Chi-sq Test, Corr, Cov

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
- χ^2 (Chi-squared)
- F

Closely related to Sampling Distribution of **Means**

- Closely related to Sampling Distribution of **Variances**
- Derived from Normal Distribution

χ^2 DISTRIBUTION

The distribution of Chi square (X^2) statistic is called Chi square (X^2) distribution

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

Chi square (X^2) hypotheis

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

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

χ^2 distribution

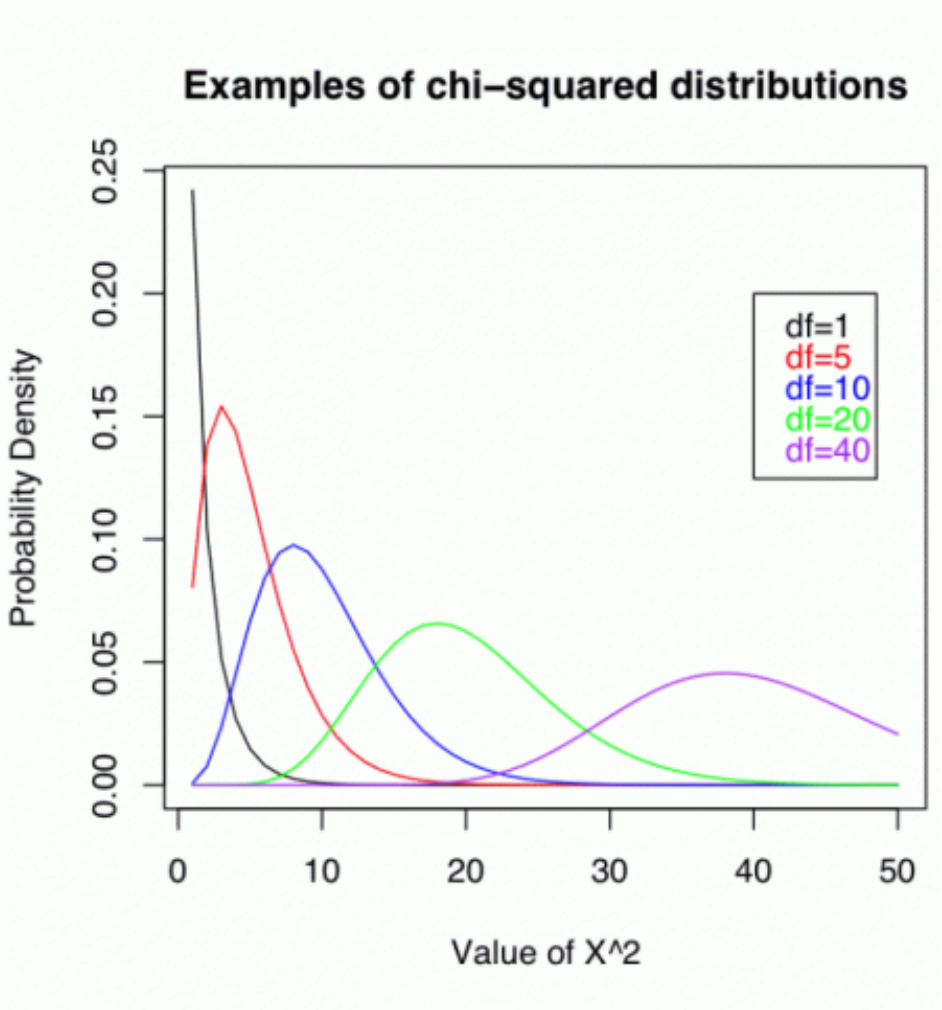
$$Recall\ z = \frac{X - \mu}{\sigma}$$

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

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

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

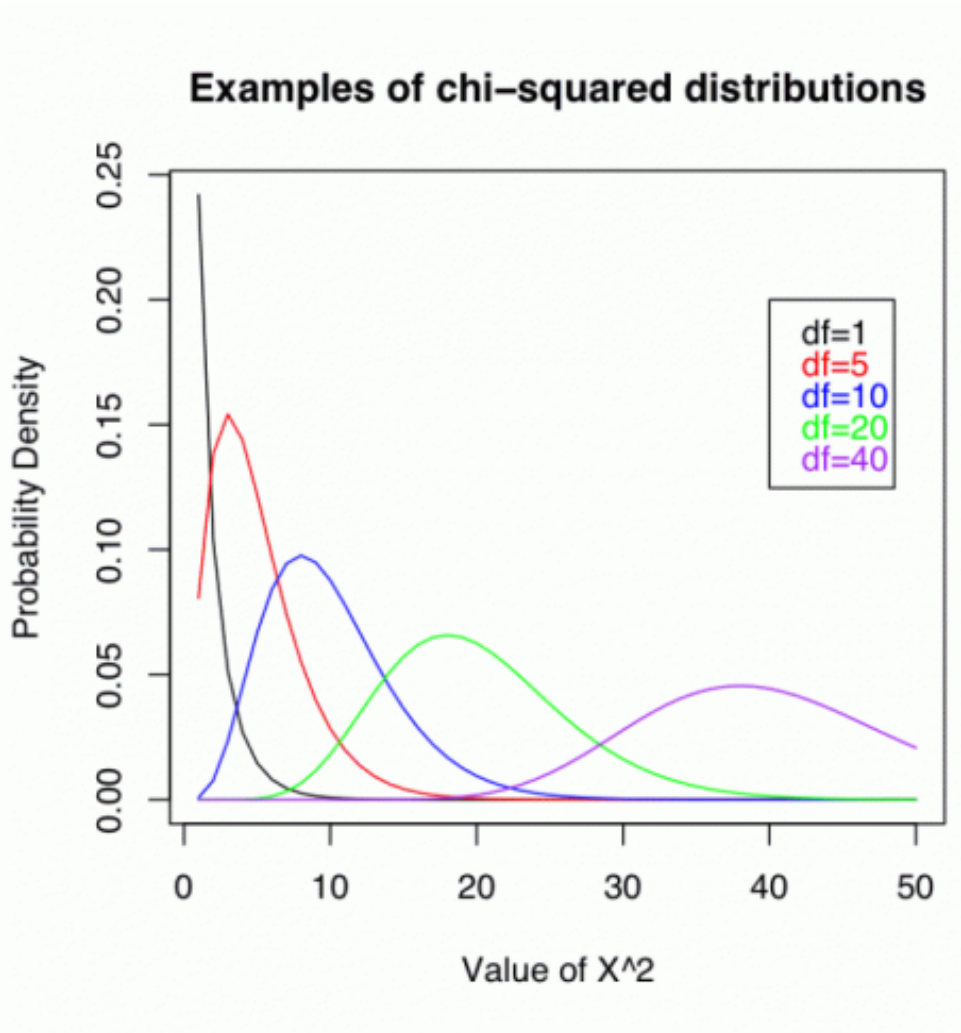
The shape depends on number of squared deviates added together.



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.

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.

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.

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?

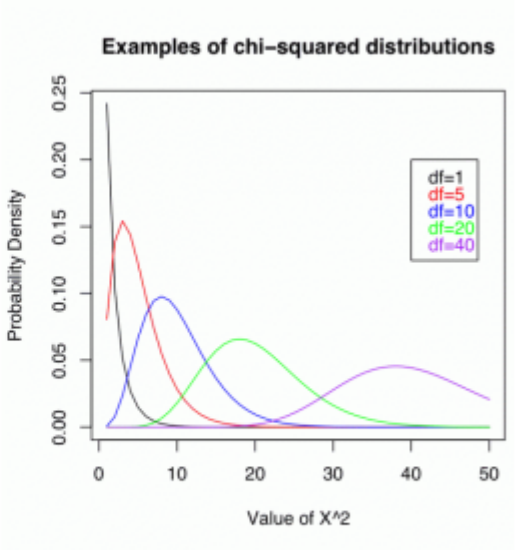
x	Observed Frequency	Expected Frequency
-2	965	977
23	10	8
48	9	8
73	9	6
98	7	1

$X^2 = 38.272$

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 5th. Therefore, there are 5-1=4 degrees of freedom.

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

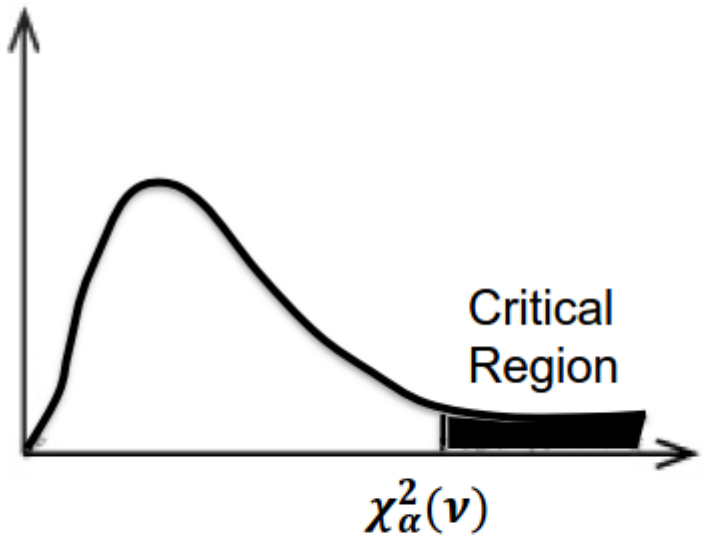
ν = (number of classes) – 1 – (number of parameters being estimated from sample data)

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.



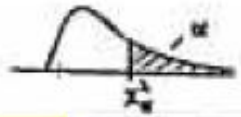
What are the expected frequencies and degrees of freedom?

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	21.026	23.337	24.054	26.217	28.300	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.

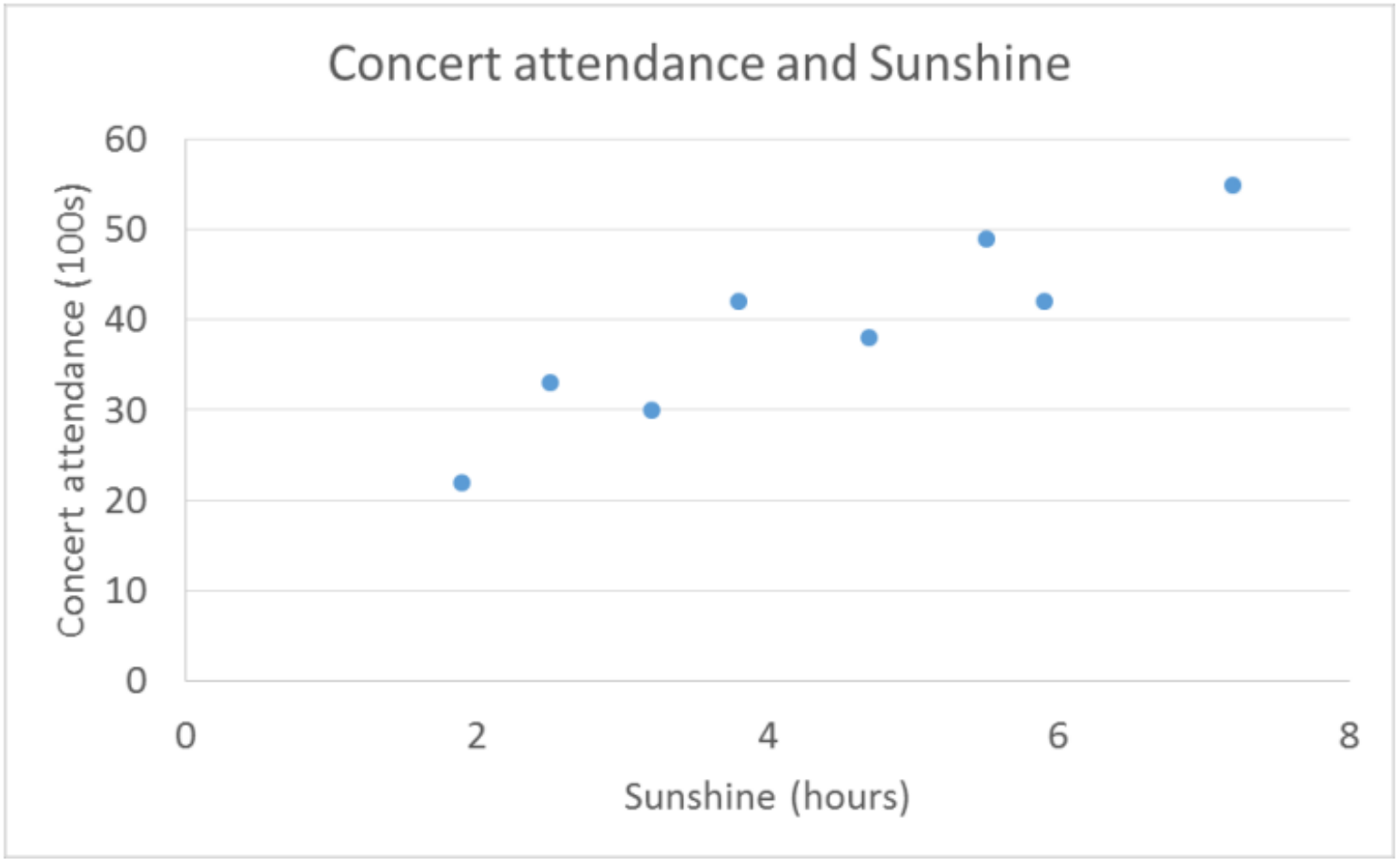
Based on predicted hours of sunshine, can we predict ticket sales?

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

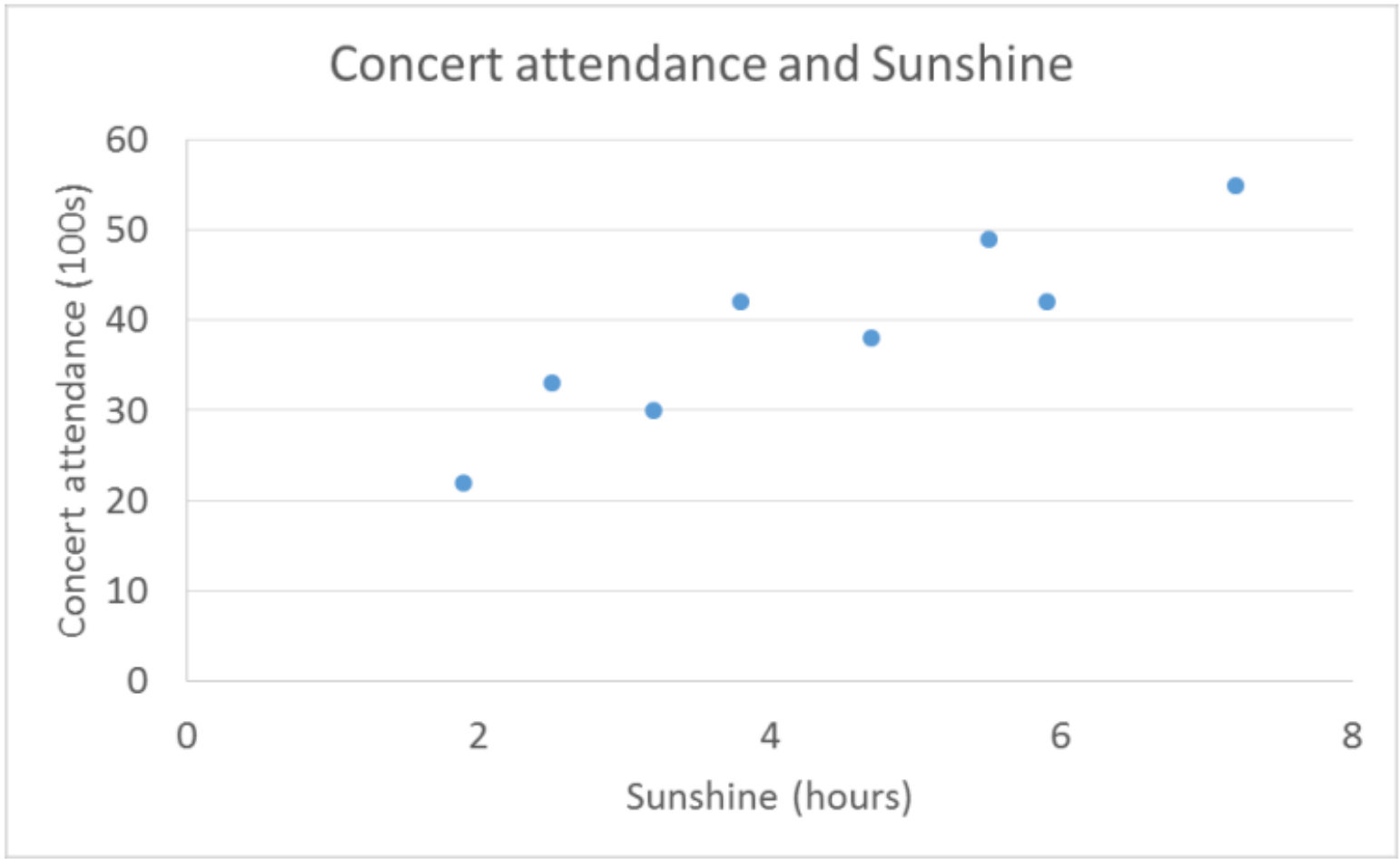
Dependent variable (response) – Concert attendance – Plotted on Y-axis



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

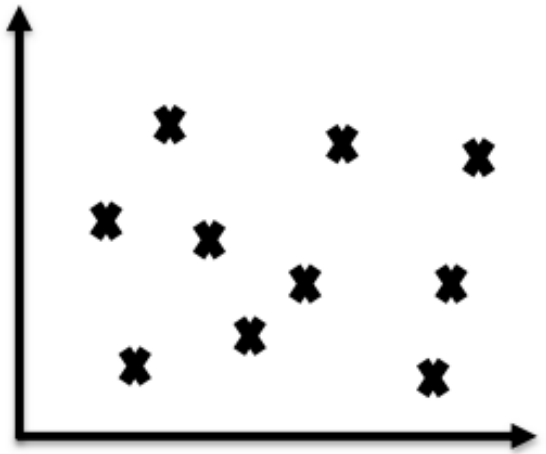
Dependent variable (response) – Concert attendance – Plotted on Y-axis



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

Positive Linear Correlation

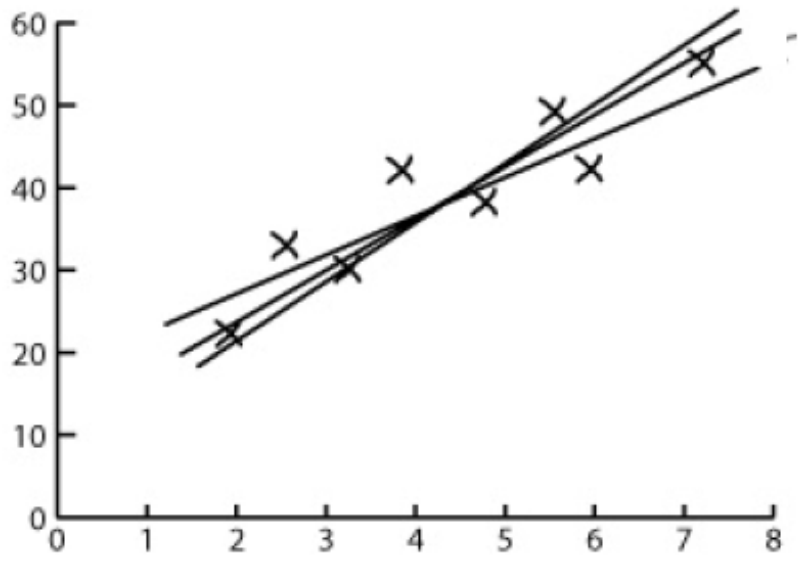
Negative Linear Correlation

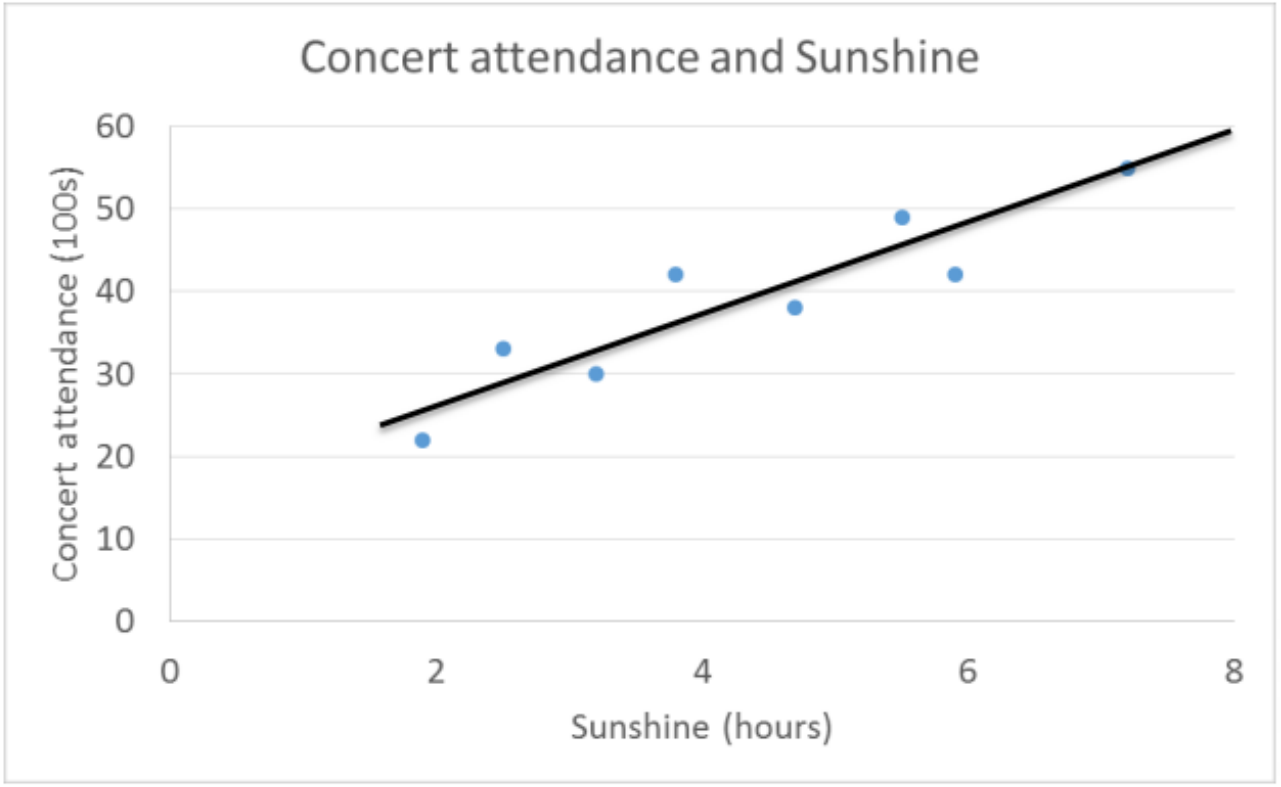


No Correlation

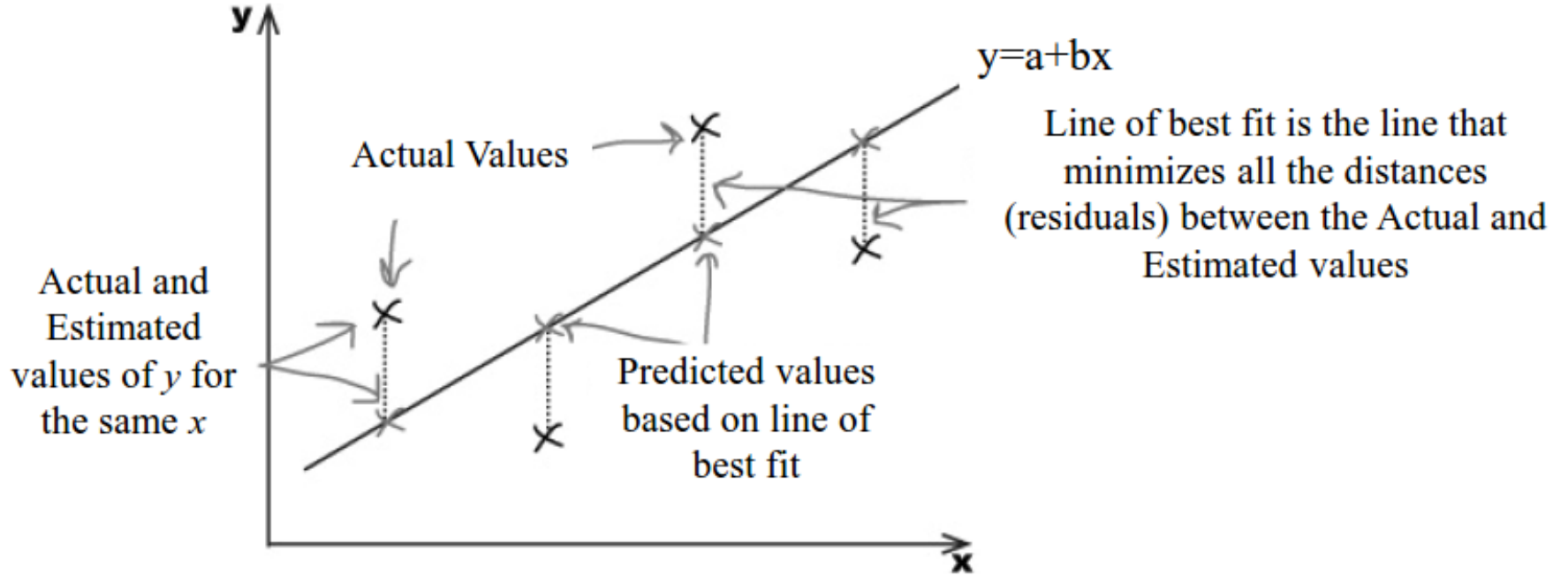
Sunshine (hours)	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
Concert attendance (100s)	22	33	30	42	38	49	42	55

- 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

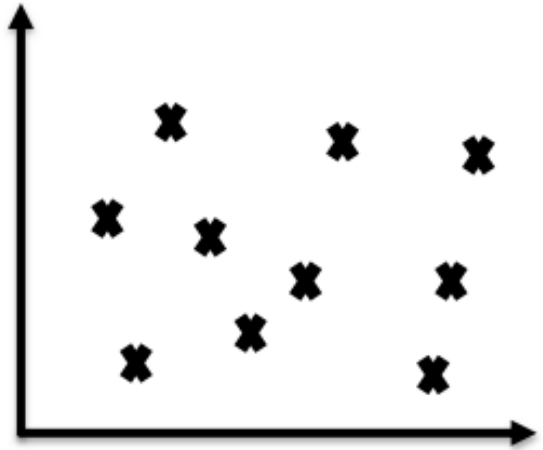
$$b = \frac{\sum((x - \bar{x})(y - \bar{y}))}{\sum(x - \bar{x})^2}$$

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



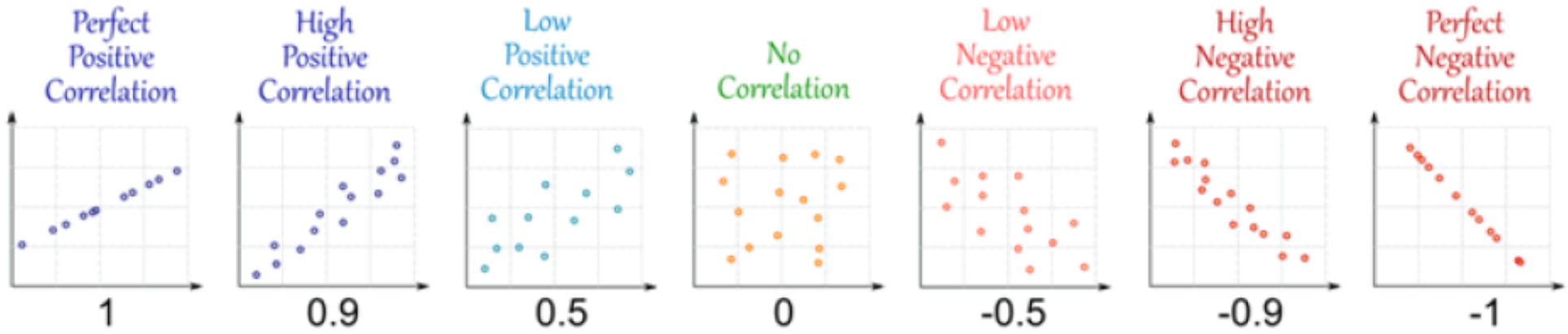
No Linear Correlation

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.
- Cov(x,x)=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			
Mean	**7.74**	**227.08**			
StDev	**0.22**	**6.07**			

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

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

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

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/

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

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

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

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

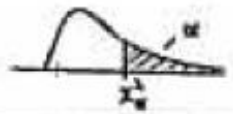
	Croupier A	Croupier B	Croupier C
Win			
Draw			
Lose			

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 $(m\text{-}1)(n\text{-}1)$ 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

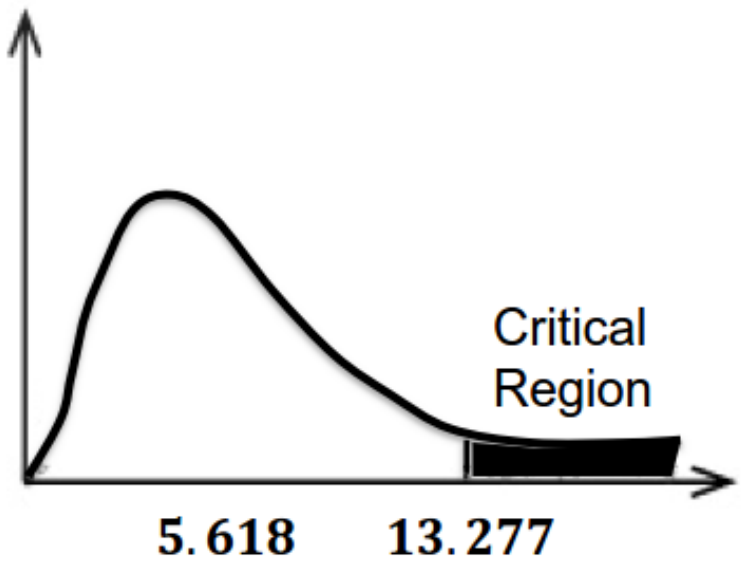


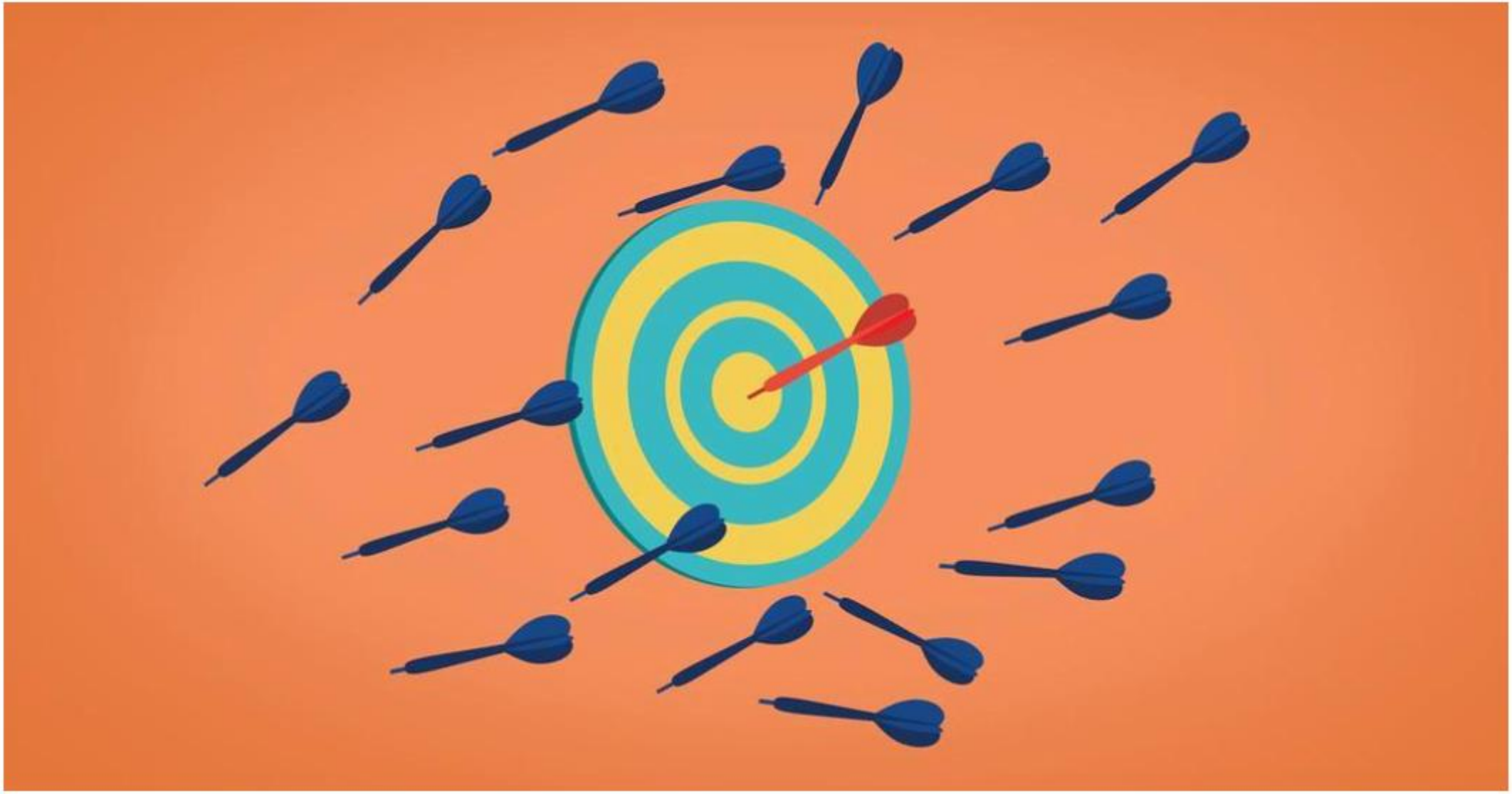
ν \ α	0.995	0.99	0.98	0.975	0.95	0.90	0.80	0.20	0.10	0.05	0.025	0.02	0.01	0.005	0.001
1	0.0^4393	0.0^3157	0.0^3628	0.0^3982	0.00393	0.0158	0.0642	1.642	2.706	3.841	5.024	5.412	6.635	7.879	10.827
2	0.0100	0.0201	0.0404	0.0506	0.103	0.211	0.446	3.219	4.605	5.991	7.378	7.824	9.210	10.597	13.815
3	0.0717	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	3.074	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 $X^2 > 13.277$.

χ^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