

Paper -5

Estimation

Population vs sample

Aggregate of anything under study is called population.

Ex:- aggregate of person in india is population of persons in india apart of population is called samples if every unit of population has the same probability to be included in the samples gradually we consider random sample for the study of the population.

Parameter and statistic

Any calculation based on population is called parameter.

Ex- if we calculate arithmetic mean variance etc from the population values they will be called parameter. The calculation based on sample values is called statistic i.e from sample values they will be statistic.

When population is large or infinite the calculation of parameter is difficult we take help of statistic for this purpose and infer about parameter we consider statistic as estimate when statistic is used for population parameter we consider statistic as estimate when statistic is used for population parameter the statistic become estimate and the process is known as estimation .

There are mainly two types of estimation

1. Point estimation
2. Interval estimation

For the differentiation of point estimation and interval estimation we can use an example . if somebody ask what is distance between patna and gaya one can answer in two ways .

1. The distance between patna and gaya is 101km
2. The distance between patna and gaya is to 90 km to 101km.

Both cases are based on estimation in the 1st case is point estimation and the 2nd is interval estimation clearly the interval estimation is better than point estimation .

Now we shall discuss first method of point estimation prof. RA Fisher discuss both point estimation and interval estimation in detail .he started point estimation by describing the four properties of a good estimator they are-

1 unbiasedness 2 consistency 3 efficiency 4 sufficiency

Unbiased estimator

An estimator of a given parameter is said to be unbiased if its expected value is equal to the true value of the parameter.

In other words, an estimator is unbiased if it produces parameter estimates that are on average correct.

An unbiased estimator is an accurate statistic that's used to approximate a population parameter. "Accurate" in this sense means that it's neither an overestimate nor an underestimate. If an overestimate or underestimate does happen, the mean of the difference is called a "bias."

In more mathematical terms, an estimator is unbiased if:

$$\langle \hat{\theta} \rangle = \theta.$$

That's just saying if the estimator (i.e. the sample mean) equals the parameter (i.e. the population mean), then it's an unbiased estimator.

You might also see this written as something like "An unbiased estimator is when the mean of the statistic's sampling distribution is equal to the population's parameter." **This essentially means the same thing: if the statistic equals the parameter, then it's unbiased.**

consistency

A consistent estimate has insignificant errors (variations) as sample sizes grow larger. More specifically, the probability that those errors will vary by more than a given amount approaches zero as the sample size increases. In other words, the more data you collect, a consistent estimator will be

close to the real population parameter you're trying to measure. The [sample mean](#) and [sample variance](#) are two well-known consistent estimators.

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

Efficiency

an efficient estimator is one that has small [variances](#) (the estimator with the smallest possible variance is also called the “best” estimator). In other words, the estimator deviates as little as possible from the “true” value you are trying to estimate

sufficiency

a statistic Y is said to be a sufficient estimator for some parameter θ if the [conditional distribution](#) of Y: $T(X_1, X_2, \dots, X_n)$ doesn't depend on θ . While this definition is fairly simple, actually finding the conditional distribution is the tough part. In fact, most statisticians consider it extremely difficult. One, slightly easier, way to find the conditional distribution is to use the Factorization Theorem.

Maximum likelihood and their application

The MLE is an example of a point estimate because it gives a single value for the unknown parameter (later our estimates will involve intervals and probabilities). Two advantages of the MLE are that it is often easy to compute and that it agrees with our intuition in simple examples

Definition: Maximum Likelihood Estimators

Suppose that the data x_1, \dots, x_n has joint density function

$$f(x_1, \dots, x_n; \theta_1, \dots, \theta_p) = f(\bar{x}, \bar{\theta})$$

Then the **Likelihood function** is defined to be

$$\begin{aligned} L(\bar{\theta}) &= L(\theta_1, \dots, \theta_p) \\ &= f(x_1, \dots, x_n; \theta_1, \dots, \theta_p) \end{aligned}$$

and the **Maximum Likelihood estimators** of the parameters $\theta_1, \dots, \theta_p$ are the values that maximize

$$L(\bar{\theta}) = L(\theta_1, \dots, \theta_p)$$

Now we shall derive maximum likelihood estimation of some important distribution.

- 1.) Binomial distribution
- 2.) Poisson distribution
- 3.) Normal distribution
- 4.) Exponential distribution

Binomial distribution

Let $X_1, X_2, \dots, X_N \in \mathbb{R}$ be samples obtained from a Binomially Distribution.

Binomial Distribution is used to model 'x' successes in 'n' Bernoulli trials. Its p.d.f. is given by

$$f(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

$$L(p) = \prod_{i=1}^n f(x_i) = \prod_{i=1}^n \left(\frac{n!}{x_i!(n-x_i)!} p^{x_i} (1-p)^{n-x_i} \right) = \left(\prod_{i=1}^n \frac{n!}{x_i!(n-x_i)!} \right) p^{\sum_{i=1}^n x_i} (1-p)^{n-\sum_{i=1}^n x_i}$$

$$\ln L(p) = \sum_{i=1}^n x_i \ln p + \left(n - \sum_{i=1}^n x_i \right) \ln(1-p)$$

$$\frac{d \ln L(p)}{dp} = \frac{1}{p} \sum_{i=1}^n x_i - \left(n - \sum_{i=1}^n x_i \right) \frac{1}{1-p} = 0$$

$$\frac{(1-\hat{p}) \sum_{i=1}^n x_i - \left(n - \sum_{i=1}^n x_i \right) \hat{p}}{\hat{p}(1-\hat{p})} = 0$$

$$\sum_{i=1}^n x_i - \hat{p} \sum_{i=1}^n x_i - n\hat{p} + \sum_{i=1}^n x_i \hat{p} = 0$$

$$\hat{p} = \frac{\sum_{i=1}^n x_i}{n} = \frac{k}{n}$$

Poisson distribution

Let $X_1, X_2, \dots, X_n \in \mathbb{R}$ be a random sample from a Poisson distribution

The p.d.f. of a Poisson Distribution is :

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}; \text{ where } x = 0, 1, 2, \dots$$

The likelihood function is:

$$L(\lambda) = \prod_{i=1}^n \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = e^{-\lambda n} \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}$$

The log-likelihood is:

$$\ln L(\lambda) = -\lambda n + \sum_{i=1}^n x_i \ln(\lambda) - \ln\left(\prod_{i=1}^n x_i!\right)$$

Setting its derivative with respect to λ to zero, we have:

$$\frac{d}{d\lambda} \ln L(\lambda) = -n + \sum_{i=1}^n x_i \cdot \frac{1}{\lambda} = 0$$

giving,

$$\hat{\lambda} = \frac{\sum_{i=1}^n x_i}{n} = \bar{X}$$

which is the maximum likelihood estimate

Normal distribution

For Uniformly Distributed random variables $X_1, X_2, \dots, X_n \in \mathbb{R}$, the p.d.f is given by:

$$f(x_i) = \frac{1}{\theta}; \text{ if } 0 \leq x_i \leq \theta$$

$$f(x) = 0; \text{ otherwise}$$

If the uniformly distributed random variables are arranged in the following order

$$0 \leq X_1 \leq X_2 \leq X_3 \dots \leq X_n \leq \theta,$$

The likelihood function is given by:

$$L(\theta) = \prod_{i=1}^n f(x_i) = \prod_{i=1}^n \frac{1}{\theta} = \theta^{-n}$$

The log-likelihood is:

$$\ln L(\theta) = -n \ln(\theta)$$

Setting its derivative with respect to parameter θ to zero, we get:

$$\frac{d}{d\theta} \ln L(\theta) = \frac{-n}{\theta}$$

which is < 0 for $\theta > 0$

Hence, $L(\theta)$ is a decreasing function and it is maximized at $\theta = x_n$

The maximum likelihood estimate is thus,

$$\hat{\theta} = X_n$$

EXPONENTIAL DISTRIBUTION

Let $X_1, X_2, \dots, X_n \in \mathbb{R}$ be a random sample from the exponential distribution with p.d.f.

$$f(x) = (1/\theta) * \exp(-x/\theta)$$

The likelihood function $L(\theta)$ is a function of $x_1, x_2, x_3, \dots, x_n$, given by:

$$L(\theta) = (1/\theta) * \exp(-x_1/\theta) * (1/\theta) * \exp(-x_2/\theta) * \dots * (1/\theta) * \exp(-x_n/\theta)$$

$$L(\theta) = (1/\theta^n) * \exp(-\sum_{i=1}^n x_i/\theta)$$

We need to maximize $L(\theta)$. The logarithm of this function will be easier to maximize.

$$\ln [L(\theta)] = -n \cdot \ln(\theta) - (1/\theta) \sum_{i=1}^n x_i$$

Setting its derivative with respect to the parameter (θ) to zero, we have:

$$(d/d\theta) \ln[L(\theta)] = (-n/\theta) + \sum_{i=1}^n (-x_i/\theta^2) = 0$$

which implies that

$$\hat{\theta} = \frac{\sum_{i=1}^n x_i}{n} = \bar{X}$$

= Mean of $x_1, x_2, x_3, \dots, x_n$

Properties of maximum likelihood estimation

- i. M.l.e are not necessarily unbiased
- ii. M.l.e are consistent
- iii. The distribution of mle's tends to normality for large sample size
- iv. The mles's are normality distribution when n is large
- v. M.l.e are most efficient M.l.e are sufficient if sufficient estimate are exist If $\hat{\theta}$ is the m.l.e of θ , $f(\hat{\theta})$ is also m.l.e of $f(\theta)$

interval estimation

Interval estimation is the use of sample data to calculate an interval of possible *orprobable* values of an unknown population parameter, in contrast to point estimation, which is a single number.

Formula

$$\mu = \bar{x} \pm Z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Where –

- \bar{x} = mean
- $Z_{\frac{\alpha}{2}}$ = the confidence coefficient
- α = confidence level
- σ = standard deviation
- n = sample size

Confidence interval

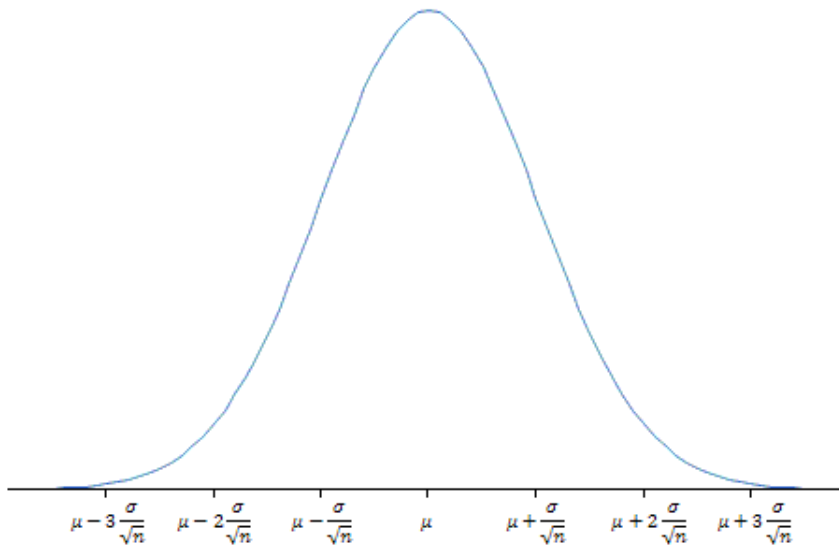
A confidence interval is an interval that will contain a population parameter a specified proportion of the time. The confidence interval can take any number of probabilities, with the most common being 95% or 99%.

A confidence interval is the probability that a value will fall between an upper and lower bound of a [probability distribution](#). For example, given a 99% confidence interval, stock XYZ's return will fall between -6.7% and +8.3% over the next year. In layman's terms, you are 99% confident that the returns of holding XYZ stock over the next year will fall between -6.7% and +8.3%.

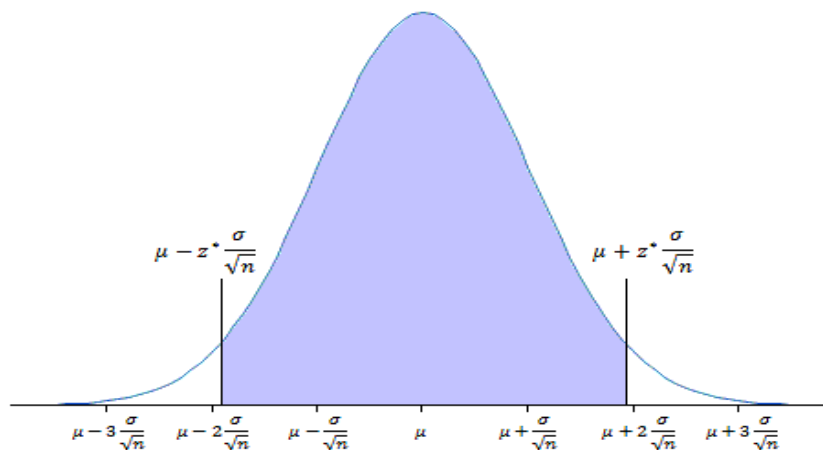
Statisticians use confidence intervals to measure uncertainty. A higher probability associated with the confidence interval means that there is a greater degree of certainty that the parameter falls within the bounds of the interval. Therefore, a higher confidence level indicates that the parameters must be broader to ensure that level of confidence.

Confidence interval for means of a normal population

The distribution of sample means for samples of size n can be illustrated as follows:



For any given value of z^* the probability that a sample mean lies within z^* standard deviations of the mean can be calculated using ordinary left-tail probability tables. Let's call this probability C .



$$P\left(\mu - z^*\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + z^*\frac{\sigma}{\sqrt{n}}\right) = C$$

Notice, in particular, that this probability tells us something about the sample means but nothing about the population mean. Now let's consider the inequality:

$$\mu - z^*\frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + z^*\frac{\sigma}{\sqrt{n}}$$

Subtract μ from all three terms:

$$-z^*\frac{\sigma}{\sqrt{n}} \leq \bar{x} - \mu \leq z^*\frac{\sigma}{\sqrt{n}}$$

Subtract \bar{x} from all three terms:

$$-\bar{x} - z^* \frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{x} + z^* \frac{\sigma}{\sqrt{n}}$$

Multiply all three terms by -1 remembering to reverse the inequalities:

$$\bar{x} + z^* \frac{\sigma}{\sqrt{n}} \geq \mu \geq \bar{x} - z^* \frac{\sigma}{\sqrt{n}}$$

Write the resulting inequality in an alternate form:

$$\bar{x} - z^* \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z^* \frac{\sigma}{\sqrt{n}}$$

Substituting this result into our original probability we obtain:

$$P\left(\bar{x} - z^* \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z^* \frac{\sigma}{\sqrt{n}}\right) = C$$

In this form, C is called the confidence level and indicates how confident we are that the population mean lies within the indicated confidence interval. For example, if $C = 0.95$ then $z^* = 1.96$. We say that we are 95% confident that the population mean lies within the interval:

$$\bar{x} - 1.96 \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96 \frac{\sigma}{\sqrt{n}}$$

Confidence interval for variance of a normal population

One variance

Theorem. If X_1, X_2, \dots, X_n are normally distributed and $a = \chi_{1-\alpha/2, n-1}^2$ and $b = \chi_{\alpha/2, n-1}^2$, then a $(1-\alpha)\%$ confidence interval for the population variance σ^2 is:

$$\left(\frac{(n-1)s^2}{b} \leq \sigma^2 \leq \frac{(n-1)s^2}{a} \right)$$

And a $(1-\alpha)\%$ confidence interval for the population standard deviation σ is:

$$\left(\frac{\sqrt{(n-1)}}{\sqrt{b}} s \leq \sigma \leq \frac{\sqrt{(n-1)}}{\sqrt{a}} s \right)$$

Proof. We learned previously that if X_1, X_2, \dots, X_n are normally distributed with mean μ and population variance σ^2 , then:

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

with $(a = \chi_{1-\alpha/2}^2)$ and $(b = \chi_{\alpha/2}^2)$, we can write the following probability statement:

$$P \left[a \leq \frac{(n-1)S^2}{\sigma^2} \leq b \right] = 1 - \alpha$$

Now, as always it's just a matter of manipulating the quantity in the parentheses. That is:

$$a \leq \frac{(n-1)S^2}{\sigma^2} \leq b$$

Taking the reciprocal of all three terms, and thereby changing the direction of the inequalities, we get:

$$\frac{1}{a} \geq \frac{\sigma^2}{(n-1)S^2} \geq \frac{1}{b}$$

Now, multiplying through by $(n-1)S^2$, and rearranging the direction of the inequalities, we get the confidence interval for σ^2 :

$$\frac{(n-1)S^2}{b} \leq \sigma^2 \leq \frac{(n-1)S^2}{a}$$

as was to be proved. And, taking the square root, we get the confidence interval for σ :

$$\frac{\sqrt{(n-1)S^2}}{\sqrt{b}} \leq \sigma \leq \frac{\sqrt{(n-1)S^2}}{\sqrt{a}}$$

two variance

Theorem. If $X_1, X_2, \dots, X_n \sim N(\mu_X, \sigma_X^2)$ and $Y_1, Y_2, \dots, Y_m \sim N(\mu_Y, \sigma_Y^2)$ are independent random samples, and:

$$(1) c = F_{1-\alpha/2}(m-1, n-1) = \frac{1}{F_{\alpha/2}(n-1, m-1)} \text{ and}$$

$$(2) d = F_{\alpha/2}(m-1, n-1),$$

then a $(1-\alpha)$ 100% confidence interval for σ_X^2/σ_Y^2 is:

$$\left(\frac{1}{F_{\alpha/2}(n-1, m-1)} \frac{s_X^2}{s_Y^2} \leq \frac{\sigma_X^2}{\sigma_Y^2} \leq F_{\alpha/2}(m-1, n-1) \frac{s_X^2}{s_Y^2} \right)$$

Proof. Because $X_1, X_2, \dots, X_n \sim N(\mu_X, \sigma_X^2)$ and $Y_1, Y_2, \dots, Y_m \sim N(\mu_Y, \sigma_Y^2)$, it tells us that:

$$\frac{(n-1)S_X^2}{\sigma_X^2} \sim \chi_{n-1}^2 \quad \text{and} \quad \frac{(m-1)S_Y^2}{\sigma_Y^2} \sim \chi_{m-1}^2$$

Then, by the independence of the two samples, we well as the definition of an F random variable, we know that:

$$F = \frac{\frac{(m-1)S_Y^2}{\sigma_Y^2} / (m-1)}{\frac{(n-1)S_X^2}{\sigma_X^2} / (n-1)} = \frac{\sigma_X^2}{\sigma_Y^2} \cdot \frac{S_Y^2}{S_X^2} \sim F(m-1, n-1)$$

Therefore, the following probability statement holds:

$$P \left[F_{1-\frac{\alpha}{2}}(m-1, n-1) \leq \frac{\sigma_X^2}{\sigma_Y^2} \cdot \frac{S_Y^2}{S_X^2} \leq F_{\frac{\alpha}{2}}(m-1, n-1) \right] = 1 - \alpha$$

Finding the $(1-\alpha)100\%$ confidence interval for the ratio of the two population variances then reduces, as always, to manipulating the quantity in parentheses. Multiplying through the inequality by:

$$\frac{S_X^2}{S_Y^2}$$

and recalling the fact that:

$$F_{1-\frac{\alpha}{2}}(m-1, n-1) = \frac{1}{F_{\frac{\alpha}{2}}(n-1, m-1)}$$

the $(1-\alpha)100\%$ confidence interval for the ratio of the two population variances reduces to:

$$\frac{1}{F_{\frac{\alpha}{2}}(n-1, m-1)} \frac{S_X^2}{S_Y^2} \leq \frac{\sigma_X^2}{\sigma_Y^2} \leq F_{\frac{\alpha}{2}}(m-1, n-1) \frac{S_X^2}{S_Y^2}$$

Minimum variance unbiased estimation

If an unbiased estimator has the variance equal to the CRLB, it must have the minimum variance amongst all unbiased estimators. We call it the **minimum variance unbiased estimator** (MVUE) of ϕ .

Sufficiency is a powerful property in finding unbiased, minimum variance estimators. If $T(\mathbf{Y})$ is an unbiased estimator of ϑ and S is a statistic sufficient for ϑ , then there is a function of S that is also an unbiased estimator of ϑ and has no larger variance than the variance of $T(\mathbf{Y})$. The following theorem formalizes this statement.

Theorem 2.5. *Rao-Blackwell theorem.*

Let $\mathbf{Y} = (Y_1, Y_2, \dots, Y_n)^T$ be a random sample, $\mathbf{S} = (S_1, \dots, S_p)^T$ be jointly sufficient statistics for $\vartheta = (\vartheta_1, \dots, \vartheta_p)^T$ and $T(\mathbf{Y})$ (which is **not** a function of \mathbf{S}) be an unbiased estimator of $\phi = g(\vartheta)$. Then, $U = E(T|\mathbf{S})$ is a statistic such that

- (a) $E(U) = \phi$, so that U is an unbiased estimator of ϕ , and
- (b) $\text{var}(U) < \text{var}(T)$.

Proof. First, we note that U is a statistic. Indeed, since \mathbf{S} are jointly sufficient for ϑ , the conditional distribution $\mathbf{Y}|\mathbf{S}$ does not depend on the parameters and so the conditional distribution of a function $T(\mathbf{Y})$ given \mathbf{S} , $T|\mathbf{S}$, does not depend on ϑ either. Thus, $U = E(T|\mathbf{S})$ is a function of the random sample only, not a function of ϑ , therefore it is a statistic.

Next, we will use the known facts about the conditional expectation and variance given Exercise 1.15. Since T is an unbiased estimator of ϕ , we have

$$E(U) = E[E(T|\mathbf{S})] = E(T) = \phi.$$

So U is also an unbiased estimator of ϕ , which proves (a). Finally, we get

$$\begin{aligned} \text{var}(T) &= \text{var}[E(T|\mathbf{S})] + E[\text{var}(T|\mathbf{S})] \\ &= \text{var}(U) + E[\text{var}(T|\mathbf{S})]. \end{aligned}$$

However, since T is not a function of \mathcal{S} we have $\text{var}(T|\mathcal{S}) > 0$, thus, it follows that $E[\text{var}(T|\mathcal{S})] > 0$, and hence (b) is proved. \square

It means that, if we have an unbiased estimator, T , of ϕ , which is not a function of the sufficient statistics, we can always find an unbiased estimator which has smaller variance, namely $U = E(T|S_1, \dots, S_p)$ which is a function of \mathcal{S} . We thus have the following result.

Concept of sufficient statistics

A *statistic* is a function $T = r(X_1, X_2, \dots, X_n)$ of the random sample X_1, X_2, \dots, X_n . Examples are

$$\begin{aligned} \bar{X}_n &= \frac{1}{n} \sum_{i=1}^n X_i, && \text{(the sample mean)} \\ s^2 &= \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2, && \text{(the sample variance)} \\ T_1 &= \max\{X_1, X_2, \dots, X_n\} \\ T_2 &= 5 \end{aligned} \tag{1}$$

The last statistic is a bit strange (it completely ignores the random sample), but it is still a statistic. We say a statistic T is an estimator of a population parameter if T is usually close to θ . The sample mean is an estimator for the population mean; the sample variance is an estimator for the population variation.

Obviously, there are lots of functions of X_1, X_2, \dots, X_n and so lots of statistics. When we look for a good estimator, do we really need to consider all of them, or is there a much smaller set of statistics we could consider? Another way to ask the question is if there are a few key functions of the random sample which will by themselves contain all the information the sample does. For example, suppose we know the sample mean and the sample variance. Does the random sample contain any more information about the population than this? We should emphasize that we are always assuming our population is described by a given family of distributions (normal, binomial, gamma or ...) with one or several unknown parameters. The answer to the above question will depend on what family of distributions we assume describes the population.

We start with a heuristic definition of a *sufficient statistic*. We say T is a sufficient statistic if the statistician who knows the value of T can do just as good a job of estimating the unknown parameter θ as the statistician who knows the entire random sample.

The mathematical definition is as follows. A statistic $T = r(X_1, X_2, \dots, X_n)$ is a sufficient statistic if for each t , the conditional distribution of X_1, X_2, \dots, X_n given $T = t$ and θ does not depend on θ .

To motivate the mathematical definition, we consider the following “experiment.” Let $T = r(X_1, \dots, X_n)$ be a sufficient statistic. There are two statisticians; we will call them A and B. Statistician A knows the entire random sample X_1, \dots, X_n , but statistician B only knows the value of T , call it t . Since the conditional distribution of X_1, \dots, X_n given θ and T does not depend on θ , statistician B knows this conditional distribution. So he can use his computer to generate a random sample X'_1, \dots, X'_n which has this conditional distribution. But then his random sample has the same distribution as a random sample drawn from the population (with its unknown value of θ). So statistician B can use his random sample X'_1, \dots, X'_n to compute whatever statistician A computes using his random sample X_1, \dots, X_n , and he will (on average) do as well as statistician A. Thus the mathematical definition of sufficient statistic implies the heuristic definition.

It is difficult to use the definition to check if a statistic is sufficient or to find a sufficient statistic. Luckily, there is a theorem that makes it easy to find sufficient statistics.

Theorem 1. (*Factorization theorem*) Let X_1, X_2, \dots, X_n be a random sample with joint density $f(x_1, x_2, \dots, x_n | \theta)$. A statistic $T = r(X_1, X_2, \dots, X_n)$ is sufficient if and only if the joint density can be factored as follows:

$$f(x_1, x_2, \dots, x_n | \theta) = u(x_1, x_2, \dots, x_n) v(r(x_1, x_2, \dots, x_n), \theta) \quad (2)$$

where u and v are non-negative functions. The function u can depend on the full random sample x_1, \dots, x_n , but not on the unknown parameter θ . The function v can depend on θ , but can depend on the random sample only through the value of $r(x_1, \dots, x_n)$.

Rao Blackwell theorem

The Rao-Blackwell theorem (sometimes called the *Rao-Blackwell-Kolmogorov theorem* or *Rao-Blackwellization*) is a way to improve the efficiency of initial estimators. [Estimators](#) are observable [random variables](#) used to estimate quantities. For example, the (observable) [sample mean](#) is an estimator for the (unknown) [population mean](#).

Theorem 1.1 Let $\mathbf{X} \sim f_{\mathbf{X}}(\mathbf{x}, \theta)$ and T be sufficient for θ , $\mathbf{x} \in \mathfrak{X}$ and $t \in \mathfrak{T}$. Let U be any unbiased estimator for $g(\theta)$. Define $V_t = E(U|T = t)$. Then V is an unbiased estimator for $g(\theta)$ and $\text{Var}(V) \leq \text{Var}(U)$ with equality iff $V = U$ with probability one.

Proof 1.1 Since $U = U(\mathbf{X})$ is an estimator, it is also a statistic. And, since T is sufficient for θ we have

$$V = E(U|T = t) \quad (1)$$

$$= \int_{\mathfrak{X}} u(x) f_{X|T}(x|T = t) dx \quad (2)$$

By Fisher, and noting that $u(x)$ is a function of x and not θ , we see that V is θ -free. Thus, V is a statistic as well.

Further,

$$E(U) = g(\theta) \quad (3)$$

$$= \int_{\mathfrak{X}} u(x) f_X(x, \theta) dx \quad (4)$$

$$= \int_{\mathfrak{T}} \left[\int_{X \in T=t} u(x) f_{X|T}(x|T = t) dx \right] f_T(t, \theta) dt \quad (5)$$

$$= \int_{\mathfrak{T}} v(t) f_T(t, \theta) dt \quad (6)$$

$$= E(V) \quad (7)$$

So, V is unbiased.

Now,

$$\text{Var}(U) = E(U - E(U))^2 \quad (8)$$

$$= E(U - E(V))^2 \quad (9)$$

$$= E((U - V)^2) + E((V - E(V))^2) + 2E((U - V)(V - E(V))) \quad (10)$$

Since we know that $E(U) = E(V)$ by above,

$$E((U - V)(V - E(V))) = \int_{\mathfrak{X}} (V - E(V))(U - V) f_X(x, \theta) dx \quad (11)$$

$$= \int_{\mathfrak{T}} (V - E(V)) \left[\int_{X \in T=t} (U - V) f_{X|T}(x|T = t) dx \right] f_T(t, \theta) dt \quad (12)$$

$$= \int_{\mathfrak{T}} (V - E(V)) [0] f_T(t, \theta) dt \quad (13)$$

$$= 0 \quad (14)$$

and thus

$$\text{Var}(U) = E((U - V)^2) + E((V - E(V))^2) \quad (15)$$

$$\geq E((V - E(V))^2) \quad (16)$$

$$\geq \text{Var}(V) \quad (17)$$

with equality iff $E((U - V)^2) = 0$ or $V = U$ with probability one.