STAT 5265 - Introductory Theory of Statistics Notes

Libao Jin

May 7, 2020

Contents

1 Overview

1.1 Independence

Definition 1.1 (Independence)**.** If X and Y are independent, then we have

$$f(x, y) = f(x) \cdot f(y).$$

Definition 1.2 (Random Samples)**.** If x_i are i.i.d., then

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

Definition 1.3 (Gamma Function)**.**

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y}\, dy, \text{for } \alpha > 0.$$

Definition 1.4 (Gamma Distribution)**.** A random variable X has a *gamma distribution* and it is referred to as a gamma random variable if and only if its probability density is given by

$$g(x; \alpha, \beta) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & \text{for } x > 0, \\ 0 & \text{otherwise,} \end{cases}$$

where $\alpha > 0$ and $\beta > 0$.

When $\alpha = 1$ and $\beta = \theta$, we have the exponential distribution.

Definition 1.5 (Exponential Distribution)**.** A random variable X has a *exponential distribution* and it is referred to as an exponential random variable if and only if its probability density is given by

$$g(x; \theta) = \begin{cases} \frac{1}{\theta} e^{-x/\theta} & \text{for } x > 0, \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta > 0$.

When $\alpha = \nu/2$ and $\beta = 2$, we have the chi-square distribution.

Definition 1.6 (Chi-Square Distribution)**.** A random variable X has a *chi-square distribution* and it is referred to as a chi-square random variable if and only if its probability density is given by

$$f(x, v) = \begin{cases} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{(\nu-2)/2} e^{-x/2} & \text{for } x > 0, \\ 0 & \text{otherwise.} \end{cases}$$

The parameter ν is referred to as the *number of degrees of freedoms*, or simply the *degrees of freedom*.

Definition 1.7 (Normal Distribution)**.** A random variable X has a *normal distribution* and it is referred to as a normal random variable if and only if its probability density is given by

$$n(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)} \text{ for } -\infty < x < \infty$$

where $\sigma > 0$.

1.2 Moment-Generating Functions

Definition 1.8 (Moment Generating Function)**.** The *moment generating function* of a random variable X, where it exists, is given by

$$M_X(t) = E(e^{tX}) = \sum_x e^{tX} \cdot f(x)$$

when X is discrete, and

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} \cdot f(x)\, dx$$

where X is continuous.

1.3 Product Moments

Definition 1.9 (Product Moments About the Origin)**.** The rth and sth product moment about the origin of the random variables X and Y, denoted by $\mu'_{r,s}$ is the expected value of X^rY^s; symbolically,

$$\mu'_{r,s} = E(X^rY^s) = \sum_x \sum_y x^r y^s \cdot f(x,y)$$

for $r = 0, 1, 2, \cdots$ and $s = 0, 1, 2, \ldots$ when X and Y are discrete, and

$$\mu'_{r,s} = E(X^rY^s) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^r y^s \cdot f(x,y)\, dxdy$$

when X and Y are continuous.

Definition 1.10 (Product Moments About the Mean)**.** The rth and sth product moment about the means of the random variables X and Y, denoted by $\mu_{r,s}$ is the expected value of $(X-\mu_X)^r(Y-\mu_Y)^s$; symbolically,

$$\mu_{r,s} = E[(X-\mu_X)^r(Y-\mu_Y)^s] = \sum_x \sum_y (x-\mu_X)^r(y-\mu_Y)^s \cdot f(x,y)$$

for $r = 0, 1, 2, \cdots$ and $s = 0, 1, 2, \ldots$ when X and Y are discrete, and

$$\mu'_{r,s} = E[(X-\mu_X)^r(Y-\mu_Y)^s] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x-\mu_X)^r(y-\mu_Y)^s \cdot f(x,y)\, dxdy$$

when X and Y are continuous.

1.4 Transformation Technique

Theorem 1.1. Let $f(x)$ be the value of the probability density of the continuous random variable X at x. If the function given by $y = u(x)$ is differentiable and either increasing or decreasing for all values within range of X fro which $f(x) \neq 0$, then, for these values of x, the equation $y = u(x)$ can be uniquely solved for x to give $x = w(y)$, and for the corresponding values of y the probability density of $y = u(X)$ is given by

$$g(y) = f[w(y)] \cdot |w'(y)| \quad \text{provided } u'(x) \neq 0$$

Elsewhere, $g(y) = 0$.

Theorem 1.2. Let $f(x_1, x_2)$ be the value of the joint probability density of the continuous random variables X_1 and X_2 at (x_1, x_2). If the functions given by $y_1 = u_1(x_1, x_2)$ and $y_2 = u_2(x_1, x_2)$ are partially differentiable with respect to both x_1 and x_2 and represent a one-to-one transformation for all values within the range of X_1 and X_2 for which $f(x_1, x_2) \neq 0$, then, for these values of x_1 and x_2, the equation $y_1 = u_1(x_1, x_2)$ and $y_2 = u_2(x_1, x_2)$ can be uniquely solvers for x_1 and x_2 to give $x_1 = w_1(y_1, y_2)$ and $x_2 = w_2(y_1, y_2)$, and for the corresponding values of y_1 and y_2, the join probability density of $Y_1 = u_1(X_1, X_2)$ and $Y_2 = u_2(X_1, X_2)$ is given by

$$g(y_1, y_2) = f[w_1(y_1, y_2), w_2(y_1, y_2)] \cdot |J|.$$

Here, J, called the *Jacobian* of the transformation, is the determinant

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} \\ \frac{\partial x_2}{\partial y_1} & \frac{\partial x_2}{\partial y_2} \end{vmatrix},$$

Elsewhere, $g(y_1, y_2) = 0$.

2 Sampling Distributions

2.1 Introduction

Definition 2.1 (Statistics)**.** The distributions of certain functions of the random variables whose values make up the sample, called *statistics*, e.g., sample mean.

Definition 2.2 (Population)**.** A set of numbers from which a sample is drawn is referred to as a *population*. The distribution of the numbers constituting a population is called the *population distribution*.

Definition 2.3 (Random Sample)**.** If $X_1, X_2, \ldots, X_n$ are independent and identically distributed random variables, we say that they constitute a *random sample* from the infinite population given by their common distribution.

If $f(x_1, x_2, \ldots, x_n)$ is the value of the joint distribution of such a set of random variables at $(x_1, x_2, \ldots, x_n)$, by virtue of independence we can write

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

where $f(x_i)$ is the value of the population distribution at x_i. Statistical inferences are usually based on *statistics*, that is, on random variables that are functions of a set of random variables $X_1, X_2, \ldots, X_n$, constituting a random sample.

Definition 2.4 (Sample Mean and Sample Variance)**.** If $X_1, X_2, \ldots, X_n$ constitute a random sample, then the *sample mean* is given by

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

and the *sample variance*

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2.$$

It is common practice also to apply the terms "random sample", "statistic", "sample mean", and "sample variance" to the values of the random variables instead of the random variables themselves. Intuitively, this makes more sense and it conforms with colloquial usage. Thus, we might calculate

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

for observed sample data and refer to these statistics as the sample mean and the sample variance. Here, the x_i, $\overline{x}$, and s^2 are values of the corresponding random variables X_i, $\overline{X}$, and S^2.

2.2 The Sampling Distribution of the Mean

Definition 2.5 (Sampling Distribution)**.** The distribution of sampling statistics is called *sampling distribution*.

Theorem 2.1. If $X_1, X_2, \ldots, X_n$ constitute a random sample from an infinite population with the mean μ and the variance σ^2, then

$$E(\overline{X}) = \mu, \quad \text{var}(\overline{X}) = \frac{\sigma^2}{n}.$$

It is customary to write $E(\overline{X})$ as $\mu_{\overline{X}}$ and $\text{var}(\overline{X})$ as $\sigma^2_{\overline{X}}$ and refer to $\sigma_{\overline{X}} = \sigma/n$ as the *standard error of the mean*.

Theorem 2.2. For any positive constant c, the probability that $\overline{X}$ will take on a value between $\mu - c$ and $\mu + c$ is at least

$$1 - \frac{\sigma^2}{nc^2}.$$

When $n \to \infty$, this probability approaches 1.

Theorem 2.3 (Central Limit Theorem)**.** If $X_1, X_2, \ldots, X_n$ constitute a random sample from an infinite population with the mean μ, the variance σ^2, and the moment-generating function $M_X(t)$, then the limiting distribution of

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

as $n \to \infty$ is the standard normal distribution.

Theorem 2.4. If $\overline{X}$ is the mean of a random sample of size n from a normal population with the mean μ and the variance σ^2, its sampling distribution is a normal distribution with the mean μ and the variance σ^2/n.

2.3 The Sampling Distribution of the Mean: Finite Populations

If an experiment consists of selecting one or more values from a finite set of numbers $\{c_1, c_2, \ldots, c_N\}$, this set is referred to as a *finite population of size* N.

Definition 2.6 (Random Sample – Finite Population)**.** If X_1 is the first value drawn from a finite population of size N, X_2 is the second value drawn, ..., X_n is the nth value drawn, and the joint probability distribution of these n random variables is given by

$$f(x_1, x_2, \ldots, x_n) = \frac{1}{N(N-1)\cdots(N-n+1)}$$

for each ordered n-tuple of values of these random variables, then $X_1, X_2, \ldots, X_n$ are said to constitute a *random sample* from the given finite population.

Definition 2.7 (Sample Mean and Variance - Finite Population)**.** The *sample mean* and the *sample variance* of the finite population $\{c_1, c_2, \ldots, c_N\}$ are

$$\mu = \sum_{i=1}^{N} c_i \cdot \frac{1}{N}, \quad \sigma^2 = \sum_{i=1}^{N} (c_i - \mu)^2 \cdot \frac{1}{N}.$$

Theorem 2.5. If X_r and X_s are the rth and sth random variables of a random sample of size n drawn from the finite population $\{c_1, c_2, \ldots, c_N\}$, then

$$\text{cov}(X_r, X_s) = -\frac{\sigma^2}{N-1}.$$

Theorem 2.6. If $\overline{X}$ is the mean of a random sample of size n taken without replacement from a finite population of size N with the mean μ and the variance σ^2, then

$$E(\overline{X}) = \mu, \quad \text{var}(\overline{X}) = \frac{\sigma^2}{n} \cdot \frac{N-n}{N-1}.$$

2.4 The Chi-Square Distribution

Definition 2.8 (Chi-Square Distribution)**.** If a random variable X has the chi-square distribution (denoted by χ^2 distribution) with ν degrees of freedom if its probability density is given by

$$f(x) = \begin{cases} \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2} & \text{for } x > 0, \\ 0 & \text{elsewhere.} \end{cases}$$

Theorem 2.7. If X has the standard normal distribution, then X^2 has the chi-square distribution with $\nu = 1$ degree of freedom.

Theorem 2.8. If $X_1, X_2, \ldots, X_n$ are independent random variables having standard normal distribution, then

$$Y = \sum_{i=1}^{n} X_i^2$$

has the chi-square distribution with $\nu = n$ degrees of freedom.

Theorem 2.9. If $X_1, X_2, \ldots, X_n$ are independent random variables having chi-square distribution with $\nu_1, \nu_2, \ldots, \nu_n$ degrees of freedom, then

$$Y = \sum_{i=1}^{n} X_i$$

has the chi-square distribution with $\nu_1 + \nu_2 + \cdots + \nu_n$ degrees of freedom.

Theorem 2.10. If X_1 and X_2 are independent random variables, X_1 has a chi-square distribution with ν_1 degrees of freedom, and $X_1 + X_2$ has a chi-square distribution with $\nu > \nu_1$ degrees of freedom, then X_2 has a chi-square distribution with $\nu - \nu_1$ degrees of freedom.

Theorem 2.11. If $\overline{X}$ and S^2 are the mean and the variance of a random sample of size n from a normal population with the mean μ and the standard deviation σ, then

(a) $\overline{X}$ and S^2 are independent;
(b) the random variable $\frac{(n-1)S^2}{\sigma^2}$ has a chi-square distribution with $n-1$ degrees of freedom.

2.5 The t Distribution

Theorem 2.12. If Y and Z are independent random variables, Y has a chi-square distribution with ν degrees of freedom, and Z has the standard normal distribution, then the distribution of

$$T = \frac{Z}{\sqrt{Y/\nu}}$$

is given by

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\Gamma(\nu/2)} \cdot \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

for $-\infty < t < \infty$ and it is called the t distribution with ν degrees of freedom.

Theorem 2.13. If $\overline{X}$ and S^2 are the mean and the variance of a random sample of size n from a normal population with the mean μ and the variance σ^2, then

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has the t distribution with $n - 1$ degrees of freedom.

2.6 The F Distribution

Theorem 2.14. If U and V are independent random variables having chi-square distributions with ν_1 and ν_2 degrees of freedom, then

$$F = \frac{U/\nu_1}{V/\nu_2}$$

is a random variable having an F distribution, that is, a random variable whose probability density is given by

$$g(f) = \frac{\Gamma\frac{\nu_1+\nu_2}{2}}{\Gamma(\nu_1/2)\Gamma(\nu_2/2)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} \cdot f^{\nu_1/2-1} \left(1 + \frac{\nu_1}{\nu_2} f\right)^{-(\nu_1+\nu_2)/2}$$

for $f > 0$ and $g(f) = 0$ elsewhere.

Theorem 2.15. If S_1^2 and S_2^2 are the variance of independent random samples of sizes n_1 and n_2 from normal populations with the variances σ_1^2 and σ_2^2, then

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}$$

is a random variable having an F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom.

2.7 Order Statistics

In an effort to deal with the problem of small samples in cases where it may be unreasonable to assume a normal population, statisticians have developed *nonparametric statistics*, whose sampling distributions do not depend upon any assumptions about the population from which the sample is taken. Statistical inferences based upon such statistics are called *nonparametric inference*.

Definition 2.9 (Order Statistics)**.** Consider a random sample of size n from an infinite population with a continuous density, and suppose that we arrange the values of $X_1, X_2, \ldots, X_n$ according to size. If we look upon the smallest of the x's as a value of the random variable Y_1, the next largest as a value of the random variable Y_2, the next largest after that as a value of the random variable Y_3, ..., and the largest as a value of the random variable Y_n, we refer to these Y's as *order statistics*.

Theorem 2.16. For random samples of size n from an infinite population that has the value $f(x)$ at x, the probability density of the rth order statistic Y_r is given by

$$\begin{aligned} g_r(y_r) &= \frac{n!}{(r-1)!\,(n-r)!}\left[\int_{-\infty}^{y_r} f(x)\,dx\right]^{r-1} f(y_r)\left[\int_{y_r}^{\infty} f(x)\,dx\right]^{n-r} \\ &= \frac{n!}{(r-1)!\,(n-r)!}\left[F(y_r)\right]^{r-1} f(y_r)\left[1-F(y_r)\right]^{n-r} \end{aligned}$$

for $-\infty < y_r < \infty$.

Theorem 2.17. For large n, the sampling distribution of the median for random samples of size $2n+1$ is approximately normal with the mean $\tilde{\mu}$ and the variance $1/\{8[f(\tilde{\mu})]^2 n\}$.

3 Decision Theory

3.1 The theory of Games

A game between manufacturer and nature, each of the "players" has the choice of two moves: the manufacturer has the choice between actions a_1 and a_2 (to expand his plant capacity now or to delay expansion for at least a year), and nature controls the choice between θ_1 and θ_2 (whether economic conditions are to remain good or whether there is to be a recession). Depending on the choice of their moves, there are the "payoffs" shown in the following table: The amounts $L(a_i, \theta_j), i, j = 0, 1$

Table 1: Payoff Matrix

	Player A a_1	Player A a_2
Player B θ_1	$L(a_1, \theta_1)$	$L_(a_2, \theta_1)$
Player B θ_2	$L(a_1, \theta_2)$	$L_(a_2, \theta_2)$

are referred to as the values of the *loss function* that characterizes the particular "game"; in other words, $l(a_i, \theta_j)$ is the loss of Player A (the amount he has to pay Player B) when he chooses alternative a_i and Player B chooses alternative θ_j. Although it does not really matter, we shall assume here that these amounts are in dollars.

Definition 3.1 (Zero-Sum Two-Person Game)**.** "Two person" means that there are two players (or, more generally, two parties with conflicting interests), and "zero-sum" means that whatever one player loses the other player wins. Thus, in a zero-sum game there is no "cut for the house" as in professional gambling, and no capital is created or destroyed during the course of play.

Games are classified according to the number of *strategies* (moves, choices, or alternatives) that each player has at his disposal. The *payoffs*, the amounts of money or other considerations that change hands when the players choose their respective strategies, are usually shown in Table 1. It is always assumed in the theory of games that each player must choose a strategy without knowing what the opponent is going to do and that once a player has made a choice it cannot be changed.

Definition 3.2 (Payoff Matrix)**.** A *payoff* in game theory is the amount of money (or other numerical consideration) that changes hands when the players choose their respective strategies. Positive payoffs represent losses of Player A and negative payoffs represent losses of player B. A *strategy* is a choice of actions by either player. The matrix giving the payoff to a given player for each choice of strategy by both players is called the *payoff matrix*.

The objectives of the theory of games are to determine *optimum strategies* (that is, strategies that are most profitable to the respective players) and the corresponding payoff, which is called the *value* of the game.

Definition 3.3 (Minimax Strategy)**.** A strategy that minimizes the maximum loss of a player is called a *minimax strategy*. The choice of a minimax strategy to make a decision is called the minimax criterion.

Definition 3.4 (Saddle Point)**.** A *saddle point* of a game is a pair of strategies for which the corresponding entry in the payoff matrix is the smallest value of its row and the greatest value of its column. A game that has a saddle point is said to be *strictly determined*.

Definition 3.5 (Randomized Strategy)**.** If a player's choice of strategy is left to chance, the overall strategy is called a *randomized strategy*, or a *mixed strategy*. By contrast, in a game where each player makes a definite choice of a given strategy, each strategy is called a *pure strategy*.

3.2 Statistical Games

Definition 3.6 (Decision Function)**.** The function that tells the statistician which decision to make for each action of nature is called the *decision function* of a statistical game. The values of this function are given by $d_i(x)$, where d_i refers to the ith decision made by the statistician and x is a value of the random variable X whose values give the actions that can be taken by nature.

Definition 3.7 (Risk Function)**.** The function that gives the expected loss to which each value of the decision function leads for each action of nature is called the *risk function*. This function is given by

$$R(d_i, \theta_j) = E\{L[d_i(X), \theta_j]\}$$

where the expectation is taken with respect to the random variable X.

3.3 Decision Criteria

Definition 3.8 (Bayes Risk)**.** If Θ is assumed to be a random variable having a given distribution, the quantity

$$E[R(d, \Theta)]$$

where the expectation is taken with respect to Θ, is called the *Bayes risk*. Choosing the decision function d for which the Bayes risk is a minimum is called the *Bayes criterion*.

4 Point Estimation

4.1 Introduction

Traditionally, problems of statistical inference are divided into *problems of estimation* and *tests of hypotheses*.

Definition 4.1 (Point Estimation)**.** Using the value of a sample statistic to estimate the value of a population parameter is called *point estimation*. We refer to the value of the statistic as a *point estimate*, and we refer to the statistics as *point estimators*.

4.2 Unbiased Estimators

Definition 4.2 (Unbiased Estimator)**.** A statistic $\hat{\Theta}$ is an *unbiased estimator* of the parameter θ of a given distribution if and only if $E(\hat{\Theta}) = \theta$ for all possible values of θ.

Definition 4.3 (Aassymptotically Unbiased Estimator)**.** Letting $b_n(\theta) = E(\hat{\Theta}) - \theta$ express the *bias* of an estimator $\hat{\Theta}$ based on a random sample of size n from a given distribution, we say that $\hat{\Theta}$ is an *asymptotically unbiased estimator* of θ if and only if

$$\lim_{n\to\infty} b_n(\theta) = 0.$$

Theorem 4.1. If S^2 is the variance of a random sample from an infinite population with the finite variance σ^2, then $E(S^2) = \sigma^2$.

4.3 Efficiency

Definition 4.4 (Minimum Variance Unbiased Estimator)**.** The estimator for the parameter θ of a given distribution that has the smallest variance of all unbiased estimators for θ is called the *minimum variance unbiased estimator*, or the *best unbiased estimator* for θ.

Proposition 4.1 (Cramér-Rao Inequality)**.** If $\hat{\Theta}$ is an unbiased estimator of θ, the variance must satisfy the inequality

$$\mathrm{Var}(\hat{\Theta}) \geq \frac{1}{n \cdot E\left[(\partial \ln f(X)/\partial\theta)^2\right]}.$$

Theorem 4.2. If $\hat{\Theta}$ is an unbiased estimator of θ and

$$\mathrm{Var}(\hat{\Theta}) = \frac{1}{n \cdot E\left[(\partial \ln f(X)/\partial\theta)^2\right]}.$$

where the quantity in the denominator is referred to as the *information* about θ that is supplied by the sample, then $\hat{\Theta}$ is a minimum variance unbiased estimator of θ.

Definition 4.5 (Relative Efficiency)**.** If $\hat{\Theta}_1$ and $\hat{\Theta}_2$ are two unbiased estimators of the parameter θ of a given population and the variance of $\hat{\Theta}_1$ is less than the variance of $\hat{\Theta}_2$, we say that $\hat{\Theta}_1$, we say that $\hat{\Theta}_1$ is *relatively more efficient* than $\hat{\Theta}_2$. We use the ratio

$$\frac{\mathrm{Var}(\hat{\Theta}_1)}{\mathrm{Var}(\hat{\Theta}_2)}.$$

as a measure of the efficiency of $\hat{\Theta}_2$ relative to $\hat{\Theta}_1$.

Proposition 4.2. If $\hat{\Theta}$ is not an unbiased estimator of a given parameter θ, we judge its merits and make efficiency comparisons on the basis of the *mean square error* $E[(\hat{\Theta} - \theta)^2]$ instead of the variance of $\hat{\Theta}$.

4.4 Consistency

The variance or mean square error of an estimator may not provide good criteria for good indication of the chance fluctuations.

Definition 4.6 (Consistent Estimator)**.** The statistic $\hat{\Theta}$ is a *consistent estimator* of the parameter θ of a given distribution if and only if for each $c > 0$

$$\lim_{n\to\infty} P(|\hat{\Theta} - \theta| < c) = 1.$$

Theorem 4.3. If $\hat{\Theta}$ is an unbiased estimator of the parameter θ and $\text{var}(\hat{\Theta}) \to 0$ as $n \to \infty$, then $\hat{\Theta}$ is a consistent estimator of θ.

4.5 Sufficiency

Definition 4.7 (Sufficient Estimator)**.** The statistic $\hat{\Theta}$ is a *sufficient estimator* of the parameter θ of a given distribution if and only if for each value of $\hat{\Theta}$ the conditional probability distribution or density of the random sample $X_1, X_2, \ldots, X_n$, given $\hat{\Theta} = \theta$, is independent of θ.

Theorem 4.4. The statistic $\hat{\Theta}$ is a sufficient estimator of the parameter θ if and only if the joint probability distribution or density of the random sample can be factored so that

$$f(x_1, x_2, \ldots, x_n; \theta) = g(\hat{\theta}, \theta) \cdot h(x_1, x_2, \ldots, x_n)$$

where $g(\hat{\theta}, \theta)$ depends only on $\hat{\theta}$ and θ, and $h(x_1, x_2, \ldots, x_n)$ does not depend on θ.

4.6 Robustness

Definition 4.8. An estimator is said to be *robust* if its sampling distribution is not seriously affected by violations of assumptions.

4.7 The Method of Moments

Definition 4.9 (Sample Moments)**.** The kth sample moments of a set of observations $x_1, x_2, \ldots, x_n$ is the mean of their kth powers and it is denoted by m'_k; symbolically,

$$m'_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k.$$

The method of moments consists equating the first few moments of a population to the corresponding moments of a sample, thus getting as many as equations as needed to solve for the unknown parameters of the population. Thus, if a population has r parameters, the method of moments consists of solving the system of equations

$$m'_k = \mu'_k \quad k = 1, 2, \ldots, r$$

for the r parameters.

4.8 The Method of Maximum Likelihood

Definition 4.10 (Maximum Likelihood Estimator)**.** If $x_1, x_2, \ldots, x_n$ are the values of a random sample from a population with the parameter θ, the *likelihood function* of the sample is given by

$$L(\theta) = f(x_1, x_2, \ldots, x_n; \theta)$$

for values of θ within a given domain. Here, $f(x_1, x_2, \ldots, x_n; \theta)$ is the value of the joint probability distribution or the joint probability density of the random variables $X_1, X_2, \ldots, X_n$ at $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$. We refer to the value of θ that maximizes $L(\theta)$ as the *maximum likelihood estimator* of θ.

4.9 Bayesian Estimation

In Bayesian estimation the parameters are looked upon as random variables having *prior distributions*, usually reflecting the strength of one's belief about the possible values that they can assume. Bayesian estimation is to combining prior feelings about a parameter with direct sample evidence by determining $\varphi(\theta|x)$, the conditional density of Θ given $X = x$. The conditional distribution (which reflects the direct sample evidence) is called the *posterior distribution* of Θ. In general, if $h(\theta)$ is the value of the prior distribution of Θ at θ and we want to combine the information that it conveys with direct sample evidence about Θ, for instance, the value of a statistic $W = u(X_1, X_2, \ldots, X_n)$, we determine the posterior distribution of Θ by means of the formula

$$\varphi(\theta|w) = \frac{f(\theta, w)}{g(w)} = \frac{h(\theta) \cdot f(w|\theta)}{g(w)},$$

where $f(w|\theta)$ is the value of the sampling distribution of W given $\Theta = \theta$ at w, $f(\theta, w)$ is the value of the joint distribution of Θ and W at θ and w, and $g(w)$ is the value of the marginal distribution of W at w.

Theorem 4.5. If X is a binomial random variable and the prior distribution of Θ is a beta distribution with the parameter α and β, then the posterior distribution of Θ given $X = x$ is a beta distribution with the parameters $x + \alpha$ and $n - x + \beta$.

Theorem 4.6. If $\overline{X}$ is the mean of a random sample of size n from a normal population with the known variance σ^2 and the prior distribution of M (capital Greek μ) is a normal distribution with mean μ_0 and the variance σ_0^2, then the posterior distribution of M given $\overline{X} = \overline{x}$ is a normal distribution with the mean μ_1 and the variance σ_1^2, where

$$\mu_1 = \frac{n\overline{x}\sigma_0^2 + \mu_0\sigma^2}{n\sigma_0^2 + \sigma^2} \text{ and } \frac{1}{\sigma_1^2} = \frac{n}{\sigma^2} + \frac{1}{\sigma_0^2}.$$

5 Interval Estimation

5.1 Introduction

Definition 5.1 (Confidence Interval)**.** If $\hat{\theta}_1$ and $\hat{\theta}_2$ are values of the random variables $\hat{\Theta}_1$ and $\hat{\Theta}_2$ such that

$$P(\hat{\Theta}_1 < \theta < \hat{\Theta}_2) = 1 - \alpha$$

for some specified probability $1 - \alpha$, we refer to the interval

$$\hat{\theta}_1 < \theta < \hat{\theta}_2$$

as a $(1-\alpha)100\%$ *confidence interval* for θ. The probability $1 - \alpha$ is called the *degree of confidence*, and the endpoints of the interval are called the lower and upper *confidence limits*.

5.2 The Estimation of Means

Ths sampleing distribution of $\overline{X}$ for random samples of size n from a normal population with the mean μ and the variance σ^2 is a normal distribution with $\mu_{\overline{X}} = \mu$ and $\sigma_{\overline{X}}^2 = \sigma^2/n$. Then we can write

$$P(|Z| < z_{\alpha/2}) = P(|\overline{X} - \mu| < z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}) = P(\overline{X} - < z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} - < z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}) = 1 - \alpha,$$

where $Z = (\overline{X} - \mu)/(\sigma/\sqrt{n})$.

Theorem 5.1. If $\overline{X}$, the mean of a random sample of size n from a normal population with the known variance σ^2, is to be used as an estimator of the mean of the population, the probability is $1-\alpha$ that the error willl be less than $z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

Theorem 5.2. If $\overline{x}$ is the value of the mean of a random sample of size n from a normal population with the known variance σ^2, then

$$\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

is a $(1-\alpha)100\%$ confidence interval for the meanof the population.

When we are dealing with a random sample from a normal population, $n < 30$, and σ is known, we make use of the fact that

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

is a random variable having the t distribution with $n-1$ degrees of freedom.

Theorem 5.3. If $\overline{x}$ and s are the values of the mean and the standard deviation of a random sample of size n from a normal population, then

$$\overline{x} - t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}} < \mu\overline{x} + t_{\alpha/2,n-1} \cdot \frac{s}{\sqrt{n}}$$

is a $(1-\alpha)100\%$ confidence interval that for the mean of the poluation.

5.3 The Estimation of Differenes Between Means

For independent random samples from normal popluations

$$Z = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n_2}}}.$$

has the standard normal distribution.

Theorem 5.4. If $\overline{x}_1$ and $\overline{x}_2$ are the values of the means of independent random samples of sizes n_1 and n_2 from normal popluations with the knowns variances σ_1^2 and σ_2^2, then

$$(\overline{x}_1 - \overline{x}_2) - z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\overline{x}_1 - \overline{x}_2) + z_{\alpha/2} \cdot \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

is a $(1-\alpha)100\%$ confidence interval for the difference between the two population means.

Theorem 5.5. If $\overline{x}_1, \overline{x}_2$, s_1, and s_2 are the values of the means and the standard deviation of independent random samples of sizes n_1 and n_2 from normal populations with equal variances, then

$$(\overline{x}_1 - \overline{x}_2) - t_{\alpha/2,n_1+n_2-2} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 + \mu_2 < (\overline{x}_1 - \overline{x}_2) + t_{\alpha/2,n_1+n_2-2} \cdot s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

6 Hypothesis Testing

6.1 Introduction

Definition 6.1 (Statistical hypothesis)**.** An assertion or conjecture about the distribution of one or more random variables is called a *statistical hypothesis*. If a statistical hypothesis completely specifies the distribution, it is called a *simple* hypothesis; if not, it is referred to as a *composite hypothesis*.

To be able to construct suitable criteria for testing statistical hypotheses, it is necessary that we also formulate *alternative hypotheses*. Frequently, statisticians formulate as their hypotheses the exact opposite of what they may want to show. The view of the assumption of "no difference," hypotheses led to the term *null hypothesis*, but nowadays this term is applied to any ypothesis that we may want to test. Symbolically, we shall use the symbol H_0 for the null hypothesis that we want to test and H_1 or H_A for the alternative hypothesis.

6.2 Testing a Statistical Hypothesis

The testing of a statistical hypothesis is the application of an explicit set of rules for deciding on the basis of a random sample whether to accept the null hypothesis or to reject it in favor of the alternative hypothesis. Suppose that a statistician wants to test the null hypothesis $\theta = \theta_0$ against the alternative hypothesis $\theta = \theta_1$. Then generate sample data by conducting an experiment and then compute the value of a *test statistic*, which will decide what action to take for each possible outcome of the sample space. The test procedure, therefore, partitions the possible values of the test statistic into two subsets: an *acceptance region* for H_0 and a *rejection region* for H_0.

Definition 6.2 (Type I and Type II Errors)**.**

(a) Rejection of a null hypothesis when it is true is called a *type I error*. The probability of committing a type I error is denoted by α.
(b) Acceptance of the null hypothesis when it is false is called a *type II error*. The probability of committing a type II error is denoted by β.

Table 2: Type I and Type II Errors

	H_0 is true	H_0 is false
Accept H_0	No error	Type II error probability $= \beta$
Reject H_0	Type I error probability $= \alpha$	No error

Definition 6.3 (Critical Region)**.** It is customary to refer to the rejection region for H_0 as the *critical region* of a test. The probability of obtaining a value of the test statistic inside the critical region when H_0 is true is called the *size* of the critical region. Thus, the size of the critical region is just the probability α of committing a type I error. This probability is also called the *level of significance* of the test.

6.3 Losses and Risks

The concepts of loss functions and risk functions also play an important part in the theory of hypothesis testing. In the decision theory approach to testing the null hypothesis that a population

parameter θ equals θ_0 against the alternative that it equals θ_1, the statistician either takes the action a_0 and accepts the null hypothesis, or takes the action a_1 and accepts the alternative hypothesis.

6.4 The Neyman-Pearson Lemma

Definition 6.4 (The Power of a Test)**.** When testing the null hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta = \theta_1$, the quantity $1 - \beta$ is referred to as the *power* of the test at $\theta = \theta_1$. A critical region for testing a simple null hypothesis $H_0 : \theta = \theta_0$ against a simple alternative hypothesis $H_1 : \theta = \theta_1$ is said to be a *best critical region* or a *most powerful critical region* if the power of the test is a maximum at $\theta = \theta_1$.

Theorem 6.1 (Neyman-Pearson Lemma)**.** If C is a critical region of size α and k is a constant such that

$$\frac{L_0}{L_1} \leq k \text{ inside } C$$

and

$$\frac{L_0}{L_1} \geq k \text{ outside } C$$

then C is a most powerful critical region of size α for testing $\theta = \theta_0$ against $\theta = \theta_1$.

6.5 The Power Function of a Test

Definition 6.5 (Power function)**.** The *power function* of a test of a statistical hypothesis H_0 against an alternative hypothesis H_1 is given by

$$\pi(\theta) = \begin{cases} \alpha(\theta) & \text{for values of } \theta \text{ assumed under } H_0, \\ 1 - \beta(\theta) & \text{for values of } \theta \text{ assumed under } H_1. \end{cases}$$

Definition 6.6 (Uniformly most powerful critical region (test))**.** If, for a given problem, a critical region of size α is uniformly more powerful than any other critical region of size α, it is said to be a *uniformly most powerful critical region*, or a *uniformly most powerful test*.

6.6 Likelihood Ratio Tests

Definition 6.7 (Likelihood ratio test)**.** If ω and ω' are complementary subsets of the parameter space Ω and if the *likelihood ratio statistic*

$$\lambda = \frac{\max L_0}{\max L}$$

where $\max L_0$ and $\max L$ are the maximum values of the *likelihood function* for all values of θ in ω and Ω, respectively, then the critical region

$$\lambda \leq k$$

where $0 < k < 1$, defines a *likelihood ratio test* of the null hypothesis $\theta \in \omega$ against the alternative hypothesis $\theta \in \omega'$.

Theorem 6.2. For large n, the distribution of $-2 \cdot \ln \Lambda$ approaches, under very general conditions, the chi-square distribution with 1 degree of freedom.