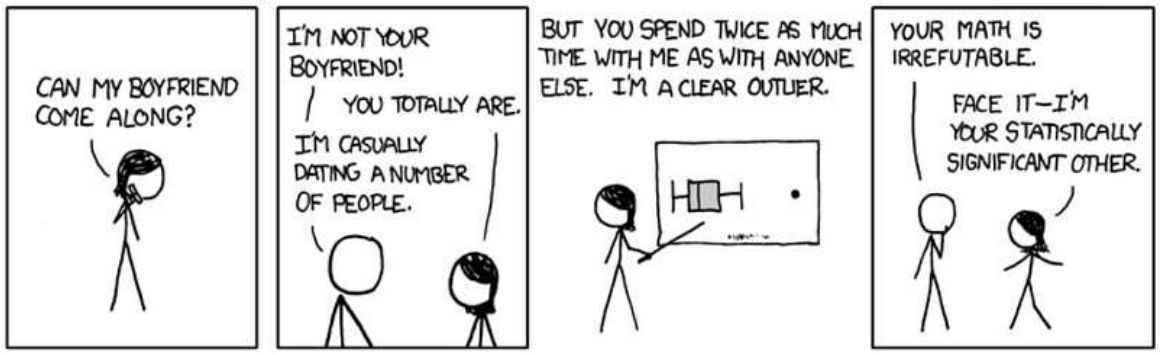
Mohit kumar

Summary

The provided web content offers a comprehensive guide to common data

Data Science Interview questions — Statistics



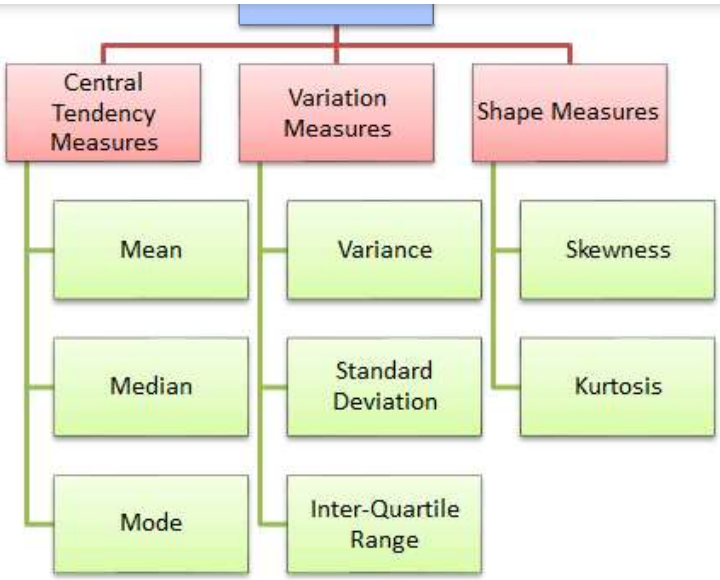
"People worry that computers will get too smart and take over the world, but the real problem is that they're too stupid and they've already taken over the

Are you preparing for Data Science Interviews? but getting no ideas how to start. What kind of problems can be asked? What topics need to be covered?

From broad mathematical discipline — Statistics, In this post I have listed top 10 Data Science interview questions based on the current Interview trend and my past 4 company's (Check out the Linkedin Profile here) interview experience:

- ***What are the different measures used to summarize the distribution?***

Ans: There are 3 types of measures used to summarize the distribution in descriptive statistics shown below in the picture:



Make sure you are well verse with all the three measures before entering into Interview room. Important Points need to remember:

- Median v/s Mean for imputation of missing values. Median should be used when the data has outlier as it is robust to outliers otherwise use mean for missing value imputation.
- To calculate Sample Variance, n-1 is used in the denominator whereas n is used to calculate Population Variance. Read more here to know why.
- Range = Maximum — Minimum Value, is also used as a measure of dispersion.
- When variable is skewed, Log transformation can be helpful to make it more symmetric.

- For Left Skewed Distribution — Mean < Median < Mode

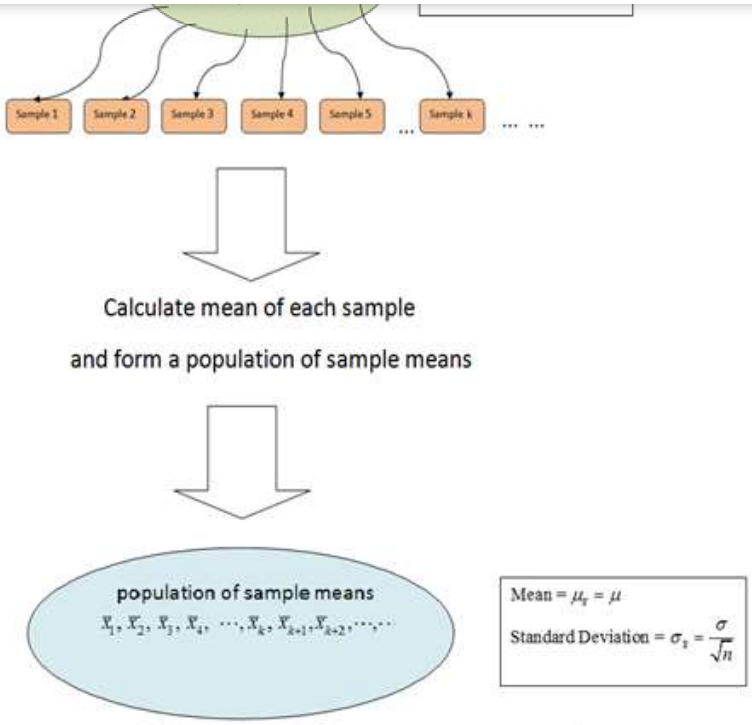
2. What is Central Limit Theorem (CLT) and its applications?

Ans: Before getting into CLT, Let's understand what is Sampling Distribution:

Sampling Distribution: A sampling distribution is a probability distribution of a statistic obtained through a large number of samples drawn from a specific population. Now we have a good understanding of Sampling Distribution.So , Let's explore CLT Theorem:

> **Central Limit Theorem:** For a population with any distribution, the Sampling distribution of sample means approaches Normal distribution as the sample size increases.

Below picture beautifully explains CLT theorem:



CLT Theorem

Try out following code in R to understand CLT Theorem.

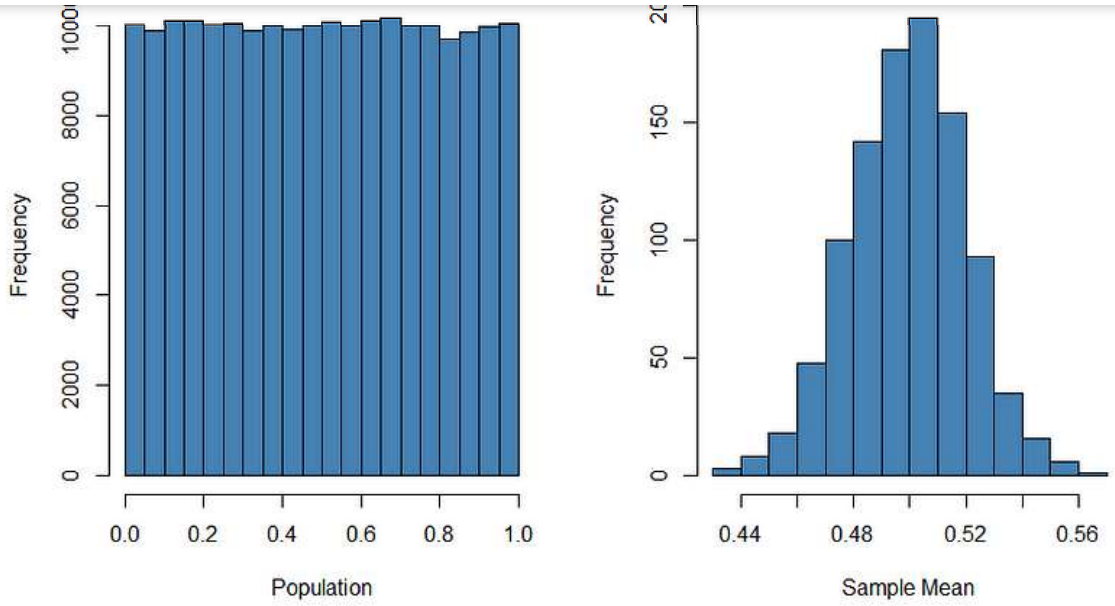
```
#Population with Uniform Distribution
samps <- runif(1000*200)  ## uniform distribution [0,1]
```

```
samp.means <- rowMeans(matrix(samps,nrow=1000,ncol=200))
```

```
# generate histogram for Population and Sampling Distribution of Sample Mean
par(mfrow = c(1,2))
```

```
hist(samps, col = "steelblue", main = "Population Distribution", xlab = "Popu
```

```
hist(samp.means, col = "steelblue", main = "Sampling Distribution of Sample M
```



3. What is the difference between Point Estimates and Confidence Interval?

Ans: Point Estimation gives us a particular value as an estimate of Population parameter. Method of Moments and Maximum Likelihood estimator methods are used to derive Point Estimators for population parameters.

θ	$\hat{\theta}$	θ
μ	$\bar{X} = \frac{\Sigma X_i}{n}$	$\bar{x}$
σ^2	$S^2 = \frac{\Sigma(X_i - \bar{X})^2}{n-1}$	s^2
p	$\hat{P} = \frac{X}{n}$	$\hat{p}$
$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2 = \frac{\Sigma X_{1i}}{n_1} - \frac{\Sigma X_{2i}}{n_2}$	$\bar{x}_1 - \bar{x}_2$
$p_1 - p_2$	$\hat{P}_1 - \hat{P}_2 = \frac{X_1}{n_1} - \frac{X_2}{n_2}$	$\hat{p}_1 - \hat{p}_2$

Point Estimator Examples

Confidence interval gives us a range of values which is likely to contain the population parameter. Confidence interval is generally preferred, as it tells us how likely this interval is to contain the population parameter. This likeliness or probability is called Confidence Level or Confidence coefficient and represented by 1 — alpha, where alpha is the level of significance.

4. How do we calculate Confidence Interval?

Ans: A confidence interval is an interval generated on the basis that a specified proportion of the confidence intervals include the true parameter in repeated sampling. How frequently the confidence interval contains the parameter is determined by the confidence level. 95% is commonly used confidence level which means that in repeated sampling 95% of the confidence intervals include the parameter. Here picture depicts the components of a Confidence interval.

$P(L \le \theta \le U) = 1-\alpha \Rightarrow$ *Confidence level* $= 1-\alpha$

•The general formula for all confidence intervals is:

Point Estimate ± (Critical Value) (Standard Error)

Lower Confidence Limit

Point Estimate

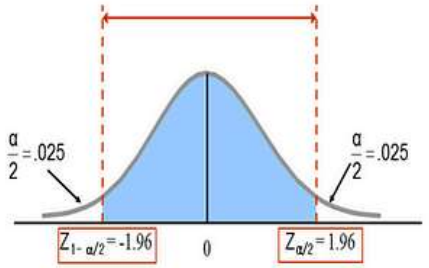
Upper Confidence Limit

L=Point Estimate -Critical Value*S.E

Width of confidence interval (=2*Critical Value*S.E.)

U=Point Estimate +Critical Value*S.E

Consider a 95% confidence interval: $1-\alpha = 0.95$



Here following pic gives the formulas use to calculate Confidence interval for population mean under 2 scenarios.

Case 1. Variance is known σ^2

- 100(1-α)% (two-sided) Confidence Interval for μ:

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

or

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

(, where $Z_{\alpha/2}$ is the standardized normal distribution critical value for a probability of α/2 in each tail)

- 100(1-α)% Upper-Confidence Bound for μ

$$\mu \le \bar{x} + z_{\alpha}\frac{\sigma}{\sqrt{n}}$$

- 100(1-α)% Lower-Confidence Bound for μ

$$\mu \ge \bar{x} - z_{\alpha}\frac{\sigma}{\sqrt{n}}$$

Case 2. Variance is unknown

- 100(1-α)% Confidence Interval for μ :

$$\bar{X} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

or

$$\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

(,where $t_{\alpha/2,n-1}$ is the critical value of the t distribution with n-1 d.f. and an area of α/2 in each tail)

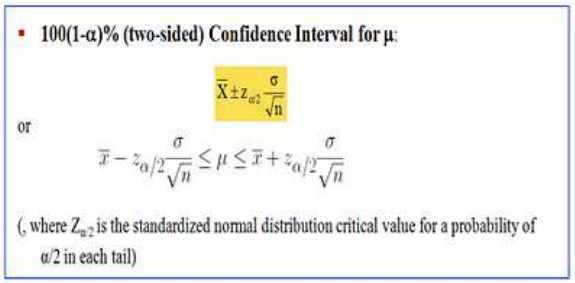
- 100(1-α)% Upper-Confidence Bound for μ

$$\mu \le \bar{x} + t_{\alpha,n-1}\frac{s}{\sqrt{n}}$$

- 100(1-α)% Lower-Confidence Bound for μ

$$\mu \ge \bar{x} - t_{\alpha,n-1}\frac{s}{\sqrt{n}}$$

> A sample of 11 circuits from a normal population has a mean resistance of 2.20 ohms. We know from past testing that the population standard deviation is 0.35 ohms. Determine a 95% confidence interval for the true mean resistance of the population.



$$\overline{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 2.20 \pm 1.96(0.35/\sqrt{11})$$
$$= 2.20 \pm .2068$$

Confidence Interval = [1.9932, 2.4068]

5. What is p-value and what does it signify?

Ans: Most of the candidates feel frustrated when it comes to explain the most widely term used in statistics “P-Value”. The fundamental problem with p-value is that no one can easily explain what exactly p-value is without using statistical jargons. Let me try to explain it as simple as possible:

Just imagine a scenario, Modi ji went to buy mangoes from a vendor. As usual, Vendor guy is claiming he is having sweet mangoes. Now as our Modi ji is statistician **(don’t be serious)**, he wants to investigate Vendor’s Hypothesis.

> Ha: Mangoes are not sweet

Now from the population of Mangoes, Modi ji picked 1 sample and investigated it. He found that Mango sample is not sweet and thoughtfully said the probability of getting mango as sweet as this one or more sweeter than this is very less (p-value), say less than 5%. So, Modi ji rejected the vendor's claim and went to another vendor.

> In general hypothesis testing procedure, we will have some hypothesis about the population parameter and we investigate it using a sample extracted from the population. P-value is nothing but the probability of observing such sample from the population given that null hypothesis is true, if the probability is too small, we doubt on the accuracy of null hypothesis and reject it, otherwise we accept Null hypothesis by saying we don't have enough evidence to reject the null hypothesis.

The *p*-value reflects the strength of evidence against the null hypothesis.

Let's get back to actual definition:

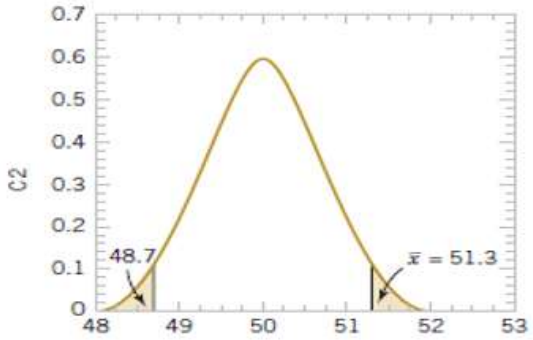
> *p*-value is defined as the probability that the data would be at least as extreme as those observed, if the null hypothesis were true.

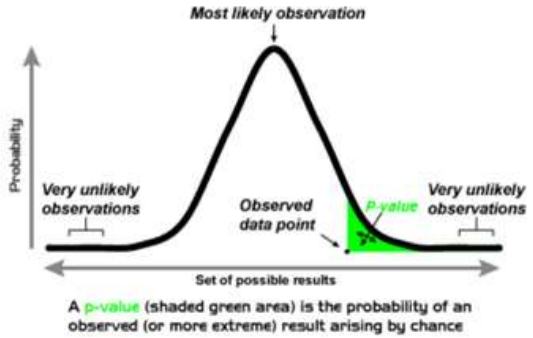
Let's understand it with example:

burning rate of solid propellant is 50 cm/s.

$$H_0: \mu = 50 \text{ cm/s}$$
$$H_1: \mu \neq 50 \text{ cm/s}$$

To investigate the producer claim, Imagine we have collected a sample of propellants and found sample mean as 51.3 and the symmetric value 48.7.





$$\begin{aligned} P\text{-value} &= 1 - P(48.7 < \bar{X} < 51.3) \\ &= 1 - P\left(\frac{48.7 - 50}{2.5/\sqrt{16}} < Z < \frac{51.3 - 50}{2.5/\sqrt{16}}\right) \\ &= 1 - P(-2.08 < Z < 2.08) \\ &= 1 - 0.962 = 0.038 \end{aligned}$$

P-Value Calculation

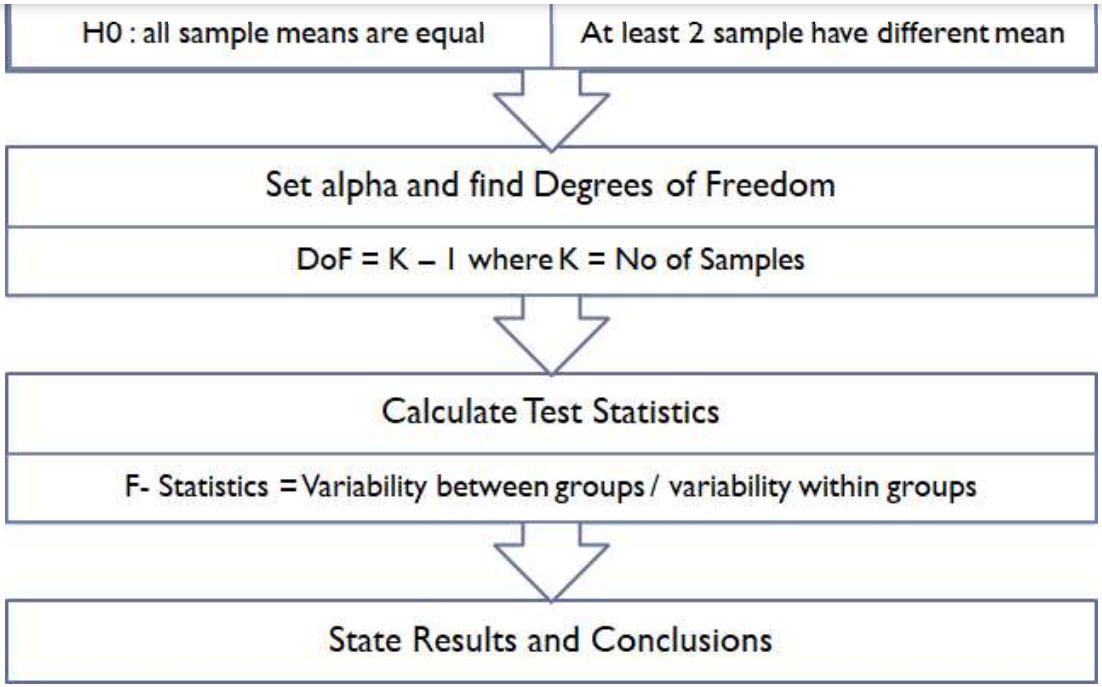
far from 50 as 51.3 and 48.7 is 0.038. Therefore, an observed sample mean of 51.3 is a rarely event if the null hypothesis is true. So, we reject the null hypothesis at 5% level.

So keep in mind: P-value helps the statistician to draw conclusions on Null hypothesis and is always between 0 and 1.

- P- Value > 0.05 denotes weak evidence against the null hypothesis which means the null hypothesis cannot be rejected.
- P-value < 0.05 denotes strong evidence against the null hypothesis which means the null hypothesis can be rejected.
- P-value=0.05 is the marginal value indicating it is possible to go either way.

6. Explain ANOVA and it's applications.

Ans: Analysis of Variance (abbreviated as **ANOVA**) is an extremely useful technique which is used to compare the means of multiple samples. Whether there is a significant difference between the mean of 2 samples, can be evaluated using z-test or t-test but in case of more than 2 samples, t-test can not be applied as it accumulates the error and it will be cumbersome as the number of sample will increase (for example: for 4 samples — 12 t-test will have to be performed). The ANOVA technique enables us to perform this simultaneous test. Here is the procedure to perform ANOVA:



Let's see with example: Imagine we want to compare the salary of Data Scientist across 3 cities of india — Bengaluru, Pune and Mumbai. In order to do so, we collected data shown below.

8	7	12
10	5	9
7	10	13
14	9	12
11	9	14
Total = 50 Mean = 10	40 8	60 12

Data Scientist salary Data (in 100k)

Following picture explains the steps followed to get the Anova results.

2. Calculate the **Grand average**
3. Take the difference between means of various samples & grand average.
4. Square these deviations & obtain total which will give sum of squares between samples **(SSC)**
5. Divide the total obtained in step 4 by the degrees of freedom to calculate the mean sum of square between samples **(MSC).**

2. Take the deviations of the various items in a sample from the mean values of the respective samples.
3. Square these deviations & obtain total which gives the sum of square within the samples **(SSE)**
4. Divide the total obtained in 3rd step by the degrees of freedom to calculate the mean sum of squares within samples **(MSE).**

$$MSC = \frac{SSC}{k-1} = \frac{40}{2} = 20$$

$$MSE = \frac{SSE}{n-k} = \frac{60}{12} = 5$$

$$\text{F- statistic} = \frac{MSC}{MSE} = 20/5 = 4$$

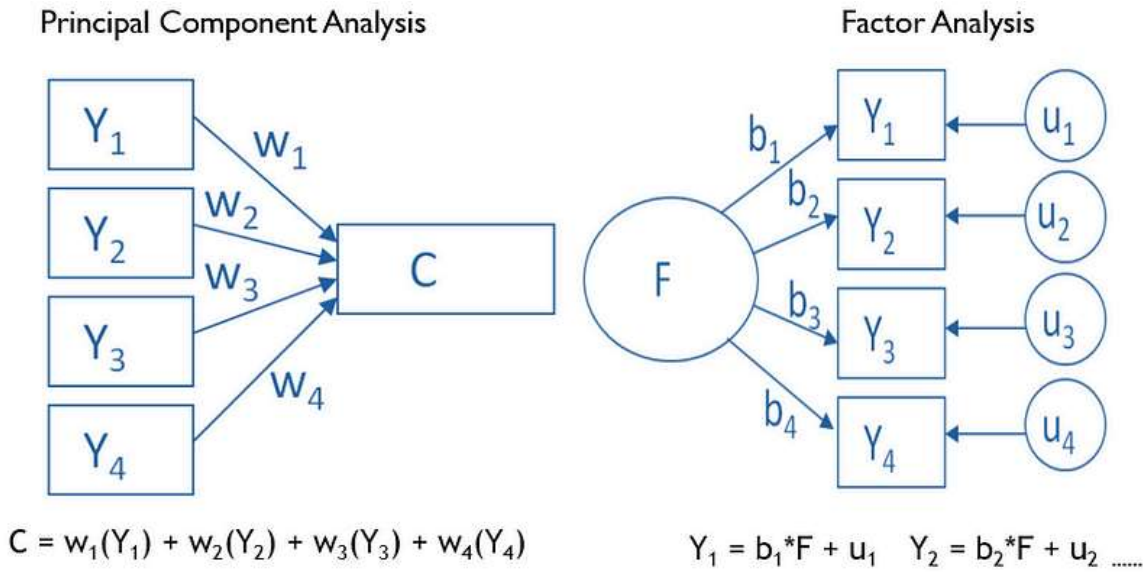
The Table value of F at 5% level of significance for d.f 2 & 12 is 3.88
The calculated value of F > table value
H0 is rejected. Hence there is significant difference in sample means

There is a limitation of ANOVA that it does not tell which pair is having significant difference. In above example, It is clear that there is a significant difference between the means of Data Scientist salary among these 3 cities but it does not provide any information on which pair is having the significant difference. This problem is being solved by Tukey HSD. If interested about it, read more here.

7. What is the difference between factor analysis and principal Component Analysis?

Ans: Principal Component analysis and factor analysis, both techniques can be used to reduce the dimensions in the data. But, generally Statisticians use

approaches are different



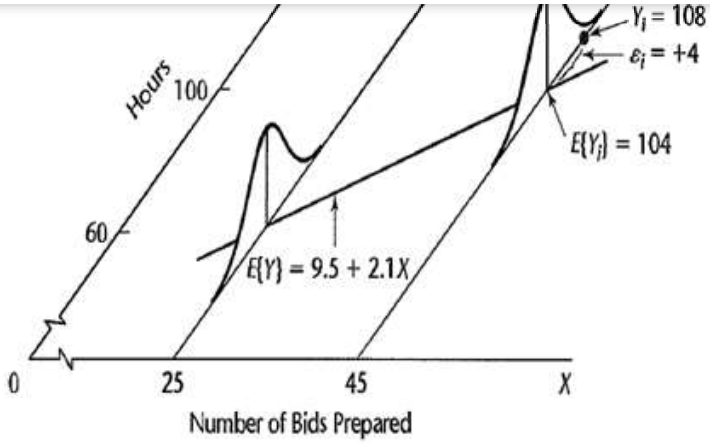
Factor analysis seeks linear combinations of variables, called factors, that represent underlying fundamental quantities of which the observed variables are expressions. More precisely, the manifest variables are linear combinations of the factors, plus unique (or specific) factors. From the above picture, It is clear the Factor F causes the responses on the 4 measured Y variables.

PCA on the other hand summarizes common variation in many variables using just a few variables. You can see in above picture from the direction of the arrows that the Y variables contribute to the component variable.

Ans: This is one of the frequently asked and simplest question which can help you in creating strong impression on the interviewer. Let's crack it:

There are fundamentally 3 Assumptions of Linear Regression

- ***Linearity:*** There is a linear relationship between dependent and independent variable. This is one of the crucial assumption as if there is non-linear relationship among dependent and independent variables, Linear regression model will be under-fitted and predictions will be quite far from the actual.
- ***Normality:*** Residuals error are assumed to be normally distributed. Let's see what does it mean with example: 1. See the following graph, We are trying to analyze the relationship between number of bids(X) requested by construction contractors and time (Y) required to prepare the bids.



Linear Regression

Regression analysis assumes that at every level of X, there is a probability distribution of Y, whose mean have a systematic relation with X. This Systematic relation is called regression function of Y on X. The response Y exceeds or falls short of the value of regression function by the error term.

Regression Function = E(Y/X)

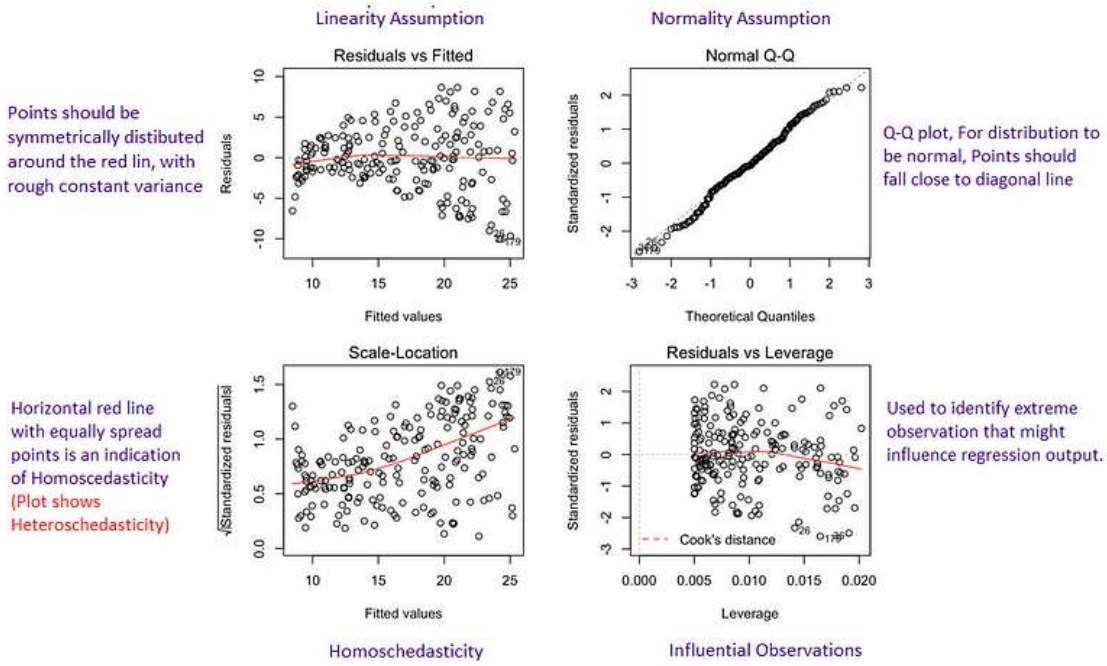
$$Y_i = 9.5 + 2.1X_i + \varepsilon_i$$

Error Terms

dependent variable also follows normal distribution with mean as Regression function and variance same as of error sigma².

3. Homoscedasticity: Residuals(Errors) are assumed to have constant variance across the level of X.

Diagnostics: Following picture is the standard result of linear regression model performed in R. It clearly explains which plot has to be used and how to diagnose the assumptions of linear regression. This process we call it as Residual Analysis.



which are critical for validation like:

- Independence of Error Terms — Check by analyzing auto-correlation plot.
- Independent variables should not be correlated with each other (Multi-collinearity) — Generally diagnosed by Variance Inflation Factor(VIF) or directly plotting Correlation plots.

9. What is the difference between Correlation and Covariance?

Ans: Correlation and Covariance are statistical concepts which are generally used to determine the relationship and measure the dependency between two random variables. Actually, Correlation is a special case of covariance which can be observed when the variables are standardized. This point will become clear from the formulas :

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Correlation Formula

Here listed key differences between covariance and correlation:

COMPARISON		
Meaning	Covariance is a statistical measure of the extent that 2 variables move in tandem relative to their respective mean (or average) values.	Correlation is a statistical measure that indicates how strongly two variables are related.
Values	[-∞, +∞]	[-1, +1]
Impact of scaling	Yes	No
Unit free measure	No	Yes

10. Can we create Logistic regression using Excel without plugins?

Ans: Yes absolutely. At this particular question candidates get stumbled if they hear it for the first time. This question tests knowledge of Logistic Regression, Excel functions and Excel solver options.

Download the excel file here explaining different steps involved to perform Logistic regression in Excel using Solver.

Steps to be performed:

- Identify the dependent and independent variable. (Imagine 2 Independent variable and 1 dependent variable.)
- Let's start with some random values in 3 coefficients as 0.1.

$$z = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2$$

4. Create another column by calculating Logit of z as:

$$Logit(z) = \frac{e^z}{1 + e^z}$$

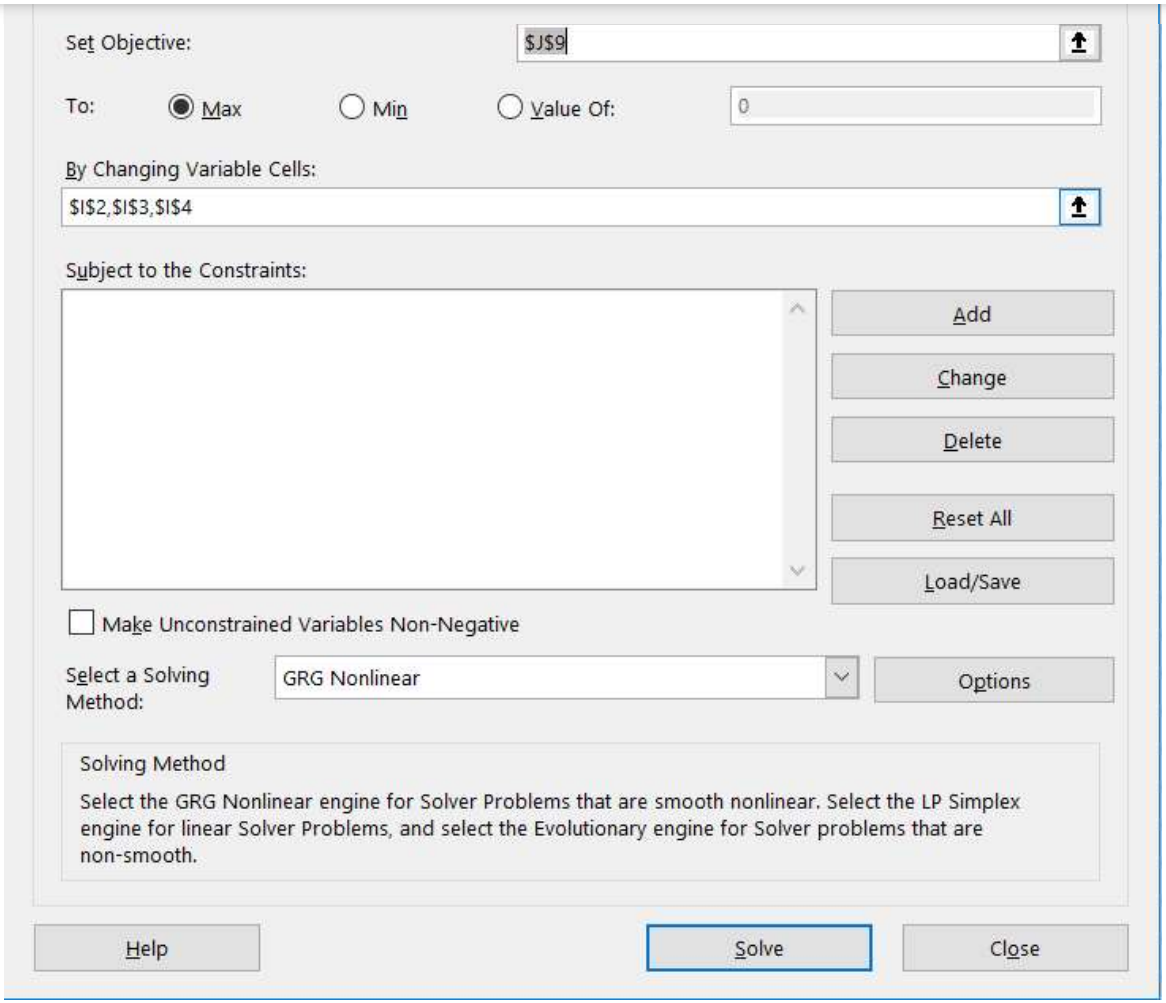
5. Calculate log-likelihood for every observation using following formula

$$LL = y * \ln(P(Y = 1) + (1 - y) * \ln(1 - P(Y = 1))$$

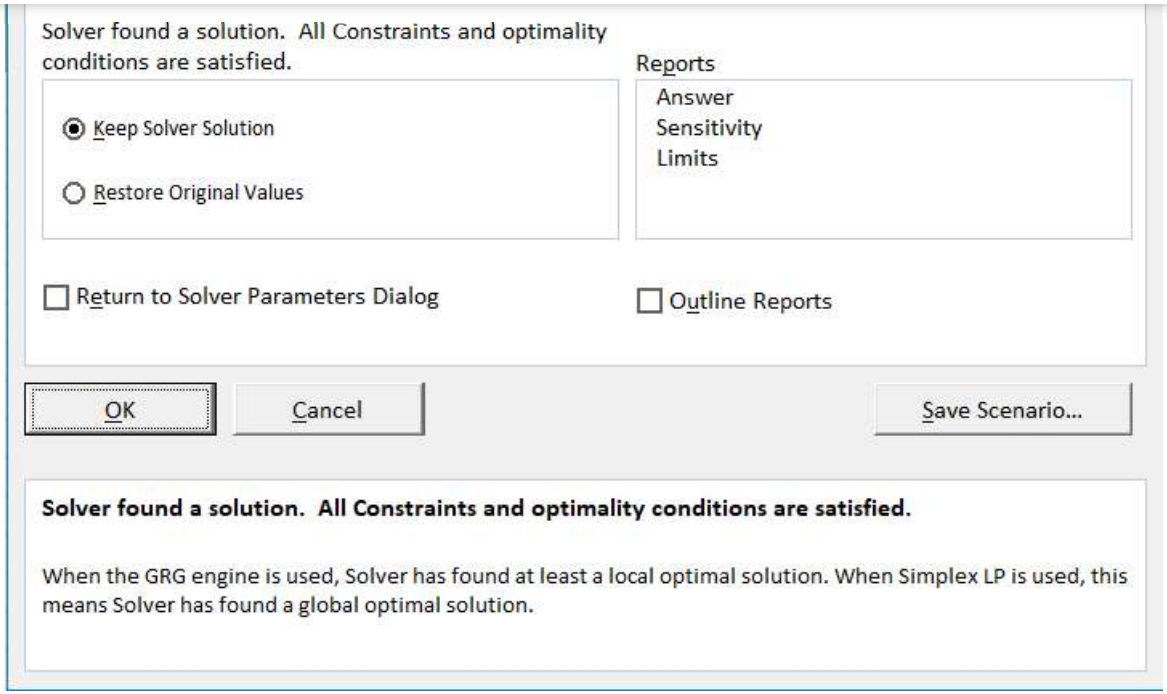
Above steps are depicted using following snapshot:

2	57	4	1	6.2	492.749	0.997975	-0.00203	beta0	0.1	
3	73	5	0	7.9	2697.282	0.999629	-7.90037	beta1	0.1	
4	22	5	1	2.8	16.44465	0.942676	-0.05903	beta2	0.1	
5	59	4	0	6.4	601.845	0.998341	-6.40166			
6	15	4	1	2	7.389056	0.880797	-0.12693			
7	36	2	1	3.9	49.40245	0.98016	-0.02004			
8	68	5	0	7.4	1635.984	0.999389	-7.40061			
9	49	5	0	5.5	244.6919	0.99593	-5.50408		Log-Likelihood	-66.5235
10	27	7	0	3.5	33.11545	0.970688	-3.52975			
11	59	3	1	6.3	544.5719	0.998167	-0.00183			
12	10	6	1	1.7	5.473947	0.845535	-0.16779			
13	78	8	0	8.7	6002.912	0.999833	-8.70017			
14	22	6	1	2.9	18.17415	0.947846	-0.05356			
15	36	4	1	4.1	60.34029	0.983698	-0.01644			
16	57	7	0	6.5	665.1416	0.998499	-6.5015			
17	73	8	0	8.2	3640.95	0.999725	-8.20027			
18	38	5	1	4.4	81.45087	0.987872	-0.0122			
19	71	7	0	7.9	2697.282	0.999629	-7.90037			
20	35	4	0	4	54.59815	0.982014	-4.01815			
21	44	5	1	5	148.4132	0.993307	-0.00672			

Now open the solver from Data Tab in excel and estimate the coefficients that maximize the log-likelihood function given in J9.



After providing the input to the solver, mentioned in the above snapshot, click on Solve and if solver will be able to find the solution, a new window will pop up which will look like this:



Just click OK. It will display the estimated coefficients in the respective cells.

beta0	12.48288092
beta1	-0.117031462
beta2	-1.469145969

Solver Output

apart of estimated coefficients, Solver will show the maximized Log-Likelihood values which is -6.65456 in the current case using estimated coefficients.

Apart of questions explained above, I will recommend following topics one must prepare before sitting in Data Science/Data Analyst/Business Analyst Interview.

- Chi-Squared Distribution and Chi-square Goodness of Fit Test
- Continuous Probability Distributions — Normal and T- distribution
- Discrete Probability Distributions: Binomial, Geometric, Poisson and Negative Binomial Distribution.
- Probability Mass Function, Probability Density function and Cumulative Distribution Function.
- Hypothesis Testing
- Sampling Techniques — Simple Random **Sampling** (SRS), Stratified **Sampling,** Cluster **Sampling**.

Here are few good resources to read more:-

- Applied Statistics and Probability for Engineers — By Montgomery
- Applied Linear Statistical Models — by Michael H. Kutner,

****

Do you have any question for me or interested in Data Science Career, Join us at “How to break into Data Science World !” webinar organized by MnG-

Data Science Webinar

What's your story? Did this guide help you better prepare for your next interview? Do you have any other statistics question which you want to be in the post? Let us know in the comments below!

If you found this interesting or helpful Please help others find it by sharing and clapping.

On Linkedin? So am I. Feel free to keep in touch — Mohit 👐.

A special thanks to Gunjan Thareja for her significant contributions and feedback.

Recommended from ReadMedium

Sanjay Kumar PhD

Interview Questions on Descriptive Statistics

1. Define mean, median, and mode. How do they differ?

11 min read

Be 10x Engineer

Staff Backend Engineer @ Google | Interview Experience

It was a crisp Tuesday morning when I first saw the email. My hands trembled slightly as I read the subject line: “Google Interview...

3 min read

Amit Yadav

Best Laptops For Data Science in 2024

I’ve found that laptops with Intel i7 or i9 processors, as well as AMD Ryzen 7 or 9, offer the best performance. But are they good overall?

11 min read

The Most Expensive Data Science Mistake I've Witnessed in My Career

Why true success in machine learning goes beyond optimizing a single metric

6 min read


Vikash Singh

Frequently Asked Hypothesis Testing Questions for Data Scientist Interviews

If you are preparing for a data science or statistical modelling role, brushing up on your hypothesis testing knowledge is of paramount...

7 min read


Ritesh Gupta

Can You Handle These 25 Toughest Data Science Interview Questions?

The role of a Data Scientist demands a unique blend of skills, including statistics, machine learning, data analysis, and programming. In...

7 min read