Statistical Analysis (II)

Sadhana Dash

November 7, 2022

Outline

Sampling Distributions
Central Limit Theorem
Parameter Estimation (Point and Interval)

The Sampling Distributions:

Random Sample : The random variables X_1, $X_2 \cdots X_n$, are a random sample of size n if all the X_is are independent random variables and every X_i has the same probability distribution.
Statistic : A statistic can be any function of the observations in a random sample.
Sampling Distribution : The probability distribution of a statistic is called a sampling distribution.
The sample Mean
Let a random sample of size n be taken from a normal population with mean, μ and variance σ^2.
Each observation in the considered sample, say, X_1,X_2, ...,X_n, is a normally and independently distributed random variable.
The sample mean is given by

$$\overline{X} = \frac{X_1 + X_2 \cdots X_n}{n} \tag{1}$$

We know that the linear functions of independent, normally distributed random variables are also normally distributed. Therefore, the distribution of sample mean will have a normal distribution.
The mean of such a distribution would be

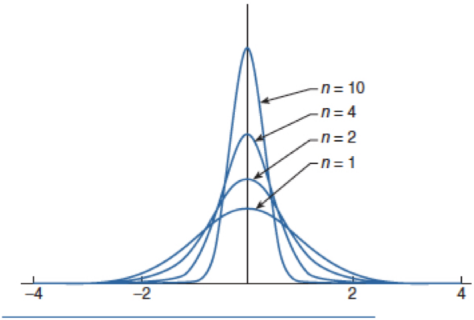
$$E[\overline{X}] = \mu_{\overline{X}} = \frac{\mu_1 + \mu_2 \cdots \mu_n}{n} = \mu \tag{2}$$

The variance would be

$$Var[\overline{X}] = \sigma^2_{\overline{X}} = \frac{\sigma_1^2 + \sigma_2^2 \cdots \sigma_n^2}{n^2} = \frac{\sigma^2}{n} \tag{3}$$

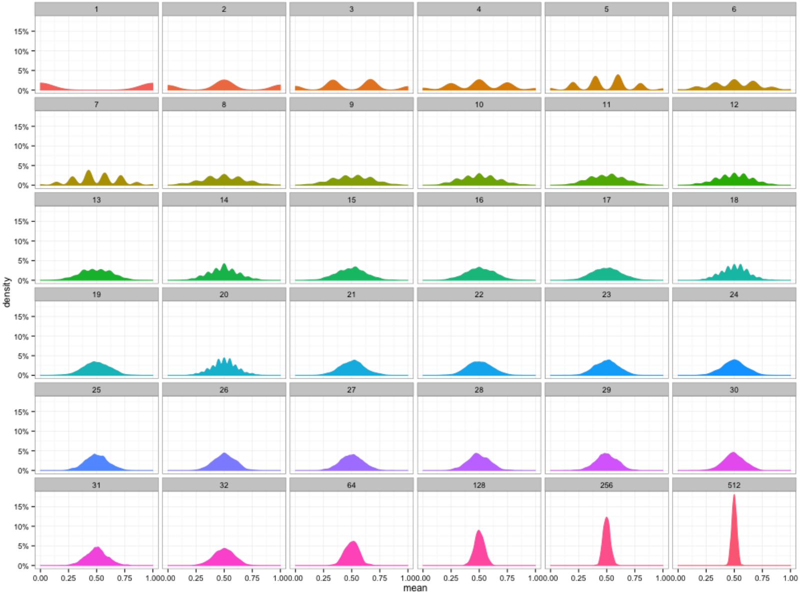
The expected value of the sample mean is the population mean, μ whereas its variance is $1/n$ times the population variance.

Distribution of sample mean from a normal population for different sample sizes.



The Central Limit Theorem :

If X_1,X_2,$\cdots$, X_n be a sequence of independent and identically distributed random variables each having mean, μ and finite variance, σ^2 , then for large n, the distribution of X_1+X_2+$\cdots$ + X_n is approximately normal with mean $n\mu$ and variance $n\sigma^2$.
It follows from the the Central Limit Theorem, CLT
$\frac{X_1+X_2+...+X_n-n\mu}{\sigma\sqrt{n}}$
is approximately a standard normal random variable.
OR
$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \sim Z$ (for $n \to \infty$)



1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
64
128
256
512
density
15%
10%
5%
0%
0.00
0.25
0.50
0.75
1.00
mean

The Point Estimation :

If X is a random variable with probability distribution f(x), characterized by the unknown parameter θ and if X_1,X_2, ...,X_n is a random sample of size n from X , the statistic
$\hat{\Theta} = h\ (X_1, X_2, \cdots, X_n)$ used to estimate the best possible value of θ is called a **point estimator** of θ .
After the sample has been selected, $\hat{\Theta}$ takes on a particular numerical value $\hat{\theta}$.
$\hat{\theta}$ is called the point estimate of θ.
A **point estimate** of a population parameter θ is a single numerical value , commonly denoted as $\hat{\theta}$ of a statistic $\hat{\Theta}$.
The statistic $\hat{\Theta}$ is called the point estimator.

The Mean Squared Error of an Estimator

The mean squared error of an estimator $\hat{\Theta}$, of the parameter θ, MSE($\hat{\Theta}$) is defined as

MSE($\hat{\Theta}$) $= \text{E}[(\hat{\Theta} - \theta)^2]$

It can be shown that

MSE($\hat{\Theta}$) $= \text{E}[(\hat{\Theta} - \theta)^2] = \text{E}[\ \hat{\Theta}$ - $\text{E}(\hat{\Theta})]^2 + [\theta - E(\hat{\Theta})]^2$

$= \text{Var}(\hat{\Theta}) + (bias)^2$

For unbiased estimators, MSE($\hat{\Theta}$) is equal to variance of $\hat{\Theta}$

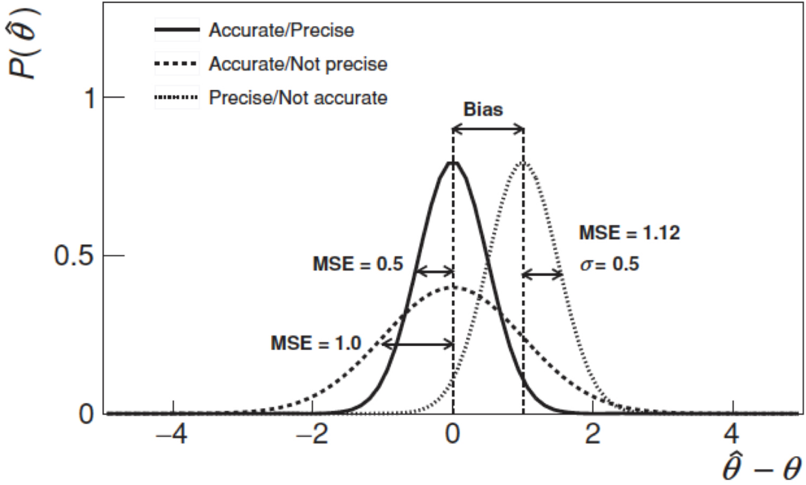
Standard Error

The standard error of an estimator $\hat{\Theta}$ is its standard deviation given by

$\sigma_{\hat{\Theta}} = \sqrt{Var(\hat{\Theta})}$.

If the standard error involves unknown parameters that can be estimated, one has to substitute those values into $\sigma_{\hat{\Theta}}$

The error is then called an estimated standard error, denoted by $\hat{\sigma}_{\hat{\Theta}}$.



$P(\hat{\theta})$
Accurate/Precise
Accurate/Not precise
Precise/Not accurate
Bias
MSE = 1.12
σ = 0.5
MSE = 0.5
MSE = 1.0
1
0.5
0
–4
–2
0
2
4
$\hat{\theta} - \theta$

The sample variance, S^2 is given by

$$\mathbf{S^2} = \frac{\sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{X_i} - \overline{\mathbf{X}})^2}{(\mathbf{n}-\mathbf{1})} \tag{4}$$

We know that

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

From equation (7),

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

$(n-1)E[S^2] = nE[X_1^2] - nE[\overline{X}^2]$

$= nVar(X_1) + n(E[X_1])^2 - nVar(\overline{X}) - n(E[\overline{X}])^2$

$= n\sigma^2 + n\mu^2 - n(\frac{\sigma^2}{n}) - n\mu^2$

$= (n-1)\sigma^2$

Thus,

$E[S^2] = \sigma^2$

The Method of Maximum Likelihood

Let us suppose that X is a random variable with probability distribution $f(x;\theta)$, where θ is a single unknown parameter.
Let X_1,$X_2, ...,$$X_n$ be a random sample of size n.
Then the likelihood function of the sample is defined as
$L(x_1, x_2, ..., x_n|\theta) = f(x_1;\theta)f(x_2;\theta)...f(x_n;\theta)$
The maximum likelihood estimator of θ is the value of θ that maximizes the likelihood function L(θ).

The M L E of Exponential Parameter

Let X be an Exponential random variable with parameter λ.
The probability density function is
$f(x;\lambda) = \lambda e^{-\lambda x}$
The likelihood function of the random sample of size n is

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

$$= \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

The log-likelihood function is

$$lnL(\lambda) = nln(\lambda) - \lambda \sum_{i=1}^{n} x_i$$

$$\frac{dlnL(\lambda)}{d\lambda} = (n/\lambda) - \sum_{i=1}^{n} x_i$$

Equating to zero

$$\hat{\lambda} = n/(\sum_{i=1}^{n} x_i)$$

The Confidence Interval :

An interval estimate for a population parameter is called a confidence interval.
We can find an interval (or range) of values that contains the actual unknown population parameter.
We can estimate lower L and upper U values between which the population parameter falls:
$L < \theta < U$
$P(L \leq \mu \leq U) = 1 - \alpha$, where $0 \leq \alpha \leq 1$
1 - α is called the confidence coefficient.
The typical confidence coefficients are 0.90, 0.95, and 0.99, with corresponding confidence levels 90%, 95%, and 99%, respectively. The greater the confidence level, the more confident we can be that the confidence interval contains the actual population parameter.

Confidence Interval on the mean of a normal population (variance known)

Let X_1,X_2, ...,X_n be a random sample of size n from a normal distribution with unknown mean μ and known variance σ^2.
We know that the sample mean, $\overline{X}$ is normally distributed with mean μ and variance, σ^2/n.
We can always construct a Z-statistic , $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$.
A **confidence interval estimate** for μ is an interval of the form $l \leq \mu \leq u$, where the endpoints l and u are computed from the sample data.
Let L and U be the random variables which correspond to lower and upper limits, we define
$P(L \leq \mu \leq U) = 1 - \alpha$, where $0 \leq \alpha \leq 1$
Since $\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ has a standard normal distribution, we can write
P ($-z_{\alpha/2} \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}) = 1 - \alpha$

Continued :

P (-$z_{\alpha/2} \leq \frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}) = 1-\alpha$

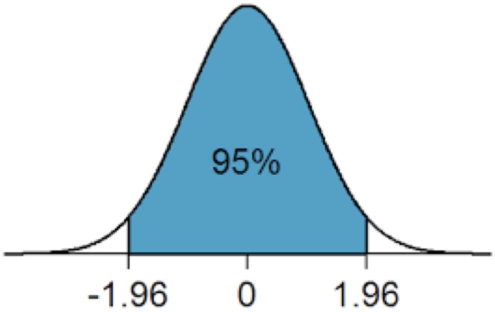
P ($\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$

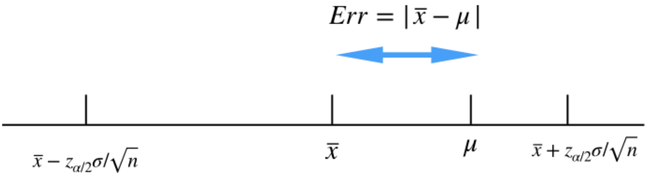
If $\overline{x}$ is the sample mean of a random sample of size n from a normal population with known variance σ^2, a 100(1 - α)% confidence interval on μ is given by

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

where $z_{\alpha/2}$ is the upper 100 $(\alpha/2)$ percentage point of the standard normal distribution.

Interpretation : If an infinite number of random samples are collected and a 100(1 - α)% confidence interval for μ is computed from each sample, 100(1 - α)% of these intervals will contain the true value of μ.


95%
-1.96
0
1.96


$Err = |\bar{x} - \mu|$
$\bar{x} - z_{\alpha/2}\sigma/\sqrt{n}$
$\bar{x}$
μ
$\bar{x} + z_{\alpha/2}\sigma/\sqrt{n}$

Choice of sample size :

Length of confidence interval :

The length of a confidence interval is a measure of the precision of estimation . It is given by

$2 \times z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

One can define the error, $Err = |\overline{x} - \mu|$

The error is less than or equal to $z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ with confidence 100(1 - α). If $\overline{x}$ is used as an estimate of μ, we can be 100(1 -α)% confident that the error will not exceed a specified amount Err when the sample size is

$n = (\frac{z_{\alpha/2}\sigma}{Err})^2$

We can observe that

- As the desired length of the confidence interval decreases, the required sample size n increases for a fixed value of and specified confidence.
- As σ increases, the required sample size n increases for a fixed desired length and specified confidence.
- As the level of confidence increases, the required sample size n increases for fixed desired length and standard deviation .