STA 2200 PROBABILITY AND STATISTICS II

Purpose At the end of the course the student should be able to handle problems involving probability distributions of a discrete or a continuous random variable.

Objectives

By the end of this course the student should be able to;

(1) Define the probability mass, density and distribution functions, and to use these to determine expectation, variance, percentiles and mode for a given distribution
(2) Appreciate the form of the probability mass functions for the binomial, geometric, hypergeometic and Poisson distributions, and the probability density functions for the uniform, exponential gamma , beta and normal, functions, and their applications
(3) Apply the moment generating function and transformation of variable techniques
(4) Apply the principles of statistical inference for one sample problems.

DESCRIPTION

Random variables: discrete and continuous, probability mass, density and distribution functions, expectation, variance, percentiles and mode. Moments and moment generating function. Moment generating function and transformation Change of variable technique for univariate distribution. Probability distributions: hypergeometric, binomial, Poisson, uniform, normal, beta and gamma. Statistical inference including one sample normal and t tests.

Pre-Requisites: STA 2100 Probability and Statistics I, SMA 2104 Mathematics for Science

Course Text Books:

1) RV Hogg, JW McKean & AT Craig *Introduction to Mathematical Statistics*, 6th ed., Prentice Hall, 2003 ISBN 0-13-177698-3
2) J Crawshaw & J Chambers *A Concise Course in A-Level statistics, with worked examples*, 3rd ed. Stanley Thornes, 1994 ISBN 0-534- 42362-0

Course Journals:

1) Journal of Applied Statistics (J. Appl. Stat.) [0266-4763; 1360-0532]
2) Statistics (Statistics) [0233-1888]

Further Reference Text Books And Journals:

a) HJ Larson *Introduction to Probability Theory and Statistical Inference(Probability and Mathematical Statistics)* 3rd ed., Wiley, 1982
b) Uppal, S. M. , Odhiambo, R. O. & Humphreys, H. M. *Introduction to Probability and Statistics*. JKUAT Press, 2005
c) I Miller & M Miller *John E Freund's Mathematical Statistics with Applications,* 7th ed., Pearsons Education, Prentice Hall, New Jersey, 2003 ISBN: 0131246461
d) Statistical Science (Stat. Sci.) [0883-4237]
e) Journal of Mathematical Sciences
f) The Annals of Applied Probability

3. Let X be the random variable the number of fours observed when two dice are rolled together once. Show that X is a discrete random variable.
4. The pmf of a discrete random variable X is given by $P(X = x) = kx$ for $x = 1,2,3, 4,5,6$ Find the value of the constant k, $P(X < 4)$ and $P(3 \le X < 6)$
5. A fair coin is flip until a head appears. Let N represent the number of tosses required to realize a head. Find the pmf of N
6. A discrete random variable Y has a pmf given by $P(Y = y) = c\left(\frac{3}{4}\right)^x$ for $y = 0,1,2,.....$ Find the value of the constant c and $P(X < 3)$
7. Verify that $f(x) = \dfrac{2x}{k(k+1)}$ for $x = 0,1,2,.....k$ can serve as a pmf of a random variable X.
8. For each of the following determine c so that the function can serve as a pmf of a random variable X.

   a) $f(x) = cx$ for $x = 1,2,3,4,5$

   b) $f(x) = cx^2$ for $x = 0,1,2,.....k$

   c) $f(x) = c\left(\frac{1}{6}\right)^x$ for $x = 0,1,2,3.....$

   d) $f(x) = c2^{-x}$ for $x =$ for $x = 0,1,2,....$
9. A coin is loaded so that heads is three times as likely as the tails. For 3 independent tosses of the coin find the pmf of the total number of heads realized and the probability of realizing at most 2 heads.

1.3 Continuous Random Variables and Probability Density Function

A **continuous** random variable can assume any value in an interval on the real line or in a collection of intervals. The sample space is uncountable. For instance, suppose an experiment involves observing the arrival of cars at a certain period of time along a highway on a particular day. Let T denote the time that lapses before the 1st arrival, the T is a continuous random variable that assumes values in the interval $[0,\infty)$

Definition: A random variable *X* is *continuous* if there exists a nonnegative function f so that, for every interval *B*, $P(X \in B) = \int_B f(x)\,dx$. The function $f = f(x)$ is called the *probability density function* of *X*.

Definition: Let X be a continuous random variable that assumes values in the interval $(-\infty,\infty)$, The f(x) is said to be a probability density function (pdf) of X if it satisfies the following conditions

$$f(x) \ge 0 \text{ for all x }, \; p(a \le x \le b) = \int_a^b f(x)\,dx = 1 \text{ and } \int_{-\infty}^{\infty} f(x)\,dx = 1$$

The support of a continuous random variable is the smallest interval containing all values of x where f(x) >= 0.

Remark A crucial property is that, for any real number x, we have $P(X = x) = 0$ (implying there is no difference between $P(X \le x)$ and $P(X < x)$); that is it is not possible to talk about the probability of the random variable assuming a particular value. Instead, we talk about the probability of the random variable assuming a value within a given interval. The probability of the random variable assuming a value within some given interval from $x = a$ to $x = b$ is defined to be the area under the graph of the probability density function between $x = a$ and $x = b$.

Example 1

Let X be a continuous random variable. Show that the function

For a 1-1 relationship between X and Y eg $Y = 2X + 3$, $f(x)$ and $g(y)$ yields exactltly the same probabilities only the random variable and the set of values it can assume changes.

Example 1

Give the pmf of a random variable X as $f(x) = \begin{cases} x/6 & \text{for } x = 1,2,3 \\ 0, & elsewhere \end{cases}$ find the pmf of $Y = X^2$

Solution

The only values of Y with non zero probabilities are $Y = 1$, $Y = 4$ and $Y = 9$. Now

$P(Y=1) = P(X^2 = 1) = P(X=1) = \frac{1}{6}$ $\quad$ $P(Y=4) = P(X^2 = 4) = P(X=2) = \frac{1}{3}$ and

$P(Y=9) = P(X^2 = 9) = P(X=3) = \frac{1}{2}$

In some cases several values of X will give rise to the same value of Y. The procedure is just the same as above but it is necessary to add the several probabilities that are associated with each value x that provides a unique value y.

Example 2

Give the pmf of a r.v X as $f(x) = \begin{cases} \dfrac{x+1}{15} & \text{for } x = 0,1,2,3,4 \\ 0, & elsewhere \end{cases}$ find the pmf of $Y = (X-2)^2$

Solution

x	0	1	2	3	4
y	4	1	0	1	4

$P(Y=0) = P(X=2) = \frac{1}{5}$ $\quad$ $P(Y=1) = P(X=1) + P(X=3) = \frac{2}{15} + \frac{4}{15} = \frac{2}{5}$

$P(Y=1) = P(X=0) = P(X=4) = \frac{1}{15} + \frac{1}{3} = \frac{2}{5}$ Therefore the pmf of Y can be written as

y	0	1	4
$P(Y=y)$	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{2}{5}$

Exercise

1. Suppose the pmf of a r.v X is given by $f(x) = \begin{cases} 1/6 & \text{for } x = 1,2,3,4,5,6 \\ 0, & elsewhere \end{cases}$, Obtain the pmf of $Y = 2X^2$ and $Z = X - 3$
2. Let the pmf of a r.v X be given by $f(x) = \begin{cases} \dfrac{x^2+1}{18} & \text{for } x = 0,1,2,3 \\ 0, & elsewhere \end{cases}$, determine the pmf of $Y = X^2 + 1$
3. Suppose the pmf of a r.v X is given by $f(x) = \begin{cases} x/10 & \text{for } x = 0,1,2,3,4 \\ 0, & elsewhere \end{cases}$, Obtain the pmf of $Y = |X - 2|$

Standard deviation $\sigma = \sqrt{E(X^2) - \mu^2} = \sqrt{{}^{11}/_6 - ({}^7/_6)^2} = \sqrt{{}^{17}/_6} \approx 1.6833$

Now $E(Y) = 12E(X) + 6 = 12({}^7/_6) + 6 = 20$

$Var(Y) = Var(12X + 6) = 12^2 \times Var(X) = 144 \times \sqrt{{}^{17}/_6} \approx 242.38812$

Example 3

A continuous random variable X has a pdf given by $f(x) = \begin{cases} \frac{1}{2}x, & 0 \le x \le 2 \\ 0, & elsewhere \end{cases}$, find the mean and standard of X

Solution

$$E(x) = \int_{-\infty}^{\infty} xf(x)dx = \int_0^2 \tfrac{1}{2}x^2 dx = \left[\frac{x^3}{6}\right]_0^2 = \frac{4}{3} \text{ and } E(x^2) = \int_{-\infty}^{\infty} x^2 f(x)dx = \int_0^2 \tfrac{1}{2}x^3 dx = \left[\frac{x^4}{8}\right]_0^2 = 2$$

Standard deviation $\sigma = \sqrt{E(X^2) - \mu^2} = \sqrt{2 - ({}^4/_3)^2} = \frac{\sqrt{2}}{3}$

Exercise

1. Suppose X has a probability mass function given by the table below

x	2	3	4	5	6
P(X=x)	0.01	0.25	0.4	0.3	0.04

Find the mean and variance of; X

2. Suppose X has a probability mass function given by the table below

x	11	12	13	14	15
P(X=x)	0.4	0.2	0.2	0.1	0.1

Find the mean and variance of; X

3. Let X be a random variable with $P(X = 1) = 0.2$, $P(X = 2) = 0.3$, and $P(X = 3) = 0.5$. What is the expected value and standard deviation of; a) X *b)* $Y = 5X - 10$?
4. A random variable W has the probability distribution shown below,

w	0	1	2	3
P(W=w)	2d	0.3	d	0.1

Find the values of the constant d hence determine the mean and variance of W. Also find the mean and variance of $Y = 10X + 25$

5. A random variable X has the probability distribution shown below,

x	1	2	3	4	5
P(X=x)	7c	5c	4c	3c	c

Find the values of the constant c hence determine the mean and variance of X.

6. The random variable Z has the probability distribution shown below,

z	2	3	5	7	11
P(Z=z)	⅙	⅓	¼	x	y

If $E(Z) = 4\frac{2}{3}$, find the values of x and y hence determine the variance of Z

7. A discrete random variable M has the probability distribution $f(m) = \begin{cases} \frac{m}{36}, & m = 1,2,3,...,8 \\ 0, & elsewhere \end{cases}$, find the mean and variance of M
8. For a discrete random variable Y the probability distribution is $f(y) = \begin{cases} \frac{5-y}{10}, & y = 1,2,3,4 \\ 0, & elsewhere \end{cases}$, calculate $E(Y)$ and var(Y)
9. Suppose X has a pmf given by $f(x) = \begin{cases} kx & \text{for } x = 1,2,3,4 \\ 0, & elsewhere \end{cases}$, find the value of the constant k hence obtain the mean and variance of X

Exercise

1. Over a very long period of time, it has been noted that on Friday's 25% of the customers at the drive-in window at the bank make deposits. What is the probability that it takes 4 customers at the drive-in window before the first one makes a deposit.
2. It is estimated that 45% of people in Fast-Food restaurants order a diet drink with their lunch. Find the probability that the fourth person orders a diet drink. Also find the probability that the first diet drinker of th e day occurs before the 5th person.
3. What is the probability of rolling a sum of seven in fewer than three rolls of a pair of dice? Hint (The random variable, X, is the number of rolls before a sum of 7.)
4. In New York City at rush hour, the chance that a taxicab passes someone and is available is 15%. a) How many cabs can you expect to pass you for you to find one that is free and b) what is the probability that more than 10 cabs pass you before you find one that is free.
5. An urn contains N white and M black balls. Balls are randomly selected, one at a time, until a black ball is obtained. If we assume that each selected ball is replaced before the next one is drawn, what is;
   a) the probability that exactly n draws are needed?
   b) the probability that at least k draws are needed?
   c) the expected value and Variance of the number of balls drawn?
6. In a gambling game a player tosses a coin until a head appears. He then receives $\$2n$, where n is the number of tosses.
   a) What is the probability that the player receives $8.00 in one play of the game?
   b) If the player must pay $5.00 to play, what is the win/loss per game?
7. An oil prospector will drill a succession of holes in a given area to find a productive well. The probability of success is 0.2.
   a) What is the probability that the 3rd hole drilled is the first to yield a productive well?
   b) If the prospector can afford to drill at most 10 well, what is the probability that he will fail to find a productive well?
8. A well-travelled highway has itstraffic lights green for 82% of the time. If a person travelling the road goes through 8 traffic intersections, complete the chart to find a) the probability that the first red light occur on the nth traffic light and b) the cumulative probability that the person will hit the red light on or before the nth traffic light.
9. An oil prospector will drill a succession of holes in a given area to find a productive well. The probability of success is 0.2.
   a) What is the probability that the 3rd hole drilled is the first to yield a productive well?
   b) If the prospector can afford to drill at most 10 well, what is the probability that he will fail to find a productive well?

2.1.5 The negative binomial distribution

Suppose a Bernoulli trial is performed until the t^{th} success is realized. Then the random variable "the number of trials until the t^{th} success is realized" has a negative binomial distribution

Definition: A random variable X has the negative binomial distribution, also called the Pascal distribution, denoted $X \sim NB(r, p)$, if there exists an integer $n \geq 1$ and a real number $p \in (0, 1)$ such that $P(X = r + x) = {}_{r+x-1}C_x \times p^r (1-p)^x = 1, 2, 3, \ldots$

If r=1 the negative binomial distribution reduces to a geometric distribution.

2.1.6 Hyper geometric Distribution

Hyper geometric experiments occur when the trials are not independent of each other and occur due to sampling without replacement hyper-geometric probabilities involve the

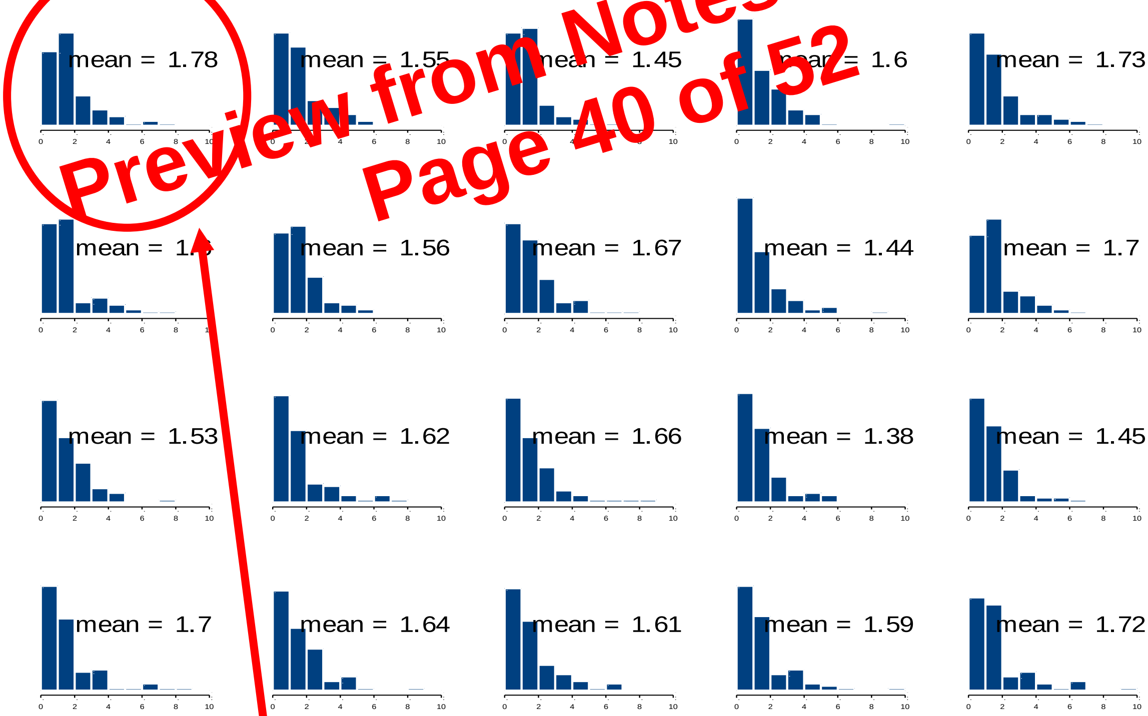
Exercise

1. If $X \sim N(65, 28)$ and $Y \sim N(85, 36)$ are 2 independent r.v, Find (a) $P(X + Y \leq 142)$ (b) $P(134 \leq X + Y \leq 166)$ (c) $P(Y - X > 4)$ (d) $P(12 \leq Y - X \leq 24)$
2. Each day Mr. Njoroge walks to the library bto read a newspaper. Total time spent walking is normally distributed with mean 15 minutes and standard deviation 2 minutes. Total time spent in the library is also normally distributed with mean 25 minutes and standard deviation $\sqrt{12}$ minutes. Find the probability that on one day;
   a) he is away from his home for more than 45 minutes.
   b) he spends more time walking than in the library

4.6 Sampling Distributions

In many investigations the data of interest can take on many possible values and it is often of interest to estimate the population mean, μ. A common estimator for μ is the sample mean $\bar{x}$. Consider the following set up: We observe a sample of size *n* from some population and compute the mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ .Since the particular individuals included in our sample are *random*, we would observe a different value of $\bar{x}$ if we repeated the procedure. That is, $\bar{x}$ *is also a random quantity*. Its value is determined partly by which people are randomly chosen to be in the sample. If we repeatedly drew samples of size *n* and calculated $\bar{x}$, we could ascertain the sampling distribution of $\bar{x}$.

Many possible samples, many possible $\bar{x}$'s



We only see one!

We will have a better idea of how good our one estimate is if we have good knowledge of how $\bar{x}$ behaves; that is, if we know the probability distribution of $\bar{x}$.

The sampling distribution is simply this probability distribution defined over all possible samples of size n from the population of size N. In the real world problems N will be large (e.g. 200 million US population) and n will be also be large (e.g., 1000 people surveyed) and ${}^{N}c_{n}$ will be astronomical number. Then the sampling distribution can only be imagined. We have chosen a simple example of N=6, n=2 so that the entire sampling distribution can be explicitly computed and visualized. Now the random variable is $\bar{x}$, it is no longer just X.

Definitions
Central Limit Theorem:- Stats that as the sample size increases, the sampling distribution of the sample means will become approximately normally distributed.
Sampling Distribution of the Sample Means:- Distribution obtained by using the means computed from random samples of a specific size.
Sampling Error :- Difference which occurs between the sample statistic and the population parameter due to the fact that the sample isn't a perfect representation of the population.
Standard Error or the Mean:- The standard deviation of the sampling distribution of the sample means. It is equal to the standard deviation of the population divided by the square root of the sample size.

4.6.2 The Mean and Standard Deviation of $\bar{x}$

What are the mean and standard deviation of $\bar{x}$?
Let's be more specific about what we mean by a sample of size *n*. We consider the sample to be a collection of *n independent and identically distributed* (or *iid*)random variables $X_1, X_2, \ldots, X_n$ with common mean μ and common standard deviation σ.

Thus, $E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(x_i) = \frac{1}{n}\sum_{i=1}^{n}\mu = \frac{1}{n}(n\mu) = \mu$

$$Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(x_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}(n\sigma^2) = \frac{\sigma^2}{n} \quad \Rightarrow SD(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

4.6.3 The Central Limit Theorem

Now we know that $\bar{x}$ has mean μ and standard deviation $\sigma/\sqrt{n}$, but what is its distribution?

If $X_1, X_2, \ldots, X_n$ are *normally distributed*, then $\bar{x}$ is also normally distributed. Thus, $X_i \sim N(\mu, \sigma^2) \Rightarrow \bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. If $X_1, X_2, \ldots, X_n$ are *not* normally distributed, then the Central Limit Theorem tells us that $\bar{x}$ is *approximately N*ormal.
In brief if $X_1, X_2, \ldots, X_n$ are *iid* random variables with mean μ and finite standard deviation σ. Then for a sufficiently large n, the sampling distribution of $\bar{X}$ is approximately Normal with mean μ and variance $\frac{\sigma^2}{n}$.

Remarks

- Central limit theorem involves two different distributions: the distribution of the original population and the distribution of the sample means
- The formula for a z-score when working with the sample means is: $z = \dfrac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ ~ N(0,1)

5.3 Approaches to Hypothesis Testing

There are three approaches to hypothesis testing namely Classical Approach, p vale approach and the confidence interval approach

5.3.1 The Classical Approach

The Classical Approach to hypothesis testing is to compare a test statistic and a critical value. It is best used for distributions which give areas and require you to look up the critical value (like the Student's t distribution) rather than distributions which have you look up a test statistic to find an area (like the normal distribution).

The Classical Approach also has three different decision rules, depending on whether it is a left tail, right tail, or two tail test.

One problem with the Classical Approach is that if a different level of significance is desired, a different critical value must be read from the table.

5.3.2 P-Value Approach

The P-Value Approach, short for Probability Value, approaches hypothesis testing from a different manner. Instead of comparing z-scores or t-scores as in the classical approach, you're comparing probabilities, or areas.

The level of significance (alpha) is the area in the critical region. That is, the area in the tails to the right or left of the critical values.

The p-value is the area to the right or left of the test statistic. If it is a two tail test, then look up the probability in one tail and double it.

If the test statistic is in the critical region, then the p-value will be less than the level of significance. It does not matter whether it is a left tail, right tail, or two tail test. This rule always holds.

Reject the null hypothesis if the p-value is less than the level of significance.

You will fail to reject the null hypothesis if the p-value is greater than or equal to the level of significance.

The p-value approach is best suited for the normal distribution when doing calculations by hand. However, many statistical packages will give the p-value but not the critical value. This is because it is easier for a computer or calculator to find the probability than it is to find the critical value.

Another benefit of the p-value is that the statistician immediately knows at what level the testing becomes significant. That is, a p-value of 0.06 would be rejected at an 0.10 level of significance, but it would fail to reject at an 0.05 level of significance. Warning: Do not decide on the level of significance after calculating the test statistic and finding the p-value.

Here are a couple of statements to help you keep the level of significance the probability value straight.

The Level of Significance is pre-determined before taking the sample. It **does not** depend on the sample at all. It is the area in the critical region,that is the area beyond the **c**ritical values. It is the probability at which we consider something unusua**l**.

The Probability-Value can only be found after taking the sample. It depends on the sample. It is the area beyond the test statistic**.** It is the probability of getting the results we obtained if the null hypothesis is true.