, 8, 9, 10

Stat 231 - Statistics **Spring 2017**

Lecture 7, 8, 9, 10: May 15th - 23rd, 2017

Lecturer: Suryapratim Banerjee *Notes By: Harsh Mistry*

7.1 Empirical C.D.F

Consider : Data Set = $\{y_1, \ldots, y_n\}$

Definition 7.1 *The empirical c.d.f F(y) is defined as*

$$F(y) = \frac{\textbf{\textit{Number of observations}} \ \leq y}{n}$$

The graph $(y, F(y))$ *is the empirical C.D.F graph*

7.2 Box-Plot

We use box plots to compare two or bore data sets to each other.

- Lower Side is Q_1
- Upper Side is Q_3
- Median is also marked: Q_2
- Upper whisker stops at the maximum value of your data set which is less than or equal to $Q_3 + 1.5IQR$
- The lower whisker stops at the minimum value which is higher than or equal to $Q_1 - 1.5IQR$
- Any observations outside the whiskers are marked individually and are outliers

7.3 Scatter Plots

We use scatter plots to find whether there is an association between X and Y.
A scatter plot = $(x_1, y_1$ for $i = 1, \ldots, n$

7.4 Types of Inference Problems

An inference is when we use descriptive statistics to make statements about the population. The different types of inference problems are :

- Estimation
- Testing of Hypothesis
- Prediction (Forecasting)

7.4.1 Estimation

Definition 7.2 *The method by which we "guess" the population attributes using sample values.*

Notations

- Variables that are known are represented by Greek letters with a hat
- Variables that do not have a hat, represent unknown values

7.4.2 Hypothesis Testing Problems

Definition 7.3 *A hypothesis is a claim made about the population*

7.4.3 Prediction Problems

Definition 7.4 *Prediction problems often consist of a data set and involve forecasting the future values*

$$\hat{y}_{n+1} = ? \quad \hat{y}_{n+2} = ?$$

7.5 Statistical Models

Notation : θ represents the population we are interested in.

A model in statistics is the Identification of the distribution of the random variable Y_i from which y_i is drawn. Moreover θ is a parameter of this distribution.

$$Y \sim f(y;\theta)$$

Example 7.5 *Trudeau's approval rating : α - unknown*

To determine the rating a sample of 100 voters and the number of voters who approve is found to be $y = 120$. In this case y is drawn from $Y \sim Bin(200, \alpha)$

7.6 The Theory of Estimation

θ is the population attribute that we are interested in, otherwise know as the **Parameter of Interest** and will never be known unless we have the entire population.

The objective is to fine an "estimate" of θ,using our sample observations $\hat{\theta}$.

$$\hat{\theta}(y_1, \dots, y_n) = \text{ Known \# once the sample is known}$$

1. Set up the statistical model.
2. "Identify" the random variable from which your data set is drawn.

Note : θ is a parameter of the random variable

7.7 Discrete Random Variables

The likelihood function $L(\theta; y_1, \dots, y_n)$ = Probability of observing your sample as a function of the unknown parameter θ .

The maximum Likelihood Estimated (MLE) $\hat{\theta}$ is the value maximizes $L(\theta)$

$$L(\theta, y_1, \dots, y_n) = P(Y_1 = y_1, Y_2, Y_2, \dots, Y_n = y_n)$$

If the sample is independent and identically (i.i.d.)

$$\begin{aligned} L(\theta; y_1, \dots, y_n) &= P(Y_1 = y_1, Y_2 = y_2, \dots, Y_n = y_n) \\ &= P(Y_1 = y_i) \dots P(Y_n = y_n) \\ &= f(y_1) f(y_2) \dots f(y_n) \\ &= \prod_{i=1}^{n} f(y_i; \theta) \end{aligned}$$

Geometric Model

$$P(Y = y) = (1 - \theta)^y \cdot \theta$$
$$L(\theta) = (1 - \theta)^{y_1} \theta \dots (1 - \theta)^{y_n} \theta = (1 - \theta)^{\sum y_i} \theta^n$$
$$l(\theta) = \sum y_i \cdot \ln(1 - \theta) + n \ln(\theta)$$

Binomial Model

$$L(\lambda) = {}^nC_y \lambda^y (1 - \lambda)^{n-y}$$
$$l(\lambda) = \ln({}^nC_y) + y \ln(\lambda) + (n - y) \ln(1 - pi)$$

Notes about the Likelihood Functions

- One can think of the sample mean not just as a number, but also as an outcome of some r.v
- $L(\theta) = \prod_{i=1}^{n} f(y_i; \theta)$: The value of the likelihood function is really small for large n.

7.7.1 Relative Likelihood Function

$$R(\theta) = \frac{L(\theta)}{L(\hat{\theta})}, \text{ Where } \hat{\theta} = MLE$$

$$R(\theta) \geq 0, \forall\theta$$

$$R(\theta) = 1, \text{ if } \theta = \hat{\theta}$$

Note : The relative likelihood function also tells the "reasonable" values of θ (The values that are "close" to $\hat{\theta}$