1. Statistics

Statistical computing is the process through which data scientists take raw data and create predictions and models backed by the data. Without an advanced knowledge of statistics it is difficult to succeed as a data scientist – accordingly it is likely a good interviewer will try to probe your understanding of the subject matter with statistics-oriented data science interview questions. Be prepared to answer some fundamental statistics questions as part of your data science interview.

Here are examples of rudimentary statistics questions we’ve found:

1. What is the Central Limit Theorem and why is it important?
   * [*Answer*](https://spin.atomicobject.com/2015/02/12/central-limit-theorem-intro/)

CLT states that if we sample from a population using a sufficiently large sample size, the mean of the samples (also known as the sample population) will be normally distributed (assuming true random sampling). What’s especially important is that this will be true regardless of the distribution of the original population.

1. What is sampling? How many sampling methods do you know?
2. What is the difference between Type I vs Type II error?
3. What is linear regression? What do the terms P-value, coefficient, R-Squared value mean? What is the significance of each of these components?
   * [*Answer*](http://www.statisticssolutions.com/assumptions-of-linear-regression/)*,*[*Answer*](http://blog.minitab.com/blog/adventures-in-statistics/how-to-interpret-regression-analysis-results-p-values-and-coefficients)
4. What are the assumptions required for linear regression?*There are four major assumptions:* 
   * *1. There is a linear relationship between the dependent variables and the regressors, meaning the model you are creating actually fits the data,*
   * *2. The errors or residuals of the data are normally distributed and independent from each other,*
   * *3. There is minimal multicollinearity between explanatory variables, and*
   * *4. Homoscedasticity. This means the variance around the regression line is the same for all values of the predictor variable.*
5. What is a statistical interaction?

Statistical Interactions

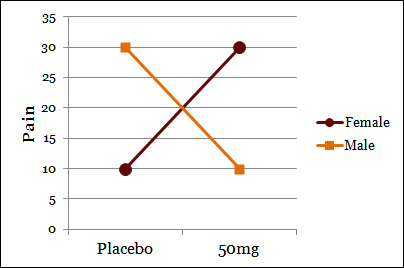
Basically, an interaction is when the effect of one factor (input variable) on the dependent variable (output variable) differs among levels of another  factor.

For example, in a pain relief drug trial, one factor is “dose” and another factor is “gender”. The dependent variable is “headache pain” (measured on a scale of 0 to 50). We can look at the findings by showing the means for each group in a table like this:

|  |  |  |
| --- | --- | --- |
|  | Male | Female |
| Placebo | 30 | 10 |
| 50mg | 10 | 30 |

If compare the “marginal means”, we can see that the average pain score for women was 20 and the average pain score for men was 20. There is no difference in headache pain between men and women. Likewise, the drug appears to have no effect on headache pain.

However, if we had stopped there, we would be missing some important findings. There is an interaction between gender and dose on headache pain. If we graph the means of each group, we can see it clearly: men have more headache pain than women unless they take the drug. There are simple effects of dose for both men and women, but the effects “wash out” one another. The result is no *main effect* of dose or gender. If you did not examine gender, it would appear that your drug did not work; if you did not study drug dose, you would see no difference between men and women. Examining the interaction allows us to see that the drug works – for men – AND that it *causes* pain in women.

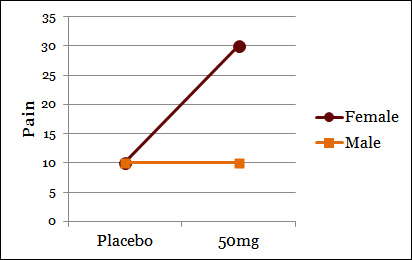
1. [](http://icbseverywhere.com/blog/wp-content/media/2012/05/Pure-Interaction.jpg)

This is called a “pure interaction” because there are no main effects, but there is an interaction.

More commonly seen, however, when effects occur in both (or all) conditions, but they are stronger in one condition than another. Also fairly common is an effect in one condition (or for one group), but no effect at all in another. For example, what if our same pain study resulted in the following means:

|  |  |  |
| --- | --- | --- |
|  | Male | Female |
| Placebo | 10 | 10 |
| 50mg | 10 | 30 |

1. For men, there is no effect of the drug on headache pain. The drug causes pain in women, who would otherwise have the same amount of pain as men.

[](http://icbseverywhere.com/blog/wp-content/media/2012/05/Interaction.jpg)

To recap, an interaction is when the effect of one variable differs among levels of another variable. When interactions are seen among continuous variables (variables with a range of values, as opposed to categories), they look a little different, but the meaning is basically the same. In the last example, difference in pain between men and women (the male average was 10, the female average was 20) are *driven by* the interaction with the drug.

1. What is selection bias?

**Selection bias** means that the sample you have chosen is not representative of the population you want to look at.

1. How to avoid Selection Bias ?
2. What is an example of a dataset with a non-Gaussian distribution?
3. What is the Binomial Probability Formula?

*Examples of similar data science interview questions found from Glassdoor:*

