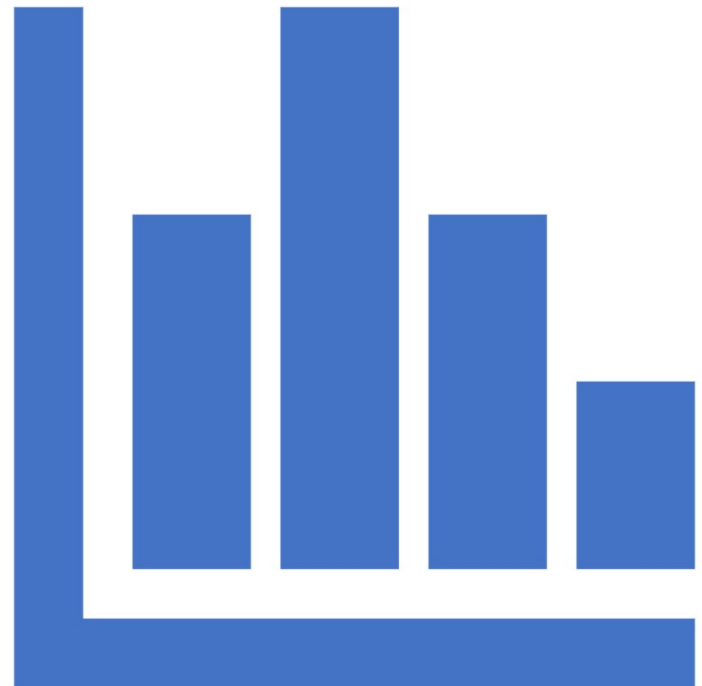


Inferential Statistics

Anshu Pandey



Empirical Rule

- The empirical rule is an important rule of thumb that is used to state the approximate percentage of values that lie within a given number of standard deviations from the mean of a set of data if the data are normally distributed.

Distance from the Mean	Values Within Distance
$\mu \pm 1\sigma$	68%
$\mu \pm 2\sigma$	95%
$\mu \pm 3\sigma$	99.7%

*Based on the assumption that the data are approximately normally distributed.

Chebyshev's Theorem

- The empirical rule applies only when data are known to be approximately normally distributed. What do researchers use when data are not normally distributed or when the shape of the distribution is unknown? Chebyshev's theorem applies to all distributions regardless of their shape and thus can be used whenever the data distribution shape is unknown or is nonnormal.

Within k standard deviations of the mean, $\mu \pm k\sigma$, lie at least

$$1 - \frac{1}{k^2}$$


proportion of the values.

Assumption: $k > 1$



Z Score

- Scores A z score represents the number of standard deviations a value (x) is above or below the mean of a set of numbers when the data are normally distributed. Using z scores allows translation of a value's raw distance from the mean into units of standard deviations.

$$z = \frac{x - \mu}{\sigma}$$


Analysing Continuous Data

ANOVA

- **ANOVA** refers to **analysis of variance** and is a statistical procedure used to test the degree to which two or more groups vary or differ in an experiment.
- Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other.
- ANOVA checks the impact of one or more factors by comparing the means of different samples

One-way ANOVA

- A one way ANOVA is used to compare two means from two independent (unrelated) groups using the F-distribution.
-
- A one way ANOVA will tell you that at least two groups were different from each other. But it won't tell you what groups were different

Calculation:- ANOVA

- ANOVA is measured using a statistic known as F-Ratio. It is defined as the ratio of Mean Square (between groups) to the Mean Square (within group).
- Mean Square (between groups) = Sum of Squares (between groups) / degree of freedom (between groups)
- Mean Square (within group) = Sum of Squares (within group) / degree of freedom (within group)

ANOVA- rep of term

<i>Source</i>	<i>SS</i>	<i>df</i>	<i>MS</i>	<i>F</i>
Between	SS_b	$k-1$	MS_b	MS_b/MS_w
Within	SS_w	$N-k$	MS_w	
Total	$SS_b + SS_w$	$N-1$		

Formula for calculation

$$\sum_{j=1}^p \sum_{i=1}^{n_j} (X_{ij} - \bar{X}_j)^2 = SS_{w/in}$$

$$\sum_{j=1}^p n_j (\bar{X}_j - \bar{X})^2 = SS_{Betw}$$

$$\sum_{j=1}^p \sum_{i=1}^{n_j} (X_{ij} - \bar{X})^2 = SS_{Tot}$$

ANOVA

- Between groups : If there are k groups in ANOVA model, then $k-1$ will be independent. Hence, $k-1$ degree of freedom.
-
- Within groups : If N represents the total observations in ANOVA ($\sum n$ over all groups) and k are the number of groups then, there will be k fixed points. Hence, $N-k$ degree of freedom.



Test One Way ANOVA

- First calculate critical values of F-factor using table [F(df(between), df(within))]
- If you found the F-value less than the critical F-value then you will not be able to reject the null hypothesis.

One way ANOVA

In [1]:

```
1 import numpy
2 from scipy import stats
```

```
1 H0 = Means of two related groups are same
2 Ha = Means of two related groups are different
```

In [2]:

```
1 france= [45,56,78,45,56,12,45,65,23,45,78]
2 spain=[23,45,12,23,56,23,45,23,56,45,23]
3 germany=[78,56,45,89,56,23,45,56,78,89,56,23]
4 #h0 = means of three groups are similar
5 #ha = means of three groups are different
6 stats.f_oneway(france,germany,spain)
```

Out[2]: F_onewayResult(statistic=4.291053283851126, pvalue=0.022639834442945588)



Analysing Categorical Data

$$\chi^2_c = \sum \frac{(O_i - E_i)^2}{E_i}$$

Chi-Square Hypothesis

- A Chi-square test is designed to analyze categorical data.
- It means that the data has been counted and divided into categories.
- Chi-Square Statistic
- The subscript “c” are the degrees of freedom. “O” is your observed value and E is your expected value.

Test a Chi Square Hypothesis

- Test the chi-square hypothesis with the following characteristics:
- Degrees of Freedom (Degrees of freedom equals the number of categories minus 1)
- Chi square test statistic value
- Now use Chi Square table and find Chi Square P-Value.
- Based on P-value we can decide that particular hypothesis will select or reject.

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of χ^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00
17	6.408	8.672	10.085	12.792	16.338	20.49	24.77	27.59	33.41
18	7.015	9.390	10.865	13.675	17.338	21.60	25.99	28.87	34.80
19	7.633	10.117	11.651	14.562	18.338	22.72	27.20	30.14	36.19
20	8.260	10.851	12.443	15.452	19.337	23.83	28.41	31.41	37.57
22	9.542	12.338	14.041	17.240	21.337	26.04	30.81	33.92	40.29
24	10.856	13.848	15.659	19.037	23.337	28.24	33.20	36.42	42.98
26	12.198	15.379	17.292	20.843	25.336	30.43	35.56	38.89	45.64
28	13.565	16.928	18.939	22.657	27.336	32.62	37.92	41.34	48.28
30	14.953	18.493	20.599	24.478	29.336	34.80	40.26	43.77	50.89
40	22.164	26.509	29.051	33.660	39.335	45.62	51.80	55.76	63.69
50	27.707	34.764	37.689	42.942	49.335	56.33	63.17	67.50	76.15
60	37.485	43.188	46.459	52.294	59.335	66.98	74.40	79.08	88.38

Testing of Hypothesis

Find the critical chi-square value. With $c-1$ degrees of freedom.

Compare the H value to the critical chi-square value.

If the critical chi-square value is less than the H statistic, reject the null hypothesis that the medians are equal.

Covariance



Covariance

A covariance refers to the measure of how two random variables will change together and is used to calculate the correlation between variables. In a finance context, covariance is the term used to describe how two stocks will move together.

Positive, Negative and Zero Covariance

A positive covariance indicates that both tend to move upwards in value and downwards in value at the same time.

Negative covariance means they will move counter to each other, when one rises, the other falls.

Zero covariance means that two random variable(in which we want to find covariance) are independent to each other. .


Calculation of Covariance

$$\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x - \bar{x})(y - \bar{y})$$

Correlation

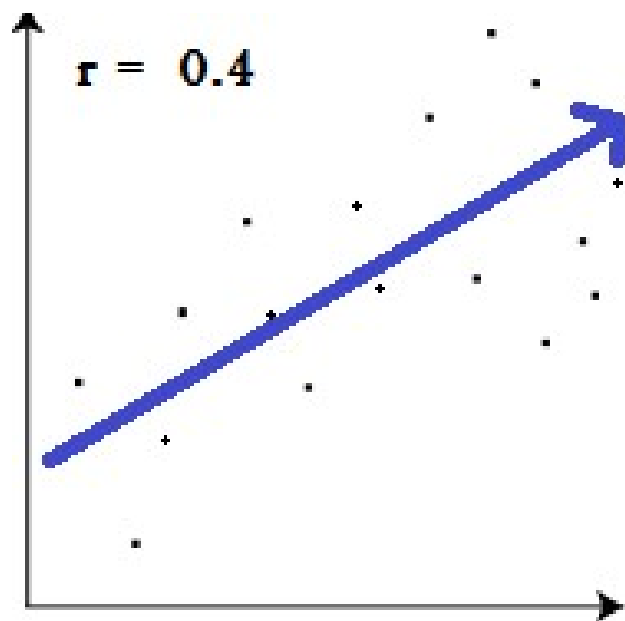


Correlation

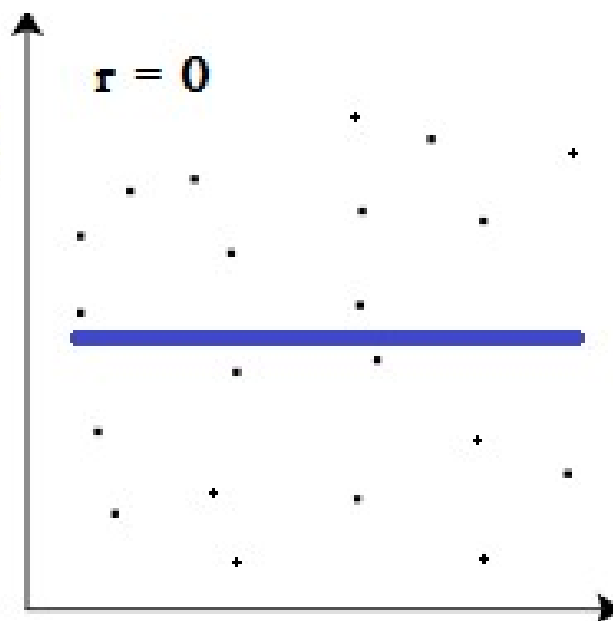
- Correlation is used to test relationships between quantitative variables or categorical variables. In other words, it's a measure of how things are related.
 - The study of how variables are correlated is called correlation analysis.
 - Correlations are useful because if you can find out what relationship variables have, you can make predictions about future behavior.
- 

Correlation

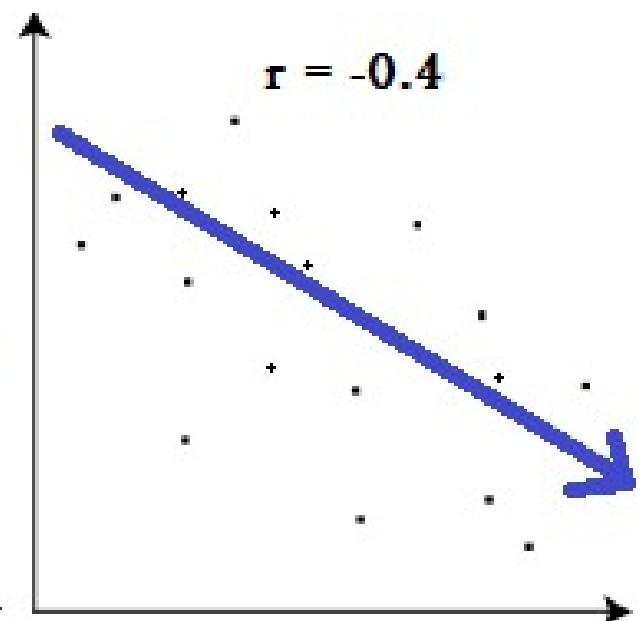
- A correlation coefficient is a way to put a value to the relationship. Correlation coefficients have a value of between -1 and 1.
- A “0” means there is no relationship between the variables at all.
-
- while -1 or 1 means that there is a perfect negative or positive correlation.



Positive Correlation



No correlation



Negative

Correlation

Calculate:- Correlation

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Thank you

Anshu Pandey



Anshu Pandey