Sampling Theory

Population: A large collection of individuals or attributes or numerical data can be regarded as population or universe. It is an aggregate of objects, animate or inanimate, under study. The population may be finite or infinite.

If the population is large, complete enumeration is not possible most of the times because of the cost involved, time consumed and also in some cases units are destroyed in the course of inspection (e.g. inspection of crackers). So we take help of sampling.

A finite subset of the population is known as **sample.**

Size of the population N is the number of objects or observations in the population. Population is said to be finite or infinite depending on the size N being finite or infinite.

Size of the sample is denoted by n. **Sampling** is the process of drawing samples from a given $population.$

Large sampling: If n≥ 30, the sampling is said to be large sampling.

Small Sampling: If n<30, the sampling is said to be small sampling.

Examples:

1. Population of India is population where as the population of Karnataka is sample.

2. Cars produced in India are the population where as the mantis cars produced in India is sample.

The statistical constants of the population such as mean (μ), Standard deviation (σ) etc are called the **parameters**. Similarly the constants for the sample drawn from the given population i.e. Mean (x) standard deviation (S) etc are called statistics.

Random sampling :

The selection of an item from the population in such a way that each has the same chance of being selected is called random sampling.

Suppose we take a sample of size n from the finite population of size N. Random sampling is a technique in which each element has an equal chance of being selected.

Sampling where each member of a population may be chosen more than once is called ***sampling with replacement*** i.e. here the items are drawn one by one and are put back to the population before the next draw. If N is the size of the finite population and n is the sample size then we have N^n samples.

Sampling where if a member cannot be chosen more than once it is called ***sampling without replacement***. Here the items are drawn one by one and are not put back to the population before the next draw. In this case there will be ${}^{N}C_{n}$ samples.

Sampling distribution :

Given a population, suppose we consider a set of samples of a certain size drawn from the population. For each sample, suppose we compute a statistics such as the mean, standard deviation etc., these statistics will vary from the sample to the other sample, suppose we group these statistics according to their frequencies and form a frequency distribution. The frequency distribution so formed is called a sampling distribution.

The standard deviation of sampling distribution is called its Standard error.

The reciprocal of the standard errors is called precision.

Sampling distribution of Means :

Consider a population for which the mean is μ and the standard deviation is σ suppose we draw a set of samples of a certain size n, from this population and find the mean $\bar{x}$ of each of these population. The frequency distribution of these means is called a sampling distribution of means. Let the mean and the standard deviation of sampling distribution of means be $\mu_{\bar{X}}$ and $\sigma_{\bar{X}}$ respectively.

Suppose the population is finite with size N or random sampling without the replacement i.e. the items drawn one by one and are not put back to the population before the next draw. In this case there will be ${}^{N}C_{n}$ samples and we have

$$\mu_{\bar{X}} = \mu \text{ and}$$

$$\sigma^2_{\bar{X}} = \frac{\sigma^2}{n} \left[\frac{N-n}{N-1}\right]$$

$$\sigma^2_{\bar{X}} = c\,\frac{\sigma^2}{n} \quad \text{where } c = \left[\frac{N-n}{N-1}\right] \text{ is called the finite population correction factor.}$$

If N is very large i.e. if the population is infinite or the sampling is finite with replacement then

$$c = 1 \text{ as } N \to \infty$$

$$\therefore\ \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$$

So, the mean of sampling distribution is equal to population mean and the corresponding standard error is $\frac{\sigma}{\sqrt{n}}$ where σ is the standard deviation of the population.

If the population is distributed normally with mean μ and S.D. σ, then the mean of all positive random samples of size n are also distributed normally with mean μ and S.E. $\frac{\sigma}{\sqrt{n}}$.

Central Limit Theorem:

This is a very important theorem regarding the distribution of the mean of a sample if the parent population is non-normal and the sample size is large.

If the variable X has a non-normal distribution with mean μ and standard deviation σ, then the limiting distribution of

Z=$\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ as $n \to \infty$,is the standard normal distribution(i.e, with mean 0 and unit S D)

There is no restriction upon the distribution of X except that it has a finite mean and variance. This theorem holds well for a sample of 30 or more which is regarded as large.

Sampling theory - 2

Statitistical Estimation is the method in which the parameters are estimated with the aid of the corresponding statistics. An estimate of the unknown true or exact value of the parameter or an interval in which the parameter is to be determined on the basis of sample data from the population.

1. **Confidence interval :**

Consider sampling distribution of a statistic S. Suppose S follows normal distribution. Let μ_s and σ_s be the mean and s.d. of the normal distribution.

a. The probability that μ_s lies in the interval (s-σ_s, s+σ_s) is 68.26%
b. The Probability that μ_s lies in the interval (s- $2\sigma_s$,s+$2\sigma_s$) is 95.44%
c. The probability that μ_s lies in the interval (s- 3σ_s , s +3σ_s) is 99.74%

The confidence intervals for μ_s indicated above are of the form

(s - $z_c\sigma_s$, s+ $z_c\sigma_s$)

Z_c=1 at 68.26% confidence level

Zc=2 at 95.44% confidence level

Z_c=3 at 99.74% confidence level

figure

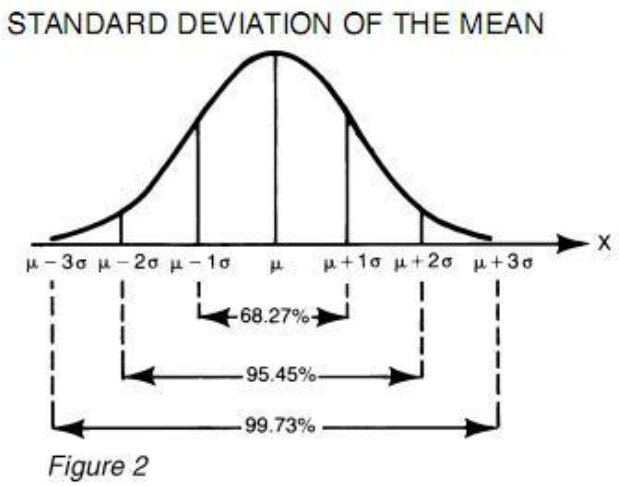

Figure 2

Percent	99.73%	99%	95.45%	95%	90%	80%	68.27%
No. of ± σ's	3.00	2.58	2.00	1.96	1.645	1.28	1.00

Z% Confidence interval :

The Z% confidence interval for μ_s if P { (s - $z_c\sigma_s$, s+ $z_c\sigma_s$) } = z%

i.e. $\frac{Z}{100}$ = P { - $z_c\sigma_s \leq \mu_s - s \leq z_c\sigma_s$ }

= P { $\left|\frac{s-\mu_s}{\sigma_s}\right| \leq z_c$ }

= P { $|Z| \leq z_c$ }, where Z = $\frac{S-\mu_s}{\sigma_s}$

then $\frac{Z}{100}$ = P { $-z_c \leq Z \leq z_c$ }

= 2 P { $0 \leq Z \leq z_c$ }

= 2 Ø (z_c)

Z = 2 Ø (z_c) × 100

Thus Z% confidence interval of μ_s is the interval (S - $z_c\sigma_s$, S+ $z_c\sigma_s$), where z_c is the positive real number.

Confidence limit : The interval (S - $z_c\sigma_s$, S+ $z_c\sigma_s$) is the Z% confidence interval for μ_s , then the quantities (S $\pm$ $z_c\sigma_s$) are called Z% confidence limits. The member $z_c$is called the corresponding confidence coefficient or the critical value confidence.

The length of the confidence interval (S - $z_c\sigma_s$, S+ $z_c\sigma_s$) ie 2l = 2 $z_c\sigma_s$ is called the error in the confidence level.

Table for the confidence coefficients Z_c for various values of Z

Z	Z_c	Z	Z_c
50	.6745	90	1.645
55	.7639	95	1.96
60	.843	95.44	2
65	.9259	96	2.05
68.26	1	97	2.195
70	1.041	98	2.33
75	1.15	99	2.58
80	1.277	99.5	2.81
85	1.445	99.74	3

The confidence interval for the population mean is $\left(\overline{X} - Z_c\frac{s}{\sqrt{N}}, \overline{X} + Z_c\frac{s}{\sqrt{N}}\right)$

1. A random sample of size N=100 is taken from a population with standard deviation $\sigma = 5.1$. Given that the sample mean is $\bar{X}$ = 21.6. Obtain the 95% confidence interval for the population mean μ

   N=100, $\sigma = 5.1$, $\overline{X}$ = 21.6

   Confidence limits for the population mean are

   $\overline{X} \pm Z_c\frac{s}{\sqrt{N}} = 21.6 \pm Z_c\frac{5.1}{\sqrt{100}}$

   For 95% confidence level , Z_c = 1.96

   $\therefore \overline{X} \pm Z_c\frac{s}{\sqrt{N}} = 21.6 \pm (1.96)\frac{5.1}{\sqrt{100}} = 21.6 \pm .9996$

Statistical Decision

Introduction:

For reaching statistical decisions, we start with some assumptions or guesses about the populations involved. Such assumptions / guesses, which may or may not be true, are called Statistical Hypotheses.

In many instances we formulate a statistical hypothesis for the sole purpose of rejecting or nullifying it. Such Hypotheses are called Null hypotheses and are generally denoted by H_0 .

A statistical hypothesis which differs from a given hypothesis is called an alternative hypothesis. A hypothesis alternative to the null hypothesis is generally denoted by H_1 .

Example: i) *Suppose we wish to reach the decision that a certain coin is biased (that is, the coin shows more heads tails or vice versa). To reach this decision, we start with the hypothesis that the coin is faired (not biased) with the sole purpose of rejecting it (at the end). This hypothesis is a null hypothesis.*

ii) *Consider the situation where the probability of an event is, say, 1/3, according to some hypothesis. For arriving at some decision, if we make the hypothesis that the probability is, say, ¼, then the hypothesis we have made is an alternative hypothesis.*

Tests of hypothesis and significance:

Procedures which enable us to decide whether to accept or reject a hypothesis or to determine whether observed samples differ significantly from expected results are called tests of hypothesis, tests of significance, or rules of decision**.**

By an error of judgement, suppose we reject a hypothesis, when it should have been accepted. Then we say that an error of Type I have been made. On the other hand, suppose we accept a hypothesis when it should be rejected; in this case, we say that an error of Type II has been made.

Type I error:

In a hypothesis test, a type I error occurs when the null hypothesis is rejected when it is in fact true; that is, H_0 is wrongly rejected.

For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug; i.e.

H_0: there is no difference between the two drugs on average.

A type I error would occur if we concluded that the two drugs produced different effects when in fact there was no difference between them.

The following table gives a summary of possible results of any hypothesis test:

		Decision	
		Reject H0	**Don't reject H0**
Truth	**H0**	Type I Error	Right decision
	H1	Right decision	Type II Error

A type I error is often considered to be more serious, and therefore it is more important to avoid than a type II error. The hypothesis test procedure is therefore adjusted so that there is a guaranteed 'low' probability of rejecting the null hypothesis wrongly; this probability is never 0. This probability of a type I error can be precisely computed as

P(type I error) = significance level = α

The exact probability of a type II error is generally unknown.

If we do not reject the null hypothesis, it may still be false (a type II error) as the sample may not be big enough to identify the falseness of the null hypothesis (especially if the truth is very close to hypothesis).

For any given set of data, type I and type II errors are inversely related; the smaller the risk of one, the higher the risk of the other.

A type I error can also be referred to as an error of the first kind.

Type II Error:

In a hypothesis test, a type II error occurs when the null hypothesis H0, is not rejected when it is in fact false. For example, in a clinical trial of a new drug, the null hypothesis might be that the new drug is no better, on average, than the current drug; i.e.

H0: there is no difference between the two drugs on average.

A type II error would occur if it was concluded that the two drugs produced the same effect, i.e. there is no difference between the two drugs on average, when in fact they produced different ones.

A type II error is frequently due to sample sizes being too small.

The probability of a type II error is generally unknown, but is symbolised by β and written

P(type II error) = β

A type II error can also be referred to as an error of the second kind.

Levels of significance:

Suppose that, under a given hypothesis H, the sampling distribution of a statistic S is a normal distribution with mean μ_s and standard deviation σ_s. Then

$$z = \frac{S-\mu_s}{\sigma_s} \text{ ---------- (1)}$$

is the standard normal variate associated with S, so that for the distribution of Z the mean is zero and the standard deviation is 1.

Accordingly, for the distribution of z, the z% confidence interval is ($-z_c$,z_c). This means that we can be Z% confident that, if the hypothesis H is true, then the value of z will lie between $-z_c$ and z_c . This is equivalent to saying that there is (100 – z)% chance that the hypothesis H is true but the value of Z lies outside the interval ($-z_c$, z_c). If we reject the hypothesis H on the grounds that the value of Z lies outside the interval ($-z_c$,z_c), we would be making a type I error and the probability of making the error is (100-Z)%. Here, we say that the hypothesis is rejected at a (100 - Z)% level of significance. Thus, a level of significance is the probability level below which we reject a hypothesis.

In practice, only two levels of significance are employed: one corresponding to Z = 95 and the other corresponding to Z= 99.

The value of the normal variate Z, determined by using the formula (1) is usually called the z – score of the statistic S. It is this score that determines the "fate" of a hypothesis H and is called the test statistic.

Rule of decision:

"Reject a hypothesis H at a (100 – Z)% level of significance if the z – score of the statistic S, determined on the basis of H, is outside the interval ($-z_c$, z_c). Do not reject the hypothesis otherwise".

Here the interval ($-z_c$,z_c) is called the interval of test.

Critical Region: The region in which a sample value falling is rejected, is known as critical region.

Normally, the test statistic, we consider follows normal distribution. Let us look into the normal curve.

Sampling Distribution

Suppose we want to study a large no. of individuals or items to make some conclusions about them, so it becomes practically impossible to examine every individual or the entire group (K/n as population).

Therefore, we prefer to examine a small part of this population K/n as a sample with the motive of drawing some conclusion about the entire population based on the information/result revealed by the sample.

This entire process K/n as statistical inference aims at in ascertaining max informa^n about the popula^n with min effort & time. [eg poll prediction is a good eg for statistical inference]

1) Population (universe) - A large collection of numerical data or individuals. [The size of popula^n is N]

2) Sample - A finite subset of the ~~universe~~ population.

[The no. of individuals in a sample is called sample size

If n = sample size is less than or equal to 30

ie If $n \leq 30$, the sample is said to be small

If $n > 30$, the sample is a large sample.

3) Sampling - The process of selecting a sample from the popula^n.

4) Random sampling - If the process of selection of an item from the pop^n is done in such a way that each has the same chance of being selected.

* n = Sample size
N = finite population size, $\Rightarrow$ possible samples = ^NC_n

Random sampling is a technique in which each of the ^NC_n sample has an equal chance of being selected.

5) Sampling with replacement – Member of the population may be selected more than once.

6) Sampling without replacement – Member cannot be chosen more than once.

* Statistic – this word is often used for the random variable or for its values.

Sampling Distributions –

Suppose we have different samples of size 'n' drawn from a population.

→ For each & every sample of size 'n' we compute mean, standard deviation etc. (obviously, they will not be the same)

Suppose we group these characteristics according to their frequencies, the frequency distributions so generated are called sampling distributions of mean and std deviation respect.

(These can be distinguished as sampling distribution of mean, standard deviation etc. The sampling distribⁿ of large samples is assumed to be a normal distribution.)

→ Standard Error – The standard deviation of the sampling distribution is called the standard error (S.E.)

Thus, S.E of the sampling distribⁿ of means is called S.E. of means.

(The S.E. is used to assess the difference b/w the expected and observed values.)

→ The reciprocal of the S.E. is called Precision.

→ If $n \geq 30$, a sample is called large otherwise small.

We consider/discuss all possible random sample of size 'n' (Popula^n size = N) and determine the mean of each one of these samples.

↳ For two cases we discuss the sampling distribution of the sample means → i) with ii) without } replacements.

Case 1 Random sampling with replacement

Here, items are drawn one by one & are put back to the pop^n before the next draw.

Popula^n size = N, sample size = n ⇒ we have N^n samples.

$\mu_{\bar{x}}$ = mean of the freq distribu^n of the sample means

$$\boxed{\mu_{\bar{x}} = \mu}$$

μ = popula^n mean

$\sigma^2_{\bar{x}}$ = Variance of the freq distribu^n of the sample means

$$\boxed{\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}}$$

σ^2 = variance of pop^n

$\sigma_{\bar{x}}$ = S.E. of the means.

Case 2 Random sampling without replacement

Here, items are drawn one by one and are not put back to the popula^n before the next draw.

Popula^n size = N, sample size = n ⇒ we have ^NC_n samples.

$$\boxed{\mu_{\bar{x}} = \mu}$$

$$\sigma^2_{\bar{x}} = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n}$$

$$\boxed{\sigma^2_{\bar{x}} = C\,\frac{\sigma^2}{n}}$$

where $C = \frac{N-n}{N-1}$ is called the finite popula^n correction factor

* If N is very large, then C is closer to 1. ($\because \lim_{N\to\infty} \frac{N-n}{N-1} = 1$)

Q1 Population $\{1,2,3\}$ i.e. $N=3$. Form the sampling distribution of the sample means in the case of (1) random samples of size 2 with replacement (2) random samples of size 2 without replacement.

Sol^n $N=3$, $\mu = \frac{1+2+3}{3} = 2$

$$\sigma^2 = \frac{1}{3}\{(1-2)^2 + (2-2)^2 + (3-2)^2\} = \frac{2}{3}$$

Case i) $n=2$ (with replacement)

$(1,1), (1,2), (1,3); (2,1), (2,2), (2,3); (3,1), (3,2), (3,3)$

These are $N^n = 3^2 = 9$ samples.

Their respective means are: $\frac{(1+1)}{2}, \frac{(1+2)}{2}, \frac{(1+3)}{2}, \frac{(2+1)}{2}, \frac{(2+2)}{2}, \frac{(2+3)}{2}, \frac{(3+1)}{2}, \frac{(3+2)}{2}, \frac{(3+3)}{2}$

$1, 1.5, 2, 1.5, 2, 2.5, 2, 2.5, 3$

The freq distribution of these means where x is the variate and f is the frequency.

x	1	1.5	2	2.5	3
f	1	2	3	2	1

$$\mu_{\bar{x}} = \frac{\Sigma fx}{\Sigma f} = \frac{1+3+6+5+3}{9} = 2 \ (= \mu)$$

$$\sigma_{\bar{x}}^2 = \frac{\Sigma f(x-\mu_{\bar{x}})^2}{\Sigma f} = \frac{1}{9}\{1(1-2)^2 + 2(1.5-2)^2 + 3(2-2)^2 + 2(2.5-2)^2 + 1(3-2)^2\}$$

$$= \frac{1}{9}\{1 + 0.5 + 0 + 0.5 + 1\} = \frac{1}{3} \left(= \frac{\sigma^2}{n} = \frac{2/3}{2} = \frac{1}{3}\right)$$

Case ii) $n=2$ (without replacement)

We have ${}^3C_2 = 3$ samples. $(1,2), (2,3), (3,1)$

Associated means are $1.5, 2.5, 2$

$$\mu_{\bar{x}} = \frac{1.5+2.5+2}{3} = 2 \quad (=\mu)$$

$$\sigma_{\bar{x}}^2 = \frac{1}{3}\{(1.5-2)^2 + (2.5-2)^2 + (2-2)^2\} = \frac{0.5}{3} = \frac{1}{6}$$

Note: Here $\left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n} = \left[\frac{3-2}{3-1}\right]\frac{2/3}{2} = \frac{1}{6} = \sigma_{\bar{x}}^2$.

★ i.e. $\sigma_{\bar{x}}^2 = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n}$

Suppose $N=500$ (large), $n=5$ (small) then $C = \frac{N-n}{N-1} = \frac{500-5}{500-1} = \frac{495}{499} = 0.992 \approx 1$.

Q2. Certain tubes manufactured by a company have mean life time of 800 hours and S.D. of 60 hours. Find the prob that a random sample of 16 tubes taken from the group will have a mean life time

a) b/w 790 hrs and 810 hrs
(b) less than 785 hrs
c) more than 820 hrs
d) b/w 770 hrs and 830 hrs.

Sol^n - $\mu = 800$, $\sigma = 60$, $n = 16$

$\therefore \sigma_{\bar{x}}^2 = \frac{\sigma^2}{n} = \frac{3600}{16} \Rightarrow \sigma_{\bar{x}} = \frac{60}{4} = 15$

We assume that the popula^n is a normal popula^n & hence the sampling distribu^n of means is also taken to be distributed normally.

The Std normal variate $z = \frac{x - \mu}{\sigma}$ in the case of sampling distribu^n of means is equivalent form $z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} \Rightarrow z = \frac{\bar{x} - 800}{15} \rightarrow (\because \mu = \mu_{\bar{x}})$

a) $P(790 < \bar{x} < 810)$ -

If $\bar{x} = 790 \Rightarrow z = -0.67$
If $\bar{x} = 810 \Rightarrow z = 0.67$
$\therefore P(790 < \bar{x} < 810) = P(-0.67 < z < 0.67)$

$\therefore P(-0.67 < z < 0.67) = 2P(0 < z < 0.67)$
$= 2\phi(0.67)$
$= 2(0.2486) = 0.4972$

b) $P(\bar{x} < 785)$ -

If $\bar{x} = 785 \Rightarrow z = -1$ $\therefore P(\bar{x} < 785) = P(z < -1)$
$= P(z > 1)$
$= P(z > 0) - P(0 < z < 1)$
$= 0.5 - \phi(1) = 0.1587$

c) $P(\bar{x} > 820)$ -

If $\bar{x} = 820 \Rightarrow z = 1.33$ $\therefore P(\bar{x} > 820) = P(z > 1.33)$
$= P(z > 0) - P(0 < z < 1.33)$
$= 0.5 - \phi(1.33) = 0.0918.$

d) $P(770 < \bar{x} < 830)$ -

If $\bar{x} = 770 \Rightarrow z = -2$ $\therefore P(-2 < z < 2) = 2P(0 < z < 2)$
If $\bar{x} = 830 \Rightarrow z = 2$ $= 2\phi(2) = 0.9544$

2. *A population consists of 4 numbers 3, 7, 11, 15.*

(a) *Find the mean and S.D of the sampling distribution of means by considering samplings of size 2 with replacement.*

(b) *If N,n denotes respectively the population size and sample size,* σ *and* $\sigma_{\bar{x}}$ *respectively denotes population S.D and S.D of the sampling distribution of means without replacement verify that*

(i) $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}\left[\frac{N-n}{N-1}\right]$

(ii) $\mu_{\bar{x}} = \mu$ *where* $\mu_{\bar{x}}$ *is the mean of this distribution and* μ *is the population mean.*

» Population mean $\mu = \frac{1}{4}(3+7+11+15) = 9$

Population variance $\sigma^2 = \frac{1}{4}\left\{(3-9)^2+(7-9)^2+(11-9)^2+(15-9)^2\right\} = 20$

Thus $\mu = 9$ **and** $\sigma = \sqrt{20}$

(a) Let us consider samples of size 2 with replacement. They are as follows.

(3, 3) (3, 7) (3, 11) (3, 15)

(7, 3) (7, 7) (7, 11) (7, 15)

(11, 3) (11, 7) (11, 11) (11, 15)

(15, 3) (15, 7) (15, 11) (15, 15)

Sampling means are as follows.

(3, 5, 7, 9); (5, 7, 9, 11); (7, 9, 11, 13): (9, 11, 13, 15)

The frequency distribution of the sampling means is as follows.

x	3	5	7	9	11	13	15
f	1	2	3	4	3	2	1

$$\mu_{\bar{x}} = \frac{\Sigma f x}{\Sigma f} = \frac{144}{16} = 9$$

$$\sigma_{\bar{x}}^2 = \frac{\Sigma f x^2}{\Sigma f} - [\mu_{\bar{x}}]^2 = \frac{1456}{16} - (9)^2 = 10$$

Thus $\mu_{\bar{x}} = 9$ **and** $\sigma_{\bar{x}} = \sqrt{10}$

Remark : $\mu_{\bar{x}} = \mu$ *and* $\sigma_{\bar{x}}^2 = \sigma^2/n$ *where* $\sigma^2 = 20$, $n = 2$

(b) Let us consider samples without replacement. They are as follows.

$(3, 7)\ (3, 11)\ (3, 15)\ (7, 11)\ (7, 15)\ (11, 15)$

The sampling means are $5, 7, 9, 9, 11, 13$

$$\therefore \qquad \mu_{\bar{x}} = \frac{1}{6}(5+7+9+9+11+13) = 9 \ ; \text{Thus } \mu_{\bar{x}} = \mu$$

$$\sigma^2_{\bar{x}} = \frac{1}{6}\left\{(5-9)^2+(7-9)^2+\cdots+(13-9)^2\right\} = \frac{40}{6} = \frac{20}{3}$$

Consider $\sigma^2_{\bar{x}} = \dfrac{\sigma^2}{n}\left[\dfrac{N-n}{N-1}\right]$

$$\text{RHS} = \frac{20}{2}\left[\frac{4-2}{4-1}\right] = 10 \times \frac{2}{3} = \frac{20}{3} = \sigma^2_{\bar{x}} = \text{LHS}$$

3. *The weights of 1500 ball bearings are normally distributed with a mean of 635 gms. and S.D of 1.36 gms. If 300 random samples of size 36 are drawn from this population, determine the expected mean and S.D of the sampling distribution of means if sampling is done*

(a) *with replacement* ***(b)*** *without replacement.*

>> Here $N = 1500$, $\mu = 635$, $\sigma = 1.36$, $n = 36$

(a) Expected mean $\mu_{\bar{x}} = \mu = \mathbf{635}$

Expected S.D $\sigma_{\bar{x}} = \sqrt{\sigma^2/n} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{1.36}{6} = \mathbf{0.227}$

(b) Expected mean $\mu_{\bar{x}} = \mu = 635$

Expected variance $\sigma_{\bar{x}}^2 = \dfrac{\sigma^2}{n}\left[\dfrac{N-n}{N-1}\right] = \dfrac{(1.36)^2}{36}\left[\dfrac{1500-36}{1500-1}\right] = 0.05$

Thus $\sigma_{\bar{x}} = \sqrt{0.05} = \mathbf{0.224}$

4. *Consider the data as in the previous problem. In the case of random sampling with replacement find how many random samples would have their mean* ***(a)*** *between 634.76 gms and 635.24 gms* ***(b)*** *greater than 635.5 gms* ***(c)*** *less than 634.2 gms* ***(d)*** *less than 634.5 gms or more than 635.24 gms.*

>> We assume that the population is a normal population and hence the sampling distribution of means is also taken to be distributed normally.

The standard normal variate $z = \dfrac{x-\mu}{\sigma}$ in the case of sampling distribution of means is in the equivalent form

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} \text{ where we have } \mu_{\bar{x}} = 635,\ \sigma_{\bar{x}} = 0.227$$

Hence we have $z = \dfrac{\bar{x} - 635}{0.227}$ ... (1)

(a) Probability of a sample having mean between 634.76 and 635.24 is represented by $P(634.76 < \bar{x} < 635.24)$

Now from (1), if $\bar{x} = 634.76$, $z = \dfrac{-0.24}{0.227} = -1.06$

if $\bar{x} = 635.24$, $z = \dfrac{0.24}{0.227} = 1.06$

Hence we have to find $P(-1.06 < z < 1.06)$

ie., $= 2P(0 < z < 1.06)$

$= 2\phi(1.06) = 2(0.3554)$ by using tables.

$= 0.7108$

Thus we have corresponding to 300 samples, the expected number of samples having their mean between 634.76 gms and 635.24 gms is given by

$300 \times 0.7108 = 213.24 \approx$ **213 samples**

(b) To find $P(\bar{x} > 635.5) \times 300$

If $\bar{x} = 635.5$ then $z = \dfrac{635.5 - 635}{0.227}$ from (1). That is $z = 2.203$

$\therefore$ $P(z > 2.203) = P(z > 0) - P(0 < z < 2.203)$

$= 0.5 - \phi(2.2)$

$= 0.5 - 0.4861 = 0.0139$

Thus $P(\bar{x} > 635.5) \times 300 = 4.17 \approx$ **4 samples**

(c) To find $P(\bar{x} < 634.2) \times 300$

If $\bar{x} = 634.2$ then $z = \dfrac{634.2 - 635}{0.227}$ from (1). That is, $z = -3.52$

$$\therefore \quad P(z < -3.52) = P(z > 3.52)$$

$$= P(z > 0) - P(0 < z < 3.52)$$

$$= 0.5 - \phi(3.52)$$

$$= 0.5 - 0.4998 = 0.0002$$

Thus $P(\bar{x} < 634.2) \times 300 = 0.06 \approx$ **0 samples**

(d) To find $\left[P(\bar{x} < 634.5) + P(\bar{x} > 635.24) \right] \times 300$

If $\bar{x} = 634.5$ then $z = -2.2$

$\bar{x} = 635.24$ then $z = 1.06$; by using (1).

$$\therefore \quad P(z < -2.2) + P(z > 1.06)$$

$$= P(z > 2.2) + P(z > 1.06)$$

$$= \left\{ P(z > 0) - P(0 < z < 2.2) \right\} + \left\{ P(z > 0) - P(0 < z < 1.06) \right\}$$

$$= \left\{ 0.5 - \phi(2.2) \right\} + \left\{ 0.5 - \phi(1.06) \right\}$$

$$= 1 - \left\{ \phi(2.2) + \phi(1.06) \right\}$$

$$= 1 - (0.4861 + 0.3554) = 0.1585$$

Multiplying this value by the sample size 300 we get $47.55 \approx$ **48 samples.**

Student's t - distribution.

Consider a small sample of size n, drawn from a normal popn with mean μ & s.d σ. If $\bar{x}$ & σ_s be the sample mean & s.d then the statistic 't' is defined as

$$t = \frac{\bar{x} - \mu}{\sigma} \sqrt{n} \quad \text{or} \quad t = \frac{\bar{x} - \mu}{\sigma_s} \sqrt{n-1}$$

where $\nu = n-1$ denotes the d.f of t.

If we calculate t for each sample. The sampling distn for t is known as Student's 't' distribution & is given by

$$y = \frac{y_0}{(1 + t^2/\nu)^{(\nu+1)/2}}$$

where y_0 is constant s.t the area under the curve is unity.

The 't' distribution is often used in tests of hypothesis about the mean when popn s.d. σ is unknown

We compute $t = \frac{\bar{x} - \mu}{s} \sqrt{n}$ & consider $|t|$

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

If $|t| > t_{0.05}$. the diff bet^n $\bar{x}$ & μ is said to be significant at 5% los.

Chi-Square test

The magnitude of discrepancy between the theoretical & observed values is given by the qty χ^2

If $\chi^2 = 0$, theory agrees with observation completely

As the χ^2 increases, the discrepancy between the observed & theoretical frequencies increases

Defn: If $O_1, O_2 \cdots O_n$ be a set of observed frequencies & $E_1, E_2 \cdots E_n$ be the corresponding set of expected (theoretical frequencies, then χ^2 is defined by the relation

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \quad \text{with } n-1 \text{ d.f.}$$

Chi-Sq. distⁿ: If $x_1, x_2 \cdots x_n$ are n independent normal variates with 0 mean & s.d unity then it can be shown that

$x_1^2 + x_2^2 + \cdots x_n^2$ is a rnd variate with χ^2 distⁿ with n d.f.

Note: If $E_i < 10$, we group them suitably for computing the value of χ^2.

Test of Goodness of fit

It is possible to test the hypo about the association of two attributes. It is easily possible to find the theoretical frequencies from the distⁿ of fit.

Chi-square test helps us to test the goodness of fit of these distributions.

If the calculated value of χ^2 is less than the table value of χ^2 at a specified level of significance, the hypo is accepted o.w rejected.

• Degrees of freedom

The number of degrees of freedom $(d \cdot f)$ usually denoted by ν is the number of values in a set which may be assigned arbitrarily. It can be interpreted as the number of independent values generated by a sample of small size for estimating a population parameter.

Examples : Let us suppose that we need to find 3 numbers whose sum is 25. That is to find a, b, c such that $a+b+c = 25$. We can arbitrarily assign values to any *two* of the variables a, b, c and hence these are the degrees of freedom. That is to say that $d \cdot f(\nu) = 2$. If there are n observations $d \cdot f$ is equal to $(n-1)$.

Suppose that we are finding the mean of a sample of size n comprising values $x_1, x_2, \cdots x_n$. We use all the n values to compute the sample mean $\bar{x}$. Then $\bar{x}$ is said to have n degrees of freedom.

Suppose we are finding the sample variance, we use the n values
$(x_1-\bar{x})^2, (x_2-\bar{x})^2, \cdots (x_n-\bar{x})^2$.

But these values do not have n degrees of freedom as they all depend on a fixed value $\bar{x}$ which has already been computed. Hence the sample variance is said to have $(n-1)\ d \cdot f$. If we compute another statistic based on the sample mean and variance, that statistic is said to have $(n-2)\ d \cdot f$ and so on. In general the number of degrees of freedom $\nu = n-k$ where n is the number of observations in the sample and k is the number of constraints / number of values which are pre determined.

5.16 Student's t Distribution

Sir william Gosset under the pen name *'Student'* derived a theoretical distribution to test the significance of a sample mean where the small sample is drawn from a normal population.

Let $x_i\,(i = 1, 2, \cdots n)$ be a random sample of size n drawn from a normal population with mean μ and variance σ^2. The statistic t is defined as follows.

$$t = \frac{\bar{x} - \mu}{(s/\sqrt{n})} = \frac{\bar{x} - \mu}{s}\sqrt{n}$$

Here $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean.

$s^2 = \frac{1}{(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is the sample variance.

$\nu = (n-1)$ denote the number of degrees of freedom of t.

The statistic t follows the Student's t distribution with $(n-1)$ $d\cdot f$ having the probability density function

$$y = f(t) = \frac{y_0}{\left[1 + t^2/\nu\right]^{(\nu+1)/2}}$$

where y_0 is a constant such that the area under the curve is unity.

Note : 1. *Statistic t is also defined as follows.*

$$t = \frac{\bar{x} - \mu}{\sigma}\sqrt{n-1}$$

2. *The constant y_0 present in p.d.f is given by*

$$y_0 = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\;\Gamma(\nu/2)}$$ *so that the p.d.f of the Student's t distribution with ν degrees of freedom is given by*

$$y = f(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\;\Gamma(\nu/2)}\left[1 + \frac{t^2}{\nu}\right]^{-(\nu+1)/2} \quad ; -\infty < t < \infty$$

3. *If ν is large $(\nu \geq 30)$ the graph of $f(t)$ closely approximates standard normal curve. In other words we can say that t is normally distributed for large samples.*

• Student's t test for a sample mean

We need to test the hypothesis, whether the sample mean $(\bar{x})$ differs significantly from the population mean (μ) / hypothetical value (μ).

We compute $t = \dfrac{\bar{x} - \mu}{s}\sqrt{n}$ and consider $|t|$.

We also take a note of the value of t for the given $d \cdot f$ from the table of standard values.

If $|t| > t_{.05}$ the difference between $\bar{x}$ and μ is said to be significant at 5% level of significance.

If $|t| > t_{.01}$ the difference is said to be significant at 1% level of significance.

If $|t|$ is less than the table value at a certain level of significance, the data is said to be conformal / consistent with the hypothesis that μ is the mean of the population.

• Confidence limits for the population mean μ

If $t_{.05}$ is the tabulated value of t for $(n-1)\,d \cdot f$ at 5% level of significance, it implies that.

$$P[|t| > t_{.05}] = 0.05$$

$$\Rightarrow \quad P[|t| \leq t_{.05}] = 1 - 0.05 = 0.95$$

Now consider $|t| \leq t_{.05}$

i.e., $$\left|\frac{\bar{x} - \mu}{s}\sqrt{n}\right| \leq t_{.05}$$

or $$\left|\frac{\bar{x} - \mu}{(s/\sqrt{n})}\right| \leq t_{.05}$$

i.e., $$-t_{.05} \leq \frac{x - \mu}{(s/\sqrt{n})} \leq t_{.05}$$

i.e., $$\frac{-s}{\sqrt{n}}\,t_{.05} \leq \bar{x} - \mu \leq \frac{s}{\sqrt{n}}\,t_{.05}$$

$$\Rightarrow \quad \mu \leq \bar{x} + \frac{s}{\sqrt{n}}\,t_{.05} \text{ and } \bar{x} - \frac{s}{\sqrt{n}}\,t_{.05} \leq \mu$$

Combining these two results we can write in the form

$$\bar{x} - \frac{s}{\sqrt{n}}\,t_{.05} \leq \mu \leq \bar{x} + \frac{s}{\sqrt{n}}\,t_{.05}$$

Thus we have 95% confidence limits for μ given by $\bar{x} \pm \frac{s}{\sqrt{n}} t_{.05}$

Similarly 99% confidence limits for μ are given by $\bar{x} \pm \frac{s}{\sqrt{n}} t_{.01}$

Note : *Confidence limits are also called Fiducial limits.*

$$Z = \frac{0.1}{\sqrt{4/9 \times 5/9\,(1/1000 + 1/800)}} = 4.243$$

$$Z = 4.243 > \begin{cases} Z_{.05} = 1.96 \\ Z_{.01} = 2.58 \end{cases}$$

Thus the hypothesis H_0 is rejected both at 5% and 1% levels of significance.

Student's t distribution / test

31. *Find the student's t for the following variable values in a sample of eight :* $-4, -2, -2, 0, 2, 2, 3, 3$, *taking the mean of the universe to be zero.*

>> $$t = \frac{\bar{x} - \mu}{s}\sqrt{n}$$

By data $\mu = 0$ and we have $n = 8$

$$\bar{x} = \frac{1}{8}(-4-2-2+0+2+2+3+3) = \frac{1}{4} = 0.25$$

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

$$= \frac{1}{7}\{(-4.25)^2 + (-2.25)^2 + (-2.25)^2 + (-0.25)^2 + (1.75)^2 + (1.75)^2 + (2.75)^2 + (2.75)^2\}$$

$$s^2 = \frac{1}{7}(49.5) = 7.07 \qquad \therefore \quad s = 2.66$$

Thus $$t = \frac{0.25 - 0}{2.66}\sqrt{8} = \mathbf{0.266}$$

Note : *The expression for* s^2 *can also be put in the following form.*

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

$$= \frac{1}{n-1}\left\{\sum_{1}^{n}x_i^2 - 2\sum_{1}^{n}x_i \cdot \frac{1}{n}\sum_{1}^{n}x_i + n\left(\frac{1}{n}\sum x_i\right)^2\right\}$$

$$= \frac{1}{n-1}\left\{\sum_{1}^{n}x_i^2 - \frac{2}{n}(\sum x_i)^2 + \frac{1}{n}(\sum x_i)^2\right\}$$

$$\therefore \quad s^2 = \frac{1}{n-1}\left\{\sum_{1}^{n}x_i^2 - \frac{1}{n}(\sum x_i)^2\right\}$$

According to this formula we have in the given example

$$s^2 = \frac{1}{7}\left\{50 - \frac{1}{8}(2)^2\right\} = \frac{1}{7}(49.5) = 7.07$$

Remark : *We can employ this formula when* $\bar{x}$ *is not an integer.*

32. *A machine is expected to produce nails of length 3 inches. A random sample of 25 nails gave an average length of 3.1 inch with standard deviation 0.3. Can it be said that the machine is producing nails as per specification?* ($t_{0.05}$ *for 24 d.f is 2.064*)

» By data we have,

$$\mu = 3,\ \bar{x} = 3.1,\ n = 25,\ s = 0.3$$

$$t = \frac{\bar{x} - \mu}{s}\sqrt{n} = \frac{0.1}{0.3}\sqrt{25} = 1.67 < 2.064$$

Thus the **hypothesis** that the machine is producing nails as per specification is **accepted at 5% level of significance.**

33. *Ten individuals are chosen at random from a population and their heights in inches are found to be 63, 63, 66, 67, 68, 69, 70, 70, 71, 71. Test the hypothesis that the mean height of the universe is 66 inches.* ($t_{.05} = 2.262$ *for 9 d.f*)

» We have $\mu = 66,\ n = 10$

$$\bar{x} = \frac{\sum x}{n} = \frac{678}{10} = 67.8$$

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

$$s^2 = \frac{1}{9}\left[(63 - 67.8)^2 + \cdots + (71 - 67.8)^2\right] = 9.067 \quad \therefore \quad s = 3.011$$

We have $t = \dfrac{\bar{x} - \mu}{s}\sqrt{n} = \dfrac{(67.8 - 66)}{3.011}\sqrt{10} = 1.89 < 2.262$

Thus the hypothesis is accepted at 5% level of significance.

34. *A sample of 10 measurements of the diameter of a sphere gave a mean of 12cm and a standard deviation 0.15cm. Find the 95% confidence limits for the actual diameter.*

» By data $n = 10,\ \bar{x} = 12,\ s = 0.15$

Also $t_{.05}$ for $9\ d \cdot f = 2.262$

Confidence limits for the actual diameter is given by

$$\bar{x} \pm \left[\frac{s}{\sqrt{n}}\right] t_{.05} = 12 \pm \frac{0.15}{\sqrt{10}}(2.262) = 12 \pm 0.1073$$

Thus 11.893cm to 12.107cm is the confidence limits for the actual diameter.

35. *A certain stimulus administered to each of the 12 patients resulted in the following change in blood pressure.* 5, 2, 8, −1, 3, 0, 6, −2, 1, 5, 0, 4. *Can it be concluded that the stimulus will increase the blood pressure?* ($t_{.05}$ *for* 11 $d \cdot f = 2.201$)

>>
$$\bar{x} = \frac{\sum x}{n} = \frac{31}{12} = 2.5833$$

$$s^2 = \frac{1}{n-1}\sum (x-\bar{x})^2 = \frac{1}{n-1}\left\{\sum x^2 - \frac{1}{n}(\sum x)^2\right\}$$

$$s^2 = \frac{1}{11}\left\{185 - \frac{1}{12}(31)^2\right\} = 9.538 \quad \therefore \quad s = 3.088$$

We have, $t = \dfrac{\bar{x}-\mu}{s}\sqrt{n}$

Let us suppose that the stimulus administration is not accompanied with increase in blood pressure, we can take $\mu = 0$

$$\therefore \quad t = \frac{2.5833 - 0}{3.088}\sqrt{12} = 2.8979 \approx 2.9 > 2.201$$

Hence the hypothesis is rejected at 5% level of significance. We conclude with 95% confidence that the stimulus in general is accompanied with increase in blood pressure.

37. *A sample of 11 rats from a central population had an average blood viscosity of 3.92 with a standard deviation of 0.61. On the basis of this sample, establish 95% fiducial limits for* μ *the mean blood viscosity of the central population* ($t_{.05} = 2.228$ *for 10 d.f*)

>> By data $\bar{x} = 3.92$, $s = 0.61$, $n = 11$

95% fiducial limits for μ are $\bar{x} \pm \frac{s}{\sqrt{n}} t_{.05}$

i.e., $= 3.92 \pm \frac{0.61}{\sqrt{11}} (2.228)$

$= 3.92 \pm 0.41 = 3.51$ and 4.33

Thus 95% confidence limits for μ are 3.51 and 4.33.

5.17 Chi - Square distribution

Chi - Square distribution provides a measure of correspondence between the theoretical frequencies and observed frequencies.

If $0_i\ (i = 1, 2, \ldots n)$ and $E_i\ (i = 1, 2, \ldots n)$ respectively denotes a set of observed and estimated frequencies, the quantity **chi - square** denoted by χ^2 is defined as follows.

$$\chi^2 = \sum_{i=1}^{n} \frac{(0_i - E_i)^2}{E_i} \quad ; \quad \text{degrees of freedom} = n - 1$$

Note : *If the expected frequencies are less than 10, we group them suitably for computing the value of chi. square.*

• Chi - Square test as a test of goodness of fit

It is possible to test the hypothesis about the association of two attributes. We have already discussed the fitting of Binomial distribution, Normal distribution, Poisson distribution to a given data. It is easily possible to find the theoretical frequencies from the distribution of fit.

Chi - Square test helps us to test the goodness of fit of these distributions.

If the calculated value of χ^2 is less than the table value of χ^2 at a specified level of significance the hypothesis is accepted, otherwise the hypothesis is rejected.

Chi-Square distribution

41. *A die is thrown 264 times and the number appearing on the face (x) follows the following frequency distribution.*

x	1	2	3	4	5	6
f	40	32	28	58	54	60

Calculate the value of χ^2.

>> The frequencies in the given data are the observed frequencies. Assuming that the dice is unbiassed the expected number of frequencies for the numbers 1, 2, 3, 4, 5, 6 to appear on the face is $\frac{264}{6} = 44$ each. Now the data is as follows :

No. on the dice	1	2	3	4	5	6
Observed frequency (0_i)	40	32	28	58	54	60
Expected frequency (E_i)	44	44	44	44	44	44

$$\chi^2 = \sum \frac{(0_i - E_i)^2}{E_i}$$

$$= \frac{(40-44)^2}{44} + \frac{(32-44)^2}{44} + \cdots + \frac{(60-44)^2}{44}$$

$$= \frac{1}{44}\left[16 + 144 + 256 + 196 + 100 + 256\right] = \frac{968}{44} = 22$$

Thus $\chi^2 = 22$

---

42. *Five dice were thrown 96 times and the numbers 1, 2 or 3 appearing on the face of the dice follows the frequency distribution as below.*

No. of dice showing 1,2 or 3	5	4	3	2	1	0
Frequency	7	19	35	24	8	3

Test the hypothesis that the data follows a binomial distribution. ($\chi^2_{0.05} = 11.07$ *for 5 d.f*)

>> The data gives the observed frequencies and we need to calculate the expected frequencies.

Probability of a single dice throwing 1, 2 or 3 is $p = 3/6 = 1/2 \quad \therefore \quad q = 1 - p = 1/2$

The binomial distribution of fit is, $N(q+p)^n = 96\left(1/2 + 1/2\right)^5$

The theoretical frequencies of getting 5, 4, 3, 2, 1, 0 successes with 5 dice are respectively the successive terms of the binomial expansion.

They are respectively $96 \times \frac{1}{2^5}, 96 \times 5C_1 \times \frac{1}{2^5}, \ldots 96 \times \frac{1}{2^5}$ or 3, 15, 30, 30, 15, 3.

We have the table of observed and expected frequencies.

0_i	7	19	35	24	8	3
E_i	3	15	30	30	15	3

$$\chi^2 = \Sigma \frac{(0_i - E_i)^2}{E_i}$$

$$= \frac{16}{3} + \frac{16}{15} + \frac{25}{30} + \frac{36}{30} + \frac{49}{15} + \frac{0}{3} = 11.7$$

$$\chi^2 = 11.7 > \chi^2_{0.05} = 11.07$$

Thus the hypothesis that the data follows a binomial distribution is rejected.

---

43. *A sample analysis of examination results of 500 students was made. It was found that 220 students had failed, 170 had secured third class 90 had secured second class and 20 had secured first class. Do these figures support the general examination result which is in the ratio*

4 : 3 : 2:1 for the respective categories ($\chi^2_{0.05} = 7.81$ *for 3 d.f*)

>> Let us take the hypothesis that these figures support to the general result in the ratio 4 : 3 : 2 : 1.

The expected frequencies in the respective category are

$$\frac{4}{10} \times 500, \frac{3}{10} \times 500, \frac{2}{10} \times 500, \frac{1}{10} \times 500 \quad \text{or} \quad 200, 150, 100, 50.$$

We have the following table.

0_i	220	170	90	20
E_i	200	150	100	50

$$\chi^2 = \Sigma \frac{(0_i - E_i)^2}{E_i}$$

$$= \frac{400}{200} + \frac{400}{150} + \frac{100}{100} + \frac{900}{50}$$

$$\chi^2 = 23.67 > \chi^2_{0.05} = 7.81$$

Thus the hypothesis is rejected.

$N = 1000$

$n = 4$

$4C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{4-x}$

$\frac{4}{24}$ $4C_x$.