Quantitative Methods

Lecture-12

BITS Pilani
Pilani Campus

innovate achieve lead

BITS Pilani
Pilani Campus

Chi-square tests

(Ch 11 Business Statistics, Levine et al.)

Today’s session

- Chi-square tests.
- Goodness of fit.
- Simultaneously testing the difference between of two or more proportions.
- Test of independence.

The Concept behind chi-square test

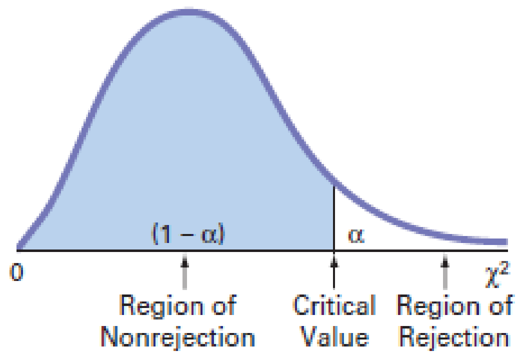
- A population “X” has certain proportions of defined attributes.
- Say, 50% of humans in the population are Females, and 50% are Male.
- You draw a representative sample (say n=50) from this population.
- What is the expected proportion of Males and Females in this sample?
- 50% each.
- What are the expected frequencies of Males and Females in this sample?
- 25 each. Why?
- Observed frequencies would be slightly different. Due to sampling error.
- However, the difference between observed and expected frequencies should be close to zero.

The concept behind chi-square test

- If you find that the differences between expected and observed frequency is close to zero.
- You can assume that the samples are from the same population.
- If you find that the differences are too large, you may conclude that they are not from the expected population.
- How large a difference is too large?
- That's where chi-square distribution comes to our help.
- $Test\ statistic$: χ^2_{STAT} (Chi-square statistic) $= \sum_{All\ cells} \frac{(f_o - f_e)^2}{f_e}$
- $\boldsymbol{f_o}$: $Observed\ frequency$, in the sample $\boldsymbol{f_e}$: $Expected\ frequency\ if\ the\ proportions\ were\ same$.
- **If null hypothesis of no difference is correct**, χ^2_{STAT} follows chi-square distribution with certain degrees of freedom (that we would discuss in each case).

The concept behind chi-square test

- $Test\ statistic$: χ^2_{STAT} (Chi-square statistic) $= \sum_{All\ cells} \frac{(f_o - f_e)^2}{f_e}$
- $\boldsymbol{f_o}$: $Observed\ frequency$, $\boldsymbol{f_e}$: $Expected\ frequency\ if\ the\ proportions\ were\ same$.
- **Null hypothesis. No difference in proportions. From same population etc.**
- **If null hypothesis of no difference is correct, χ^2_{STAT} follows chi-square distribution with df degrees of freedom**.
- If null hypothesis is true, what should be χ^2_{STAT}?
- It should be close to zero or in the region of nonrejection.
- If χ^2_{STAT} is in the region of rejection, we can reject the null of no difference.
- Notice it starts with 0 (the left most point).
- Rejection region means too far from zero.
- In Chi-square test, entire α probability region is in the right tail.



innovate achieve lead

BITS Pilani
Pilani Campus

Chi-Square test: Goodness of fit

Chi-square: Goodness of fit

Primary question

- Does the observed data fit assumed distribution.
- **Degrees of freedom: (**Number_of_values) – 1 – (no_of_estimated_parameters)
- no_of_estimated_parameters: In normal distribution you may have to estimate the mean and the standard deviation

The goodness of fit

Does the following sample represent a **"uniform distribution".** Are there equal proportion of each option/value in the population?

Option	Observed Frequency (fo)	Expected Frequency (fe)	Expected Frequency (fo-fe)^2	Expected Frequency (fo-fe)^2/fe
A	15	20	25	1.25
B	19	20	1	0.05
C	25	20	25	1.25
D	21	20	1	0.05
Total	**80**	**80**		**2.6**

$$\chi^2_{STAT} = \sum_{All\ cells} \frac{(f_o - f_e)^2}{f_e}$$

$$\chi^2_{STAT} = \mathbf{2.6}$$

Degrees of freedom:
Number of values – 1 – no_of_estimated_parameters

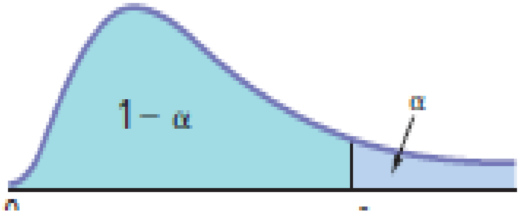
The goodness of fit

Option	Observed Frequency (fo)	Expected Frequency (fe)	Expected Frequency (fo-fe)^2	Expected Frequency (fo-fe)^2/fe
A	15	20	25	1.25
B	19	20	1	0.05
C	25	20	25	1.25
D	21	20	1	0.05
Total	**80**	**80**		**2.6**

$$\chi^2_{STAT} = \sum_{All\ cells} \frac{(f_o - f_e)^2}{f_e}$$

Critical Value:
7.815

$$\chi^2_{STAT} = 2.6$$



	Cumulative Probabilities											
	0.005	0.01	0.025	0.05	0.10	0.25	0.75	0.90	0.95	0.975	0.99	0.995
	Upper-Tail Areas (α)											
Degrees of Freedom	0.995	0.99	0.975	0.95	0.90	0.75	0.25	0.10	0.05	0.025	0.01	0.005
1			0.001	0.004	0.016	0.102	1.323	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	0.575	2.773	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	1.213	4.108	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	1.923	5.385	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	2.675	6.626	9.236	11.071	12.833	15.086	16.750

Degrees of freedom:
Number of values – 1 – no_of_estimated_parameters

Degrees of freedom:
4 – 1 – 0 = 3

Significance:
0.05

Rject H_0 if $\chi^2_{STAT} > \chi^2_{\alpha}$

Conclusion:
Data fits uniform distribution

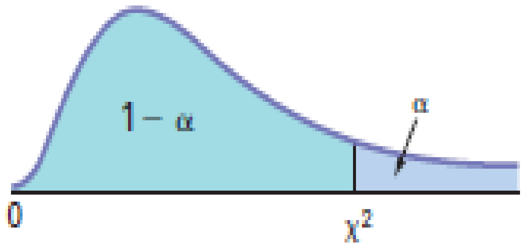
Goodness of fit examples: Fair dice?

- A dice is tossed 600 times and frequency of 1 to 6 is noted.
- Is it a fair die?
- Ho: Data comes from a fair dice population.
- How would you approach it?
- Observed frequency is given. What are the expected proportions?
- What are the expected frequencies: 100 each (why?)
- Degrees of freedom: 5 (why?)
- Calculate sample chi-square statistic and compare it with the critical value.

Rject H_0 if $\chi^2_{STAT} > \chi^2_{\alpha}$

Outcome	Frequency
1	80
2	105
3	100
4	95
5	100
6	120
Total	600

$$\chi^2_{STAT} = \sum_{All\ cells} \frac{(f_o - f_e)^2}{f_e}$$

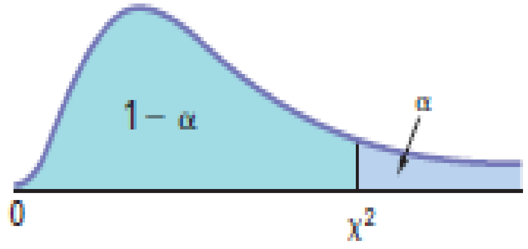


Goodness of fit examples: Normal distribution?

- Your coconut water shop recorded the time taken to serve the customers (minutes).
- Are time taken normally distributed?
- H_o: Data fits normal distribution.
- How will you approach it?
- Degrees of freedom.
- 7-1-2 = 4 (why?)
- Two parameters, μ and σ, need to be estimated to find expected frequencies.
- Find expected proportions for each bin (interval) (How?).
- Find expected frequencies for each interval (How?).
- Calculate the test statistic, critical value and test the hypothesis. (How?).

$$\chi^2{}_{STAT} = \sum_{All\ cells} \frac{(f_o - f_e)^2}{f_e}$$

minutes	Frequency
< 2	0
2-3	9
3-4	30
4-5	72
5-6	30
6-7	9
> 7	0
Total	150



Rject H_0 if $\chi^2{}_{STAT} > \chi^2{}_{\alpha}$

innovate achieve lead

BITS Pilani
Pilani Campus

Chi-square test: Difference in two proportions

Chi-square: Difference in two proportions

- You have two stores selling coconut water. You ran a satisfaction survey.
- You asked customers of each store: "Will you buy again?"
- You want to test if the proportion of customers willing to buy again is same for both the stores.
- H0: π1=π2
- H1: π1≠π2
- $Test\ statistic$: $\chi^2{}_{STAT}$ (Chi-square statistic) $= \sum_{All\ cells} \frac{(f_o - f_e)^2}{f_e}$
- **We need to answer two questions.**
- **Degrees of freedom**? 1 (Why?)
- How to calculate $\boldsymbol{f_e}$: $Expected\ frequency\ if\ the\ proportions\ were\ same$ at both the stores.

Difference in two proportions

Row	Group-1 (Store-1)	Group-2 (Store-2)	Total
Items of interest (yes)	X1	X2	X
Items of not interest (no)	n1 – X1	n2 – X2	n-X
Total sample	**n1**	**n2**	**n**

Row	Store-1	Store-2	Total
Buy-again? (Yes)	163	154	**317**
Buy-again? (No)	64	108	**172**
Total sample	**227**	**262**	**489**

- How to calculate f_e?
- Under null hypothesis, proportion of items of interest would be same.
- We can take proportion from store-1, store-2, or from the total. Which estimate is better?
- Overall estimated proportion of the item of interest: $\bar{p} = \frac{X}{n}$;
- X: Total number of items of interest, n: Total sample size.
- f_e = column_total* $\bar{p}$, for items of interest, and column_total* $(1 - \bar{p})$ for the items of not interest.

Chi-square: Difference in two proportions

Expected proportion $\bar{p} = \frac{X}{n} = \frac{total\_items\_of\_interest}{Total\_samplesize} = \frac{317}{489} = .648,\ and\ (1 - \overline{p}) = .352$

Row	Store-1 Observed	Store-1 Expected	Store-2 Observed	Store-2 Expected	Total
Buy-again? (Yes)	163	147.16	154	169.84	**317**
Buy-again? (No)	64	79.84	108	92.16	**172**
Total sample	**227**	**227**	**262**	**262**	**489**

- Store1: Total customers (sample)?:
- 227
- Expected proportion of the items of interest?
- 0.648
- Expected frequency?
- fe = 227*0.648 = 147.16

Chi-square: Difference in two proportions

Row	Store-1 Observed	Store-1 Expected	Store-2 Observed	Store-2 Expected	Total
Buy-again? (Yes)	163	227*.648 = 147.16	154	169.84	**X=317** $\bar{p}$**=.648**
Buy-again? (No)	64	227*.352 = 79.84	108	92.16	**172** $(1-\overline{p}) = .352$
Total sample	**227**	**227**	**262**	**262**	**489**

fo	fe	fo-fe	(fo-fe)^2	(fo-fe)^2/fe
163.00	147.16	15.84	250.91	1.70
154.00	169.84	-15.84	250.91	1.48
64.00	79.84	-15.84	250.91	3.14
108.00	92.16	15.84	250.91	2.72
			Total	**9.05**

Chi-square: Difference in two proportions

- $Test\ statistic$: χ^2_{STAT} (Chi-square statistic) $= \sum_{All\ cells} \frac{(f_o - f_e)^2}{f_e}$
- $\boldsymbol{f_o}: Observed\ frequency, \boldsymbol{f_e}: Expected\ frequency\ if\ the\ proportions\ were\ same.$
- **If null hypothesis is correct, $\boldsymbol{\chi^2_{STAT}}$ follows chi-square distribution with 1 degree of freedom**.
- Degrees of freedom = (nrows-1)*(ncols-1) if the data is presented in a contingency table.
- Row and Column totals are fixed and hence one degree of freedom is lost in each.
- Mechanics of hypothesis testing remains the same.
- Calculate the test statistic for our data (χ^2_{STAT}) for our data.
- Compare it with the cutoff values of χ^2 distribution with 1 degree of freedom for a given α.
- Reject null of the χ^2_{STAT} is in the region of rejection, or p-value is less than alpha.

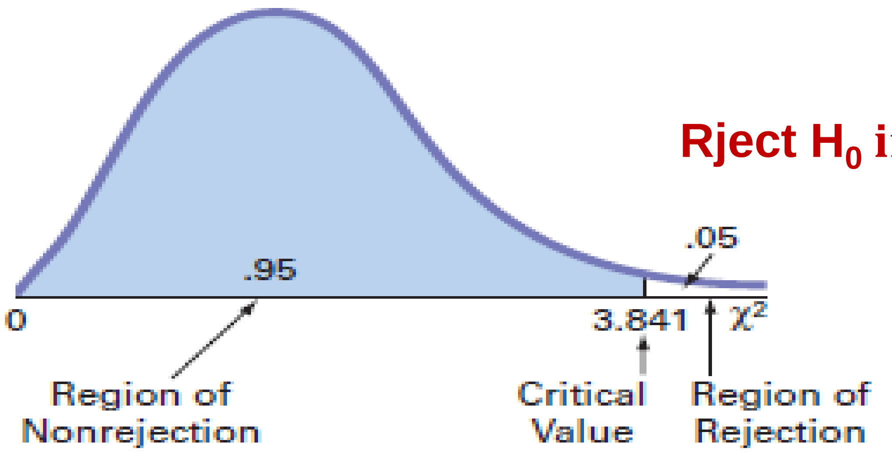
Chi-square: Difference in two proportions

- $\chi^2{}_{STAT} = \sum_{All\ cells} \frac{(f_o - f_e)^2}{f_e}$

fo	fe	fo-fe	(fo-fe)^2	(fo-fe)^2/fe
163.00	147.16	15.84	250.91	1.70
154.00	169.84	-15.84	250.91	1.48
64.00	79.84	-15.84	250.91	3.14
108.00	92.16	15.84	250.91	2.72
			Total	**9.05**

Should we reject the null hypothesis?

	Cumulative Probabilities						
	.005	.01	. . .	.95	.975	.99	.995
	Upper-Tail Area						
Degrees of Freedom	.995	.99	. . .	.05	.025	.01	.005
1			. . .	3.841	5.024	6.635	7.879
2	0.010	0.020	. . .	5.991	7.378	9.210	10.597
3	0.072	0.115	. . .	7.815	9.348	11.345	12.838
4	0.207	0.297	. . .	9.488	11.143	13.277	14.860
[illegible]	[illegible]	[illegible]		[illegible]	[illegible]	15.086	16.750

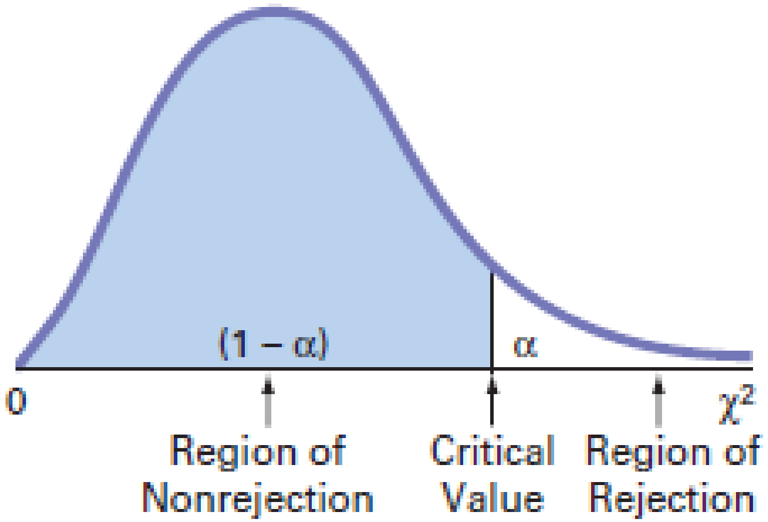


Rject H_0 if $\chi^2{}_{STAT} > \chi^2{}_{\alpha}$

Conclusion: Proportion of customers who are willing to buy again is different in each store.

Chi-square: Difference in two proportions

- Rject H_0 if $\chi^2_{STAT} > \chi^2_{\alpha}$
- Otherwise, do not reject the null hypothesis.
- If null is true, expected and actual frequencies would be equal.
- χ^2_{STAT} should be close to zero.
- Notice (the left boundary of the distribution is zero).
- The entire α probability region is in the right tail.
- Too far from zero means the proportions are different.



innovate achieve lead

BITS Pilani
Pilani Campus

Chi-square test: Difference in more than two proportions

Chi-square: Difference in more than two proportions

- H0: π1 = π2 = π3 =... = πc
- H1: at least one proportion is different from the others.

- Now the contingency table has two rows and more than two columns ('c' columns).
- $Test\ statistic$: χ^2_{STAT} (Chi-square statistic) $= \sum_{All\ cells} \frac{(f_o - f_e)^2}{f_e}$
- $\boldsymbol{f_o}$: $Observed\ frequency$, $\boldsymbol{f_e}$: $Expected\ frequency\ if\ the$ null hypothesis is true.
- Overall estimated proportion of items of interest $\bar{p} = \frac{X}{n}$, X = total number of items of interest, n = total sample size.
- **If null hypothesis is correct, χ^2_{STAT} follows chi-square distribution with c-1 degrees of freedom (why?).**
- Mechanics of the hypothesis testing remains the same.

Chi-square: Difference in more than two proportions

- Assuming we have three stores selling coconut water.

H0: π1 = π2 = π3

(Proportoin of customers willing to buy again is same in all the stores)

Row	Store-1 Observed	Store-2 Observed	Store-3 Observed	Total
Buy-again? (Yes)	128	199	186	**X=513**
Buy-again? (No)	88	33	66	**187**
Total sample	**216**	**232**	**252**	**700**

$$Expected\ proportion: \bar{p} = \frac{X}{n} = \frac{total\_items\_of\_interest}{Total\_samplesize} = \frac{513}{700} = .733,$$

$$and\ (1 - \overline{p}) = .267$$

Frequency table and chi-square statistic

fo	fe	(fo-fe)^2	(fo-fe)^2/fe
128	158.30	917.92	5.80
88	57.70	917.92	15.91
199	170.02	839.67	4.94
33	61.98	839.68	13.55
186	184.68	1.74	0.01
66	67.32	1.74	0.03
			Total (χ^2) 40.23

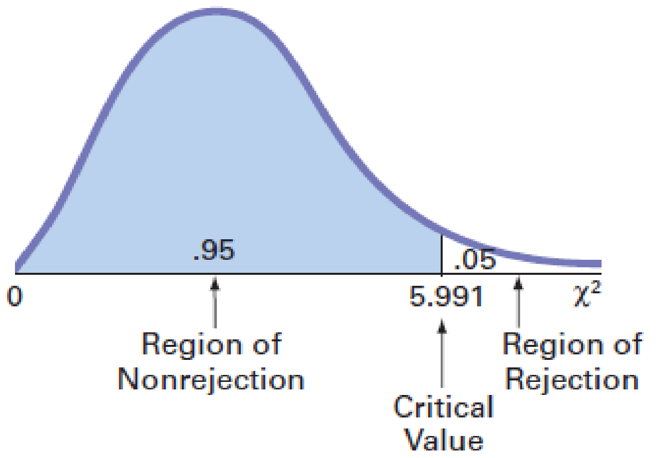
$Test\ statistic$: χ^2_{STAT} (Chi-square statistic) $= \sum_{All\ cells} \frac{(f_o - f_e)^2}{f_e}$

Chi-square: Difference in more than two proportions

Row	Store-1 Observed	Store-1 Expected	Store-2 Observed	Store-2 Expected	Store-3 Observed	Store3 Expected	Total
Buy-again? (Yes)	128	158.30	199	170.02	186	184.68	**X=513**
Buy-again? (No)	88	57.70	33	61.98	66	92.16	**187**
Total sample	**216**	**216**	**232**	**232**	**252**	**252**	**700**

- Degrees of freedom?
- (n_rows-1)*(n_cols-1) = 2
- $\chi_{STAT}{}^{2}$= 40.23
- Critical value $\chi_{\alpha=.05,df=2}{}^{2}$= 5.991   - XLS formula: CHISQ.INV.RT(α, df)
- **Should we reject the null hypothesis?**

Rject H_0 if $\chi^2{}_{STAT}$ > $\chi^2{}_{\alpha}$



Chi-square more than two proportions: Important notes

- For accurate results:
- At least 20% of the cells with expected frequency of 5 or more.
- Expected frequency in each cell must be at least 1.
- Columns can be merged if this criteria is not met.

Chi-square test: Test of independence

Chi-square: Test of independence

- Your coconut stores survey also had one additional question.

If you do not plan to purchase again, please let us know why?

1. Price?
2. Location?
3. Staff Behavior?
4. Others?

- Now you would like test if the reason for not buying again is independent of the store?
- In other words, are the proportion of these reasons same across the stores?

Chi-square: Test of independence

- H0: Two categorical variables are independent (there is no relationship between them).
- H1: Two categorical variables are dependent. (there is a relationship between them).

In the test of proportions

- We have one factor with two or more levels (stores).
- We test whether proportion of items of interest is same across all levels (stores).

In the test of independence

- We have two factors (“stores” and “the reason for not buying again”). Each factor has two or more levels.

Chi-square: Test of independence

- Assuming we have three stores selling coconut water.

Why will not buy again?	Store-1 Observed	Store-2 Observed	Store-3 Observed	Total
Price	23	7	37	**67**
Location	39	13	8	**60**
Staff behavior	13	5	13	**31**
Others	13	8	8	**29**
Total	88	33	66	**187**

- H0: "Reasons to not buy again" is independent of the stores (there is no relationship).
- H1: "Reasons to not buy again" is dependent on the stores (there is a relationship between them).

Chi-square: Test of independence

- H0: Two catergorical variables are independent (there is no relationship between them).
- H1: Two categorical variables are dependent. (there is a relationship between them).
- Now the contingency table has more than two rows and more than two columns (‘c’ columns).
- $Test\ statistic$: $\chi^2{}_{STAT}$ (Chi-square statistic) = $\sum_{All\ cells} \frac{(f_o - f_e)^2}{f_e}$
- $\boldsymbol{f_o}$: $Observed\ frequency$, $\boldsymbol{f_e}$: $Expected\ frequency\ if\ the$ null hypothesis is true.
- **If null hypothesis is correct, $\chi^2{}_{STAT}$ follows chi-square distribution,**
- With (number of rows – 1)*(number of columns – 1) degrees of freedom.
- All the other mechanics of the hypothesis testing remains the same.

Chi-square: Test of independence

Why will not buy again?	Store-1 Observed	Store-2 Observed	Store-3 Observed	Total
Price	23	7	37	**67**
Location	39	13	8	**60**
Staff behavior	13	5	13	**31**
Others	13	8	8	**29**
Total	88	33	66	**187**

Calculating the expected frequency

- Expected proportion: P(A and B) = P(A)*P(B) [Under null hypothesis of independence]
- P(Price and Store_1) = P(Price)*P(Store_1) = (67/187)*(88/187) = 0.169
- Expected frequency (first cell): 0.169*187 = 31.53

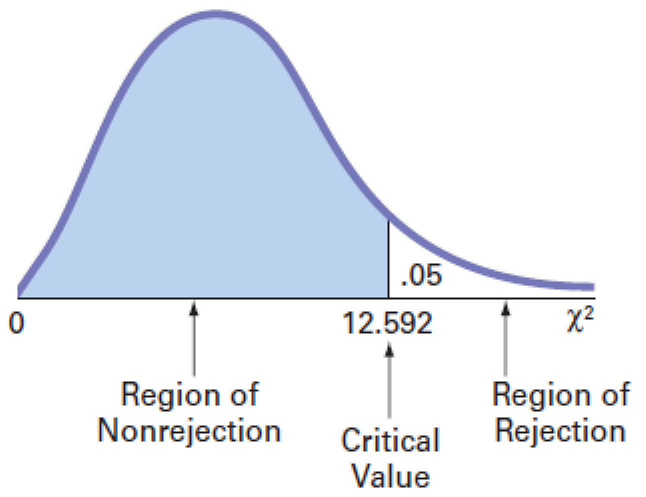
Degrees of freedom?

- Df = (nrows – 1)*(ncols – 1) = 3*2 = 6

Chi-square: Test of independence

Cell	fo	fe	(fo-fe)	(fo-fe)^2	(fo-fe)^2/fe
Store1/Price	23	31.53	-8.53	72.76	2.31
Store2/Price	7	11.82	-4.82	23.23	1.97
Store3/Price	37	23.65	13.35	178.22	7.54
Store1/Location	39	28.24	10.76	115.78	4.1
Store2/Location	13	10.59	2.41	5.81	0.55
Store3/Location	8	21.18	-13.18	173.71	8.2
Store1/Staff	13	14.59	-1.59	2.53	0.17
Store2/Staff	5	5.47	-0.47	0.22	0.04
Store3/Staff	13	10.94	2.06	4.24	0.39
Store1/Others	13	13.65	-0.65	0.42	0.03
Store2/Others	8	5.12	2.88	8.29	1.62
Store3/Others	8	10.24	-2.24	5.02	0.49
				Total	27.41

Stores and Reasons for not buying again are not independent



- Test Chi-Square Statistic: 27.41
- Chi-square cut-off (df: 6 and alpha: .05): 12.592
- **Should we reject the null hypothesis?**

Degree of freedom, df- in statistics

1. **Variance = Sum of square of errors from mean/df**
   - For population variance, df= number of observations.
   - For sample population, df=number ofobservations-1.
   - 1 df is lost since mean is computed from the sample.

2. **t distribution**
   - df= sample size-1
   - 1 df is lost since standard deviation is computed from the sample.

3. **F distribution**
   - df_N= sample $size_N$-1, df_D=sample $size_D$-1
   - 1 df is lost each in Numerator and in Denominator since means are computed from the samples.

4. **Chi square distribution:** df depends on the problem
   - df= (No of colums-1)*(No of rows-1) in contingency tables.
   - df= Number of frequencies-1, if the data contained only one row or column.

BITS Pilani
Pilani Campus

Relationship between two variables

Ch 2 (Scatter Plots) and Ch 3 (Correlations and Covariance)

Business Statistics, Levine et al.

Relationship between two variables?

- What do we mean by “relationship between two variables”?
- Information about one variable conveys some information about the other.

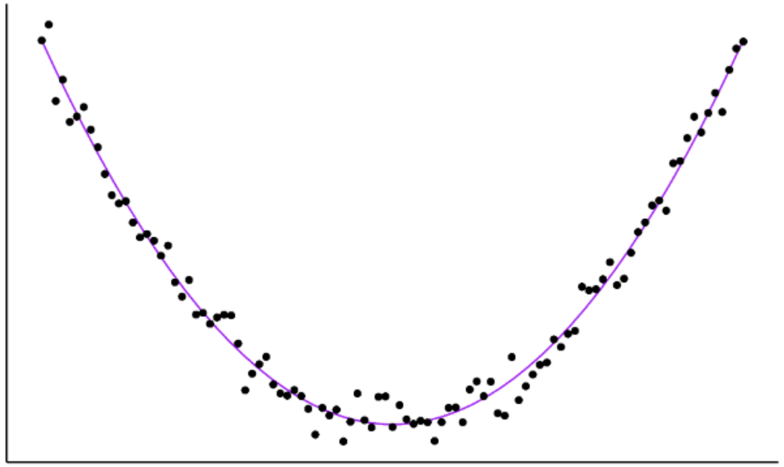
Some examples

- Does knowing darkness of the clouds give information about chance of rain?
- Yes, there is a relationship. Positive: When one goes up, other also goes up on an average.
- Is there a relationship between number of absent days and total marks?
- Yes, there is a relationship. Negative: When one goes up, other goes down on an average.
- Is there a relationship between student weight and total marks? (No relationship)
- Knowing weight of a student does not give any information about her/his potential marks in the exam.

Scatter plots and correlations

- Scatter plot: Visualizing the relationship between Var-A and Var-B.
- Correlations: A way to quantify the nature of linear realationship.
- What happens to variable on Y-Axis, when variable on X-axis increases?
- See from a distribution lens and not from the absolute value lens.
- A: Expected value: E[Y | X=400].
- B: Expected value: E ([Y | X=1200].
- Does E[Y] linearly increase with X? (Positive Correlation)
- Does E[Y] linearly decrease with Y? (Negative correlation)
- When X increases? Does E[Y] remain the same? (No correlation).
- Which type of correlation 10-Year Bond yields and Us Equity market have?

Relationship and Linear Relationship



Linear Relationship:
You can draw a trend line

Relationship but not linear.
You can not draw a trend line.
Only a relationship curve can be drawn.

Linear relationship: Covariance

Measures the strength of linear relationship between two variables

Sample covariance: cov(X,Y) = $\frac{\sum_{i=1}^{n}(X-\bar{X})(Y-\bar{Y})}{n-1}$

n: Sample size

$\bar{X}$ =? $\bar{X}$ =5 $\bar{Y}$=? $\bar{Y}$ =7 n=? n=4

X	Y	X-$\bar{X}$	Y-$\bar{Y}$	(X-$\bar{X}$)(Y-$\bar{Y}$)
2	4	-3	-3	9
4	6	-1	-1	1
6	8	1	1	1
8	10	3	3	9

$\sum_{i=1}^{n}(X-\bar{X})(Y-\bar{Y})$ = **20**

cov(X,Y) = 20/3 = 6.67

Issues with Covariance.

- cov(X,Y) = $\frac{\sum_{i=1}^{n}(X-\bar{X})(Y-\bar{Y})}{n-1}$
- The value is unit dependent.
- Our example (centimeters v/s millimeters)
- For cm values: cov(X,Y) = 6.67 cm^2
- For mm values cov(X,Y) = 666.67 mm^2
- Makes the interpretation difficult.
- Value becomes very large as the data-size increases (uncapped).

X	Y
2	4
4	6
6	8
8	10

X	Y
20	40
40	60
60	80
80	100

Linear relationship: Coefficient of Correlation

- Measures the strength of linear relationship between two variables.
- Coefficient of correlation: r or ρ = $\frac{cov(X,Y)}{S_x S_y}$ $\frac{\sum_{i=1}^{n}(X-\bar{X})(Y-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y-\bar{Y})^2}}$
- $S_x = \sqrt{\frac{\sum_{i=1}^{n}(X-\bar{X})^2}{n-1}}$
- $S_y = \sqrt{\frac{\sum_{i=1}^{n}(Y-\bar{Y})^2}{n-1}}$

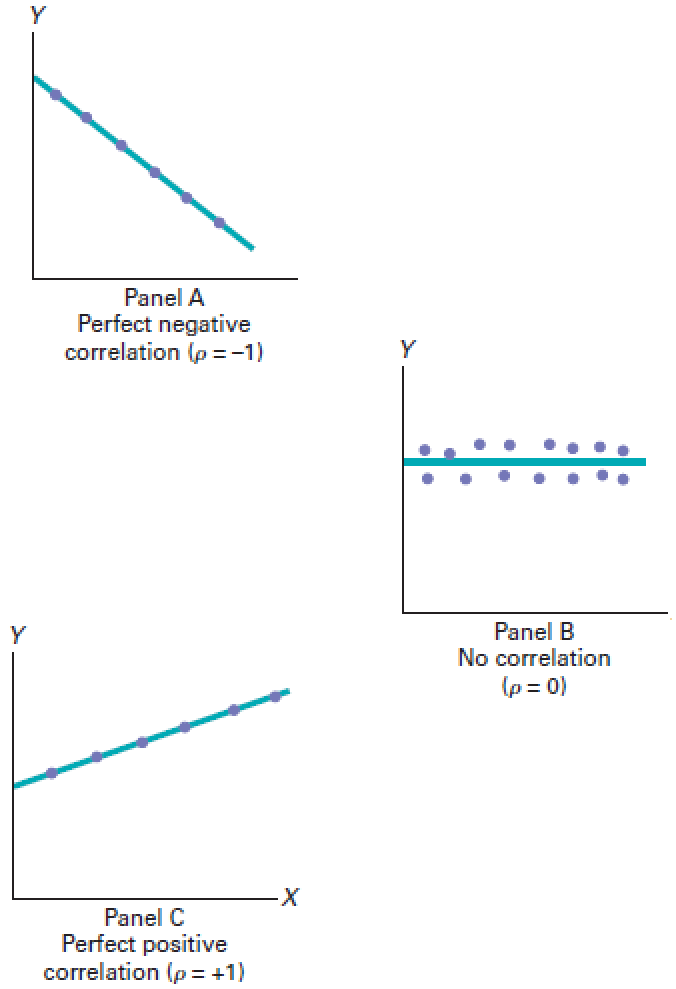
$\bar{X}$ =5 $\bar{Y}$=7

X	Y	X-$\bar{X}$	Y-$\bar{Y}$	$(X-\bar{X})^2$	$(Y-\bar{Y})^2$	$(X-\bar{X})(Y-\bar{Y})$
2	4	-3	-3	9	9	9
4	6	-1	-1	1	1	1
6	8	1	1	1	1	1
8	10	3	3	9	9	9
				Sum: 20	**Sum: 20**	**Sum: 20**

r=? **r= 20/20 = 1**

Coefficient of correlation

- Unit Independent.
- Varies between -1 and +1
- -1: Perfect negative linear correlation
- 0:   No <u>linear</u> correlation.
- +1: Perfect positive linear correlation.



Q&A