ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/296950685

Estimation and Inferential Statistics

Book · January 2015

DOI: 10.1007/978-81-322-2514-0

CITATION
1

READS
10,125

3 authors, including:

Pradip KUMAR Sahu
Bidhan Chandra Krishi Viswavidyalaya

98 PUBLICATIONS **172** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project Research Article View project

Project Adhoc Project View project

All content following this page was uploaded by Pradip KUMAR Sahu on 03 May 2016.

The user has requested enhancement of the downloaded file.

Pradip Kumar Sahu · Santi Ranjan Pal
Ajit Kumar Das

Estimation and Inferential Statistics

Springer

Contents

Introduction

In a statistical investigation, it is known that for reasons of time or cost, one may not be able to study each individual element of the population. In such a situation, a random sample should be taken from the population, and the inference can be drawn about the population on the basis of the sample. Hence, statistics deals with the collection of data and their analysis and interpretation. In this book, the problem of data collection is not considered. We shall take the data as given, and we study what they have to tell us. The main objective is to draw a conclusion about the unknown population characteristics on the basis of information on the same characteristics of a suitably selected sample. The observations are now postulated to be the values taken by random variables. Let X be a random variable which describes the population under investigation and F be the distribution function of X. There are two possibilities. Either X has a distribution function of F_θ with a known functional form (except perhaps for the parameter θ, which may be vector), or X has a distribution function F about which we know nothing (except perhaps that F is, say, absolutely continuous). In the former case, let Θ be the set of possible values of unknown parameter θ, then the job of statistician is to decide on the basis of suitably selected samples, which member or members of the family $\{F_\theta, \theta \in \Theta\}$ can represent the distribution function of X. These types of problems are called *problems of parametric statistical inference*. The two principal areas of statistical inference are the "area of estimation of parameters" and the "tests of statistical hypotheses". The problem of estimation of parameters involves both point and interval estimation. Diagrammatically, let us show components and constituents of statistical inference as in chart.

Chapter 1
Theory of Point Estimation

1.1 Introduction

In carrying out any statistical investigation, we start with a suitable probability model for the phenomenon that we seek to describe (The choice of the model is dictated partly by the nature of the phenomenon and partly by the way data on the phenomenon are collected. Mathematical simplicity is also a point that is given some consideration in choosing the model). In general, model takes the form of specification of the joint distribution function of some random variables $X_1, X_2, \ldots X_n$ (all or some of which may as well be multidimensional). According to the model, the distribution function F is supposed to be some (unspecified) member of a more or less general class $\mathbb{F}$ of distribution functions.

Example 1.1 In many situations, we start by assuming that $X_1, X_2, \ldots X_n$ are iid (independently and identically distributed) unidimensional r.v's (random variables) with a common but unspecified distribution function, F_1, say. In other words, the model states that F is some member of the class of all distribution functions of the form

$$F(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} F_1(x_i).$$

Example 1.2 In traditional statistical practice, it is frequently assumed that $X_1, X_2 \ldots X_n$ have each the normal distribution (but its mean and/or variance being left unspecified), besides making the assumption that they are iid r.v's.

In carrying out the statistical investigation, we then take as our goal, the task of specifying F more completely than is done by the model. This task is achieved by taking a set of observations on the r.v's $X_1, X_2, \ldots, X_n$. These observations are the raw material of the investigation and we may denote them, respectively, by $x_1, x_2, \ldots, x_n$. These are used to make a guess about the distribution function F, which is partly unknown.

© Springer India 2015

P.K. Sahu et al., *Estimation and Inferential Statistics*,
DOI 10.1007/978-81-322-2514-0_1

Chapter 2
Methods of Estimation

2.1 Introduction

In chapter one, we have discussed different optimum properties of good point estimators viz. unbiasedness, minimum variance, consistency and efficiency which are the desirable properties of a good estimator. In this chapter, we shall discuss different methods of estimating parameters which are expected to provide estimators having some of these important properties. Commonly used methods are:

1. Method of moments
2. Method of maximum likelihood
3. Method of minimum χ^2
4. Method of least squares

In general, depending on the situation and the purpose of our study we apply any one of the methods that may be suitable among the above-mentioned methods of point estimation.

2.2 Method of Moments

The method of moments, introduced by K. Pearson is one of the oldest methods of estimation. Let $(X_1, X_2, \ldots X_n)$ be a random sample from a population having p.d.f. (or p.m.f) $f(x,\theta)$, $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$. Further, let the first k population moments about zero exist as explicit function of θ, i.e. $\mu'_r = \mu'_r(\theta_1, \theta_2, \ldots, \theta_k)$, $r = 1, 2, \ldots, k$. In the method of moments, we equate k sample moments with the corresponding population moments. Generally, the first k moments are taken because the errors due to sampling increase with the order of the moment. Thus, we get k equations $\mu'_r(\theta_1, \theta_2, \ldots, \theta_k), = m'_r$, $r = 1, 2, \ldots, k$. Solving these equations we get the method of moment estimators (or estimates) as $m'_r = \frac{1}{n}\sum_{i=1}^{n} X_i^r$ (or $m'_r = \frac{1}{n}\sum_{i=1}^{n} x_i^r$).

© Springer India 2015

P.K. Sahu et al., *Estimation and Inferential Statistics*,
DOI 10.1007/978-81-322-2514-0_2

Chapter 3
Theory of Testing of Hypothesis

3.1 Introduction

Consider a random sample from an infinite or a finite population. From such a sample or samples we try to draw inference regarding population. Suppose the form of the distribution of the population is F_θ which is assumed to be known but the parameter θ is unknown. Inferences are drawn about unknown parameters of the distribution. In many practical problems, we are interested in testing the validity of an assertion about the unknown parameter θ. Some hypothesis is made regarding the parameters and it is tested whether it is acceptable in the light of sample observations. As for examples, suppose we are interested in introducing a high yielding rice variety. We have at our disposal a standard variety having average yield x quintal per acre. We want to know whether the average yield for the new variety is higher than x. Similarly, we may be interested to check the claim of a tube light manufacturer about the average life hours achieved by a particular brand. A problem of this type is usually referred to as a problem of testing of hypothesis. Testing of hypothesis is closely linked with estimation theory in which we seek the best estimator of unknown parameter. In this chapter, we shall discuss the problem of testing of hypothesis.

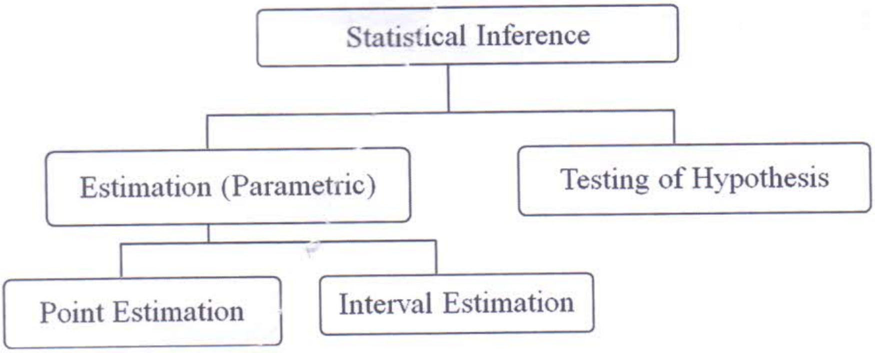
3.2 Definitions and Some Examples

In this section, some aspects of statistical hypotheses and tests of statistical hypothesis will be discussed.

Let $\rho = \{p(x)\}$ be a class of all p.m.f or p.d.f. In testing problem $p(x)$ is unknown, but ρ is known. Our objective is to provide more information about $p(x)$ on the basis of $X = x$. That is, to know whether $p(x) \in \rho^* \subset \rho$.

© Springer India 2015

P.K. Sahu et al., *Estimation and Inferential Statistics*,
DOI 10.1007/978-81-322-2514-0_3



Problem of Point Estimation

The problem of point estimation relates to the estimating formula of a parameter based on random sample of size n from the population. The method basically comprises of finding out an estimating formula of a parameter, which is called the *estimator* of the parameter. The numerical value, which is obtained on the basis of a sample while using the estimating formula, is called *estimate*. Suppose, for an example, that a random variable X is known to have a normal distribution $N(\mu, \sigma^2)$, but we do not know one of the parameters, say μ. Suppose further that a sample $X_1, X_2, \ldots, X_n$ is taken on X. The problem of point estimation is to pick a statistic $T(X_1, X_2, \ldots, X_n)$ that best estimates the parameter μ. The numerical value of T when the realization is $x_1, x_2, \ldots, x_n$ is called an *estimate* of μ, while the statistic T is called an *estimator* of μ. If both μ and σ^2 are unknown, we seek a joint statistic $T = (U, V)$ as an estimate of (μ, σ^2).

Example Let $X_1, X_2, \ldots, X_n$ be a random sample from any distribution F_θ for which the mean exists and is equal to θ. We may want to estimate the mean θ of distribution. For this purpose, we may compute the mean of the observations $x_1, x_2, \ldots, x_n$, i.e., say

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

This $\bar{x}$ can be taken as the point estimate of θ.

Example Let $X_1, X_2, \ldots, X_n$ be a random sample from Poisson's distribution with parameter λ, i.e., $P(\lambda)$, where λ is not known. Then the mean of the observations $x_1, x_2, \ldots, x_n$, i.e.,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

is a point estimate of λ.

Example Let $X_1, X_2, \ldots, X_n$ be a random sample from a normal distribution with parameters μ and σ^2, i.e., $N(\mu, \sigma^2)$, where both μ and σ^2 are unknown. μ and σ^2 are the mean and variance respectively of the normal distribution. In this case, we may take a joint statistics $(\bar{x}, s^2)$ as a point estimate of $N(\mu, \sigma^2)$, where

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \text{sample mean}$$

and

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_1 - \bar{x})^2 = \text{sample mean square.}$$

Problem of Interval Estimation

In many cases, instead of point estimation, we are interested in constructing of a family of sets that contain the true (unknown) parameter value with a specified (high) probability, say $100(1-\alpha)\%$. This set is taken to be an interval, which is known as *confidence interval* with a confidence coefficient $(1-\alpha)$ and the technique of constructing such intervals is known as *interval estimation*.

Let $X_1, X_2, \ldots, X_n$ be a random sample from any distribution F_θ. Let $\underline{\theta}(x)$ and $\bar{\theta}(x)$ be functions of $x_1, x_2, \ldots, x_n$. If $P[\underline{\theta}(x) < \theta < \bar{\theta}(x)] = 1 - \alpha$, then $(\underline{\theta}(x), \bar{\theta}(x))$ is called a $100(1-\alpha)\%$ confidence interval for θ, whereas $\underline{\theta}(x)$ and $\bar{\theta}(x)$ are, respectively, called lower and upper limits for θ.

Example Let $X_1, X_2, \ldots, X_n$ be random sample from $N(\mu, \sigma^2)$, whereas both μ and σ^2 are unknown. We can find $100(1-\alpha)\%$ confidence interval of μ. To estimate the population mean μ and population variance σ^2, we may take the observed sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and the observed sample mean square

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2$$

respectively. $100(1-\alpha)\%$ confidence interval of μ is given by

$$\bar{x} \pm t_{\frac{\alpha}{2},n-1} \frac{s}{\sqrt{n}}$$

where $t_{\frac{\alpha}{2},n-1}$ is the upper $\frac{\alpha}{2}$ point of the t-distribution with $(n-1)$ d.f.

Problem of Testing of Hypothesis

Besides point estimation and interval estimation, we are often required to decide which value among a set of values of a parameter is true for a given population distribution, or we may be interested in finding out the relevant distribution to describe a population. The procedure by which a decision is taken regarding the plausible value of a parameter or the nature of a distribution is known as the *testing of hypotheses*. Some examples of hypothesis, which can be subjected to statistical tests, are as follows:

1. The average length of life μ of electric light bulbs of a certain brand is equal to some specified value μ_0.
2. The average number of bacteria killed by tests drops of germicide is equal to some number.
3. Steel made by method A has a mean hardness greater than steel made by method B.
4. Penicillin is more effective than streptomycin in the treatment of disease X.
5. The growing period of one hybrid of corn is more variable than the growing period for other hybrids.
6. The manufacturer claims that the tires made by a new process have mean life greater than the life of a tire manufactured by an earlier process.
7. Several varieties of wheat are equally important in terms of yields.
8. Several brands of batteries have different lifetimes.
9. The characters in the population are uncorrelated.
10. The proportion of non-defective items produced by machine A is greater than that of machine B.

The examples given are simple in nature, and are well established and have well-accepted decision rules.

Problems of Non-parametric Estimation

So far we have assumed in statistical inference (parametric) that the distribution of the random variable being sampled is known except for some parameters. In practice, the functional form of the distribution is unknown. Here, we are not concerned to the techniques of estimating the parameters directly, but with certain pertinent hypothesis relating to the properties of the population, such as equalities of distribution, tests of randomness of the sample without making any assumption about the nature of the distribution function. Statistical inference under such a setup is called non-parametric.

Bayes Estimator

In case of parametric inference, we consider density function $f(x/\theta)$, where θ is a fixed unknown quantity which can take any value in parametric space Θ. In Bayesian approach, it is assumed that θ itself is a random variable and density $f(x/\theta)$ is the density of x for a given θ. For example, suppose we are interested in estimating P, the fraction of defective items in a consignment. Consider a collection of lots, called superlots. It may happen that the parameter P may differ from lot to lot. In the classical approach, we consider P as a fixed unknown parameter, whereas in Bayesian approach, we say that P varies from lot to lot. It is random variable having a density $f(P)$, say. Bayes method tries to use this additional information about P.

Example Let $X_1, X_2, \ldots X_n$ be a random sample from PDF

$$f(x;a,b) = \frac{1}{\beta(a,b)} x^{a-1}(1-x)^{b-1},\ 0<x<1; a,b>0.$$

Find the estimators of a and b by the method of moments.

Answer
We know

$$E(x) = \mu_1^1 = \frac{a}{a+b} \quad \text{and} \quad E(x^2) = \mu_2^1 = \frac{a(a+1)}{(a+b)(a+b+1)}$$

Hence

$$\frac{a}{a+b} = \bar{x},\ \frac{a(a+1)}{(a+b)(a+b+1)} = \frac{1}{n}\sum_{i=1}^{n} x_i^2$$

Solving, we get

$$\widehat{b} = \frac{(\bar{x}-1)\left(\sum x_i^2 - n\bar{x}\right)}{\left(\sum x_i - \bar{x}\right)^2} \quad \text{and} \quad \widehat{a} = \frac{\bar{x}\widehat{b}}{1-\bar{x}}$$

Example Let $X_1, X_2, \ldots X_n$ be a random sample from PDF

$$f(x;\theta,r) = \frac{1}{\theta^r\sqrt{r}} e^{-x/\theta} x^{r-|1},\ x>0; \theta>0, r>0$$

Chapter 4
Likelihood Ratio Test

4.1 Introduction

In the previous chapter we have seen that UMP or UMP-unbiased tests exist only for some special families of distributions, while they do not exist for other families. Further, computations of UMP-unbiased tests in K-parameter family of distribution are usually complex. Neyman and Pearson (1928) suggested a simple method for testing a general testing problem.

Consider $X \sim p(x|\theta)$, where θ is a real parameter or a vector of parameters, $\theta \in \Theta$.

A general testing problem is

$$H : \theta \in \Theta_0 \,\text{Against}\, K : \theta \in \Theta_1.$$

Here, H and K may be treated as the subsets of Θ. These are such that $H \cap K = \phi$ and $H \cup K \subseteq \Theta$. Given that $X = x, p(\mathrm{x}|\theta)$ is a function of θ and is called likelihood function. Likelihood test for H against K is provided by the statistic

$$L(x) = \frac{\operatorname*{Sup}_{\theta \in H} p(x|\theta)}{\operatorname*{Sup}_{\theta \in H \cup K} p(x|\theta)},$$

which is called the likelihood ratio criterion for testing H against K. It is known that

(i) $p(x|\theta) \geq 0 \forall \theta$
(ii) $\operatorname*{Sup}_{\theta \in H} p(x|\theta) \leq \operatorname*{Sup}_{\theta \in H \cup K} p(x|\theta)$.

Obviously $0 \leq L(x) \leq 1$. The numerator in $L(x)$ measures the best explanation that the observation X comes from some population under H and the denominator

© Springer India 2015

P.K. Sahu et al., *Estimation and Inferential Statistics*,
DOI 10.1007/978-81-322-2514-0_4

Chapter 5
Interval Estimation

5.1 Introduction

Inpoint estimation when a random sample $(X_1, X_2, \ldots, X_n)$ is drawn from a population having distribution function F_θ and θ is the unknown parameter (or the set of unknown parameter). We try to estimate the parametric function $\gamma(\theta)$ by means of a single value, say t, the value of a statistic T corresponding to the observed values $(x_1, x_2, \ldots, x_n)$ of the random variables $(X_1, X_2, \ldots, X_n)$. This estimate may differ from the exact value of $\gamma(\theta)$ in the given population. In other words, we take t as an estimate of $\gamma(\theta)$ such that $|t - \gamma(\theta)|$ is small with high probability. In the point estimate we try to choose a unique point in the parameter space which can reasonably be considered as the true value of the parameter. Instead of unique estimate of the parameter we are interested in constructing a family of sets that contain the true (unknown) parameter value with a specified (high) probability. In many problems of statistical inference we are not interested only in estimating the parameter or testing some hypothesis concerning the parameter, we also want to get a lower or an upper bound or both, for the real-valued parameter. Here two limits are computed from the set of observations, say t_1 and t_2 and it is claimed with a certain degree of confidence (measured in probabilistic terms) that the true value of $\gamma(\theta)$ lies between t_1 and t_2. Thus we get an interval (t_1, t_2) which we expect would include the true value of $\gamma(\theta)$. So this type of estimation is called intervalestimation. In this chapter we discuss the problem of interval estimation.

5.2 Confidence Interval

An interval depending on a random variable X is called a random interval. For example, (X, $2X$) is a random interval. Note that, $\frac{1}{2} \le X \le 1 \Leftrightarrow X \le 1 \le 2X$.

© Springer India 2015
P.K. Sahu et al., *Estimation and Inferential Statistics*,
DOI 10.1007/978-81-322-2514-0_5

Chapter 6
Non-parametric Test

6.1 Introduction

In parametric tests we generally assume a particular form of the population distribution (say, normal distribution) from which a random sample is drawn and we try to construct a test criterion (for testing hypothesis regarding parameter of the population) and the distribution of the test criterion depends upon the parent population.

In non-parametric tests the form of the parent population is unknown. We only assume that the population, from which a random sample is drawn, is continuous and try to develop a test criterion whose distribution is independent of the population distribution under the hypothesis under consideration. A non-parametric test is concerned with the form of the population but not with any parametric value.

A test procedure is said to be distribution free if the statistic used has a distribution which does not depend upon the form of the distribution of the parent population from which the sample is drawn. So in such procedure assumptions regarding the population are not necessary.

Note Sometimes the term 'distribution free' is used instead of non-parametric. But we should make some distinction between them.

In fact, the terms 'distribution free' and 'non-parametric' are not synonymous. The term 'distribution free' is used to indicate the nature of the distribution of the test statistic whereas the term 'non-parametric' is used to indicate the type of hypothesis problem investigated.

Advantages and disadvantages of non-parametric method over parametric method

Advantages

(i) Non-parametric methods are readily comprehensible, very simple and easy to apply and do not require complicated sample theory.

© Springer India 2015

P.K. Sahu et al., *Estimation and Inferential Statistics*,
DOI 10.1007/978-81-322-2514-0_6

Chapter 7
Statistical Decision Theory

7.1 Introduction

In this chapter we discuss the problems of point estimation, hypothesis testing and interval estimation of a parameter from a different standpoint.

Before we start the discussion, let us first define certain terms commonly used in statistical inerence problem and decision theory. Let $X_1, X_2, \ldots, X_n$ denote a random sample of size n from a distribution that has the p.d.f. $f(x, \theta)$, where θ is an unknown state of nature or an unknown parameter and Θ is the set of all possible values of θ, i.e. parameter space (known).

To make some inference about θ, i.e. to take some decisions or action about θ, the statistician takes an action on the basis of the sample point $(x_1, x_2, \ldots, x_n)$.

Let us define

$$\begin{aligned} Œ &= \text{the set of all possible actions for statistician (action space)} \\ &\equiv \text{to choose an action a from Œ.} \end{aligned}$$

So, θ = true state of nature and a = action taken by the statistician.

The value $L(\theta, a)$ is the loss incurred by taking action 'a' when θ is true. Equivalently, it is a measure of the degree of undesirability of choosing an action 'a' when θ is true and this gives a preference pattern over œ for given θ, i.e. the smaller the loss the better the action under θ. $L(\theta, a)$ is a real-valued function on Θ x œ = Loss function. Thus (Θ, œ, L) is the basic element in our discussion.

Example 7.1 Let θ = average life length of electric bulbs produced in a factory and $\Theta = (0, \infty)$.

Point estimation of θ

To estimate the value of $\theta \equiv$ to choose one value from $(0, \infty)$; so $a = (0, \infty)$.

Observe life lengths of some randomly selected bulbs.

Define $L(\theta, a) = (\theta - a)^2$ = squared error loss function

© Springer India 2015

P.K. Sahu et al., *Estimation and Inferential Statistics*,
DOI 10.1007/978-81-322-2514-0_7

Appendix

A.1 Exact Tests Related to Binomial Distribution

A.1.1 We have an infinite population for which π = unknown proportion of individuals having certain character, say A. We are to test $H_0 : \pi = \pi_0$.

For doing this we draw a sample of size n. Suppose x = no. of individuals in the sample have character A. The sufficient statistic x is used for testing $H_0 : \pi = \pi_0$. Suppose x_0 is the observed value of x. Then $x \sim \text{bin}(n, \pi)$.

(a) $H_1 : \pi > \pi_0; \omega_0 : P[x \geq x_0/H_0] \leq \alpha$ i.e., $\sum\limits_{x \geq x_0} \binom{n}{x} \pi_0^x (1 - \pi_0)^{n-x} \leq \alpha$

(b) $H_2 : \pi < \pi_0; \omega_0 : P[x \leq x_0/H_0] \leq \alpha$ i.e., $\sum\limits_{x \leq x_0} \binom{n}{x} \pi_0^x (1 - \pi_0)^{n-x} \leq \alpha$

(c) $H_3 : \pi \neq \pi_0$; where $\pi_0 = \frac{1}{2}$ may be of our interest.

$$\begin{aligned}
&\omega_0 : P\left[\left|x - \frac{n}{2}\right| \geq d_0/H_0\right] \leq \alpha \\
&\text{i.e., } P\left[x \geq \frac{n}{2} + d_0/H_0\right] + P\left[x \leq \frac{n}{2} - d_0/H_0\right] \leq \alpha \\
&\text{i.e., } \sum_{x \geq \frac{n}{2} + d_0} \binom{n}{x} \left(\frac{1}{2}\right)^n + \sum_{x \leq \frac{n}{2} - d_0} \binom{n}{x} \left(\frac{1}{2}\right)^n \leq \alpha \text{ where } d_0 = \left|x_0 - \frac{n}{2}\right|
\end{aligned}$$

Note

(1) For other values of π_0 the exact test cannot be obtained as binomial distribution is symmetric only when $\pi = \frac{1}{2}$.

(2) For some selected n and π the binomial probability sums considered above are given in Table 37 of Biometrika (Vol. 1)

© Springer India 2015
P.K. Sahu et al., *Estimation and Inferential Statistics*,
DOI 10.1007/978-81-322-2514-0

A.1.2 Suppose we have two infinite populations with π_1 and π_2 as the unknown proportion of individuals having character A. We are to test $H_0 : \pi_1 = \pi_2$.

To do this we draw two samples from two populations having sizes n_1 and n_2. Suppose x_1 and x_2 as the random variables denoting the no. of individuals in the 1st and 2nd samples with character A.

To test $H_0 : \pi_1 = \pi_2$ we make use of the statistics x_1 and x_2 such that $x_1 + x_2 = x$ (constant), say.

Under $H_0 : \pi_1 = \pi_2 = \pi$ (say),

$$\begin{aligned} f(x_1) &= \text{p.m.f. of } x_1 = \binom{n_1}{x_1} \pi^{x_1} (1-\pi)^{n_1 - x_1} \\ f(x_2) &= \text{p.m.f. of } x_2 = \binom{n_2}{x_2} \pi^{x_2} (1-\pi)^{n_2 - x_2} \\ f(x) &= \text{p.m.f. of } x = \binom{n_1+n_2}{x} \pi^{x} (1-\pi)^{n_1+n_2-x}. \end{aligned}$$

The conditional distribution of x_1 given x has p.m.f.

$$f(x_1/x) = \frac{\binom{n_1}{x_1}\binom{n_2}{x_2}}{\binom{n_1+n_2}{x}}, \text{ which is hypergeometric and independent of } \pi.$$

Suppose the observed values of x_1 and x are x_{10} and x_0 respectively.

(a) $H_1 : \pi_1 > \pi_2,\ \omega_0 : P[x_1 \geq x_{10}/x = x_0] \leq \alpha$

$$\text{i.e., } \sum_{x_1 \geq x_{10}} \frac{\binom{n_1}{x_1}\binom{n_2}{x_0 - x_1}}{\binom{n_1+n_2}{x_0}} \leq \alpha$$

(b) $H_2 : \pi_1 < \pi_2,\ \omega_0 : P[x_1 \leq x_{10}/x = x_0] \leq \alpha$

$$\text{i.e., } \sum_{x_1 \leq x_{10}} \frac{\binom{n_1}{x_1}\binom{n_2}{x_0 - x_1}}{\binom{n_1+n_2}{x_0}} \leq \alpha$$

(c) $H_3 : \pi_1 \neq \pi_2$, exact test is not available.

Note The above probabilities can be obtained from the tables of hypergeometric distributions (Standard University Press).

A.2 Exact Tests Related to Poisson Distribution

A.2.1 Suppose we have a Poisson population with unknown parameter λ. We draw a random sample $(x_1, x_2, \ldots, x_n)$ of size n from this population. Here, we are to test $H_0 : \lambda = \lambda_0$.

To develop a test we make use of the sufficient statistic $y = \sum_{i=1}^{n} x_i$, which is itself distributed as Poisson with parameter $n\lambda$. The p.m.f. of y under H_0 is therefore $f(y) = e^{-n\lambda_0} \frac{(n\lambda_0)^y}{y!}, y = 0, 1, 2\ldots$

Suppose y_0 is the observed value of y.

(a) $H_1 : \lambda > \lambda_0,\ \omega_0 : P[y \geq y_0/\lambda = \lambda_0] \leq \alpha$

$$\text{i.e., } \sum_{y \geq y_0} e^{-n\lambda_0} \frac{(n\lambda_0)^y}{y!} \leq \alpha.$$

(b) $H_2 : \lambda < \lambda_0,\ \omega_0 : P[y \leq y_0/\lambda = \lambda_0] \leq \alpha$

$$\text{i.e., } \sum_{y \leq y_0} e^{-n\lambda_0} \frac{(n\lambda_0)^y}{y!} \leq \alpha.$$

(c) $H_3 : \lambda \neq \lambda_0$: exact test is not available.

Note These probabilities may be obtained from Table 7 of Biometrika (Vol. 1)

A.2.2 Suppose we have two populations $P(\lambda_1)$ and $P(\lambda_2)$. We draw a random sample $(x_{11}, x_{12}, \ldots, x_{1n_1})$ of size n_1 from $P(\lambda_1)$ and another random sample $(x_{21}, x_{22}, \ldots, x_{2n_2})$ of size n_2 from $P(\lambda_2)$. We are to test $H_0 : \lambda_1 = \lambda_2 = \lambda$ (say). Here we note that $y_1 = \sum_{i=1}^{n_1} x_{1i} \sim P(n_1\lambda_1)$ and $y_2 = \sum_{i=1}^{n_2} x_{2i} \sim P(n_2\lambda_2)$.

To develop a test we shall make use of the sufficient statisticsy_1 and y_2 but shall concentrate only on those for which $y = y_1 + y_2 =$ constant. Under H_0 the p.m.f. of y_1, y_2 and y are

$$f(y_1) = e^{-n_1\lambda} \frac{(n_1\lambda)^{y_1}}{y_1!}; f(y_2) = e^{-n_2\lambda} \frac{(n_2\lambda)^{y_2}}{y_2!} \text{ and } f(y) = e^{-(n_1+n_2)\lambda} \frac{\{(n_1+n_2)\lambda\}^y}{y!}$$

The conditional distribution of y_1 given y has the p.m.f. as

$$
\begin{aligned}
f(y_1/y) &= \frac{e^{-n_2\lambda}\frac{(n_2\lambda)^{y-y_1}}{(y-y_1)!}\cdot e^{-n_1\lambda}\frac{(n_1\lambda)^{y_1}}{y_1!}}{e^{-(n_1+n_2)\lambda}\,\frac{\{(n_1+n_2)\lambda\}^{y}}{y!}} \\
&= \frac{y!}{y_1!y_2!}\frac{n_1^{y_1}n_2^{y_2}}{(n_1+n_2)^{y}} \\
&= \binom{y}{y_1}\left(\frac{n_1}{n_1+n_2}\right)^{y_1}\left(1-\frac{n_1}{n_1+n_2}\right)^{y_2} \sim \text{bin}\left(y,\frac{n_1}{n_1+n_2}\right) \text{ free of } \lambda.
\end{aligned}
$$

So this may be regarded as sufficient statistic. Suppose the observed values of y_1 and y are y_{10} and y_0 respectively. We consider the conditional p.m.f. $f(y_1/y_0)$ for testing H_0.

(a) $H_1 : \lambda_1 > \lambda_2; \omega_0 : P[y_1 \geq y_{10}/y = y_0] \leq \alpha$

$$
\text{i.e.,} \quad \sum_{y_1 \geq y_{10}} \binom{y_0}{y_1}\left(\frac{n_1}{n_1+n_2}\right)^{y_1}\left(\frac{n_2}{n_1+n_2}\right)^{y_0-y_1} \leq \alpha
$$

(b) $H_2 : \lambda_1 < \lambda_2; \;\; \omega_0 : P[y_1 \leq y_{10}/y = y_0] \leq \alpha$

$$
\text{i.e.,} \quad \sum_{y_1 \leq y_{10}} \binom{y_0}{y_1}\left(\frac{n_1}{n_1+n_2}\right)^{y_1}\left(\frac{n_2}{n_1+n_2}\right)^{y_0-y_1} \leq \alpha.
$$

(c) $H_3 : \lambda \neq \lambda_0$: exact test is not available.

A.3 A Test for Independence of Two Attributes

In many investigations one is faced with the problem of judging whether two qualitative characters, say A and B, may be said to be independent. Let us denote the forms of A by $A_i\{i = 1(1)k\}$ and the forms of B by $B_j\{j = 1(1)l\}$, and the probability associated with the cell A_iB_j in the two-way classification of the population

View publication stats