Module 2 (Statistical modelling)

Selvakkadunko Selvaratnam

University of Toronto

Discrete random variables

DEFINITION

Let Ω be a sample space. A discrete random variable is a function $X : \Omega \to \Re$ that takes on a finite number of values $a_1, a_2, \cdots, a_n$ or an infinite number of values $a_1, a_2, \cdots$

DEFINITION

The probability that X takes on the value x, $P(X = x)$, is defined as the sum of the probabilities of all sample points in Ω that are assigned the value x. We will sometimes denote $P(X = x)$ by $p(x)$.

DEFINITION

The probability distribution for a discrete variable X can be represented by a formula, a table, or a graph that provides $p(x) = P(X = x)$ for all x.

Result

For any discrete probability distribution, the following must be true:

(1) $0 \leq p(x) \leq 1$ for all x,

(2) $\sum_x p(x) = 1$, where the summation is over all values of x with nonzero probability.

Discrete random variables

Expected value of a random variable X

Let X be a discrete random variable with the probability function $p(x)$. Then the expected value of X, $E(X)$, is defined to be

$$E(X) = \sum_x xP(x),$$

where $P(x) = P(X = x)$.

Variance of a random variable X

If X is a random variable with mean $E(X) = \mu$, the variance of a random variable X is defined to the expected value of $(X - \mu)^2$. That is,

$$V(X) = E[(X - \mu)^2].$$

The standard deviation of X is the positive square root of $V(X)$.

Binomial distributions

Properties of Binomial distribution

- The experiments consists of a fixed number, n, of identical trials.
- Each trial results in one of two outcomes: success, S, or failure, F.
- The probability of success on a single trial is equal to some value p and remains the same from trial to trial. The probability of a failure is equal to $q = (1 - p)$.
- The trails are independent.

Binomial distribution

A discrete random variable X has a binomial distribution with parameters n and p, where $n = 1, 2, \cdots$ and $0 \leq p \leq 1$, if its probability mass function is given by

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x} \text{ for } x = 0, 1, 2, \cdots, n,$$

where $\binom{n}{x} = \dfrac{n!}{(n - x)!\, x!}$. We denote this distribution by $B(n, p)$. We have

1. $E(X) = np$,
2. $V(X) = np(1 - p)$,

The Poisson probability distribution

Why Poisson Distribution?

Suppose that we want to find the probability distribution of the number of automobile accidents at a particular intersection during a time period of one week. At first glance the random variable, the number of accidents, may not seem even remotely related to a binomial random variable, but we will see that an interesting relationship exists.

Poisson distribution

A random variable X is said to have a Poisson probability distribution if and only if

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!} \text{ for } x = 0, 1, 2, \cdots, \lambda > 0.$$

We write $X \sim Pois(\lambda)$

Result

If $X \sim Pois(\lambda)$, show that

1. $E(X) = \lambda$
2. $V(X) = \lambda$

Continuous random variable

Definition

A random variable that can take on any value in an interval is called **continuous**, and the purpose of this module is to study probability distribution for continuous random variables.

Definition of distribution function

Let Y denote any random variable. The **distribution function** of Y, denoted by $F(y)$, is such that $F(y) = P(Y \leq y)$ for $-\infty < y < \infty$.

Properties of Distribution Function

If $F(y)$ is a distribution function, then

(a) $F(-\infty) = \lim_{y \to -\infty} F(y) = 0$.

(b) $F(\infty) = \lim_{y \to \infty} F(y) = 1$.

(c) $F(y)$ is a nondecreasing function of y. [If y_1 and y_2 are any values such that $y_1 < y_2$, then $F(y_1) \leq F(y_2)$]

Probability density function

Definition

A random variable Y with distribution function $F(y)$ is said to be **continuous** if $F(y)$ is continuous, for $-\infty < y < \infty$.

Probability density function

Let $F(y)$ be the distribution function for a continuous random variable Y. Then, $f(y)$, given by

$$f(y) = \frac{dF(y)}{dy} = F'(y)$$

wherever the derivative exists, is called the **probability density function** for the random variable Y.

Properties of a density function

If $f(y)$ is a density function for a continuous random variable, then

(a) $f(y) \geq 0$ for all y, $-\infty < y < \infty$.

(b) $\int_{-\infty}^{\infty} f(y)dy = 1$.

The Uniform Probability Distribution

Definition

If $a < b$, a random variable Y is said to have a continuous **uniform probability distribution** on the interval (a, b) if and only if the density function of Y is

$$f(y) = \begin{cases} \dfrac{1}{(b-a)}, & a \leq y \leq b; \\ 0, & \text{elsewhere.} \end{cases}$$

Result

If $a < b$ and Y is a random variable uniformly distributed on the interval (a, b), then we have

$$\mu = E(Y) = \frac{a+b}{2} \text{ and } \sigma^2 = V(Y) = \frac{(b-a)^2}{12}.$$

The Normal Probability Distribution

The most widely used continuous probability distribution is the normal distribution, a distribution with the familiar bell shape that was discussed in connection with the empirical rule.

Definition

A random variable Y is said to have a **normal probability distribution** if and only if, for $\sigma > 0$ and $-\infty < \mu < \infty$, the density function of Y is

$$f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-\mu)^2/(2\sigma^2)}, -\infty < y < \infty.$$

Result

If Y is a normally distributed random variable with parameters μ and σ, then we have

$$E(Y) = \mu \text{ and } V(Y) = \sigma^2.$$

Exponential distribution

Definition

A random variable Y is said to have an exponential distribution with parameter $\lambda > 0$ if and only if the density of Y is $f(y) = \begin{cases} \lambda e^{-\lambda y}, & 0 \leq y < \infty; \\ 0, & \text{elsewhere.} \end{cases}$

Result

If $Y \sim \text{Exp}(\lambda)$, then, we have

1. $\mu = E(Y) = \frac{1}{\lambda}$,
2. $\sigma^2 = V(Y) = \frac{1}{\lambda^2}$,
3. $F(y) = 1 - e^{-\lambda y}; y \geq 0$.

The law of large numbers

Basics

1. We study limits in probability to understand the long-term behavior of random process and sequence of random variables.
2. The use of simulation to approximate the probability of an event A is justified by one of the most important limit results in probability - the law of large numbers.
3. A consequence of the law of the large numbers is that repeated trials of a random experiment the proportion of trials in which A occurs converges to $P(A)$, as the number of trials $(= n)$ goes to infinity.

Notations

1. We consider a sequence of random variables $X_1, X_2, X_3, \cdots, \cdots$.
2. We should think of X_i as the result of the ith repetition of a particular measurement or experiment.
3. Also, we have $X_1, X_2, X_3, \cdots, \cdots$ independent and identically distributed sequence.
4. We shall denote the distribution function of each random variable X_i by F, its expectation by μ, and the standard deviation by σ.

The law of large numbers

Expectation and variance of an average

1. The average of the first n random variables in the sequence is

$$\bar{X}_n = \frac{X_1 + X_2 + \cdots + X_n}{n}.$$

2. We have $E(\bar{X}_n) = \mu$ and $Var(\bar{X}_n) = \frac{\sigma^2}{n}$.

Chebyshev's inequality

For an arbitrary random variable X and any $\epsilon > 0$:

$$P(|X - E(X)| \geq \epsilon) \leq \frac{1}{\epsilon^2} Var(X).$$

Chebyshev's inequality

Result

Denote $Var(X)$ by σ^2 and consider the probability that X is within a few standard deviations from its expectation μ:

$$P(|X - E(X)| < k\sigma) \geq 1 - \frac{1}{k^2}.$$

where k is a small integer. For k = 2, 3, 4 the right-hand side is 3/4, 8/9, and 15/16, respectively.

e.g 1:

Calculate $P(|Y - \mu| < k\sigma)$ exactly for k = 1, 2, 3, 4 when Y has an exponential distribution with mean 1 and compare this with the bounds from Chebyshev's inequality.

The law of large numbers

Introduction

- The Strong Hotel has infinitely many rooms. In each room, a guest is flipping coins - forever. Each guest generates an infinite sequence of zeros and ones. We are interested in the limiting behavior of the sequences in each room.
- The strong law of large numbers says that in **virtually every room of the hotel** the sequence of averages will converge to 1/2. And not only will these averages get arbitrarily close to 1/2 after a very long time, but each will stay close to 1/2 for all the remaining terms of the sequence. Those sequences whose averages converge to 1/2 constitute a set of "probability 1." And those sequences whose averages do not converge to 1/2 constitute a set of "probability 0."

The law of large numbers

Weak law of large numbers

Let $X_1, X_2, \cdots$ be an i.i.d sequence of random variables with finite mean μ and variance σ^2. For $n = 1, 2, \cdots$, let $S_n = X_1 + \cdots, X_n$. Then

$$\lim_{n\to\infty} P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) = 0.$$

That is, we say that $\bar{X}_n$ converges in probability to μ and also we can write

$$\bar{X}_n \xrightarrow{P} \mu.$$

Strong law of large numbers

Let $X_1, X_2, \cdots$ be an i.i.d sequence of random variables with finite mean μ. For $n = 1, 2, \cdots$, let $S_n = X_1 + \cdots, X_n$. Then

$$P\left(\lim_{n\to\infty} \frac{S_n}{n} = \mu\right) = 1.$$

We say that S_n/n converges to μ with probability 1.

Strong law of large numbers

e.g 2

Write the code to illustrate the Strong Law of Large Numbers (SLLN) for independent and identically distributed (i.i.d) sequence for the Binomial distribution with the number of trials, $s = 12$ and the probability of success,

(a) $p = 0.5$,

(b) $p = 0.7$.

Simulate $N = 10000$ times and use ggplot to generate graphs to visualize SLLN for the above two scenarios.

Introduction

Statistical methods are particularly useful for studying, analyzing, and learning about populations of experimental units.

Experimental unit

An experimental (or observational) unit is an object (e.g., person, thing, transaction, or event) about which we collect data.

Population

A population is a set of all units (usually people, objects, transactions, or events) that we are interested in studying.
For example, populations may include

1. all employed workers in the United States,
2. all registered voters in California,
3. everyone who is afflicted with AIDS,
4. all the cars produced last year by a particular company.

Introduction

Variable

A variable is a characteristic or property of an individual experimental (or observational) unit in the population.
For example, we may be interested in the variables age, gender, and number of years of education of the people currently unemployed in the United States.

Measurement

Measurement is the process we use to assign numbers to variables of individual population units.

Sample

A sample is a subset of the units of a population.
For example, instead of polling all 145 million registered voters in the United States during a presidential election year, a pollster might select and question a sample of just 1,500 voters.

Introduction

Statistical inference

A statistical inference is an estimate, prediction, or some other generalization about a population based on information contained in a sample.

Observed sample

We consider the random variables $Y_1, Y_2, \cdots, Y_n$ observed in a random sample from a population of interest. The random variables $Y_1, Y_2, \cdots, Y_n$ are independent and have the same distribution that has a population mean μ and population variance σ^2. We are interested to estimate the population mean μ.

Sample mean

The sample mean can be calculated by

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

Introduction

Sample variance

The sample variance can be computed by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

Population parameter

A population parameter is an unknown numerical value. We are interested to conduct a statistical inference about the population parameter. For instance, the population mean μ and variance σ^2 are the examples of population parameters.

Statistic

A statistic is a function of the observable random variables in a sample and known constants. For instance, the sample mean $\bar{Y}$ and the sample variance S^2 are the examples for statistic.

Introduction

Sampling Distribution

The main objective of an experiment is to conduct the statistical inference about the population mean μ. We do not know any information about the population. But, we have a sample information. Then, how we can connect the sample information to population? A sampling distribution is a bridge between the sample information and population.

The sampling distribution of the sample mean $\bar{Y}$

Properties of the Sampling Distribution of $\bar{Y}$

A random sample of size n has been chosen from any population with mean μ and standard deviation σ. Let $Y_1, Y_2, \cdots, Y_n$ are the random sample.

1. The mean of the sampling distribution $\bar{Y}$, $\mu_{\bar{Y}} = E(\bar{Y}) = \mu$.
2. The standard deviation of the sampling distribution $\bar{Y}$, $\sigma_{\bar{Y}} = \sigma/\sqrt{n}$.
3. The sample mean $\bar{Y}$ is a point estimator for the population mean μ.

Simple Linear Regression

Deterministic Model

If we were to construct a model that hypothesized an exact relationship between variables, it would be called a deterministic model. For example, if we believe that y, the reaction time (in seconds), will be exactly one-and-one-half times x, the amount of drug in the blood, we write

$$y = 1.5x$$

This represents a deterministic relationship between the variables y and x. It implies that y can always be determined exactly when the value of x is known. There is no allowance for error in this prediction.

Probabilistic Models

Our probabilistic model will include both a deterministic component and a random error component. For example, if we hypothesize that the response time y is related to the percentage x of drug by

$$y = 1.5x + \text{Random error}$$

we are hypothesizing a probabilistic relationship between y and x.

Simple Linear Regression

Model Description

In this module, we present the simplest of probabilistic models—the straight-line model—which gets its name from the fact that the deterministic portion of the model graphs as a straight line. Fitting this model to a set of data is an example of regression analysis, or regression modeling. A First-Order (Straight-Line) Probabilistic Model

$$y = \beta_0 + \beta_1 x + \epsilon,$$

where
y = Dependent or response variable (quantitative variable to be modeled or predicted)
x = Independent or predictor variable (quantitative variable used as a predictor of y)
$\beta_0 + \beta_1 x$ = Deterministic component
Random error component ϵ is assumed to follow a $N(0, \sigma)$ distribution.
β_0 = y-intercept of the line
β_1 = Slope of the line

Estimating Model Parameters

Observed Sample

Let $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ be the observed n-pairs.

PRINCIPLE OF LEAST SQUARES

The vertical deviation of the point (x_i, y_i) from a line $y = b_0 + b_1 x$ is

$$\text{height of point - height of line} = y_i - (b_0 + b_1 x_i)$$

The sum of squared vertical deviations from the points $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ to the line is then

$$g(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2.$$

The point estimates of β_0 and β_1, denoted by $\hat{\beta}_0$ and $\hat{\beta}_1$ and called the least squares estimates, are those values that minimize $g(b_0, b_1)$. The estimated regression line or least squares regression line (LSRL) is then the line whose equation is

$$y = \hat{\beta}_0 + \hat{\beta}_1 x.$$

Examples

e.g 3:

As brick-and-mortar shops decline and online retailers like Amazon and Wayfair ascend, demand for warehouse storage space has steadily increased. Despite effectively being empty shells, warehouses still require professional appraisal. The following data on x= truss height (ft), which determines how high stored goods can be stacked, and y = sale price ($) per square foot appeared in the article "Challenges in Appraising 'Simple' Warehouse Properties" (The Appraisal J. 2001: 174–178).

Warehouse	1	2	3	4	5	6	7	8
x	12	14	14	15	15	16	18	22
y	35.53	37.82	36.90	40.00	38.00	37.50	41.00	48.50
Warehouse	9	10	11	12	13	14	15	16
x	24	26	26	27	28	30	30	33
y	46.20	50.35	49.13	48.07	50.90	54.78	54.32	57.17
Warehouse	17	18	19					
x	22	24	36					
y	47.00	47.50	57.45					

e.g 3 continue:

R output for e.g 1:

```
Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-2.35522 -0.63584 -0.08796  0.92263  3.01053

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.77215    1.11347   21.35 1.03e-13 ***
x            0.98715    0.04684   21.07 1.27e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.416 on 17 degrees of freedom
Multiple R-squared:  0.9631,    Adjusted R-squared:  0.961
F-statistic: 444.1 on 1 and 17 DF,  p-value: 1.271e-13
```

Examples

e.g 3 continue:

(a) Write the equation of the straight-line, probabilistic model.

(b) A point estimate for the true average price for all warehouses with 25-ft truss height.

(c) Compute the coefficient of determination, R^2 and interpret the result.

Examples

e.g 4:

Let $X_1, X_2, \cdots$ be an i.i.d sequence of random variables with finite mean μ and variance σ^2. For $n = 1, 2, \cdots$, let $S_n = X_1 + \cdots, X_n$. Then show that

$$\bar{X}_n \xrightarrow{P} \mu,$$

where $\bar{X}_n = \dfrac{S_n}{n}$. We can say that $\bar{X}_n$ is a consistent estimator for μ.

e.g 5:

Let $X_1, X_2, \cdots$ be an i.i.d sequence of random variables with finite mean μ and variance σ^2. Show that

$$S_n^2 \quad = \quad \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

is a consistent estimator of σ^2.

Examples

e.g 6:

We consider a data for miles per gallon (MPG) of 100 cars. The histogram for the data is shown below

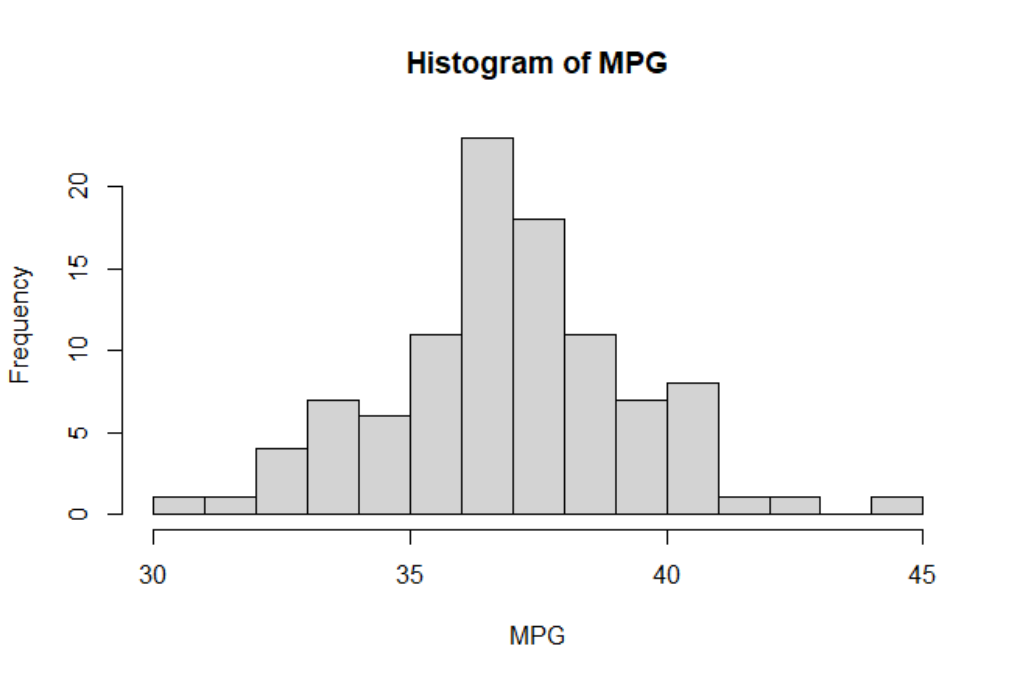

Figure 1

What is an appropriate distribution for this data?

Examples

e.g 7:

The built-in ‘rivers’ dataset in R gives the lengths in miles of 141 North American rivers compiled by the US Geological Survey. The histogram for the data is shown below

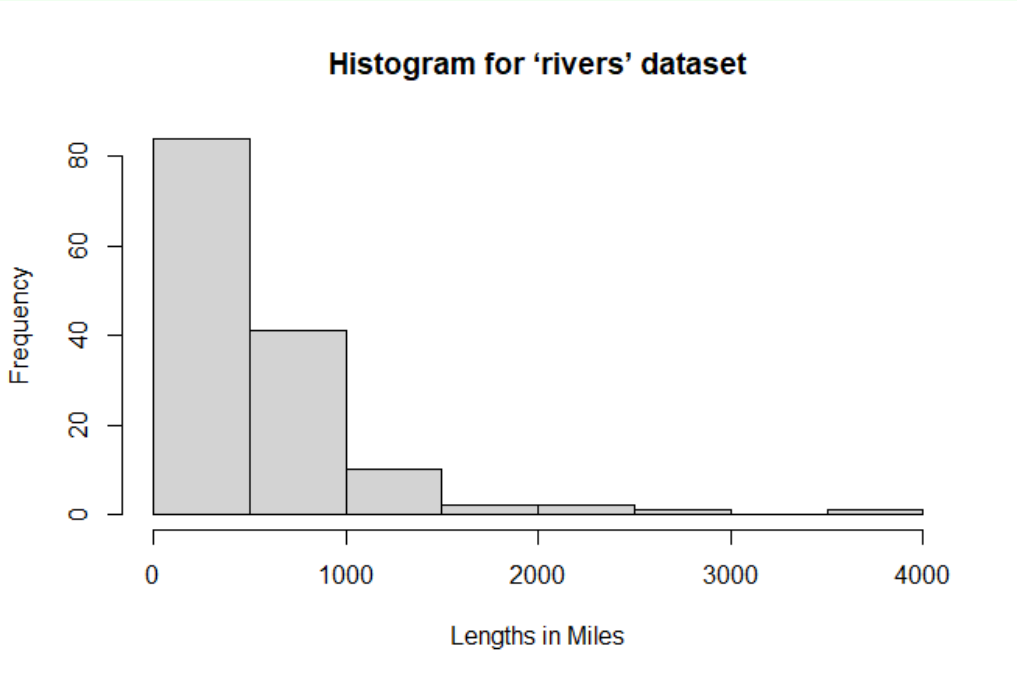

Figure 2

What is an appropriate distribution for this data?