Statistics and Data Analysis
Lecture 5

Lecture by Dr. Zohar Yakhini
Typeset by Steven Karas

2016-12-06
Last edited 17:58:29 2016-12-28

Agenda

- Matlab intro
- HWA1 - Solution
- Class:Gauss, Inference, Central Limit Theorem
- HWA2 - Comments

1 Homework

1.1 Problem 1

This is a binomial distribution. If we change the question to extend towards finding not one but more defective products, we want to find the $P(X \geq 3)$.

1.2 Problem 3

Example of solution where $E(X) = 6$ and $E(Y) = 7$:

$$X = \begin{cases} 5 & \sim P(0.5) \\ 7 & \sim P(0.5) \end{cases}$$

$$Y = \begin{cases} X = 5 & \begin{cases} \frac{41}{5} & \sim P(0.5) \\ \frac{43}{5} & \sim P(0.5) \end{cases} \\ X = 7 & \begin{cases} \frac{41}{7} & \sim P(0.5) \\ \frac{43}{7} & \sim P(0.5) \end{cases} \end{cases}$$

Note that $P(x, y) \neq P(x)P(y)$.

Birthday paradox

$$P(\text{all unique}) = P\left(\begin{matrix}\text{the } i\text{th is} \\ \text{different} \\ \text{from the} \\ i-1\end{matrix} \middle| \begin{matrix} i-1 \text{ are} \\ \text{different}\end{matrix}\right) = \sum_{i=1}^{k} \left(1 - \frac{i-1}{365}\right)$$

1.3 Problem 2

$Cost(k) =$ Represents the cost of one night @k tokens

$$E(Cost(k)) = \frac{C}{365} \cdot k + \sum_{i=k+1}^{\infty} Poi(3,i)(i-k)\frac{C}{150}$$

$$= \frac{C}{365} \cdot k - \frac{C}{150} \cdot k \underbrace{\sum_{i=k+1}^{\infty} Poi(3,1)}_{1-CDFPoi(3,k)} + \frac{C}{150} \sum_{i=k+1}^{\infty} i \cdot Poi(3,i)$$

$$E - \sum_{i=0}^{k} i \cdot Poi(3,i)$$

1.3.1 Part 2

$$\frac{c}{23}k - \frac{c}{20}k \sum_{i=4k+1}^{\infty} Poi(3000,i)$$

$$(1 - CDFPoi(3000,i))$$

$$E - \sum_{i=0}^{4k} i \cdot Poi(3000,i)$$

1.4 Problem

$$\sum_{i=0}^{n} \binom{n}{k} i^2 p^i (1-p)^{n-i} = \underbrace{np(1-p)}_{V(X)} + \underbrace{(np)^2}_{E(X)^2}$$

He said that moving from what was written in the homework to what he wrote on the board is simple algebra, also noting that $E(X^2) - (E(X))^2$ is the definition of the left side.

Extra Change the i^2 to i^4.

2 Matlab

- All IDC computer labs have Matlab installed
- In MyIDC > IDCApps
- 65 USD for student license

gmdistribution gmdistribution is a

$$f_Y(t) = \sum_{i=1}^{k} w_i f_{X_i}(t) \text{ where } X_i \sim N(\mu_i, \delta_i^2)$$

$$F_Y(t) = \sum_{i=1}^{k} w_i F_{X_i}(t)$$

3 Central Limit Theorem

Let $X_1, ..., X_n$ be random variables sampled independently from the same distribution with mean μ and nonzero variance σ^2. Let $\overline{X_n}$ be the average of $X_1, ..., X_n$.

$$\lim_{n\to\infty}\left(\frac{\sqrt{n}}{\sigma}(\overline{X_n}-\mu)\leq x\right)=\Phi(x)$$

Where $\Phi(x)$ is the standard normal density function.

Intuition For almost any distribution, if we sample it enough, the average distance of our samples from the mean is normally distributed.

3.1 Example

Let $X \sim U[0,1]$. Assume that we drew $n = 12$ samples.

What is $P\left(|\overline{X_n}-\frac{1}{2}|\leq 0.1\right)$? One way is to compute the distribution of $\overline{X_n}$. Much easier is to use the CLT as such:

For $U \sim U([0,1])$, we get $E(U) = \frac{1}{2}$ and $Var(U) = \frac{1}{12}$. Assuming that $n = 12$ is large enough, we use the CLT:

3.2 Sum of Poissons

Recall that the sum of Poissons is Poisson. Therefore, when λ is sufficiently large, we can write $Poi(\lambda)$ as the sum of λ independent copies of $Poi(1)$.

Example

$$Poi(3000)=\sum_{i=0}^{3000}\bar{X}_i$$

$$\frac{n\overline{X_n}-n\mu}{\sqrt{n}\sigma}\sim N(0,1)$$

$$CDFPoi(3000,3017)\sim\Phi\left(\frac{3017-3000}{\sqrt{3000}}\right)$$

3.3 Confidence Intervals

Given a "fair" coin that is tossed 100 times, and turns up heads 61 times. What is the probability of this event? Can we believe the "fairness" of the coin?

Using CLT:

$$P\left(\frac{\sqrt{100}}{0.5}(\overline{X_{100}}-0.5)\leq x\right)\cong\Phi(x)$$

$$P\left((\overline{X_{100}}-0.5)\leq\frac{x}{20}\right)\cong\Phi(x)$$

$$P((\overline{X_{100}}-0.5)\leq 0.11)\cong\Phi(2.2)$$

Thus, the probability of a fair coin landing heads more than 60 times is 0.023 and we can decide if we want to trust the coin, or if we reject the null model. It's important to remember that the null model has several components:

- The coin is fair
- The same coin was tossed 100 times
- The tosses are independent of each other

Continuing What if the coin turned up heads 54 times? We get this is > 0.15.

Defn Given a predetermined desired confidence level, we can find a range around μ, such that if we observe a sample outside this range, we should reject the null model. It is sometimes used in the inverse, by virtually rejecting all other options.

3.3.1 Example: Cholesterol Levels

In a survey in the early 2010s, the US HRSA found that the mean total cholesterol (TCh) level for males aged 20 and over is 211 mg/dL and the stddev is 46 mg/dL.

We are sampling 25 individuals in some specific hospital for performing a healthcare experiment. We measure the sample mean for TCh in this population and want to test whether TCh might be a confounding variable. We will reject the null hypothesis if the sampel mean falls outside an interval around $\mu = 211$ where it should fall in 90% of all cases. We use the CLT to compute the interval such that:

$$P((\overline{X_{25}} - 211) \leq 9.2x) \cong \Phi(x) = 0.95$$

We check our tables to get that x is around 1.65. Thus:

$$15.2 \leq (\overline{X_{25}} - 211) \leq 1.65 \cdot 9.2 = 15.2$$

This gives us a confidence interval of $(196, 226)$[1] for rejecting the null hypothesis.

3.4 Gene Regulation

3.5 Internet CTR

Checking success of a banner ad. First, we assume clicks are independent.

3.5.1 A

If we repeat experiments in which

[1] In scipy: 'norm.interval(0.9, loc=211, scale=46/math.sqrt(25))'

3.5.2 B

4 Log Normal Distribution

A random variable Y is said to have a log-normal distribution if its log has a normal guassian distribution.

$Y = e^{\mu+\sigma Z}$ where Z is the standard normal.

Log-Normals are useful for intrinsically positive quantities. Note that the mean, median, and mode are all different, and the distribution has a heavy tail.

5 Homework 2

Due December 17th.

5.1 Problem 3

Let $X_0 \sim U\{-5, ..., 5\}$.[2]

[2] "Same as X_0" means resample