DATA SCIENCE FOR ENGINEERS

Statistics-FAQ's

Correlation

1. In the lecture explanation, RVs (x and y) can be uncorrelated (if so, sigma-xy = 0). Doesn't it also imply that they are independent?
   Or Uncorrelation ==> Independence...?
   Explain the relation between correlation and independence?
   Answer:
   Theorem: Two independent variable are uncorrelated.
   Proof: If X and Y are independent variables, Then Cov(X,Y)=0 Hence two independent variables are uncorrelated.

   But the converse of the theorem is not true,i.e, two correlated variables may not be independent as the following example illustrates:
   X~N(0,1) and Y=X^2
   since X~N(0,1), E(X)=0
   Cov(X,Y)=E(XY)-E(X)E(Y)=E(X^3)-E(X)E(Y)=0
   Correlation coefficient r(X,Y)=0
   Hence X and Y are uncorrelated but not independent (Y=X^2)

Distributions

1. It is mentioned in the slides that the pmf of a binomial distribution is characterized by only a single parameter ie p but isn't the value of n is also an important parameter, affecting the distribution function.
   Answer:
   n is the independent Bernoulli trials.
   Both n and p are the parameters of the distribution.
   In the slide they have also mentioned that for large n it tends to a Gaussian distribution

Hypothesis testing

1. How to get the critical value in hypothesis testing for an alpha

Answer:

First you need to know the distribution (of the test statistic in hypothesis testing). Second you need to know whether it is a two-sided test or a one-sided test.

For example, if the test statistic has a standard normal distribution

(as the case when you are testing for the mean with known standard deviation), the test is two-sided, and level of significance is 5%, then the lower

critical value is value corresponding to lower tail probability of standard normal random variable z0.025 and upper critical value corresponds to z0.975.

For a one-sided hypothesis it will be z0.05 or z0.95 depending on whether the

alternative is $\mu < 0$ or $\mu > 0$. Use R functions qnorm for obtaining the critical value

2. How to calculate Alpha value for z, t, f and chi square tests without using R.

Answer:

There is a t, z, f, chi square table for the respective tests. We will be providing the table if the particular question requires the f, chi square table in the exam.

3. How does this alpha (learning parameter) is calculated in hypothesis testing?

Answer:

The learning parameter is the step size which is adjusted by trying out different values and choosing the value that helps in speeding-up the gradient descent. The initial guess is chosen randomly and different values are chosen in multiple runs in order to avoid the possibility of getting the local minimum instead of the global minimum

Outliers

1. Does one need to reorder the data points in ascending order even if there is an outlier?

Answer:

Yes. We need to reorder the data points in ascending order even if there is an outlier. But median will not change drastically with respect to outliers when compared to mean. It will still stay around where most of the data is.

Probabilty

1. P(B|A)=P(A^B)/P(A)

Using ^ for intersection symbol

In two coin toss experiment we saw that P(A^B)=P(A)*P(B)

If A=first toss is head P(A)=0.5

B=Two successive heads P(B)= 0.25

P(B|A)=P(A^B)/P(A)

Which should be equal to (P(A)*P(B))/P(A)=(0.5*0.25)/0.5

But in the lecture you've shown that P(A^B)=0.25 How it comes to be 0.25?

ANSWER:

If two events A and B are not independent, then information available about the outcome of event A can influence the predictability of event B.

Conditional probability,

P(B|A)=P(A^B)/P(A)

P(A^B) is not equal to (P(A)*P(B)) since events A and B are not independent.

P(A^B)=0.25

P(A) =0.5

P(B|A) =(0.25)/0.5

=0.5

2. Will the Subset Probability be less than the set or can it be less than equal to.

Answer:

If B is a subset (not a proper subset) of A, So, the probability of event B will be less than or equal to the probability of event A.