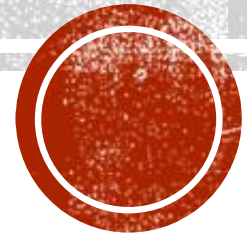


# Chi-sqr test, Corr, Cov, r-sqr

Shah Ayub Quadri

[Ayub.quadri89@gmail.com](mailto:Ayub.quadri89@gmail.com)



# Agenda

- Chi-square test
- Correlation
- Covariance
- Coefficient of determinant



# Common Test Statistics for Inferential Techniques

Inferential techniques (Confidence Intervals and Hypothesis Testing) most commonly use 4 test statistics:

- $z$
  - $t$
  - $\chi^2$  (Chi-squared)
  - $F$
- } Closely related to Sampling Distribution of **Means**
- } • Closely related to Sampling Distribution of **Variances**  
• Derived from Normal Distribution



# $\chi^2$ DISTRIBUTION

The distribution of Chi square ( $\chi^2$ ) statistic is called Chi square ( $\chi^2$ ) distribution

Chi square ( $\chi^2$ ) test represents a useful method of comparing the experimentally obtained result with those to be expected theoretically on some hypothesis.

Chi square ( $\chi^2$ ) hypothesis

*$H(0)$  No Difference between expected and observed values.*

*$H(1)$  Difference exists between observed and expected values.*



# $\chi^2$ distribution

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

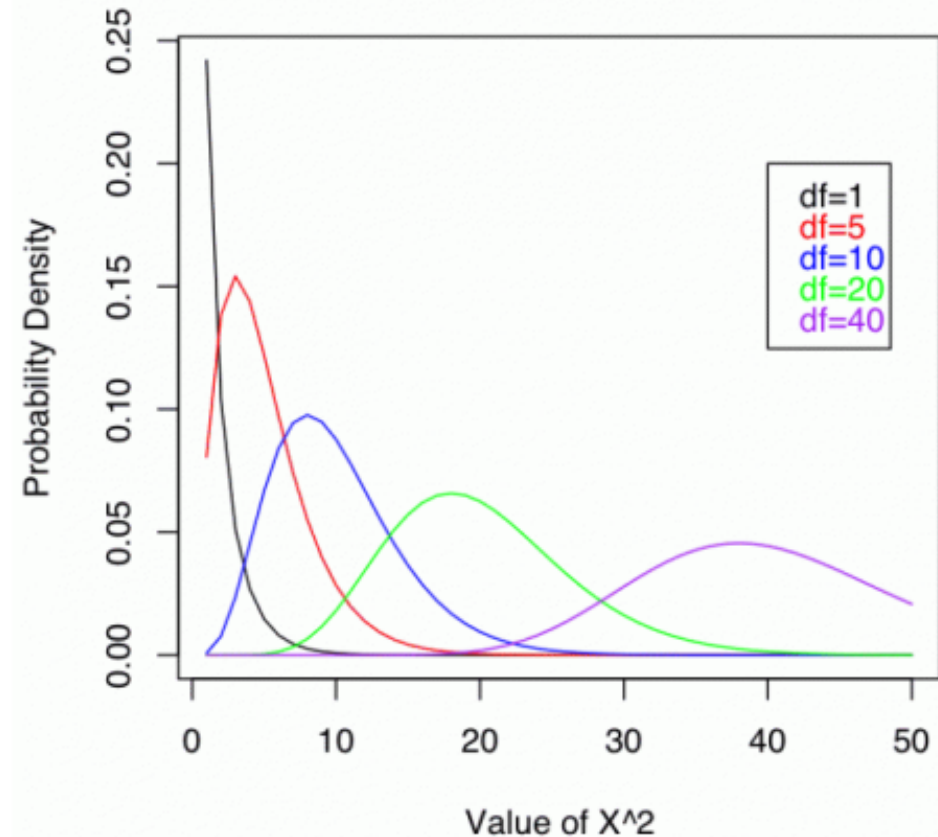
$$z^2 = \frac{(X - \mu)^2}{\sigma^2}$$

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

$\chi^2$  distribution is a distribution of the squared deviates.

The shape depends on number of squared deviates added together.

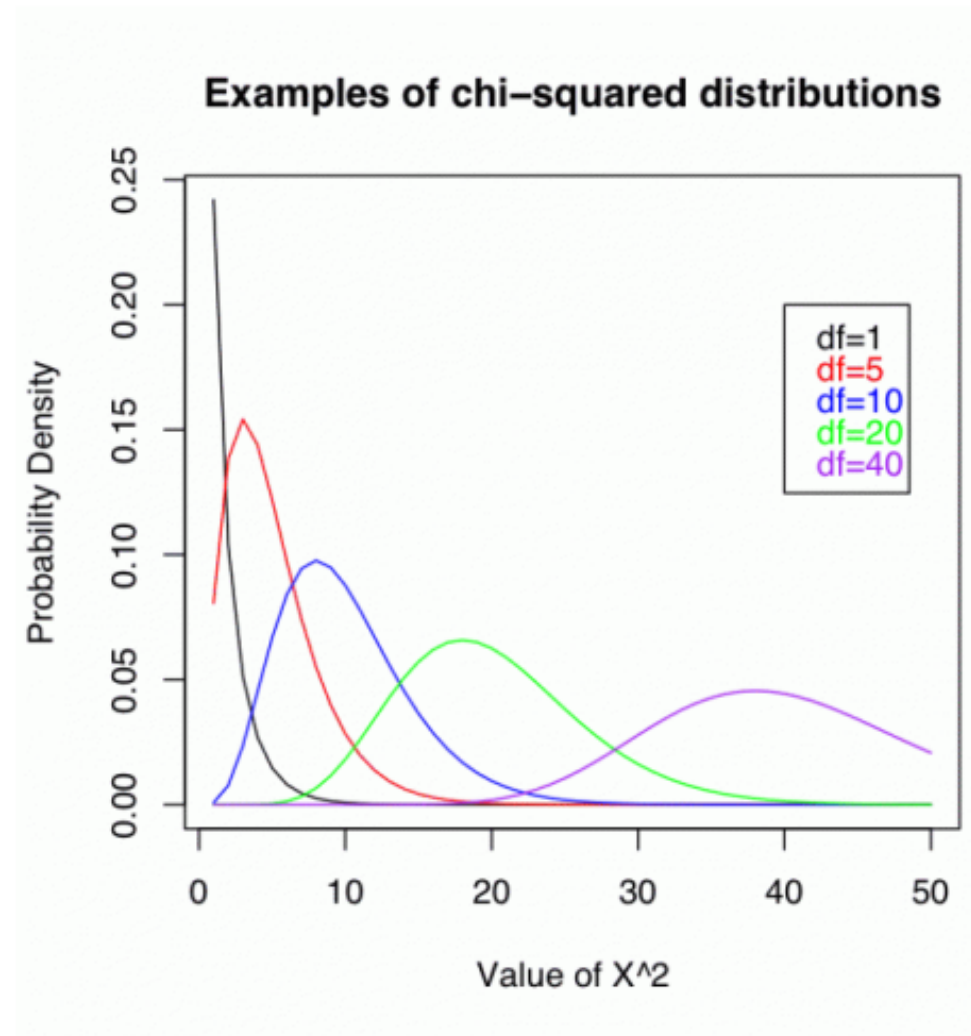
Examples of chi-squared distributions



# $\chi^2$ distribution

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

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



# 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.





## $\chi^2$ 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.





# Example – chi square test

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.

<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?



<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

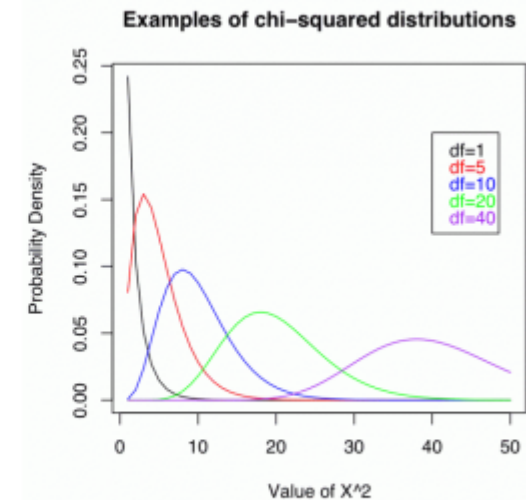
$$\chi^2 = 38.272$$

Is this high?

To find this, we need to look at the  $\chi^2$  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<sup>th</sup>. Therefore, there are  $5-1=4$  degrees of freedom.

$\nu = (\text{number of classes}) - (\text{number of restrictions}), \text{ or}$

$\nu = (\text{number of classes}) - 1 - (\text{number of parameters being estimated from sample data})$

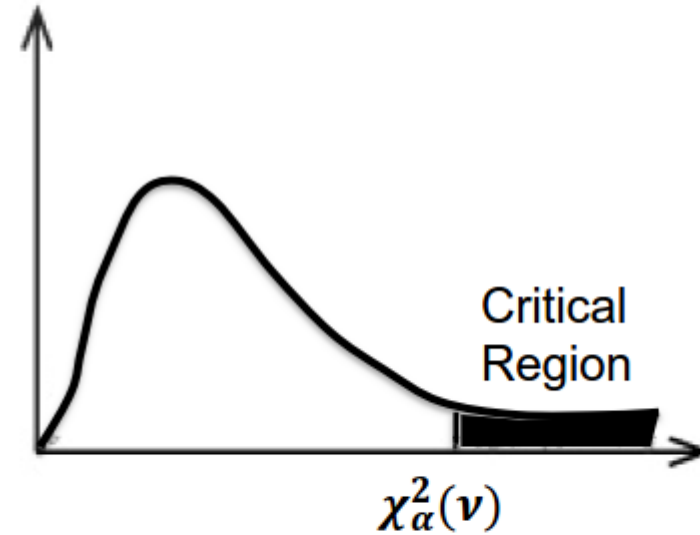


# How do we know the Significance of the difference?

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.



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

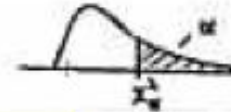
$$v = 4$$





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	0.70	0.60	0.50	0.05	0.025	0.02	0.01	0.005	0.001
$\nu$																
1	0.004393	0.00457	0.00468	0.00479	0.00490	0.00500	0.00510	0.00520	0.00530	0.00540	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	0.719	1.065	1.386	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	1.642	2.366	3.075	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	2.366	3.075	3.746	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	3.089	3.858	4.607	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	3.858	4.607	5.293	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	4.607	5.293	6.345	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	5.308	6.032	6.958	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	6.094	6.897	7.879	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	6.938	7.779	8.837	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	7.779	8.642	9.816	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	7.802	8.658	9.590	10.828	21.026	22.718	23.442	25.669	27.688	32.909

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



# **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?

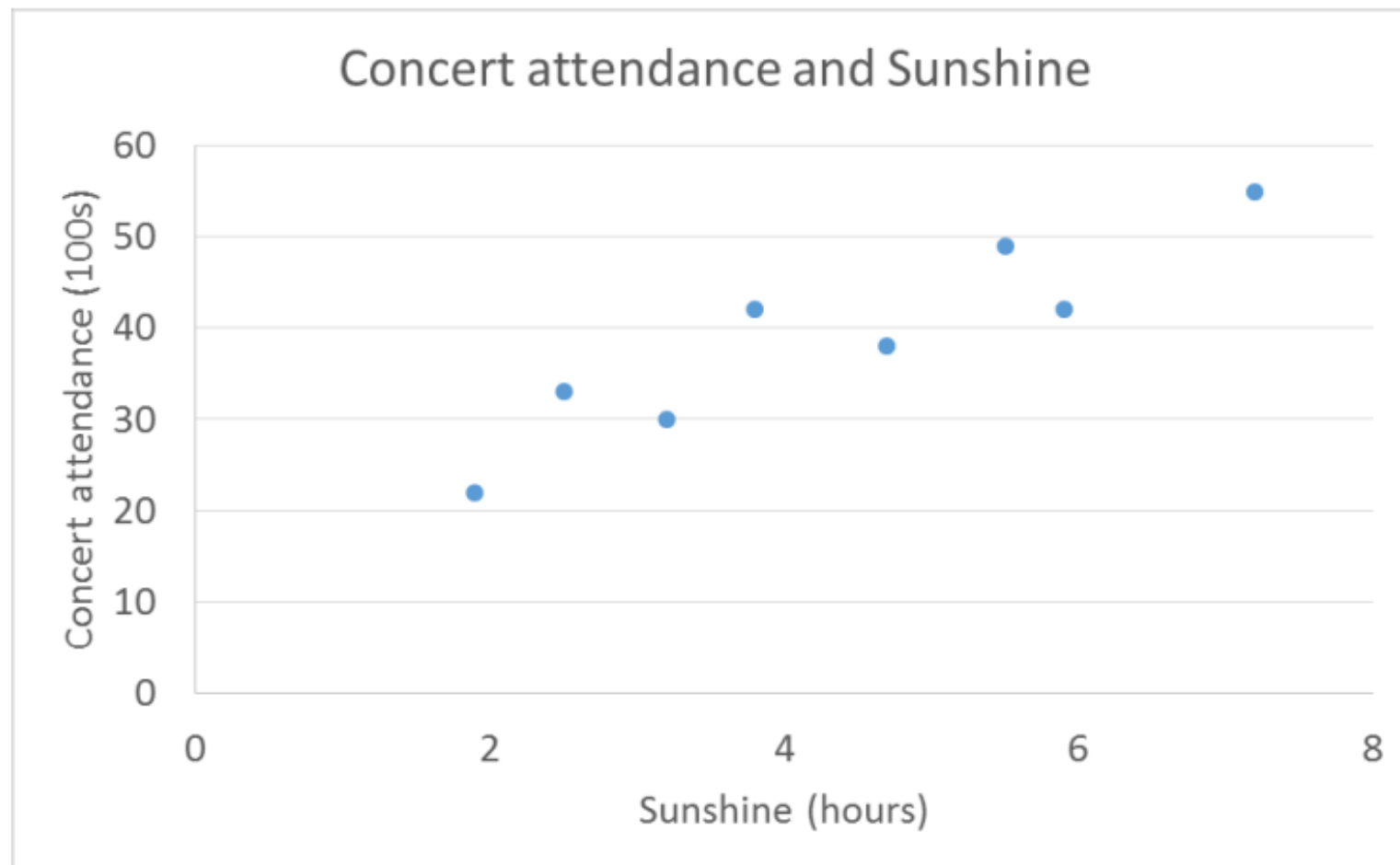
Are sunshine and concert attendance correlated?



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

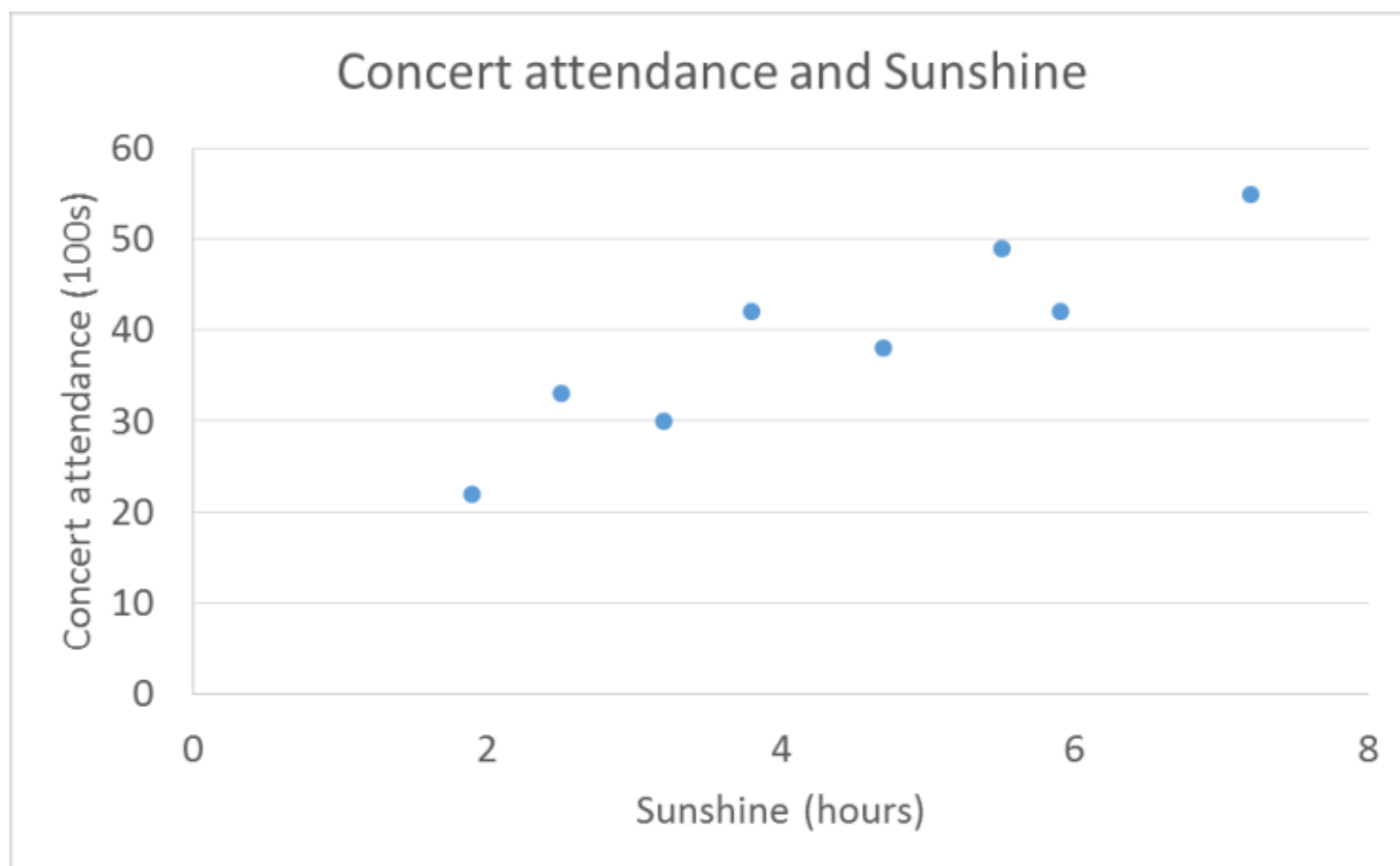
Dependent variable (response) – Concert attendance – Plotted on Y-axis



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



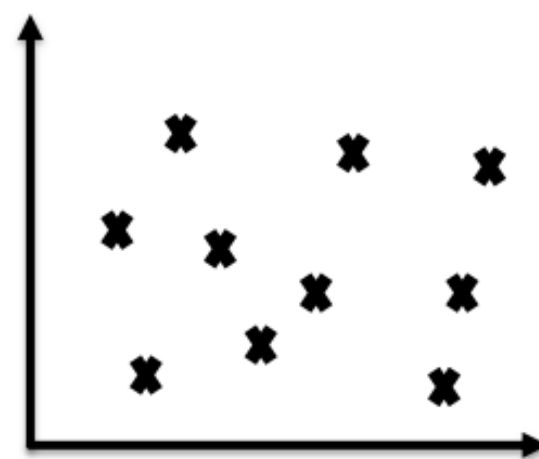
Hours of sunshine and concert attendance are correlated, i.e., in general, longer sunshine hours indicate higher attendance.



Positive Linear  
Correlation



Negative Linear  
Correlation



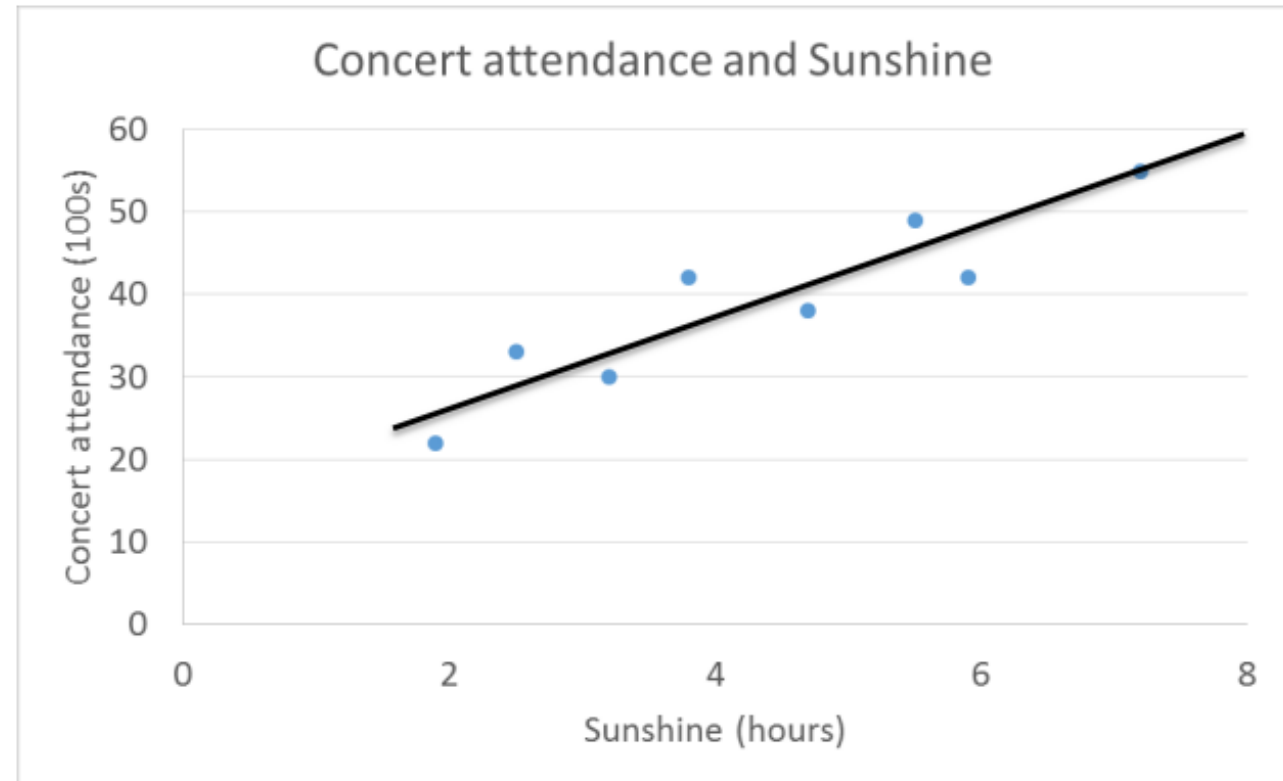
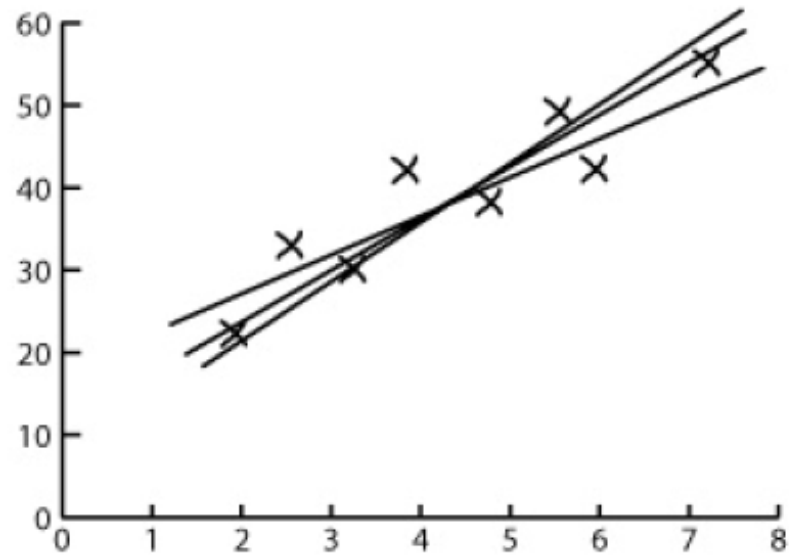
No Correlation





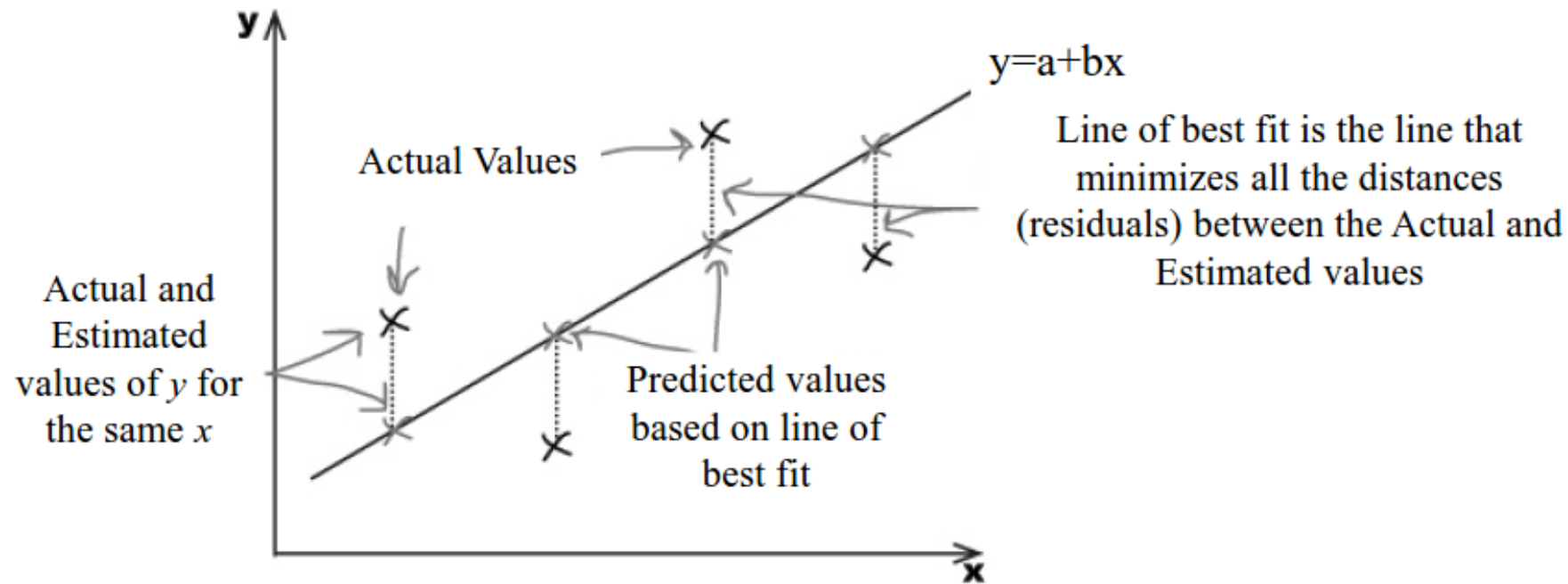
Sunshine (hours)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
Concert attendance (100s)	22	33	30	42	38	49	42	55

- Line of best fit





We need to 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.  $(y_i - \hat{y}_i)$  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

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



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  $a$ .

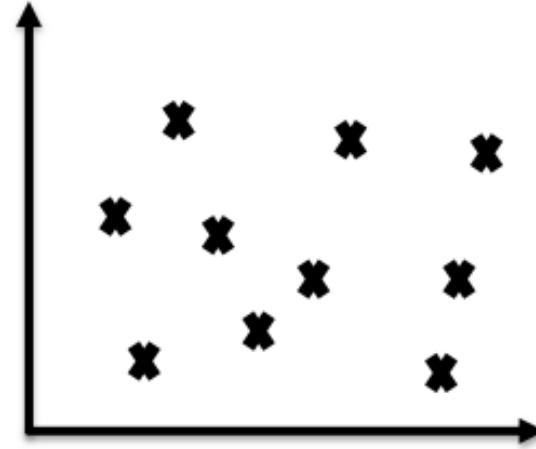
This method of fitting the line of best fit is called **least squares regression**.



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



Perfect Linear  
Correlation



No Linear  
Correlation

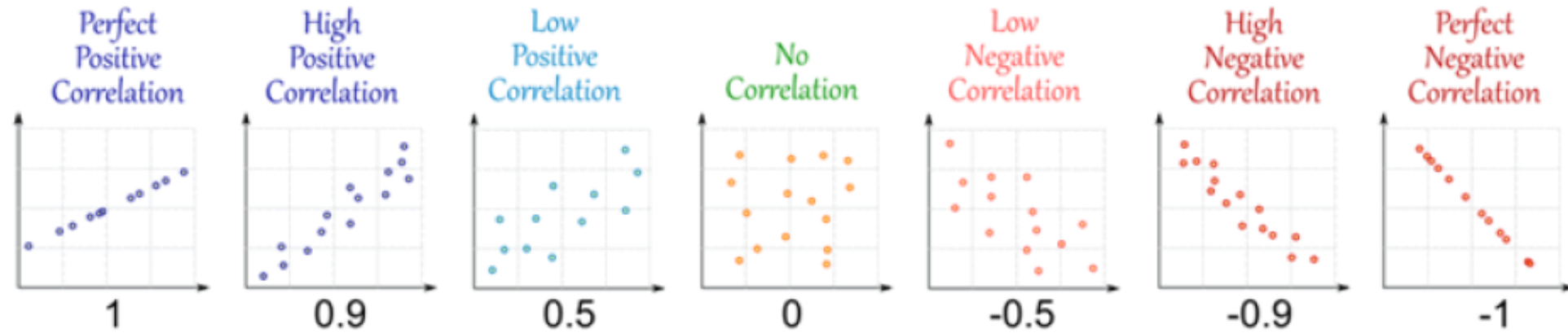
The fit of the line is given by **correlation coefficient**  $r$ .

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



# 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.



# 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.



# 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 them 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.



# Example – cov, corr, $r^2$

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

Month	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



Month	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>			

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

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

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

*Follow the R code for cov, cor*



# Resources

## Understanding Chi-square test

- <http://www.statisticshowto.com/probability-and-statistics/chi-square/>
- <http://www.yourarticlelibrary.com/project-reports/chi-square-test/chi-square-test-meaning-applications-and-uses-statistics/92394/>



# $\chi^2$ 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.



Possible Outcomes		Croupier A	Croupier B	Croupier C	Observed Results
	Win	43	49	22	
	Draw	8	2	5	
	Lose	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 of Win and A*

$$= \text{Grand Total} \times \frac{\text{Total Wins}}{\text{Grand Total}} \times \frac{\text{Total A}}{\text{Grand Total}} = \frac{\text{Total Wins} \times \text{Total A}}{\text{Grand Total}}$$

$$= \frac{\text{Row Total} \times \text{Column Total}}{\text{Grand Total}}$$



## $\chi^2$ 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	$(121 \cdot 98) / 250$	$(121 \cdot 95) / 250$	$(121 \cdot 57) / 250$



# $\chi^2$ 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
	$\sum O = 250$	$\sum E = 250$	$\sum \frac{(O - E)^2}{E} = 5.618$



## $\chi^2$ 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.

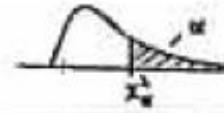




# $\chi^2$ independence test – Determine critical region

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

TABLE OF CHI-SQUARE DISTRIBUTION



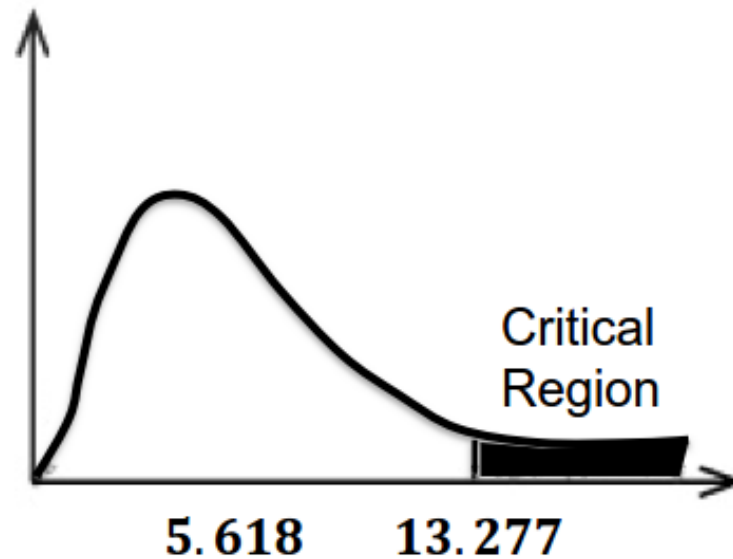
$\alpha$	0.995	0.99	0.98	0.975	0.95	0.90	0.80	0.70	0.60	0.50	0.40	0.30	0.20	0.10	0.05	0.025	0.01	0.005	0.001
$\nu$																			
1	0.004393	0.00457	0.00475	0.00492	0.00510	0.00538	0.00576	0.00625	0.00685	0.00756	0.00841	0.00943	0.01064	0.01213	0.01393	0.01607	0.01887	0.02204	0.02643
2	0.01000	0.02010	0.04004	0.05906	0.103	0.211	0.446	0.719	1.065	1.386	1.679	1.953	2.219	2.479	2.779	3.000	3.219	3.579	4.001
3	0.0717	0.115	0.185	0.216	0.352	0.584	1.005	1.642	2.365	3.075	3.841	4.605	5.405	6.251	7.378	8.535	9.841	11.345	13.815
4	0.207	0.297	0.429	0.484	0.711	1.064	1.649	2.365	3.075	3.841	4.605	5.405	6.251	7.378	8.535	9.841	11.345	12.838	16.268
5	0.412	0.554	0.752	0.831	1.145	1.610	2.343	3.000	3.819	4.601	5.401	6.251	7.289	8.535	9.841	11.345	12.838	14.860	18.465
6	0.676	0.872	1.134	1.237	1.635	2.204	3.070	3.822	4.603	5.403	6.251	7.289	8.535	9.841	11.345	12.838	14.860	16.750	20.517
7	0.989	1.239	1.564	1.690	2.167	2.833	3.822	4.603	5.403	6.251	7.289	8.535	9.841	11.345	12.838	14.860	16.750	18.548	22.457
8	1.344	1.646	2.032	2.180	2.733	3.490	4.594	5.403	6.251	7.289	8.535	9.841	11.345	12.838	14.860	16.750	18.548	20.278	24.322
9	1.735	2.088	2.532	2.700	3.325	4.168	5.380	6.251	7.289	8.535	9.841	11.345	12.838	14.860	16.750	18.548	20.278	21.955	26.125
10	2.156	2.558	3.059	3.247	3.940	4.865	6.179	7.289	8.535	9.841	11.345	12.838	14.860	16.750	18.548	20.278	21.955	23.589	27.877
11	2.603	3.053	3.609	3.816	4.575	5.578	6.989	8.183	9.590	11.017	12.592	14.566	16.581	18.559	20.367	22.161	23.919	25.588	29.588
12	3.074	3.571	4.128	4.348	5.191	6.213	7.682	8.933	10.371	11.813	13.277	15.338	17.275	19.216	21.024	22.718	24.352	26.217	31.264

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



## $\chi^2$ independence test – Decision

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







*Practice is the key to success*

