STATISTICS AND PROBABILITY

Table des matières

Chapitre 1

Introduction to statistics, charts and graph

Chapitre 2

Descriptive statistics

Chapitre 3

Probability

Chapitre 4

Discrete distributions

4.1 Definition and description of a discrete distribution

A random variable is a variable that contains the outcomes of a chance experiment. It represents a possible numerical value from a random experiment.

- Suppose an experiment is to measure the arrival X of automobiles at a tollbooth during a 30-second period. The possible outcomes ares 0 car, 1 car, 2 cars,... n cars.
- Roll a die twice. Let X be the number of times 4 comes up. (then X could be 0, 1, or 2 times)
- Toss a coin 5 times. Let X be the number of heads (then $X = 0, 1, 2, 3, 4, \text{ or }$ 5)
- Counting the number of people who arrive at a store during a 5-minute period. (then $X = 0, 1, 2, ...$)
- Measuring a day's temperature X : (then X could take any number in a range -5 to 25)

There are two categories of random variable : discrete random variables and continuous random variable

A random variable is a discrete random variable if the set of all possible values is at most a finite or countably infinite number of possible values. In most statistical situations, discrete random variables produce values that are nonnegative whole numbers.

- The number of heads in 25 tosses of coins

- The number of phone calls you receive per day

- Randomly selecting 25 people who consume soft drinks and determining how many prefer diet soft drinks

- Determining the number of defects in a batch of 50 items

- Sampling 100 registered voters and determining how many voted for the mayor in the last election

...

Let X be a discrete random variable and x be one of its possible values. The probability that random variable X takes specific value x is denoted $P(X = x)$ or $P(x)$

The probability distribution function of a random variable is a representation of the probabilities for all the possible outcomes.

Example : Toss 2 coins and X is the number of heads. Show the probability distribution.

Remark : Let X be a discrete random variable with probability distribution function $P(x)$: Then,

$0 \leq P(x) \leq 1$ for any value of x and the individual probabilities sum to 1, ie $\sum_x P(x) = 1$. (The notation indicates summation over all possible x values).

Description : We can describe a discrete distribution graphically with an histogram and numerically with the mean, variance and standard deviation.

- For the histogram, the different outcomes are represented on the x-axis and probabilities on the y-axis.
- The mean or expected value of a discrete distribution is the long-run average of occurrences.
  $\mu = E(x) = \sum x.p(x)$ where $E(x)$ is the long-run average, x an outcome and $p(x)$ the probability of that outcome.
- The variance and standard deviation of a discrete distribution are solved for by using the outcomes (x) and probabilities $[P(x)]$ in a manner similar to that of computing a mean.
  $\sigma^2 = \sum[(x-\mu)^2].P(x)$ where x is an outcome, $P(x)$ is the probability of the given outcome x and μ the mean.
  σ is the square root of the variance $\sigma = \sqrt{\sum[(x-\mu)^2].P(x)}$

Example : page 155. Question. Compute the expected value, the variance and the standard deviation of the number of crisis.

Linear Functions of Random Variables :

Let random variable X have mean μ_X and variance σ_X^2. Let a and b be any constants and let $Y = a + bX$. then :

$\mu_X = E(a + bX) = a + b\mu_X$, $\sigma_Y^2 = b^2\sigma_X^2$ and $\sigma_Y = |b|\sigma_X$

4.2 Binomial distribution

The most widely known of all discrete distribution is the binomial distribution.

- The following assumptions underlie the use of binomial distribution.

1. The experiment involves n identical trials
2. Each trial has only two possible outcomes denoted success or as failure
3. Each trial is independent of the previous trials.
4. The terms p and q remain constant throughout the experiment, where p is the probability of getting a success on any one trial and the term $q = 1 - p$ is the probability of getting a failure on any one trial.

Usually the outcome of interest is labelled *success*. The word *failure* is used only in opposition to success.

In n trials, only x success is possible where x is a whole number between 0 and n.

For example, if 5 parts are randomly selected from a batch of parts, only $0, 1, 2, 3, 4$ or 5 defective parts are possible.

p remains constant throughout the experiment. However, if the population is large in comparison with sample size, the effect of sampling without replacement is minimal. Generally the sample size for using the binomial distribution with samples taken without replacement is $n < 5\%N$.

– **Solving a binomial problem** : Consider a binomial experience that consists of n independent trials (or the number being sampled), where the probability of getting a success in one trial is p and $1-p$ the probability of getting failure in one trial. Then the probability of getting x success is given by :

$P(x) = {}_nC_x p^x q^{n-x} = \frac{n!}{x!(n-x)!} p^x q^{n-x} \quad x = 0, 1, 2, ..., n.$

The expected value is $\mu = n.p$ and the standard deviation is $\sigma = \sqrt{n.p.q}$

Examining the mean of a binomial distribution gives an intuitive feeling about the likelihood of a given outcome.

Cumulative Probability Function : The cumulative probability function, denoted $F(x_0)$, shows the probability that X does not exceed the value x_0. $F(x_0) = P(X \leq x_0)$

– **Using Binomial table**
– **Examples**

Problem 5.5, Problem 5.6, problem 5.7
Problem 5.2 page 163
Go to the end of page 167.

4.3 Poisson distribution

The binomial distribution describes a distribution of two possible outcomes designated as success and failure from a given number of trials. The Poisson distribution focuses only on the number of discrete occurrences over some interval or continuum.

– **It has the following characteristics**

- It is a discrete distribution
- It describes the occurrence of rare events.
- Each occurrence is independent of the other occurrence
- It describes discrete occurrences over continuum or interval
- The occurrences in each interval can range from zero to infinity
- The expected number of occurrence must hold constant throughout the experiment

– **Apply the Poisson Distribution when** :

–You wish to count the number of times an event occurs in a given continuous interval
–The probability that an event occurs in one subinterval is very small and is the same for all subintervals
–The number of events that occur in one subinterval is independent of the number of events that occur in the other subintervals
–There can be no more than one occurrence in each subinterval
–The expected number of events per unit is λ (lambda)

– **Solving a Poisson problem** :

Suppose that we can expect some independent event to occur "λ" times over a specified time interval. The probability of exactly "x" occurrences is equal to $p(x) = \frac{\lambda^x e^{-\lambda}}{x!}$ where :
$x = 0, 1, 2, ...$, λ is the long-run average, and e is the base of the natural logarithm system (2.71828...)
The expected value or the mean of X is λ and the variance is λ.

– **Using Poisson table**
– **Examples :**

1) Above Problem 5.7 page 174

2) Problem 5.8 page 175

3) Suppose a fast-food restaurant can expect two customers every 3 minutes, on average. What is the probability that four or fewer customers will enter a restaurant in a 9 minutes period ?

4)

– **Approximating Binomial problems by the Poisson distribution.**

Let X be the number of successes from n independent trials, each with probability of success p. The distribution of the number of successes, X , is binomial, with mean np. If the number of trials, n , is large and np is of only moderate size (preferably $np \leq 7$ and $n > 20$), this distribution can be approximated by the Poisson distribution with $\lambda = np$. The probability distribution of the approximating distribution is therefore $p(x) = \frac{(np)^x e^{-np}}{x!}$ for $x = 0, 1, 2, ...$

Example : Problem 5.10 page 180.

4.4 Hypergeometric distribution

Recall the binomial distributions applies, in theory, only to experiments in which the trials are done with replacement (independent events). Hypergeometric distribution applies only to experiments in which the trials are done without replacement. It also consists of two possible outcomes : success and failure.

– The following assumptions underlie the use of hypergeometric distribution.

1. It is a dicrete distribution.
2. Each outcome consists of either a success or a failure
3. Sampling is done without replacement
4. The population N is finite and known
5. The number of success in the population A, is known

– **Solving a Hypergeometric problem**

The hypergeometric formula is given by : $P(x) = \frac{{}_AC_x \cdot {}_{N-A}C_{n-x}}{{}_NC_n}$ where :

N is the size of the population, n is the size of the sample, A the number of sucesses in the population x the number of sucesses in the sample, sampling is done without replacement.

– **Examples**

1) Page 185, 4eme ligne. Twenty-four ...

2) 3 different computers are checked from 10 in the department. 4 of the 10 computers have illegal software

loaded. What is the probability that 2 of the 3 selected computers have illegal software loaded ?

3) problem 5.11

4) Problem 5.32

– Recall : When the sample size is less than 5% of the population, the use of binomial distribution rather than the hypergeometric distribution is acceptable when sampling is done without replacement.