Probability and Statistics

Population

The totality of all observations which are related to a particular study

The totality of all observations which are related to a particular study under consideration constitute a population. Population can be finite or infinite.
For example, the observations obtained by measuring the atmospheric pressure every day from the past on to the future constitute an infinite population, *or* all measurements on the depth of a lake from any conceivable position also constitute an infinite population.

Binomial Population, Normal Population or in General Population

Binomial Population:

A population whose observations are values of a *r.v.* having a binomial distribution.
Example: Checking of defective items in an assembly line.

Normal Population:

Life lengths of storage batteries may be a normal population.

Hence, we can speak about population mean and variance.

In the field of statistical inference the statistician is interested in arriving at conclusions concerning a population when it is impossible or impractical to observe the entire set of observations that make up the population. For example, in attempting to determine the average life length of a certain brand of electric bulb, if would be impossible to test all such bulbs. Therefore, we must depend on a subset of observations from the population to help us make inferences concerning that population. This brings us to consider the notion of sampling.

A sample is a subset of a population.

Any sampling procedure that produce inferences that consistently overestimate or consistently underestimate some characteristics of the population is said to be biased.

Sample

To eliminate bias in the sampling procedure It is desirable to choose a random sample

A sample is a subset of a population. If our inferences from the sample to the population are to be valid, we must obtain samples that are representative of the population. Any sampling procedure that produce inferences that consistently overestimate or consistently underestimate some characteristics of the population is said to be *biased.* To eliminate any possibility of bias in the sampling procedure, it is desirable to choose *a random sample* in the sense that the observations are made independently and at random.

A set of observations X1,X2, . . . ,Xn constitutes a random sample of size n from a finite population of size N

Random Sample(Finite Population):

A set of observations $X_1, X_2, \ldots, X_n$ constitutes a random sample of size n from a finite population of size N, if its values are chosen so that each subset of n of the N elements of the population has the same probability of being selected.

Random Sample(Infinite Population):

A set of observations $X_1, X_2, \ldots, X_n$ constitutes a random sample of size n from an infinite population $f(x)$ if

(i) each X_i is a $r.v.$ whose distribution is $f(x)$

(ii) these n random variables are independent.

Selection of Random Sample

(a) Using Random Number Table (In the case of finite population of small size)

When the population size is large , the use of random numbers can became very laborious and at times practically impossible. When the population size is infinite, the situation is different since we cannot physically number the elements of the population. In this case, we may able to approximate conditions of randomness by choosing one unit each half hour (in the case of selecting a sample from a production line). The proper use of artificial or mechanical devices for selecting random samples is always preferable to human judgement, as it is extremely difficult to avoid unconscious biases when making almost any kind of selection.

Population : mean μ & S.D σ

Sample : mean $\bar{x}$ & S.D s ← statistics

Random Sample:

Let $X_1, X_2, \ldots, X_n$ be n independent *r.vs*, each having the same probability distribution $f(X)$. Then $X_1, X_2, \ldots, X_n$ is said to be a random sample of size n from the population $f(x)$ and $f(x_1, x_2, \ldots, x_n) = f(x_1)f(x_2)\ldots f(x_n)$.

In the purpose of most statistical investigations is to generalize from information contained in random samples about the population from which the samples were obtained. In particular. we are usually concerned with the problem of making inferences about the *parameters* of populations, such as the mean μ ot the S.D σ. In making such inferences, we use *statistics* such as $\overline{x}$ and s, quantities calculated on the basis of sample observations.

Statistic Any function of the random variable constituting a random sample is called a statistic

Any function of the random variable constituting a random sample is called a statistic.

Some Important Statistics The probability distribution of a statistic is called a sampling distribution.

- **The Sample Mean:**
  If $X_1, X_2, \ldots, X_n$ represents a random sample of size n, then *the sample mean* is defined by the statistic

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

- **The Sample Variance**
  The variability in the sample should display how observations spread out from the average. In this context, we consider *the sample variance*, which is defined by

  sample standard deviation $$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n \text{ or } n-1}$$

  and S (*ie,* the positive square root of S^2) is called *the sample standard deviation.* The probability distribution of a statistic is called *a sampling distribution.*

The Sampling Distribution of Mean (σ Unknown)

Theorem 0.1. *If a random sample of size n is taken from a population having the mean μ and the variance σ^2, then $\overline{X}$ is a r.v. whose distribution has the mean μ. For samples from infinite population the variance of this distribution is $\frac{\sigma^2}{n}$.*

If Population : mean μ & Variance σ^2

Then $\bar{X}$ is a r.v whose Distribution Mean μ and variance of this distribution is (σ^2) / n

Proof. We have

$$\begin{aligned}\mu_{\overline{X}} &= E\left(\overline{X}\right)\\ &= E\left(\frac{X_1+X_2+\cdots+X_n}{n}\right)\\ &= \frac{1}{n}\left[E(X_1)+E(X_2)+\cdots+E(X_n)\right]\\ &= \frac{1}{n}\times n\mu\\ &= \mu\end{aligned}$$

$$\left[or\ \mu_{\overline{X}} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\sum_{i=1}^{n}\frac{x_i}{n}f(x_1,x_2,\ldots,x_n)dx_1dx_2\ldots dx_n\ and\ apply\ independence\ property\right]$$

$$\begin{aligned}Var(\overline{X}) &= Var\left(\frac{X_1+X_2+\cdots+X_n}{n}\right)\\ &= \frac{1}{n^2}\left(Var(X_1)+Var(X_2)+\cdots+Var(X_n)\right)\\ &= \frac{1}{n^2}\times n\sigma^2\\ &= \frac{\sigma^2}{n}\end{aligned}$$

□

Chebyshev's Theorem — **If a probability distribution has mean μ, and standard deviation σ, the probability of getting a value which deviates from μ by atleast kσ is at most 1/ (k^2)** [P:204, example]

Law of Large Numbers

Theorem 0.2. ***(Chebyshev's Theorem):*** *If a probability distribution has mean μ, and standard deviation σ, the probability of getting a value which deviates from μ by atleast $k\sigma$ is at most $\frac{1}{k^2}$.*
ie,

$$P\left(|X-\mu|\geq k\sigma\right) < \frac{1}{k^2}$$

or

$$P\left(|X-\mu| < k\sigma\right) \geq 1-\frac{1}{k^2}$$

mean μ and standard deviation σ

X1 , X2 ,....., Xi ,..... Xn

the probability of getting a value which deviates from μ by atleast kσ is at most 1/(k^2)

We have $\mu_{\overline{X}} = \mu$ and $\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$.

$$\therefore\ P\left(|\overline{X}-\mu| < \frac{k\sigma}{\sqrt{n}}\right) \geq 1-\frac{1}{k^2}.$$

Take $\frac{k\sigma}{\sqrt{n}} = \epsilon$. Then

$$P\left(|\overline{X}-\mu| < \epsilon\right) \geq 1-\frac{\sigma^2}{n\epsilon^2}.$$

Thus, for any given $\epsilon > 0$, the probability that $\overline{X}$ differs from μ by less than ϵ can be made arbitrarily close to 1 by choosing n sufficiently large. *ie,* the larger the sample size, the closer we can expect $\overline{X}$ to be to the mean of the population. In this sense we can say that the mean becomes more and more reliable as an estimate of μ as the sample size is increased. This result - that $\overline{X}$ becomes arbitrarily close to μ with arbitrarily high probability - is called the *law of large numbers.*
$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$ is called *the standard error of the mean*, which is treated as the measure of reliability of the mean.

LAW OF LARGE NUMBERS

As the sample size is increased ↑ estimate X becomes arbitrarily close to μ with arbitrarily high probability - is called the law of large numbers.

Note:
Theorem 0.1 provides only partial information about the theoretical sampling distribution of the mean. In general, it is impossible to determine such distribution exactly without having any knowledge about the actual form of the population. But it is possible to find the limiting distribution as $n \to \infty$ of a *r.v.* whose values are closely related to $\overline{X}$, assuming only that the population has finite variance σ^2.

Theorem 0.3. ***(Central Limit Theorem):*** *If $\overline{X}$ is the mean of a sample of size n taken from a population having the mean μ and the finite variance σ^2, then $Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ is a r.v. whose distribution function approaches that of the standard normal distribution as $n \to \infty$.*

Note:
If a random sample come from a normal population, the sampling distribution of the mean is normal regardless of the size of the sample.

Example: A random sample of size 100 is taken from an infinite population having the mean $\mu = 76$ and the variance $\sigma^2 = 256$. what is the probability that $\overline{X}$ will be between 75 and 78.
Example: Page 210.

The Sampling Distribution of the Mean(σ unknown)

P(L1 < $\overline{X}$ < L2)

Case I: If n is large, even though σ is unknown, we can assure that $Z = \dfrac{\overline{X} - \mu}{s/\sqrt{n}} \to N(0, 1)$.

Case II: If n is small, then the sampling distribution of $Z = \dfrac{\overline{X} - \mu}{s/\sqrt{n}}$ is not known exactly unless we make the assumption that the sample comes from a normal population.
ie,

Theorem 0.4. *If $\overline{X}$ is the mean of a random sample of size n taken from a normal population having the mean μ and the variance σ^2, and $S^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n-1}$, then $t = \dfrac{\overline{X} - \mu}{s/\sqrt{n}}$ is a r.v. having the t-distribution with the parameter $\nu = n - 1$.*

t-Distribution (OR) Student Distribution

(Student is the pseudonym for William Sealy Goset.)

$$f(u) = k\left(1 + \frac{u^2}{n}\right)^{-\left(\frac{n+1}{2}\right)} : -\infty < u < \infty$$

with n, the number of degrees of freedom.

$$k = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\,\Gamma(n/2)}$$

$$\lim_{n\to\infty} k\left[1 + \frac{u^2}{n}\right]^{-\frac{n}{2}}\left[1 + \frac{u^2}{n}\right]^{-1/2} = \frac{1}{e^{\frac{u^2}{2}}}$$

t-distribution is the same as the N(0, 1)

Properties of the t-Distribution

1. Mean of the distribution is zero.
2. Variance $= \dfrac{n}{n-2}$, $n \geq 2$.
3. Variance > 1. As $n \to \infty$, variance $\to 1$ so that the t-distribution is the same as the $N(0, 1)$ distribution.

Sampling Distribution of Mean:

Case : 1
P(L1 < $\overline{X}$ < L2)
By central Limit Theorem
$Z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ *is a r.v.*

Case 2: If σ is not know and sample is very large then s^2 = σ^2
P(L1 < $\overline{X}$ < L2)
$t = \dfrac{\overline{X} - \mu}{s/\sqrt{n}}$ *is a r.v.*
Apply T-Distribution (n-1) Degree of freedom

Case 3: If σ is not know and sample is Small then s^2 ≠ σ^2

P(L1 < X̄ < L2)

We can't Apply CLT and T-Distribution

Case 4: If sample Comes from Normal Population

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}} \ is \ a \ r.v.$$

P(L1 < X̄ < L2) Apply T-Distribution (n-1) Degree of freedom

Note: The t-distribution should not be used with small samples from populations that are not approximately normal.

Example: ACME corporation manufactures light bulbs, the life length of which are normally distributed. The CEO claims that an average ACME light bulb lasts 300 days. A researcher randomly selects 15 bulbs for testing. The sampled bulbs last an average of 290 days, with a standard deviation of 50 days. If the CEO's claim were true, what is the probability that 15 randomly selected bulbs would have an average life of no more than 290 days?
Given,

$$n = 15, \quad \overline{x} = 290, \quad s = 50, \quad \mu = 300$$

$$\begin{aligned} P(\overline{x} \leq 290) &= P\left(t < \frac{290 - 300}{50/\sqrt{15}}\right) \\ &= P(t < -0.7745966)_{14} \\ &= 0.226 \ (using \ t - calculator, \ online \ statistical \ table) \end{aligned}$$

Thus, if the true bulb life were 300 days, there is 22.60% chance that the average bulb life for 15 randomly selected bulbs would be less than or equal to 290 days.

Note:
The overall shape of a t-distribution is similar to that of a normal distribution. Both are bell shaped and symmetrical about the mean.

Example (Page No.214): A manufacturer of fuses claims that with 20% overload, the fuses will blow in 12.40 minutes on average. To test this claim, a sample of 20 of the fuses was subjected to a 20% overload, and the times it took them to blow had a mean 10.63 minutes and a standard deviation of 2.48 minutes. If it can be assumed that the data constitute a random sample from a normal population, do they tend to support or refute the manufacturer's claim?

Given that, $n = 20, \quad \overline{x} = 10.63, \quad s = 2.48, \quad \mu = 12.40$

$$\begin{aligned} p(\overline{x} \leq 10.63) &= P\left(t \leq \frac{10.63 - 12.40}{2.48/\sqrt{20}}\right) \\ &= P(t \leq -3.19)_{n=19} \\ &< 0.005 \end{aligned}$$

$[P(t > 2.861) = 0.005, \ \therefore P(t < -2.861) = 0.005]$.
Hence the data tend to refute the manufacturer's claim. In all likelihood, the mean blowing time of his fuses with a 20% overload is less than 12.40 minutes.

The Sampling Distribution of the Variance

P(L1 < (S^2) < L2)

Theorem 0.5. *If S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2, then*

Chi-square Distribution with (n-1) Degree of freedom

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

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

Chi-Square Variate = Sum of the squares of the Standard Normal Variates

$$\left[\textit{Hint: } \sum(X_i-\mu)^2=\sum(X_i-\overline{X}+\overline{X}-\mu)^2 \Rightarrow \frac{\sum(X_i-\mu)^2}{\sigma^2}=\frac{(n-1)s^2}{\sigma^2}+\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}\right)^2\right].$$

Example: A manufacturer of car batteries guarantees that his batteries will last on the average, 3 years with a standard deviation of 1 year. If time of these batteries have life time of 1.9, 2.4, 3, 3.5 and 4.2 years, is the manufacturer still convinced that his batteries have a standard deviation of 1 year? Assume that the battery life times follows normal distribution.

$$s^2=0.815, \quad \chi^2=\frac{4\times 0.815}{1}=3.26$$

$$P\left(0.711<\chi^2<9.488\right)=0.95-0.05=0.90$$

Chi-square Distribution

$$f(\chi^2)=\frac{1}{2^{n/2}\Gamma(n/2)}e^{-\chi^2/2}\left(\chi^2\right)^{(n/2)-1}, \quad 0<\chi^2<\infty$$

is the χ^2-distribution with n *d.o.f.*

[*ie,* a *Gamma Distribution* with $\alpha=n/2$ and $\beta=2$, $f(x)=\frac{1}{\beta^2\Gamma\alpha}x^{\alpha-1}e^{-(x/\beta)}$, $x>0$.]

Mean$=n$

Variance$=\frac{n}{2}\times 4=2n$

$m.g.f=(1-2t)^{-n/2}$

Example: An optical firm purchases glass to be ground into lenses, and it is known from past experience that the variance of the refractive index of this kind of glass is 1.26×10^{-4}. As it is important that the various pieces of glass have nearly the same index of refraction, the firm rejects such a shipment if the sample variance of 20 pieces selected at random exceeds 2.00×10^{-4}. Assuming that the sample values may be looked upon as a random sample from a normal population, what is the probability that a shipment will be rejected even though $\sigma^2=1.26\times 10^{-4}$?

$$\begin{aligned}
&\sigma^2=1.26\times 10^{-4}, \ ; \ n=20\\
P\left(S^2\geq 2\times 10^{-4}\right)&=P\left(\frac{(n-1)S^2}{\sigma^2}\geq\frac{n-1}{\sigma^2}\times 2\times 10^{-4}\right)\\
&=P\left(\chi^2\geq\frac{19}{1.26\times 10^{-4}}\times 2\times 10^{-4}\right)\\
&=P\left(\chi^2\geq 30.2\right)_{19}\\
&<0.05. \quad \left[\because P\left(\chi^2>\chi^2_{0.05}\right)=P(\chi^2>30.1)=0.05\right].
\end{aligned}$$

Hence the probability that a good shipment will erroneously be rejected is less than 0.05.

$T=\frac{Z}{\sqrt{\frac{V}{n}}}=\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ $\quad \frac{n\cdot S^2}{\sigma^2}$

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

Note:

If $Z\to N(0,1)$, and $V\to\chi^2$ then $T=Z/\sqrt{V/n}\to t_n$.

t- distribution with (n-1) degree of freedom

A problem closely related to that of finding the distribution of the sample variance is that finding the distribution of the ratio of the variances of two independent random samples. This problem is important because it arises in tests in which we want to determine whether two samples come from populations having equal variances. If they do, the two sample variances should be nearly the same.

Theorem 0.6. *If S_1^2 and S_2^2 are the variances of independent random samples of size n_1 and n_2, respectively, taken from two normal populations having the same variances σ_1^2 andσ_2^2 respectively, then $F=\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$ is a random variable having the* F-distribution *with parameters $\gamma_1=n_1-1$ and $\gamma_2=n_2-1$.*

Note: Theorem 0.6 requires the assumption that we are sampling from normal population.

F-Distribution

$$h(f) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}\left(\frac{n_1}{n_2}\right)^{n_1/2}\frac{f^{\frac{n_1}{2}-1}}{\left(1+\frac{n_1 f}{n_2}\right)^{\frac{n_1+n_2}{2}}} \quad : \quad 0 < f < \infty$$

Example: If two independent random samples of size $n_1 = 7$ and $n_2 = 13$ are taken from a normal population, what is the probability that the variance of the first sample will be atleast three times as large as that of the second sample?

$$\begin{aligned} P\left(S_1^2 \geq 3S_2^2\right) &= P\left(\frac{S_1^2}{S_2^2} \geq 3\right) \\ &= P(F \geq 3)_{(6,12)} \quad = 0.05 \end{aligned}$$

Notes:

1. If $U \to \chi^2_{n_1}$ and $V \to \chi^2_{n_2}$ then $F = \dfrac{U/n_1}{V/n_2} \to F_{n_1,n_2}$.
2. $F_{1-\alpha}(v_1, v_2) = \dfrac{1}{F_\alpha(v_2, v_1)}$.
   Example: For $\gamma_1 = 10$ and $\gamma_2 = 20$, $F_{0.95} = \dfrac{1}{F_{0.05}(20,10)} = \frac{1}{2.77} = 0.36$.

Estimation

Statistical inference can be divided into two major areas: *estimation* and *tests of hypotheses.*

Point Estimate

A point estimate of some population parameter θ is a single value $\hat{\theta}$ of a statistic $\hat{\Theta}$. For example, the value $\overline{x}$ of the statistic $\overline{X}$, computed from a sample of size n, is a point estimate of the population parameter μ.

Desirable Properties of a Good Estimate

(1) ***Unbiasedness:*** A statistic $\hat{\Theta}$ is said to be an *unbiased* estimator of the parameter θ if $E(\hat{\Theta}) = \theta$.
Example 1: $\overline{X}$ is an unbiased estimator of μ.

Unbiased Estimator $E(\hat{\Theta}) = \theta$

θ comes from population

$\hat{\Theta}$ comes from sample

Example 2: S^2 is an unbiased estimator of σ^2.

Hint:

$$\begin{aligned}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2 &= \sum_{i=1}^{n}\left[(X_i-\mu)-\left(\overline{X}-\mu\right)\right]^2\\ &= \sum_{i=1}^{n}(X_i-\mu)^2-2\left(\overline{X}-\mu\right)\left(\sum_{i=1}^{n}(X_i-\mu)\right)+n\left(\overline{X}-\mu\right)^2\\ &= \sum_{i=1}^{n}(X_i-\mu)^2-2n\left(\overline{X}-\mu\right)^2+n\left(\overline{X}-\mu\right)^2\\ &= \sum_{i=1}^{n}(X_i-\mu)^2-n\left(\overline{X}-\mu\right)^2\end{aligned}$$

$$\begin{aligned}\therefore E(S^2) &= \frac{1}{n-1}\left[\sum_{i=1}^{n}E\left(X_i-\mu\right)^2-nE\left(\overline{X}-\mu\right)^2\right]\\ &= \frac{1}{n-1}\left[n\sigma^2-n\frac{\sigma^2}{n}\right]\\ &= \sigma^2\end{aligned}$$

Hence S^2 is an unbiased estimate of σ^2.

(2) **Consistency:** A sequence of point estimates $\{T_1, T_2, \dots\}$ will be called consistent for θ if $P\left(|T_n-\theta|<\epsilon\right)\to 1$ as $n\to\infty$ where $T_n = T(X_1, X_2, \dots, X_n)$.

Example 1: $P(|\overline{X}-\mu|<\epsilon)\to 1$ as $n\to\infty$. Hence $\overline{X}$ is a consistent estimate for μ.

Example 2: Let $X_1, X_2, \dots, X_n$ be *i.i.d* $N(\mu, \sigma^2$ random variables. Then $\frac{(n-1)S^2}{\sigma^2}\sim\chi^2_{n-1}$. $\therefore$ $E\left(\frac{S^2}{\sigma^2}\right)=1$ and $Var\left(fracS^2\sigma^2\right)=\frac{2}{n-1}\Rightarrow E(S^2)=\sigma^2$ and $Var(S^2)=\frac{2\sigma^4}{n-1}$.

Consistency: Inorder to prove consistency Property Apply Chebyshev's Inequality

P (|Tn − θ| < ∈) → 1 as n →∞

$$\therefore P\left(|S^2-\sigma^2|<k\sigma^2\sqrt{\frac{2}{n-1}}\right)\geq 1-\frac{1}{k^2}.$$

Now put $k\sigma^2\sqrt{\frac{2}{n-1}}=\epsilon\Rightarrow k=\frac{\epsilon}{\sigma^2}\sqrt{\frac{n-1}{2}}$ and so $\frac{1}{k^2}\to 0$ as $n\to\infty$. Hence, S^2 is a consistent estimate of σ^2.

(3) **More Efficient Estimator:** If $\hat{\Theta}_1$ and $\hat{\Theta}_2$ are two estimators of the same population parameter Θ, and if $\sigma^2_{\hat{\Theta}_1}<\sigma^2_{\hat{\Theta}_2}$, then $\hat{\Theta}_1$ is a more efficient estimator of Θ than $\hat{\Theta}_2$. If we consider all possible unbiased estimators of some parameter Θ. the one with the smallest variance is called the *most efficient estimator of* Θ.

More Efficient Estimator:
If we consider all possible unbiased estimators of some parameter Θ.
The one with the smallest variance is called the most efficient estimator of Θ.

Here our choice for an estimator of Θ is $\hat{\Theta}_1$.

(4) **Sufficiency:** Let $X_1, X_2, \dots, X_n$ be a random sample from a population having parameter Θ. Then, a statistic $\hat{\Theta}$ is said to be a sufficient for Θ if $f(x_1, x_2, \dots, x_n|\hat{\Theta}=\hat{\theta})$ does not depend on Θ.

Sufficiency :
population having parameter Θ
statistic ˆΘ is said to be a sufficient for Θ if f(x1, x2, . . . , xn| ˆΘ = ˆθ) does not depend on Θ

b(1, p) Binomial Distribution
b(n, p) Bernoulli Distribution

i.e, if $\hat{\Theta}$ is sufficient for Θ, we need only concentrate on $\hat{\Theta}$ since it exhausts all the information that the sample has about Θ.

Binomial Distribution

Example: Let $X_1, X_2, \ldots, X_n$ be *i.i.d* $b(1,p)$ random variables. Then $T = \sum_{i=1}^{n} X_i$ is a sufficient estimate for p.

$$P\left\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n \middle| \sum_{i=1}^{n} X_i = t\right\} = \frac{P\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n, T = t\}}{P\left(\sum_{i=1}^{n} X_i = t\right)}$$

$$= \frac{p^{x_1}(1-p)^{1-x_1}p^{x_2}(1-p)^{1-x_2}\ldots p^{x_n}(1-p)^{1-x_n}}{\binom{n}{t}p^t(1-p)^{n-t}}$$

$$= \frac{1}{\binom{n}{t}}$$

independent of

This is independent of T
Hence T is a sufficient Estimate for P

Hence $T = \sum_{i=1}^{n} X_i$ is sufficient for p.

X is sufficient for μ if σ^2 is known

Example 2: If X_i, $i = 1, 2, \ldots, n$ are from $N(\mu, \sigma^2$ then $\overline{X}$ is sufficient for μ if σ^2 is known and $\sum(X_i - \mu)^2$ is sufficient for σ^2 if $\mu = \mu_\circ$ is known.

(Xi − μ)^2 is sufficient for σ^2 if μ = μ◦ is known

Theorem 0.7. *Let $X_1, X_2, \ldots, X_n$ be discrete random variables with* pmf *$P_\theta(x_1, x_2, \ldots, x_n)$. Then T is sufficient for θ if and only if*

$$P_\theta(x_1, x_2, \ldots, x_n) = h(x_1, x_2, \ldots, x_n)g_\theta(T(x_1, x_2, \ldots, x_n))$$

only, where h is a non-negative function of $x_1, x_2, \ldots, x_n$ only and does not depend on θ and g is a non-negative function of θ and T only.

Example: Let $X_1, X_2, \ldots, X_n$ be a sample from $N(\mu, \sigma^2)$ where μ and σ^2 are unknown. Then,

$$f(x_1, x_2, \ldots, x_n) = \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} e^{-\frac{\sum(X_i - \mu)^2}{2\sigma^2}}$$

$$= \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} e^{-\frac{1}{2}\left[\frac{(n-1)S^2}{\sigma^2} + \left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}}\right)^2\right]}$$

$$= \frac{1}{\left(\sqrt{2\pi}\sigma\right)^n} e^{-\left[\frac{\sum x_i^2}{2\sigma^2} + \frac{n\mu^2}{2\sigma^2} - \frac{\mu\sum x_i}{\sigma^2}\right]}$$

$$= h(x_1, x_2, \ldots, x_n)g\left(\mu, \sigma, \sum x_i^2, \sum x_i\right)$$

Thus $\left(\sum x_i, \sum x_i^2\right)$ is jointly sufficient for (μ, σ^2) or $(\overline{X}, S^2)$ is sufficient for (μ, σ^2).

If σ^2 is unknown, $\overline{X}$ is not sufficient for μ.
If σ^2 is known, $\overline{X}$ is sufficient for μ.
If μ is unknown, S^2 is not sufficient for σ^2.
If $\mu = \mu_\circ$ is known, $\sum(X_i - \mu_\circ)^2$ is sufficient for σ^2.

Method of Maximum Likelihood Estimation

Bernoulli Distribution

Example: Let $X \sim b(n,p)$. One observation on X is available, and it is known that n is either 2 or 3 and $p = \frac{1}{2}$ or $p = \frac{1}{3}$. Our object is to find an estimate of p.
$P(X = x)$ for each possible pair (n, p) :

x	$(2,\frac{1}{2})$	$(2,\frac{1}{3})$	$(3,\frac{1}{2})$	$(2,\frac{1}{3})$	Maximum Probability
0	$\frac{1}{4}$	$\frac{4}{9}$	$\frac{1}{8}$	$\frac{8}{27}$	$\frac{4}{9}$
1	$\frac{1}{2}$	$\frac{4}{9}$	$\frac{3}{8}$	$\frac{12}{27}$	$\frac{1}{2}$
2	$\frac{1}{4}$	$\frac{1}{9}$	$\frac{3}{8}$	$\frac{6}{27}$	$\frac{3}{8}$
3	0	0	$\frac{1}{8}$	$\frac{1}{27}$	$\frac{1}{8}$

Hence,

$$(\hat{n},\hat{p})(x) = \begin{cases} (2,\frac{1}{3}) & if \quad x=0 \\ (2,\frac{1}{2}) & if \quad x=1 \\ (3,\frac{1}{2}) & if \quad x=2 \\ (3,\frac{1}{2}) & if \quad x=3 \end{cases}$$

The principle of maximum likelihood essentially assumes that the sample is representative of the population and chooses as the estimate that value of the parameter that maximises the *pdf (pmf)* $f_\theta(x)$.

Definition: Let $X_1, X_2, \ldots, X_n$ be a random sample with joint *pdf* or *pmf* $f_\theta(x_1, x_2, \ldots x_n)$. Then $L(\theta : x_1, x_2, \ldots x_n) = f_\theta(x_1, x_2, \ldots x_n)$, considered as a function of θ, is called a likelihood function.

Quiz 1 End

Interval Estimator

An interval estimate of a population parameter θ is an interval of the form $\hat{\Theta}_L < \theta < \hat{\Theta}_U$, where $\hat{\Theta}_L$ and $\hat{\Theta}_U$ depends on the value of the statistic $\hat{\Theta}$ for a particular sample and also the sampling distribution of $\hat{\Theta}$.

α is called Level of Significance
(1 - α) * 100% is called Confidence Interval
(1 - α) is called Confidence Coefficient or Degree of Confidence

Since different samples will generally yield different values of $\hat{\Theta}$ and, therefore different values of $\hat{\theta}_L$ and $\hat{\theta}_U$, these end points of the interval are values corresponding random variables $\hat{\Theta}_L$ and $\hat{\Theta}_U$. From the sampling distribution of $\hat{\Theta}$, we shall be able to determine $\hat{\Theta}_L$ and $\hat{\Theta}_U$ such that $P\left(\hat{\Theta}_L < \theta < \hat{\Theta}_U\right) = 1-\alpha$ *for* $0 < \alpha < 1$, then we have a probability $1-\alpha$ of selecting a random sample that will produce an interval containing θ. The interval $\hat{\Theta}_L < \theta < \hat{\Theta}_U$ computed from selected sample is then called a $(1-\alpha)100\%$ *confidence interval*, $1-\alpha$ is called the *confidence coefficient* or *degree of confidence* and the end points $\hat{\Theta}_L$ and $\hat{\Theta}_U$, are called the *lower and upper confidence limits.* Thus, when $\alpha = 0.05$, we have a 95% confidence interval and $\alpha = 0.01$ we obtain a 99% confidence interval. The wider the confidence interval is, the more confident we can be that the given interval contains the unknown parameter. Of course, it is better to be 95% confident that the average life of a certain television transistor is between 6 and 7 years than to be 99% confident that it is between 3 and 10 years. Ideally, *we prefer a short interval with high degree of confidence.*

Estimation of Mean

$\overline{X}$ is likely to be a very accurate estimate of μ when n is large ($\because Var(\overline{x}) = \frac{\sigma^2}{n}$, which is small when n is large).

Confidence Interval of μ : σ Known

If n is large or our sample is selected from a normal population, we have $\overline{X} \to N(\mu, \frac{\sigma}{\sqrt{n}})$.

For finding a $(1-\alpha)100\%$ confidence interval for μ, consider

$$P\left(-Z_{\alpha/2} < Z < Z_{\alpha/2}\right) = 1-\alpha$$

$$i.e, \quad P\left(-Z_{\alpha/2} < \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} < Z_{\alpha/2}\right) = 1-\alpha$$

$$i.e, \quad P\left(\overline{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Thus we have,

If $\overline{x}$ is the mean of a random sample of size n from a population with unknown variance σ^2, a $(1-\alpha)100\%$ confidence interval for μ is given by $\overline{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \overline{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, where $Z_{\alpha/2}$ is the $z-$value leaving an area of $\frac{\alpha}{2}$ to the right.

Example: The average zinc concentration recovered from a sample of zinc measurements in 36 different locations is found to be 2.6*grams per milliliter*. Find the 95% and 99% confidence intervals for the mean zinc concentration in the river. Assume that the population standard deviation is 0.3.

95% confidence interval is $2.6 - 1.96\frac{0.3}{\sqrt{36}} < \mu < 2.6 + 1.96\frac{0.3}{\sqrt{36}}$

$$i.e,\ 2.5 < \mu < 2.7.$$

for α = 0.025 Z = 1.96
for α = 0.005 Z = 2.57

99% confidence interval is $2.6 - 2.575\frac{0.3}{\sqrt{36}} < \mu < 2.6 + 2.575\frac{0.3}{\sqrt{36}}$

$$i.e,\ 2.47 < \mu < 2.73.$$

i.e, a longer interval is required to estimate μ with a higher degree of confidence.

Note: The $(1-\alpha)100\%$ confidence interval provides an estimate of the accuracy of our point estimate. If μ is actually the center value of the interval, then $\overline{x}$ estimates μ without error. Most of the time, however, $\overline{x}$ will not be equal to μ and the point estimate is in error. The size of this error will be absolute value of the difference between μ and $\overline{x}$, and we can be $(1-\alpha)100\%$ confident that this will not exceed $Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

Note: If $\overline{x}$ is used as an estimate of μ, we can be $(1-\alpha)100\%$ confident that the error will not exceed a specified amount e when the sample size is $n = \left(\frac{Z_{\alpha/2}\sigma}{e}\right)^2$ $\left[Hint:\ take\ Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = e\right]$.

Example: How large a sample is required in the above example if we want to be 95% confident that our estimate of μ is off by less than 0.05?

$$n = \left(\frac{1.96 \times 0.3}{0.05}\right)^2 = 138.3$$

$\therefore$ we can be 95% confident that a random sample of size 139 will provide an estimate $\overline{x}$ differing from μ by an amount less than 0.05.

Confidence Interval of μ : σ Unknown

If we have a random sample from a normal population then $T = \frac{\overline{X} - \mu}{s/\sqrt{n}} \to t_{n-1}$. In this situation with σ unknown, T can be used to construct a confidence interval on μ.

Consider,

$$P\left(-t_{\alpha/2} < T < t_{\alpha/2}\right) = 1-\alpha$$

$$i.e,\ \ P\left(-t_{\alpha/2} < \frac{\overline{X}-\mu}{s/\sqrt{n}} < t_{\alpha/2}\right) = 1-\alpha$$

$$i.e,\ P\left(\overline{X} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right) = 1-\alpha$$

Thus, we have

If $\overline{x}$ and s are the mean and standard deviation of a random sample of size n from a normal population with unknown variance σ^2, a $(1-\alpha)100\%$ confidence interval for μ is given by $\overline{x} - t_{\alpha/2}\frac{s}{\sqrt{n}} < \mu < \overline{x} + t_{\alpha/2}\frac{s}{\sqrt{n}}$, where $t_{\alpha/2}$ is the $t-$value with $n-1$ degrees of freedom leaving an area of $\frac{\alpha}{2}$ to the right.

Large Sample Confidence Interval

Even when normally can not be assumed, σ is unknown, and $n \geq 30$, scan replace σ and the confidential interval $\overline{x} \pm Z_{\alpha/2}\frac{s}{\sqrt{n}}$ may be used. This is often referred to as a large sample confidence interval.

Example: The contents of 7 similar containers of sulphuric acid are 9.8, 10.2, 10.4, 9.8, 10, 10.2 and 9.6 litters. Find a 95% confidence interval for the mean of all such containers, assuming an approximately normal distribution.

$$\overline{x} = 10, \quad s = 0.283, \quad t_{0.025} = 2.447, \quad n = 6$$

Hence, the 95% confidence interval for μ is $9.74 < \mu < 10.26$

Confidence Interval of σ^2

Assume that a random sample of size n is drawn from a normal population with variance σ^2. Obviously, s^2 is a point estimate of σ^2. Hence S^2 is called an estimator of σ^2. An interval estimate of σ^2 can be established by using the statistic

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} \rightarrow \chi^2_{n-1}$$

. In order to construct a $(1-\alpha)100\%$ confidence interval, we may write

$$P\left(\chi^2_{1-\alpha/2} < \chi^2 < \chi^2_{\alpha/2}\right) = 1-\alpha$$

$$\Rightarrow P\left(\chi^2_{1-\alpha/2} < \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}\right) = 1-\alpha$$

$$\Rightarrow P\left(\frac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}\right) = 1-\alpha$$

Thus we have

If S^2 is the variance of a random sample of size n from a normal population, a $(1-\alpha)100\%$ confidence interval for σ^2 is $\frac{(n-1)S^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)S^2}{\chi^2_{1-\alpha/2}}$, where $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ are χ^2-value with $n-1$ degrees of freedom, leaving areas of $\alpha/2$ and $1-\alpha/2$ respectively to the right.

Example: The following are the weights, in decagrams, of 10 packages of grass seeds distributed be a certain company: 46.4, 46.1, 45.8, 47, 46.1, 45.9, 45.8, 46.9, 45.2 and 46. Find a 95% confidence interval for the variance of all such packages of grass seed distributed by this company, assuming a normal population.

$$S^2 = 0.286, \quad \alpha = 0.05, \quad \chi^2_{0.025} = 19.023, \quad \chi^2_{0.975} = 2.7$$

$\therefore$ 95% confidence interval for σ^2 is $0.135 < \sigma^2 < 0.953$.

Confidence Interval for Proportion

A point estimator of the proportion p in a binomial experiment is given by the statistic $\hat{P} = \frac{X}{n}$, where X represents the number of successes in n trials. Therefore, the sample proportion $\hat{p} = \frac{x}{n}$ will be used as the point estimate of the parameter p.

If the unknown proportion p is not expected to be too close to zero or 1, we can establish a confidence interval for p by considering the sampling distribution of $\hat{P}$. Designating a failure in each binomial trial by 0 and a success by 1, the number of successes, x, can be interpreted as the sum of n values consisting only of zeros and ones, and $\hat{p}$ is just the sample mean of these n values. Hence by the central limit theorem, for n sufficiently large, $\hat{P}$ is approximately normally distributed with mean,

$$\mu_{\hat{P}} = E(\hat{P}) = E\left(\frac{X}{n}\right) = \frac{np}{p} = p.$$

Variance,

$$\sigma^2_{\hat{P}} = \sigma^2_{X/n} = \frac{1}{n^2}\left(\sigma^2_X\right) = \frac{1}{n^2} npq = \frac{pq}{n}.$$

$$P\left(-Z_{\alpha/2} < Z < Z_{\alpha/2}\right) = 1 - \alpha$$

$$\Rightarrow P\left(-Z_{\alpha/2} < \frac{\hat{P} - p}{\sqrt{\frac{pq}{n}}} < Z_{\alpha/2}\right) = 1 - \alpha$$

$$\Rightarrow P\left(\hat{p} - Z_{\alpha/2}\sqrt{\frac{pq}{n}} < p < \hat{p} + Z_{\alpha/2}\sqrt{\frac{pq}{n}}\right) = 1 - \alpha$$

It is difficult to manipulate the inequalities so as to obtain a random interval whose end points are independent of p. the unknown parameter. When n is large, replace p in the radical sign by $\hat{p} = \frac{x}{n}$. Then we get

$$P\left(\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right) = 1 - \alpha$$

For our particular random sample of size n, the sample proportion $\hat{p} = \frac{x}{n}$ is computed, and hence we get the following result.

Result:
If $\hat{p}$ is the proportion of success in a random sample of size n. and $\hat{q} = 1 - \hat{p}$, an approximate $(1-\alpha)100\%$ confidence interval for the binomial parameter p is given by

$$P\left(\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right)$$

, where $Z_{\alpha/2}$ is the $Z-$value leaving an area of $\alpha/2$ to the right.

Note: When n is small and the unknown proportion p is believed to be close to 0 or to 1, the confidence interval procedure established here is unreliable and therefore, should not be used.

Example: In a random sample of $n = 500$ families owing television sets in a city, it is found that 340 subscribed HBO. Find a 95% confidence interval for actual proportion of families in the city who subscribe to HBO.

$$\hat{p} = \frac{340}{500} = 0.68, \quad Z_{0.025} = 1.96$$

$\therefore$ 95% confidence interval for p is $0.64 < p < 0.72$.

Notes:

(1) If $\hat{p}$ is used as an estimate of p, we can be $(1-\alpha)100\%$ confident that the error will not exceed $Z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}$.

(2) If $\hat{p}$ is used as an estimate of p, we can be $(1-\alpha)100\%$ confident that the error will be less than a specified amount e when the sample size is approximately $n = \frac{Z^2_{\alpha/2}\hat{p}\hat{q}}{e^2}$.

(3) Note (2) is somewhat misleading in that we must use $\hat{p}$ to determine the sample size n, but $\hat{p}$ is computed from the sample. If a crude estimate of p can be made without taking a sample, we could use this value of $\hat{p}$ and then determine n. Lacking such an estimate, we could take preliminary sample of size $n \geq 30$ to provide an estimate of p. Then, using (2), we can determine approximately how many observations are needed to provide the desired degree of accuracy.

Example: How large a sample is required in the above example if we want to be 95% confident that our estimate of p is within 0.02?

Treat 500 families families as preliminary sample providing p as estimate $\hat{p} = 0.68$. Then $n = 2090$.

(4) In the confidence interval for p, if we assign a particular value of $\hat{p}$, namely $\hat{p} = 1/2$, then n will turn out to be larger than necessary for the specified degree of confidence and as a result our degree confidence will increase. Thus we have,
If $\hat{p}$ is used as an estimate of p, we can be atleast $(1-\alpha)100\%$ confidence that the error will not exceed a specified amount e when the sample size is $n = \frac{Z^2_{\alpha/2}}{4e^2}$.

Example: How large a sample is required in the above example if we want to be atleast 95% confident that our estimate of p is within an error 0.02?

$$n = \frac{1.96^2}{4(0.02)^2} = 2401.$$