Statistics and Data Analysis

Lecture 6

Lecture by Dr. Zohar Yakhini
Typeset by Steven Karas

2016-12-13
Last edited 17:58:53 2016-12-28

1 Clarifications on HW2

In the first question, we should normalize the histograms when plotting them. Notably, we should also produce multiple histograms on the same graph (per category). It is strongly recommended to not produce an overlapping histogram (as produced by matplotlib, R, etc by default).

In step 4 of the same slide, the intent is to produce a graph that shows the relation. He mentioned QQ plots.

We will cover what a normal approximation is today.

2 Agenda

1. Inference - comparing two normal distributions and t-test
2. Lognormal distributions
3. Correlations

3 Inference

3.1 p-value

The p-value of an observation of a statistical quantity Q on actual data (empirical mean, max, %iles, etc) under a given Null model is the probability that, drawing instances from the Null model and computing Q for the randomly drawn data, the result would be the same as or more extreme than the given observed result.

For example, a coin that turns up heads 61 times out of 100 gives us a p-value of 0.023 that the coin is fair. In even more direct terms, it is important to note that we only ever reject previous models, and propose a new one (which is not actually tested).

3.2 Empirical std

$$\overline{\sigma_n} = \left(\frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{X_n}) \right)$$

3.3 Pay rates in randomistan

$emp\mu_M = 1394$ and $emp\sigma_M = 138$. $emp\mu_F = 1324$ and $emp\sigma_F = 105$.

A Useful Fact If $X \sim N(\mu, \sigma)$, then $Z = (X - \mu)/\sigma$ is a standard normal random variable.

To example the differences in pay, we need to know more about the size of the sample. Lets first see if we can say that $\mu(M) > \mu(F)$. Let's assume the sample size was large, and all samples were independent. As such:

$$\frac{emp\mu M - \mu(M)}{emp\sigma(M)/n_M^{1/2}} \sim N(0,1)$$

$$\frac{emp\mu F - \mu(F)}{emp\sigma(F)/n_F^{1/2}} \sim N(0,1)$$

If $\mu(M) = \mu(F)$, then:

$$emp\mu M - emp\mu F \sim N(0, \frac{1}{N}(emp\sigma^2(M) + emp\sigma^2(F)))$$

If $n = 100$, then $\Phi(10\frac{70}{173}) \leq 10^{-4}$. If $n = 12$, then $\Phi(\sqrt{12}\frac{70}{173}) \cong 0.08$.

3.4 Student t-distribution

Takes a parameter called "degrees of freedom". Not covered in the course. We may see Welch's t-test in a later lecture. Useful for small sample sizes ($n < 20$).

3.5 Normal approximation/fitting

A procedure for selecting a distribution that best fits a set of observed data; Aims to find the distribution that best describes the data I've observed.

Key questions: How do we evaluate the fit? How do we compare between two models?

We will only attempt to search in the family of normal distributions and their composition.

We want to find the empirical mean and std of the observed data.

4 Lognormal distribution

A random variable X is siad to have a lognormal distribution if its log, $\log(X) \sim N$. In other words, X is lognormal if $X = e^Y$ for some Gaussian Y; That is that $X = e^{\mu + \sigma Z}$, where Z is the standard normal. μ is referred to as the location and σ is the scale of X. However, they are **not** the mean or std of X, but rather that of $\log X$.

4.1 PDF

Let X be a standard lognormal random variable. Let $F(x)$ be the CDF, and $f(x)$ be the PDF of the standard lognormal. Let Y be standard normal and $\Gamma(x)$ be the CDF, and $\gamma(x)$ be the PDF of the standard normal.

$$F(x) = P(X \leq x) = P(e^Y \leq x) = P(Y \leq \ln x) = \Gamma(\ln x)$$

Therefore, since the PDF is the derivative of the CDF, we get that:

$$f(x) = F'(x) = \Gamma'(\ln x)\frac{1}{x} = \gamma(\ln x)\frac{1}{x}$$

4.2 Median

e^μ

4.3 Mean

4.4 Mode

The mode is the maximum of the PDF. So take the derivative of the PDF:

$$f(x) = e^{-\frac{\ln x^2}{2}}$$

$$f'(x) = -\frac{1}{x^2}e^{-\frac{\ln x^2}{2}} - \frac{1}{x}e^{-\frac{\ln x^2}{2}}$$

$$e^{-\frac{\ln x^2}{2}}\left(-\frac{1}{x^2} - \ln x\frac{1}{x^2}\right)$$

$$\ln x + 1 = 0 \Rightarrow x_0 = e^{-1} \cong 0.37$$

Generally speaking, the mode is at:

$$e^{\mu-\lambda^2}$$

4.5 Products of Normals

Let $X_i \sim U\{1, ..., n\}$. We saw that $\sum n = \sum X_i \sim N(3.5n, 2.9n)$.

$$\ln \prod n = \sum_{i=1}^{n} \ln X_i$$

Because X_i are independent and distributed identically, so are the logs and assuming that n is sufficiently large, then it follows that.

4.6 Example: Stock prices

Given the prices of a stock at time 0 as $C(0)$. Generally, denote at time t the price as $C(t)$. We want to assess the distribution of its future price $C(10)$.

Break the interval from 0 to 10 into 1000 intervals. Let $B_k = C(0.01k)$. We want to model B_{1000}.

$$B_{1000} = \frac{B_{1000}}{B_{999}}...\frac{B_{k+1}}{B_k}...\frac{B_1}{B_0} \cdot C(0)$$

$$R_k = \frac{B_k}{B_{k-1}}$$

Assume that the ratios are independent and identically distributed. Therefore we apply CLT and get:

$$B_k = R_k \cdot ... \cdot R_1$$

$$\ln B_k = \sum_{i=1}^{k} \ln R_i \quad \ln R_i \sim N(\mu, \sigma)$$

$$\ln B_k \sim N(\mu \cdot k, \sigma^2 k)$$

Concrete example Assuming that stock growth has a one hour growth distribution with $\mu = 0.001$ and $\sigma = 0.01$, what is the probability of the stock more than doubling itself in 10 days of 8 hours market trading?

4.7 Example: Gene expression levels

Not really sure what he wants to show here. Maybe just another example?

5 Correlations

Measure of relation between two variables. May cover the formal definition next week.

5.1 Pearson

5.2 Covariance

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$$

$$Cov(X, Y) = E(XY) - E(X)E(Y)$$

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sigma(X)\sigma(Y)}$$

$$empCov = \frac{\sum Cov(X, Y)}{N - 1}$$

$$\rho = \frac{Cov_{XY}}{\sigma_X \sigma_Y}$$