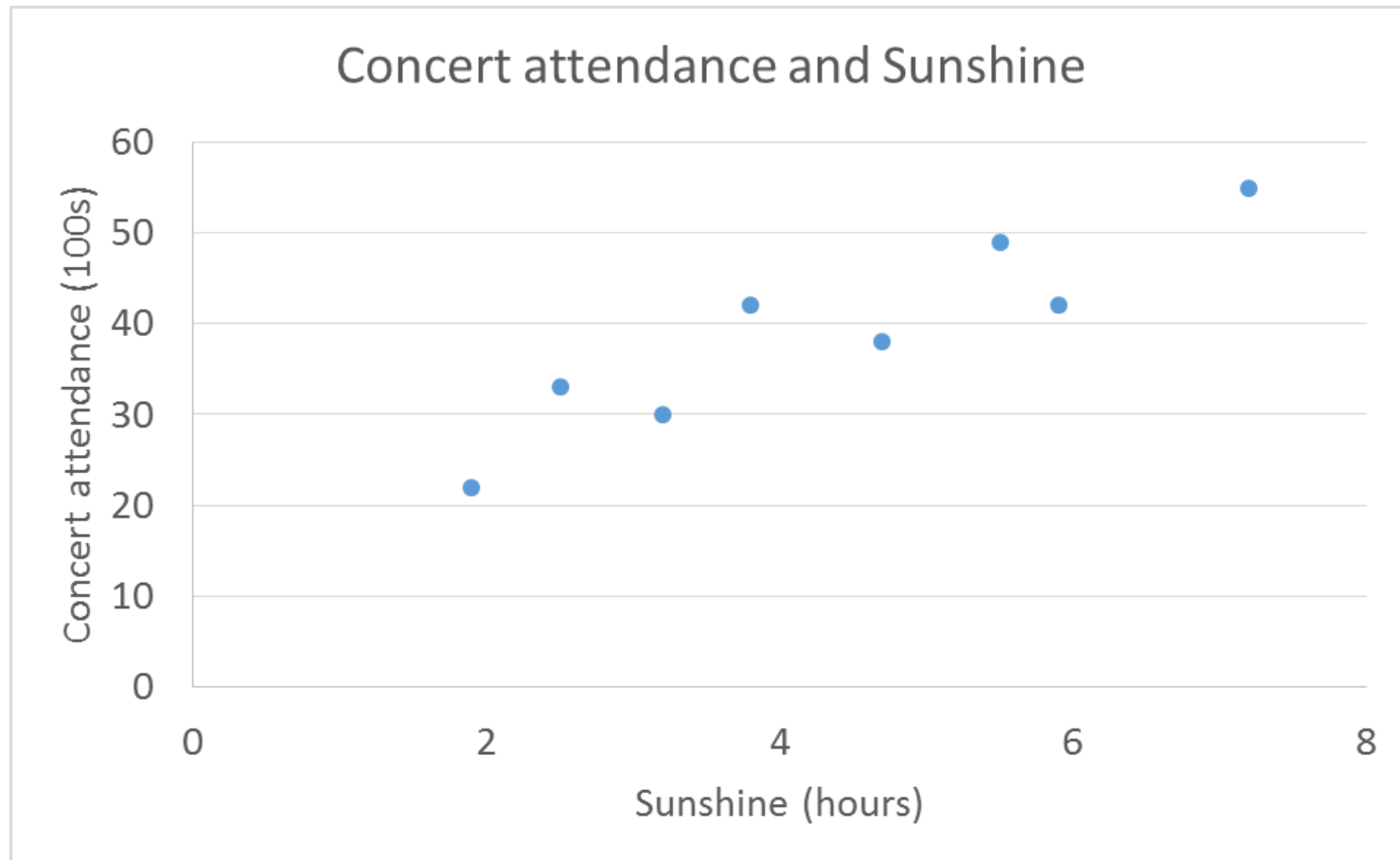
Image Source: <http://blurtonline.com/wp-content/uploads/2013/06/Shaky-Knees-1514.jpeg>;  
Last accessed: May 1, 2014

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

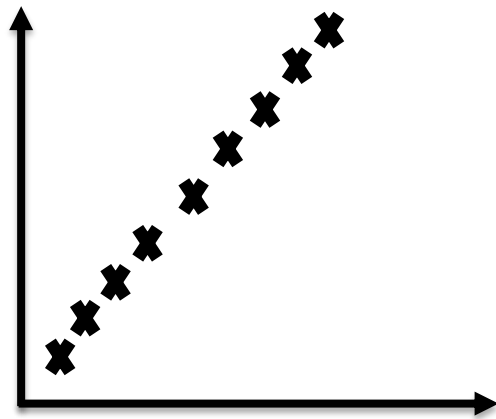
- The band makes a loss if less than 3500 people attend.
- Based on predicted hours of sunshine, can we predict ticket sales?
- Are sunshine and concert attendance correlated?

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

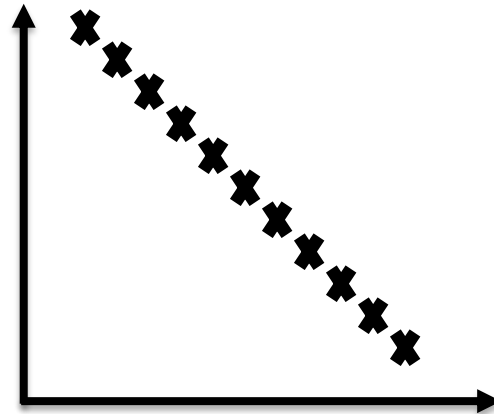
- Independent variable (explanatory) – Sunshine – Plotted on X-axis
- Dependent variable (response) – Concert attendance – Plotted on Y-axis



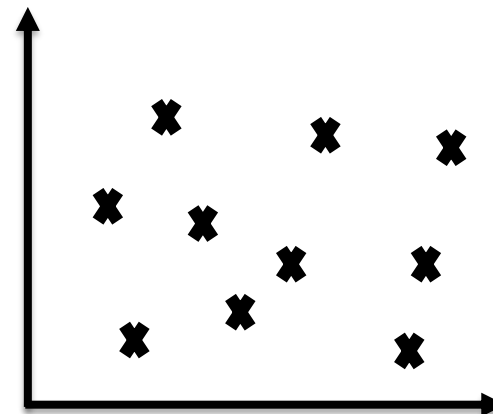
- Hours of sunshine and concert attendance are correlated, i.e., in general, longer sunshine hours indicate higher attendance.



Positive Linear  
Correlation



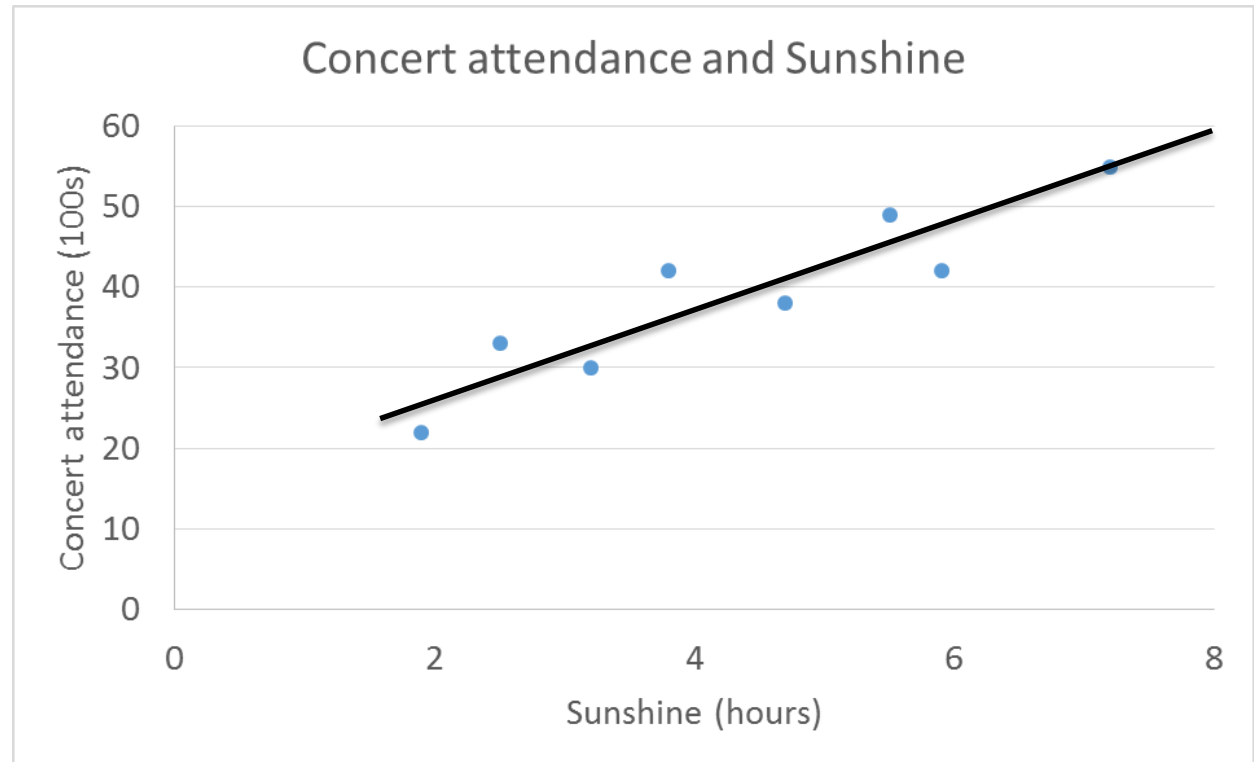
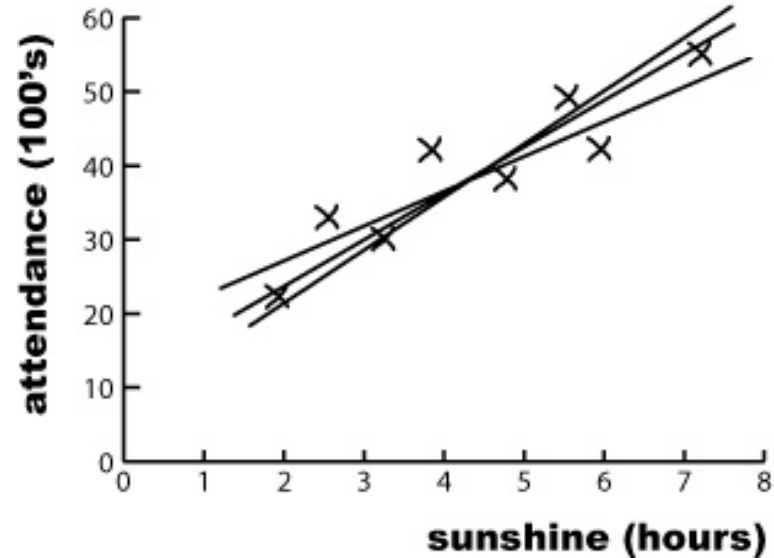
Negative Linear  
Correlation



No Correlation

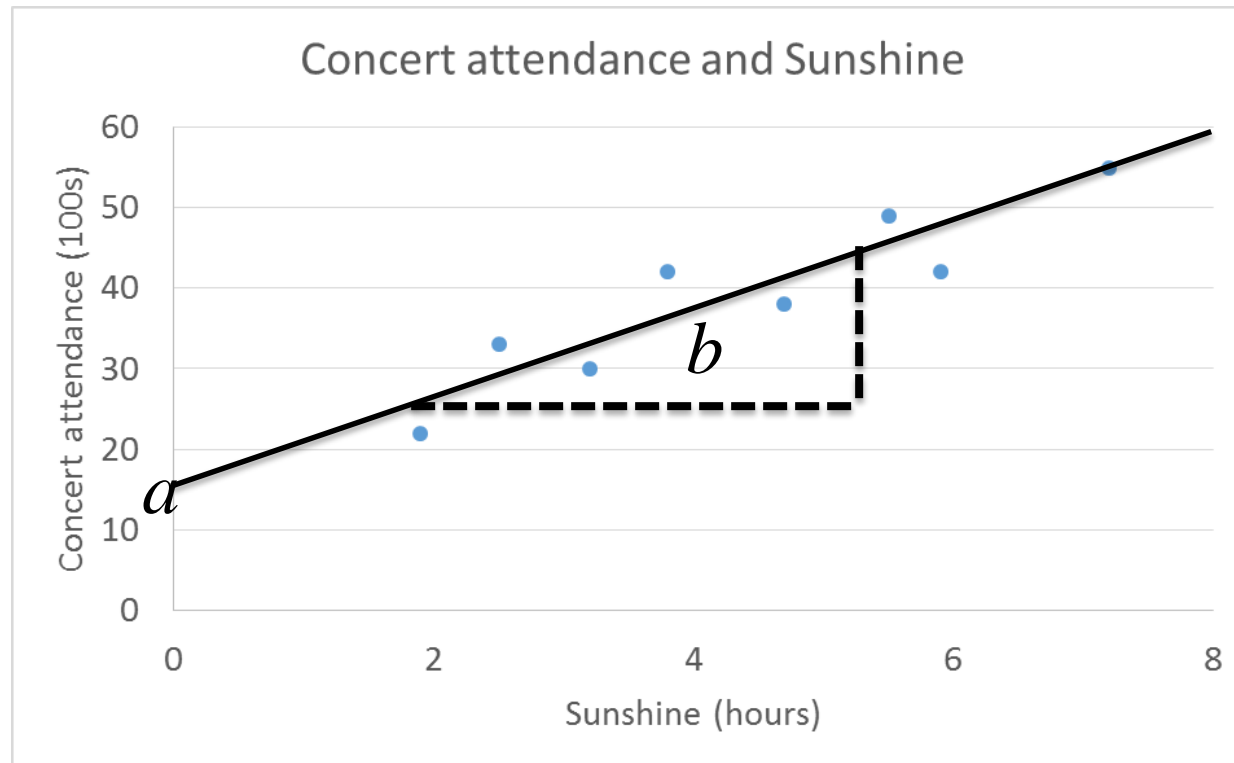
Sunshine (hours)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
Concert attendance (100s)	22	33	30	42	38	49	42	55

- Line of best fit



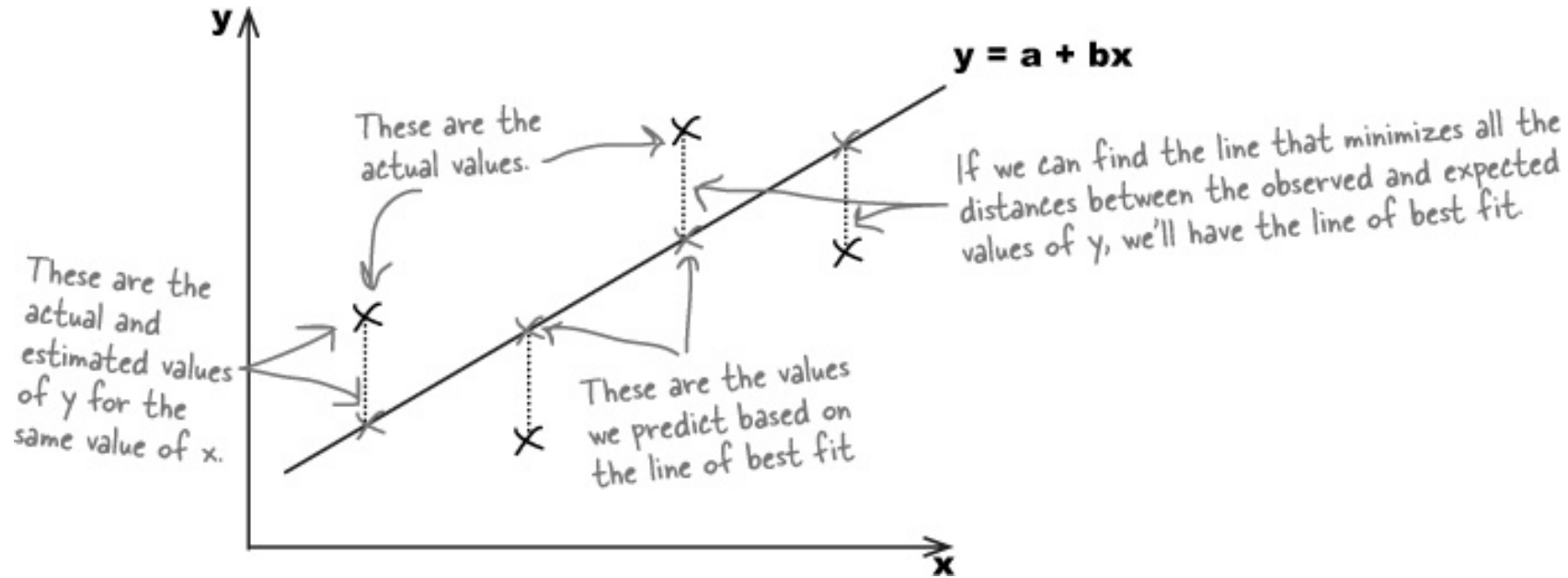


We need to find the equation of the line.



$$y = a + bx$$

We need to minimize errors.



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. But the problem, as seen earlier also, is that they will all cancel out.  $(y_i - \hat{y}_i)$  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.

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

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

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

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

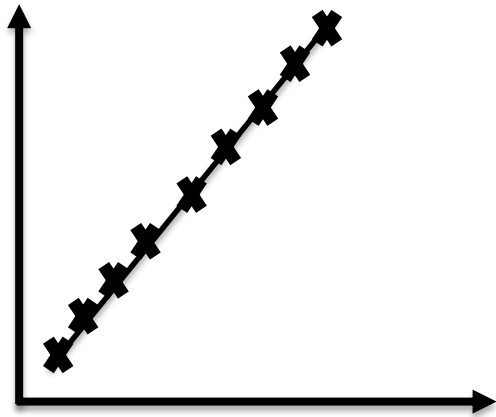
The value of  $b$ , the slope, that minimizes the SSE is given by

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

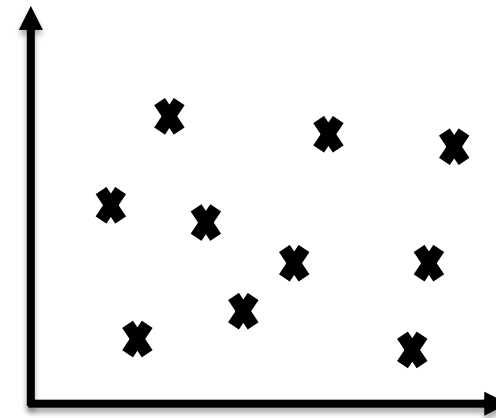
How do you calculate  $a$ . The line of best fit must pass through  $(\bar{x}, \bar{y})$ . Substituting in the equation  $y = a + bx$ , we can find  $a$ .

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

But how do you know how accurate this line is?



Accurate Linear  
Correlation

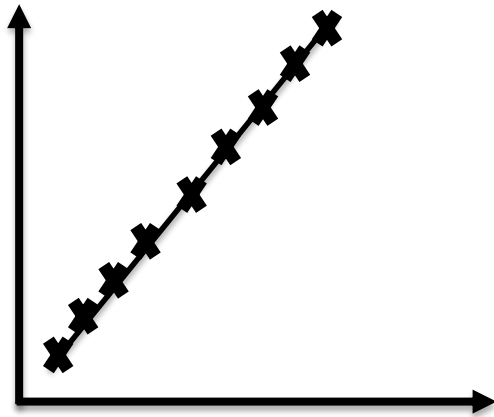


No Linear  
Correlation

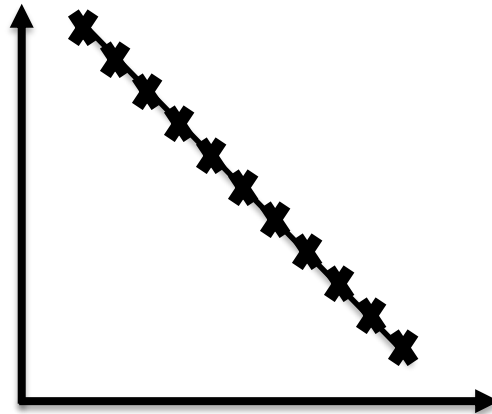
The fit of the line is given by **correlation coefficient**.

# Correlation Coefficient

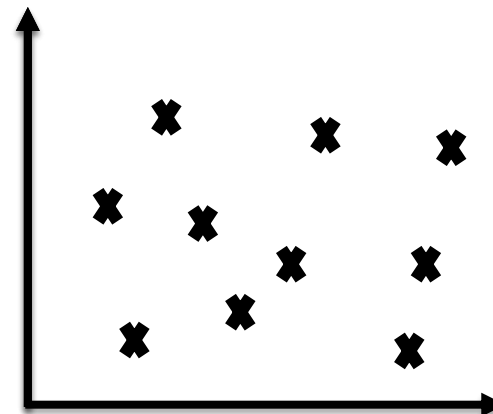
Correlation coefficient,  $r$ , is a number between -1 and 1 and tells us how well a regression line fits the data.



$r = 1$



$r = -1$



$r = 0$

# Correlation Coefficient

$r = \frac{bs_x}{s_y}$  where  $b$  is the slope of the line of best fit,  $s_x$  is the standard deviation of the  $x$  values in the sample, and  $s_y$  is the standard deviation of the  $y$  values in the sample.

$$s_x = \sqrt{\frac{\sum(x-\bar{x})^2}{n-1}} \text{ and } s_y = \sqrt{\frac{\sum(y-\bar{y})^2}{n-1}}.$$

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

Find  $r$  for this data.

# Correlation Coefficient and Covariance

$s_x^2 = \frac{\sum(x-\bar{x})^2}{n-1}$ ,  $s_y^2 = \frac{\sum(y-\bar{y})^2}{n-1}$ ,  $s_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}$ , where  $s_x^2$  is the sample variance of the  $x$  values,  $s_y^2$  is the sample variance of the  $y$  values and  $s_{xy}$  is the covariance.

$$b = \frac{s_{xy}}{s_x^2} \text{ and so, } r = \frac{s_{xy}}{s_x s_y}.$$



# Covariance

$$s_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{n-1}, r = \frac{s_{xy}}{s_x s_y}$$

- If both x and y are large distance away from their respective means, the resulting covariance will be even larger.
  - The value will be positive if both are below the mean or both are above.
  - If one is above and the other below, the covariance will be negative.
- If even one of the values is very close to the mean, the covariance will be small.
- $\text{Cov}(x,x)=\text{Var}(x)$

# Covariance and Correlation

$$s_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n - 1}, r = \frac{s_{xy}}{s_x s_y}$$

- The value of covariance itself doesn't say much. It only shows whether the variables are moving together (positive value) or opposite to each other (negative value).
- To know the strength of how the variables move together, covariance is standardized to the dimensionless quantity, correlation.

# Coefficient of Determination

The coefficient of determination is given by  $r^2$  or  $R^2$ . It is the percentage of variation in the  $y$  variable that is explainable by the  $x$  variable. For example, what percentage of the variation in open-air concert attendance is explainable by the number of hours of predicted sunshine.

If  $r^2 = 0$ , it means you can't predict the  $y$  value from the  $x$  value.

If  $r^2 = 1$ , it means you can predict the  $y$  value from the  $x$  value without any errors.

Usually,  $r^2$  is between these two extremes.

# Coefficient of Determination

$$r^2 = \left( \frac{s_{xy}}{s_x s_y} \right)^2$$

Another way of calculating the same is

$$r^2 = \frac{\sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$$

# Covariance, Correlation and $R^2$

How do the interest rates of federal funds and the commodities futures index co-vary and correlate?

Day	Interest Rate	Futures Index
1	7.43	221
2	7.48	222
3	8.00	226
4	7.75	225
5	7.60	224
6	7.63	223
7	7.68	223
8	7.67	226
9	7.59	226
10	8.07	235
11	8.03	233
12	8.00	241

# Covariance, Correlation and R<sup>2</sup>

Day	Interest Rate	Futures Index	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x}) * (y - \bar{y})$
1	7.43	221			
2	7.48	222			
3	8.00	226			
4	7.75	225			
5	7.60	224			
6	7.63	223			
7	7.68	223			
8	7.67	226			
9	7.59	226			
10	8.07	235			
11	8.03	233			
12	8.00	241			
<b>Mean</b>	<b>7.74</b>	<b>227.08</b>			
<b>StDev</b>	<b>0.22</b>	<b>6.07</b>			

$$Cov = \frac{12.216}{11} = 1.111$$

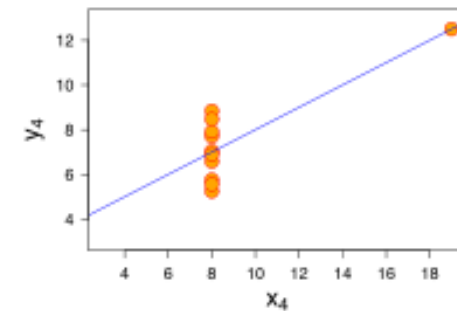
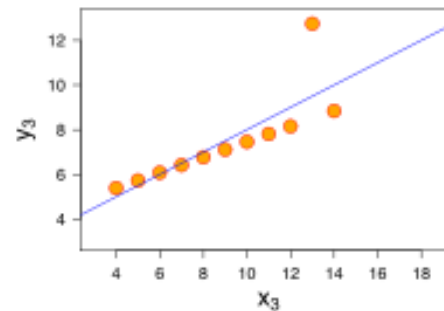
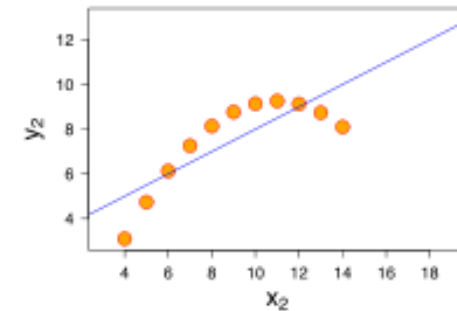
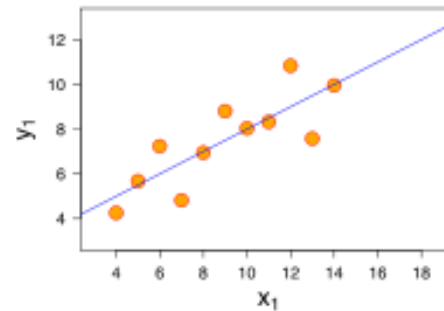
$$r = \frac{1.111}{0.22 * 6.07} = 0.815$$

$$R^2 = 0.815^2 = 0.665$$

# 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 understanding the data by getting an average value to describe it. When Mean didn't work, we went to Median and then to Mode.

We then found that along with the average, we need to understand the spread. We looked at Range, Interquartile Range, Variance and Standard Deviation.

We then looked at variety of ways this data is Distributed 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 because irrespective of our confidence, we could never be certain and need to prove our claims. Of course, there are errors in these tests too.

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

We then studied ANOVA and 2-sample z-test as a means of understanding significant differences between means.

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

## **International School of Engineering**

Plot 63/A, 1<sup>st</sup> Floor, 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>