

Testing of Hypothesis

Hypothesis

Concise, testable statement or belief about the the parameter(s) of a probability distribution.

- “IQ is independent of gender”
- “People who listen to music on iPods is more likely to have premature hearing loss”
- “Depression is related to increased fast food consumption”

Hypothesis testing:

Using sample statistics, test and draw conclusions about the statement or belief.

- Can never be 100% sure about accuracy of our findings
- Use probability to determine whether it is *likely* that we are correct or incorrect
- For example, our statistical analyses might tell us that there is only a 5% probability of getting the result by chance

Simple Hypothesis

A hypothesis that **completely specify the underlying distribution** is called a simple hypothesis. For example

$$\begin{aligned} \mu &= \mu_0, & \text{for Normal distribution, if } \sigma^2 \text{ is known,} \\ \sigma^2 &= \sigma_0^2, & \text{for Normal distribution, if } \mu \text{ is known,} \\ \lambda &= \lambda_0, & \text{for Poisson distribution, etc.} \end{aligned}$$

Composite Hypothesis

A hypothesis that **does not specify the population distribution completely** is known as a composite hypothesis. For example,

$\mu = \mu_0$, for Normal distribution if both μ and σ^2 are unknown,
 $\mu \neq \mu_0$, for Normal distribution,
 $\lambda > 1$, for Poisson distribution, etc.

Null Hypothesis

The null hypothesis, H_0 , is often a **default proposition** based on previous experience or knowledge, about the value(s) of population parameter(s) – proportion (p) or mean (μ) or standard deviation (σ). The null hypothesis is generally a proposition of “**no difference**” with a given **reference** value or the apparent difference, if any, is due to chance only. The null hypothesis is generally **assumed to be true** until evidence indicates otherwise. Outcome of hypothesis testing is thus either **reject H_0** , or **fail to reject H_0** .

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. We would write

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

We give special consideration to the null hypothesis. This is due to the fact that the null hypothesis relates to the statement of *status quo*, whereas the alternative hypothesis relates to the statement to be accepted if / when the null is rejected.

When writing the Null Hypothesis, make sure it includes an “=” symbol. It may look like one of the following:

- $H_0 : \mu = 40$
- $H_0 : \mu \leq 40$
- $H_0 : \mu \geq 40$

Alternative Hypothesis

The alternative hypothesis, H_1 , is what researchers believes to be true or hopes to prove true instead of H_0 . The alternative hypothesis is a **claim** of “a difference in the population parameters”, and the researcher/analyst tries to establish evidence **against** the null hypothesis based on sample observations.

If one is conducting a study and want to use a hypothesis test to support his claim, the claim must be so worded that it becomes the alternative hypothesis. So that by virtue of rejecting the null hypothesis, one can establish his claim, the alternative hypothesis.

For example, in the clinical trial of new drug, the **research hypothesis** might be that the new drug has a different effect, on average, compared to that of the current drug. We would write the alternate hypothesis in that case as:

H_1 : the new drug, on the average, have different effect.

The alternative hypothesis might also be that the new drug is better, on average, than the current drug. In this case we would write

H_1 : the new drug, on the average, is better than the existing drug.

When writing the Alternate Hypothesis, make sure it **should never include an “=”** symbol. It should look similar to one of the following:

- $H_1 : \mu \neq 40$ [Nondirectional (**two-tailed**) alternate]
- $H_1 : \mu > 40$ [Directional (**right-tailed, greater than type**) alternate]
- $H_1 : \mu < 40$ [Directional (**left-tailed, less than type**) alternate]

Nonstatistical Hypothesis Testing

A criminal trial is an example of hypothesis testing without a statistic. In a trial a jury must decide between following two hypotheses. The null hypothesis is

H_0 : The accused is innocent

The alternative hypothesis is

H_1 : The accused is guilty

The jury does not know which hypothesis is true. They must make a decision on the basis of evidence presented.

In the language of statistics convicting the accused is called *rejecting the null hypothesis in favor of the alternative hypothesis*. That is, the jury is saying that there is enough evidence to conclude that the accused is guilty (i.e., there is enough evidence to support the alternative hypothesis).

If the jury acquits the accused, it is stating that *there is not enough evidence to support the alternative hypothesis and consequently fail to reject the null hypothesis*. Notice that the jury is not saying that the accused is innocent, only concludes that there is not enough evidence to support the alternative hypothesis. That is why we never say that we accept the null hypothesis.

Statistical Hypothesis Testing

Similarly, statistical hypothesis testing works by collecting data and measuring the likelihood of the particular set of data, assuming that the null hypothesis is true. So, the final conclusion, once the test has been carried out, is always given in terms of the null hypothesis. We either "Reject H_0 in favour of H_1 " (**strong decision**) or "Fail to reject H_0 " (**weak decision**).

Remember that, decision/conclusion can never be "Accept H_1 ", or even "Reject H_1 ".

If we conclude "**fail to reject H_0** ", this does not necessarily mean that the null hypothesis is true, it only suggests that there is not sufficient evidence, based on the observed data, for us to prefer H_1 over H_0 . Similarly, rejecting the null hypothesis, suggests that the alternative hypothesis *may* be true.

Note: *It is important to keep in mind that the null and alternative hypotheses are statement about **population parameters**, not statements about the **sample estimators**.*

Null and Alternative Hypothesis: Applications

1. Testing Research Hypotheses

The research hypothesis should be expressed as the alternative hypothesis.

The conclusion that the research hypothesis is **true** comes from sample data that **contradicts the null hypothesis**.

2. Testing the validity of a claim

Manufacturers' claims are usually given the benefit of the doubt and stated as the null hypothesis.

The conclusion that the claim is **false/true** comes from sample data that **contradicts/accepts** the null hypothesis.

3. Testing in Decision-Making Situations

A decision maker might have to choose between two courses of action, one associated with the null hypothesis and another associated with the alternative hypothesis.

Example: Accepting a shipment of goods from a supplier or returning the shipment of goods to the supplier.

Test Statistic

A statistic is function of sample data values. To begin with we assume that the (null) hypothesis about the population parameter is true. We compare the value of the statistic with the hypothetical value of the parameter. If the difference between them is small, the hypothesis is accepted and if the difference between them is large, the hypothesis is rejected. A statistic on which the decision can be based whether to accept or reject a hypothesis is called test statistic. It may be noted that test statistic is also a random variable.

Acceptance and Rejection Region

All possible values which a test-statistic T may assume, can be divided into two mutually exclusive groups: (i) first group consisting of values which appear to be consistent with the null hypothesis, and (ii) second group having values which are unlikely to occur if H_0 is true. The first group is called the acceptance region (\bar{C}) and the second set of values forms the rejection or critical region (C) for the null hypothesis. The value(s) that separates the critical region from the acceptance region is called the critical value(s). The critical value that can be in the same units as that of parameter or in the standardized units, is to be decided by the experimenter keeping in view the degree of confidence he /she is willing to have in the null hypothesis.

Errors in Test of Hypothesis

There are two kinds of errors that can be made in test of hypothesis: (1) incorrectly rejecting a true null hypothesis and (2) incorrectly retaining/accepting a false null hypothesis. The former error is called a Type I error and the latter error is called a Type II error. These two types of errors are defined in the table.

Statistical Decision based on sample observation ↓	True state of the Null Hypothesis	
	H_0 True	H_0 False
Reject H_0	Type I error [False Positive]	Correct
Do not Reject H_0	Correct	Type II error [False Negative]

If one concludes that there is a real difference between the means when, in fact, there is not, this is called a false positive. If one instead concludes that there is no real difference between the means when, in fact, there is a difference, this is called a false negative.

The probability of a Type I error is denoted by the Greek letter alpha (α) and is called the Type I error rate, whereas the probability of a Type II error is denoted by the Greek letter beta (β), i.e.

Probability of Type I Error = $P(\text{rejecting a TRUE } H_0) = P(T \in C | H_0 \text{ is true}) = \alpha$

Probability of Type II Error = $P(\text{failing to reject a FALSE } H_0)$
 $= P(T \in \bar{C} | H_0 \text{ is false}) = \beta$

The probability of type I error, α is known as the **level of significance** of the test. A researcher who accepts this error decides to reject the **true** null hypothesis. Test of hypothesis starts with the assumption that the null hypothesis is true and intends to minimize making a Type I error (taking a clue from criminal trial). However, there is always some probability of committing this type of error, so researchers directly control the probability of a Type I error by stating an alpha (α) level. Critical region vis-à-vis acceptance region depends upon the stated value of α and thus also known as '**size of the critical region**'. Since the experimenter can choose a value of α that is relatively small, so rejecting a null hypothesis is considered as a **strong decision**.

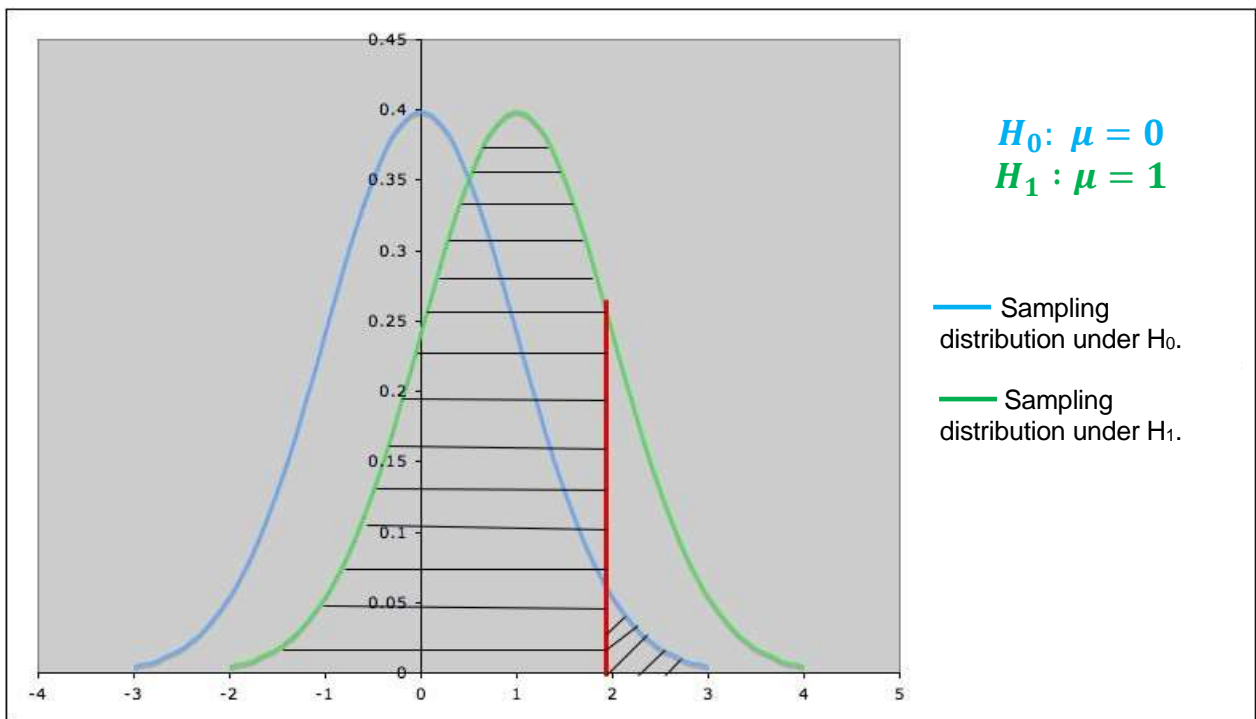
A Type II error is only an error in the sense that an **opportunity to reject the incorrect null hypothesis was lost**. In this decision, we decide to retain null hypotheses that are in fact false. But, we can always go back and conduct more studies and hence termed a **weak decision**.

Traditionally, researchers have long been more concerned about making type I error and have conventionally fixed the probability of Type I Error ≤ 0.05 . Whereas Type II error has received less attention from the researchers. Type II error are generally the result of too few sample size. Consequently in hypothesis testing, we try to control the probability of a Type I error (α) only and assume that probability of Type II error will be taken care of by large sample size.

Relationship between Type I and Type II Error

The following diagram illustrates the Type I error and the Type II error against the **specific alternative hypothesis** " $\mu = 1$ " in a hypothesis test for a population mean μ , with null hypothesis " $\mu = 0$," alternate hypothesis " $\mu > 0$ ", and significance level $\alpha = 0.05$.

- The **blue** (leftmost) curve is the sampling distribution assuming the null hypothesis " $\mu = 0$ ".
- The **green** (rightmost) curve is the sampling distribution assuming the specific alternate hypothesis " $\mu = 1$ ".
- The vertical **red** line shows the *critical value* for rejection of the null hypothesis: the null hypothesis is rejected for values of the test statistic to the *right* of the **red** line (and *not* rejected for values to the *left* of the **red** line).
- The area of the diagonally hatched region to the *right* of the **red** line and under the **blue** curve is the probability of type I error (α) [2.5%]
- The area of the horizontally hatched region to the *left* of the **red** line and under the **green** curve is the probability of Type II error (β).



Power of a Test

The probability of rejecting H_0 when it is false is known as the “power of test”, i.e. power of test for a parameter θ is

$$\begin{aligned}\gamma(\theta) &= P(\text{rejecting } H_0 | H_0 \text{ is false}) \\ &= P(T \in C | H_0 \text{ is false}) \\ &= 1 - P(T \in \bar{C} | H_0 \text{ is false}) \\ &= 1 - \beta\end{aligned}$$

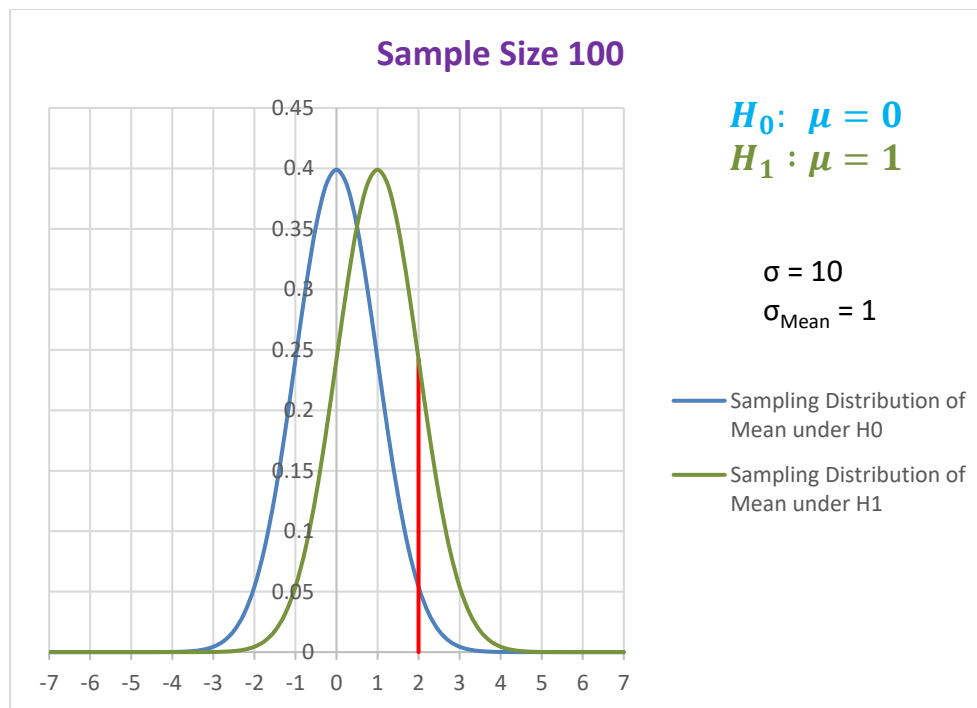
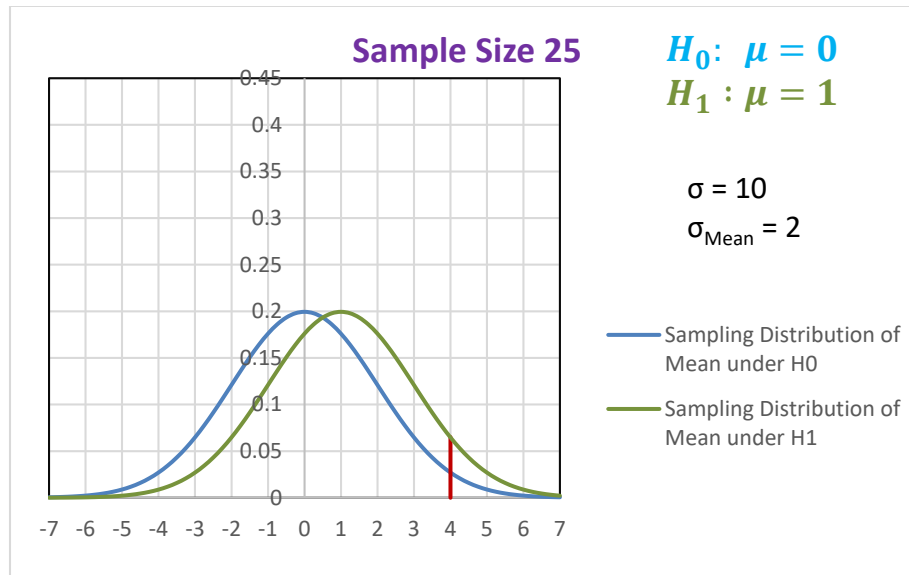
In the previous illustration, the region to the right of the red line under the green curve represents the power of the test.

Power of a test can be increased by –

- a) **increase in effect size**, where effect size = true value – hypothesized value, i.e. larger the difference between true value and hypothesized value higher will be the power,
- b) **increasing α** , or
- c) **increasing sample size**.

Effect of Sample Size on Power of Test

The pictures in the next page show the sampling distribution for the mean under the null hypothesis $\mu = 0$ together with the sampling distribution under a specific alternate hypothesis $\mu = 1$, but *for different sample sizes*.



- The first picture is for sample size $n = 25$; the second picture is for sample size $n = 100$.
- Note that both graphs are in the same scale. In both pictures, the blue curve is centered at 0 (corresponding to the null hypothesis) and the green curve is centered at 1 (corresponding to the alternative hypothesis).

- In each picture, the red line is the **critical value** for rejection with $\alpha = 0.05$ (for a one-tailed test) -- that is, in each picture, the area under the *blue* curve to the right of the red line is 0.05.
- In each picture, the area under the *green* curve to the right of the red line is the power of the test against the alternate $\mu = 1$. Note that this area is *larger* in the second picture (the one with larger sample size) than in the first picture.
- Thus, **larger sample size gives larger power**. Larger sample size increases the power by reducing the standard error.

Most Powerful Test (MPT)

Consider the test of the **simple** null hypothesis $H_0: \theta = \theta_0$ against the **simple** alternative hypothesis $H_1: \theta = \theta_1$. Let C and C_1 be two critical regions of size α , that is,

$$P(C; \theta_0) = \alpha \text{ and } P(C_1; \theta_0) = \alpha .$$

Critical region C is said to be the most powerful critical region of size α if, for every other critical region C_1 of size α , we have:

$$P(C; \theta_1) \geq P(C_1; \theta_1) \Rightarrow \text{Power}(C) \geq \text{Power}(C_1)$$

i.e. C is said to be the most powerful critical region of size α if the power of C is at least as great as the power of every other critical region C_1 of size α .

Any test based on this critical region C is called the most powerful test of level of significance α with respect to the alternate hypothesis.

Uniformly Most Powerful Test (UMPT)

Consider the test of the **simple** null hypothesis $H_0: \theta = \theta_0$ against the **composite** alternative hypothesis $H_1: \theta \neq \theta_0$. Let C and C_1 be two critical regions of size α , that is, let

$$P(C; \theta_0) = \alpha \text{ and } P(C_1; \theta_0) = \alpha$$

C is said to be the uniformly most powerful critical region of size α if, for every other critical region C_1 of size α , we have:

$$P(C; \theta) \geq P(C_1; \theta), \forall \theta \neq \theta_0 \Rightarrow \text{Power}(C) \geq \text{Power}(C_1), \forall \theta \neq \theta_0$$

that is, C is the best critical region of size α if the power of C is at least as great as the power of every other critical region C_1 of size α for any alternative $\theta \neq \theta_0$.

The resulting test is said to be **uniformly most powerful** with respect to the composite alternate hypothesis. **In other words, UMPT is a test that is simultaneously most powerful for all alternatives of interest in an experiment.**

Unbiased Test

A test is said to be **unbiased** when the probability of rejecting the null hypothesis, when it is true, is less than or equal to α , and the probability of rejecting the null hypothesis, when the alternative hypothesis is true, is greater than or equal to α , i.e. **$\text{prob}(\text{type I error}) \leq \alpha$ and $\text{power of the test} \geq \alpha$.**

Uniformly Most Powerful Unbiased Test (UMPUT)

Consider the test of the **simple** null hypothesis $H_0: \theta = \theta_0$ against the **composite** alternative hypothesis $H_1: \theta \neq \theta_0$. Let C and C_1 be two unbiased critical regions of size α , that is

$$P(C; \theta_0) = \alpha \text{ and } P(C; \theta) \geq \alpha, \forall \theta \neq \theta_0,$$

$$P(C_1; \theta_0) = \alpha \text{ and } P(C_1; \theta) \geq \alpha, \forall \theta \neq \theta_0.$$

Now, C is said to be the uniformly most powerful among unbiased critical regions of size α if, for every other critical region C_1 of size α , we have:

$$P(C; \theta) \geq P(C_1; \theta), \forall \theta \neq \theta_0$$

Any test based on this critical region C is known as UMPUT of level α against the composite alternate. **In other words, UMPUT is an unbiased test that is simultaneously most powerful for all alternatives of interest in an experiment.**

The Logic of Hypothesis Testing

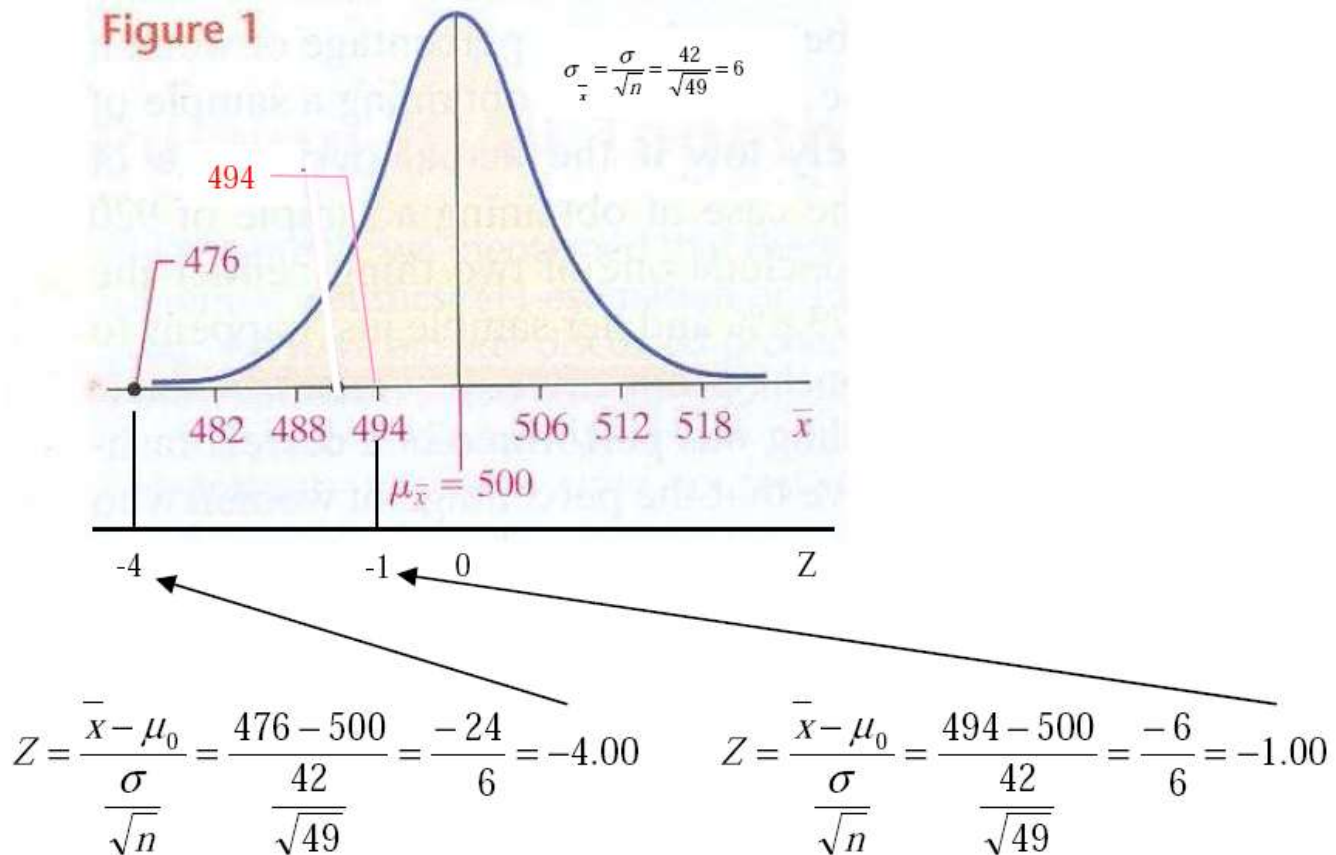
Problem

The packaging on a light bulb states that the bulb will last 500 hours under normal use. A consumer advocate would like to know if the mean lifetime of a bulb is less than 500 hours and accordingly the hypothesis to be tested will be framed as

$$H_0: \mu \geq 500 \text{ hours versus } H_1: \mu < 500 \text{ hours}$$

A random sample of 49 light bulbs (i.e. $n = 49$) is burned to determine how long a light bulb lasts **on the average**. Assume we know the population standard deviation is $\sigma = 42$.

Figure 1



If $\bar{x} = 494$, then the sample mean is one standard error (standard deviation of sample mean) below 500.

- $P(\bar{x} = 494) = 0.1587$, which would happen 16% of the time under H_0 .
- In this case, we may not reject H_0 as the probability of such occurrence is quite high.
- Note: We would only reject H_0 in the event of obtaining an “unusual” sample (i.e., **a sample that occurs with low probability under the null hypothesis**).

If $\bar{x} = 476$, then the sample mean is four standard error below 500.

- $P(\bar{x} = 476) = 0.0$, which essentially says there is almost no chance of finding a sample of mean 476 when H_0 is true.
- In this case, we would reject the null hypothesis.
- In this case, we are inclined to believe that the sample has come from a population, whose mean is less than 500.

Thus, we reject the null hypothesis if the sample mean is “too many” standard error (standard deviation of sample mean) away from the null hypothesis (H_0).

Or, stated another way, we reject the null hypothesis if the sample data result in a statistic, which is \bar{x} in this case, that is unlikely under the assumption that the null hypothesis is true.

Principles of Hypothesis Testing

Suppose, we wish to test the hypothesis

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

Where μ_0 is a specified constant. Appropriate test statistic, in such a case, is

$$Z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

where \bar{x} is the sample mean based on the random sample x_1, x_2, \dots, x_n of size n .

If the null hypothesis is true, $E(\bar{x}) = \mu_0$ and it follows that the distribution of Z_0 is $N(0,1)$. Consequently, if $H_0: \mu = \mu_0$ is true, the probability that the test statistic Z_0 will fall between $-z_{\alpha/2}$ and $z_{\alpha/2}$ is $1 - \alpha$. Hence, under null hypothesis, probability that the test statistic Z_0 will fall in the region $z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$ is α .

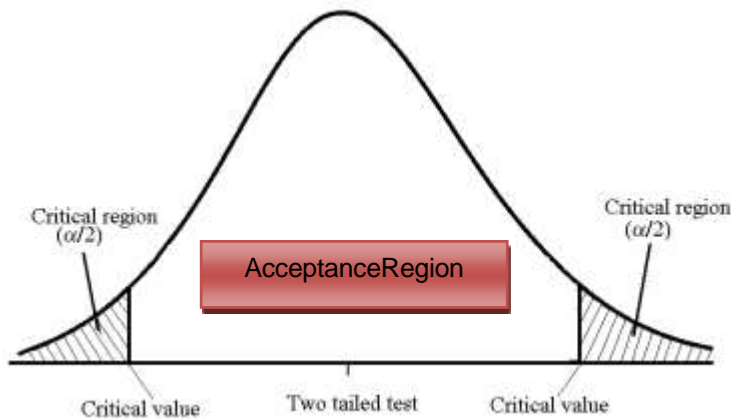
Clearly, a sample producing a value of the test statistic that falls in tails of the distribution of Z_0 would be less probable, if $H_0: \mu = \mu_0$ is true; therefore, it is an indication that H_0 is false. Thus, we should reject H_0 if the observed value of the test statistic z_0 is either

$$z_0 > z_{\alpha/2} \text{ or } z_0 < -z_{\alpha/2}$$

and we should fail to reject H_0 if

$$-z_{\alpha/2} \leq z_0 \leq z_{\alpha/2}.$$

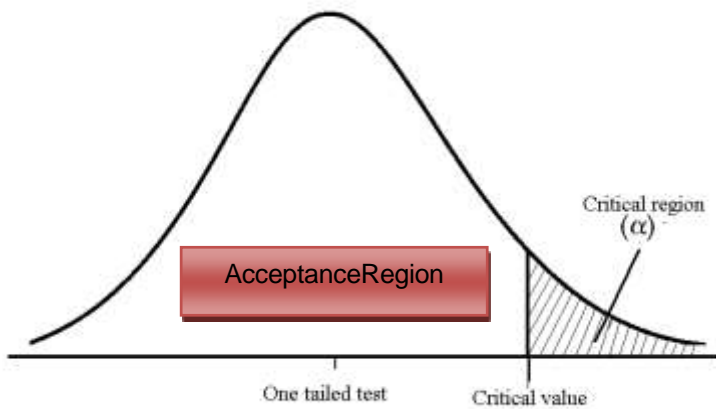
The region bounded by $-z_{\alpha/2}$ and $z_{\alpha/2}$ is the **region of acceptance**, whereas the region defined by $z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$ is the **critical** or **rejection region** for this test.



Two tailed/sided test:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu \neq \mu_0$$

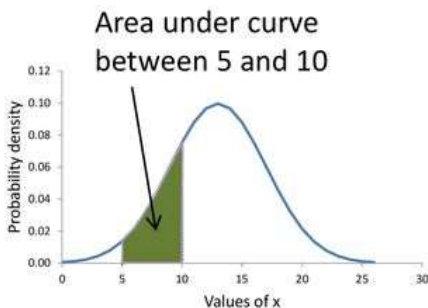


Right tailed/sided test:

$$H_0 : \mu = \mu_0$$

$$H_1 : \mu > \mu_0$$

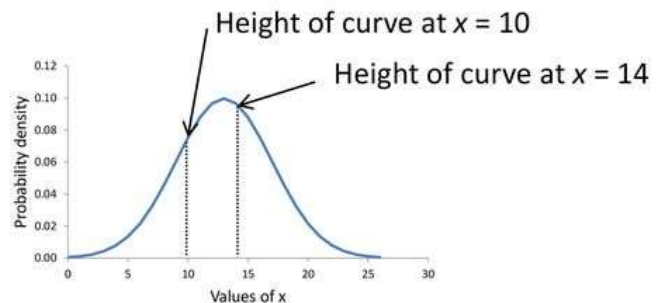
Probability



What is the *probability* that $5 \leq x \leq 10$ given a normal distribution with $\mu = 13$ and $\sigma = 4$? Answer: 0.204

What is the *probability* that $-1000 \leq x \leq 1000$ given a normal distribution with $\mu = 13$ and $\sigma = 4$? Answer: 1.000

Likelihood



What is the *likelihood* that $\mu = 13$ and $\sigma = 4$ if you observed a value of

- (a) $x = 10$ (answer: the *likelihood* is 0.075)
- (b) $x = 14$ (answer: the *likelihood* is 0.097)

Conclusion: if the observed value was 14, it is *more likely* that the parameters are $\mu = 13$ and $\sigma = 4$, because 0.097 is higher than 0.075.

Probability refers to the chance of an event, in the sample space, to occur, i.e. possibility of the event to happen given the sample distribution.

Likelihood measures how likely a sample of data follow a statistical model with given values of the unknown parameters. Likelihood function describes a hypersurface whose peak, if exists, represent the combination of model parameter values that maximize the likelihood of the sample data.

Neyman-Pearson Lemma

Let $f(x_1, x_2, \dots, x_n | \theta)$ be the joint p.d.f / p.m.f of X_1, X_2, \dots, X_n then the function of θ defined by:

$$L(\theta) = L(\theta | x_1, x_2, \dots, x_n) = f(x_1, x_2, \dots, x_n | \theta)$$

is known as the *likelihood function*.

If we compare the likelihood function at two parameter points and find that $L(\theta_1 | x) > L(\theta_2 | x)$, then the sample we actually observed is more likely to have occurred under $\theta = \theta_1$ than under $\theta = \theta_2$. This can be interpreted as θ_1 is a more plausible value for θ than θ_2 . Similarly, if we find $L(\theta_1 | x) < L(\theta_2 | x)$, then $\theta = \theta_2$ is more likely.

If there exists a critical region C of size α and a **non-negative** constant K such that

$$\frac{L(\theta_0 | x_1, x_2, \dots, x_n)}{L(\theta_1 | x_1, x_2, \dots, x_n)} < K, \text{ for } (x_1, x_2, \dots, x_n) \in C, \text{ and}$$

[i.e. **reject H_0**]

$$\frac{L(\theta_0 | x_1, x_2, \dots, x_n)}{L(\theta_1 | x_1, x_2, \dots, x_n)} \geq K, \text{ for } (x_1, x_2, \dots, x_n) \in \bar{C}.$$

[i.e. **fail to reject H_0**]

- If X falls in the critical region likelihood of θ_1 will be large compared to that of θ_0 and hence the ratio will be small (less or equal to some constant K).
- Reverse is true if X falls in the acceptance region.

then C is most powerful (MP) critical region of size α for testing $H_0: \theta = \theta_0$ against $H_1: \theta = \theta_1$.

If X_1, X_2, \dots, X_n is identically and independently distributed, then the ratio can be written as

$$\frac{L(\theta_0 | x_1, x_2, \dots, x_n)}{L(\theta_1 | x_1, x_2, \dots, x_n)} = \frac{\prod_{i=1}^n f(x_i | \theta_0)}{\prod_{i=1}^n f(x_i | \theta_1)}$$

Neyman-Pearson Lemma suggests a simple method of test construction that is closely related to the maximum-likelihood method of estimation.

Likelihood Ratio Test

Suppose that $L(\theta|x_1, x_2, \dots, x_n)$ is the likelihood function of θ corresponding to set of values x_1, x_2, \dots, x_n of random variables X_1, X_2, \dots, X_n . Suppose the hypothesis to be tested be:

$$H_0: \theta = \theta_0$$

$$H_1: \theta \neq \theta_0$$

Parameter Space: $\Theta = \{\theta: 0 < \theta < \infty\}$

Parameter Space under H_0 : $\Theta_0 = \{\theta: \theta = \theta_0\}$

Parameter Space under H_1 : $\Theta_1 = \{\theta: \theta \neq \theta_0\}$

If the null hypothesis is true, we would expect the likelihood

$$L(\theta_0|\mathbf{x}) = \prod_{i=1}^n f(x_i|\theta_0)$$

will be relatively large.

The above likelihood is compared against the maximum of the likelihood over the entire parameter space, i.e.

$$L(\hat{\theta}|\mathbf{x}) = \max_{\theta \in \Theta} \prod_{i=1}^n f(x_i|\theta), \text{ where } \hat{\theta} \text{ is the m.l.e. of } \theta.$$

However, we can not reject the null hypothesis unless $L(\theta_0|\mathbf{x})$ is much smaller than $L(\hat{\theta}|\mathbf{x})$. The **likelihood ratio test** is based on the likelihood ratio

$$\lambda = \frac{L(\theta_0|\mathbf{x})}{L(\hat{\theta}|\mathbf{x})} = \frac{\prod_{i=1}^n f(x_i|\theta_0)}{\max_{\theta \in \Theta} \prod_{i=1}^n f(x_i|\theta)}$$

and basis for rejection of the null hypothesis $H_0: \theta = \theta_0$ is the critical region $C = \{\mathbf{x} | \lambda(x_1, x_2, \dots, x_n) < \lambda_0\}$ ($0 < \lambda_0 < 1$). Note that, for rejection of the null hypothesis the likelihood ratio λ has to be small (refer NP Lemma).

For **composite null hypothesis**, the above likelihood ratio approach is slightly modified as the exact value of all the parameter(s), under null hypothesis, is not known. Consequently, we take the maximum likelihood estimates of such unknown parameter(s) over Θ_0 . So, the **generalized likelihood ratio test statistic** is defined as

$$\lambda = \frac{\max_{\theta \in \Theta_0} L(\theta|\mathbf{x})}{\max_{\theta \in \Theta} L(\theta|\mathbf{x})} = \frac{\max_{\theta \in \Theta_0} \prod_{i=1}^n f(x_i|\theta)}{\max_{\theta \in \Theta} \prod_{i=1}^n f(x_i|\theta)}$$

The constant λ_0 so chosen as to make the probability of Type I Error associated with the test equals to α , i.e. we find C and λ_0 such that $P(C, H_0) = \alpha$.

Example 1

Let x_1, x_2, \dots, x_n be a random sample from a normal distribution with **unknown mean** μ and **known variance** σ^2 . Define the critical regions for the following test

- i) $H_0 : \mu = \mu_0; H_1 : \mu \neq \mu_0$
- ii) $H_0 : \mu = \mu_0; H_1 : \mu > \mu_0$
- iii) $H_0 : \mu = \mu_0; H_1 : \mu < \mu_0$

Solution

$$L(\mu_0, \sigma^2 | x_1, x_2, \dots, x_n) = \frac{1}{(\sigma\sqrt{2\pi})^n} \cdot \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu_0)^2 \right\}$$

- i) Since under $H_1 : \mu \neq \mu_0$, MLE for μ is \bar{x} , hence we have,

$$L(\hat{\mu}, \sigma^2 | x_1, x_2, \dots, x_n) = \frac{1}{(\sigma\sqrt{2\pi})^n} \cdot \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \bar{x})^2 \right\}$$

Thus, $\lambda = \exp \left[-\frac{1}{2\sigma^2} \left\{ \sum_{i=1}^n (x_i - \mu_0)^2 - \sum_{i=1}^n (x_i - \bar{x})^2 \right\} \right]$

$$= \exp \left\{ -\frac{n}{2\sigma^2} (\bar{x} - \mu_0)^2 \right\}$$

Now, $\lambda < \lambda_0 \Rightarrow \exp \left\{ -\frac{n}{2\sigma^2} (\bar{x} - \mu_0)^2 \right\} < \lambda_0$

$$\Rightarrow -\frac{n}{2\sigma^2} (\bar{x} - \mu_0)^2 < \ln(\lambda_0),$$

$$\Rightarrow \left| \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \right| > -\sqrt{2\ln(\lambda_0)} = \lambda_1, \text{ say}$$

Hence, the likelihood ratio test has the critical region $C = \left\{ X : \left| \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \right| > \lambda_1 \right\}$, where λ_1 is such that $P(C; \theta_0) = \alpha$, i.e.

$$\begin{aligned} P\left(\left|\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right| > \lambda_1; \theta_0\right) &= \alpha. \\ \Rightarrow P\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -\lambda_1 \text{ or } \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > \lambda_1; \theta_0\right) &= \alpha \end{aligned}$$

Now, under $H_0: \mu = \mu_0$, the statistic $\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ has the $N(0,1)$ distribution. Since the distribution is symmetric about zero, λ_1 must be the upper $\alpha/2$ percent point of the distribution, i.e. $Z_{\alpha/2}$. We therefore have the critical region as \bar{x}

$$C = \left\{ X : \left| \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \right| > Z_{\alpha/2} \right\}.$$

ii) For alternative $H_1: \mu > \mu_0$, it is highly likely that $\bar{x} - \mu_0 > 0$.

$$\begin{aligned} \lambda(x_1, x_2, \dots, x_n) < \lambda_0 &\Rightarrow \exp\left\{-\frac{n}{2\sigma^2}(\bar{x} - \mu_0)^2\right\} < \lambda_0 \\ &\Rightarrow -\frac{n}{2\sigma^2}(\bar{x} - \mu_0)^2 < \ln(\lambda_0) \\ &\Rightarrow \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > -\sqrt{2\ln(\lambda_0)} = \lambda_1, \text{ say} \end{aligned}$$

So, the likelihood ratio test has the critical region $C = \left\{ X : \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > \lambda_1 \right\}$, where λ_1 is such that $P\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > \lambda_1; \theta_0\right) = \alpha$. Since, $\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$ has the $N(0,1)$ distribution, λ_1 must be equal to Z_α . Thus $C = \left\{ X : \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} > Z_\alpha \right\}$.

iii) In this case the alternate hypothesis is $H_1: \mu < \mu_0$ and hence $\bar{x} - \mu_0 < 0$.

So, the critical region will be

$$C = \left\{ X : \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} < -Z_\alpha \right\}.$$

Example 2

Suppose x_1, x_2, \dots, x_n be a random sample from a normal distribution, where **both mean μ and variance σ^2 are unknown**. Consider the problem of testing the null hypothesis $H_0 : \mu = \mu_0$ against the alternative i) $H_1 : \mu \neq \mu_0$, ii) $H_1 : \mu > \mu_0$ and iii) $H_1 : \mu < \mu_0$. Find out the critical regions of testing against the three alternatives.

Solution:

In this problem parameter spaces are

$$\Theta = \{(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}^+\} \text{ and}$$

$$\Theta_0 = \{(\mu, \sigma_0^2) : \mu = \mu_0, \sigma_0^2 \in \mathbb{R}^+\}.$$

The likelihood function under null hypothesis is:

$$L(\mu_0, \sigma_0^2 | x_1, x_2, \dots, x_n) = \frac{1}{(\sigma_0 \sqrt{2\pi})^n} \cdot \exp \left\{ -\frac{1}{2\sigma_0^2} \sum_{i=1}^n (x_i - \mu_0)^2 \right\}$$

So, the log likelihood function is

$$\ln L = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma_0^2) - \frac{1}{2\sigma_0^2} \sum (x_i - \mu_0)^2$$

Differentiating above w.r.t. σ_0^2 and equating the same to zero, we get

$$\hat{\sigma}_0^2 = \frac{1}{n} \sum (x_i - \mu_0)^2.$$

Thus, the likelihood function under null hypothesis reduces to

$$L(\mu_0, \hat{\sigma}_0^2 | \mathbf{x}) = (2\pi\hat{\sigma}_0^2)^{-\frac{n}{2}} \exp\left\{-\frac{n}{2}\right\}$$

Again **the likelihood function** over the entire parameter space Θ is:

$$L(\mu, \sigma^2 | x_1, x_2, \dots, x_n) = \frac{1}{(\sigma\sqrt{2\pi})^n} \cdot \exp\left\{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2\right\}$$

Similarly, maximizing above likelihood over Θ , by taking $\hat{\mu} = \bar{x}$, we get

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

Therefore, maximum of the likelihood over Θ is

$$L(\hat{\mu}, \hat{\sigma}^2 | \mathbf{x}) = (2\pi\hat{\sigma}^2)^{-\frac{n}{2}} \exp\left\{-\frac{n}{2}\right\}$$

Consequently, the likelihood ratio statistic is

$$\lambda(x_1, x_2, \dots, x_n) = \frac{L(\mu_0, \hat{\sigma}_0^2 | x_1, x_2, \dots, x_n)}{L(\hat{\mu}, \hat{\sigma}^2 | x_1, x_2, \dots, x_n)} = \left(\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right)^{n/2} = \left(\frac{\sum (x_i - \bar{x})^2}{\sum (x_i - \mu_0)^2}\right)^{n/2}$$

$$\begin{aligned} &= \left[\frac{\sum (x_i - \bar{x})^2}{\sum (x_i - \bar{x})^2 + n(\bar{x} - \mu_0)^2} \right]^{n/2} = \left[\frac{(n-1)s^2}{(n-1)s^2 + n(\bar{x} - \mu_0)^2} \right]^{n/2} \\ &= \left[\frac{n-1}{(n-1) + \frac{n(\bar{x} - \mu_0)^2}{s^2}} \right]^{n/2} \end{aligned}$$

$$\begin{aligned}
\text{Now, } \lambda(x_1, x_2, \dots, x_n) < \lambda_0 &\Rightarrow \left[\frac{n-1}{(n-1) + \frac{n(\bar{x} - \mu_0)^2}{s^2}} \right]^{n/2} < \lambda_0 \\
&\Rightarrow \frac{n-1}{(n-1) + \frac{n(\bar{x} - \mu_0)^2}{s^2}} < \lambda_1 \left[\lambda_1 = \lambda_0^{\frac{2}{n}} \right] \\
&\Rightarrow \frac{(n-1) + \frac{n(\bar{x} - \mu_0)^2}{s^2}}{(n-1)} > \lambda_2 \left[\lambda_2 = \frac{1}{\lambda_1} \right] \\
&\Rightarrow 1 + \frac{\frac{n(\bar{x} - \mu_0)^2}{s^2}}{(n-1)} > \lambda_2 \\
&\Rightarrow \frac{(\bar{x} - \mu_0)^2}{\frac{s^2}{n}} > \lambda_3 \quad [\lambda_3 = (n-1) \times (\lambda_2 - 1)]
\end{aligned}$$

i) **Alternate** $H_1: \mu \neq \mu_0$.

So, in this case we must have

$$\frac{|\bar{x} - \mu_0|}{s/\sqrt{n}} > \lambda_3.$$

The constant λ_3 is to be so determine that $P\left(\frac{|\bar{x} - \mu_0|}{s/\sqrt{n}} > \lambda_3; H_0\right) = \alpha$. Hence, λ_3 must be $t_{\alpha/2, n-1}$, since under H_0 , $\frac{|\bar{x} - \mu_0|}{s/\sqrt{n}} \cap t_{n-1}$.

Hence, the critical region for testing $H_0: \mu = \mu_0$ against $H_1: \mu \neq \mu_0$ is

$$C = \left\{ X : \frac{|\bar{x} - \mu_0|}{s/\sqrt{n}} > t_{\alpha/2, n-1} \right\}$$

Critical regions for other two alternatives can be similarly obtained.

Statement of Central Limit Theorem

The central limit theorem states that given a distribution with a mean μ and variance σ^2 , the sampling distribution of the mean approaches a normal distribution with mean (μ) and variance σ^2/n as n , the sample size, increases.

Statement :

If x_1, x_2, \dots, x_n is a random sample of size n taken from a population (finite or infinite) with mean μ and variance σ^2 , and if \bar{x} is the sample mean, then distribution of

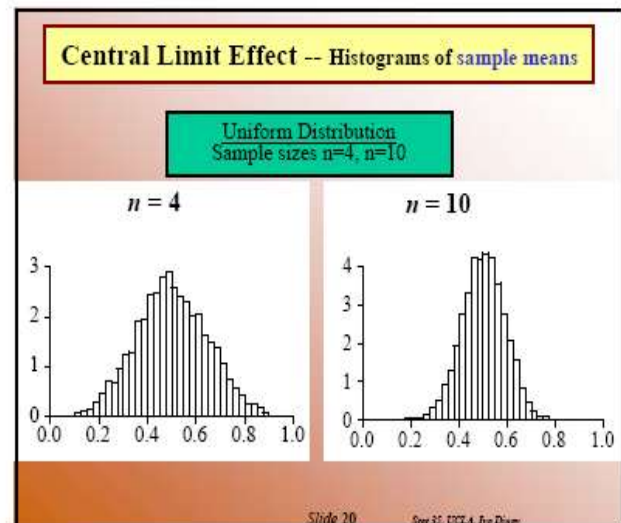
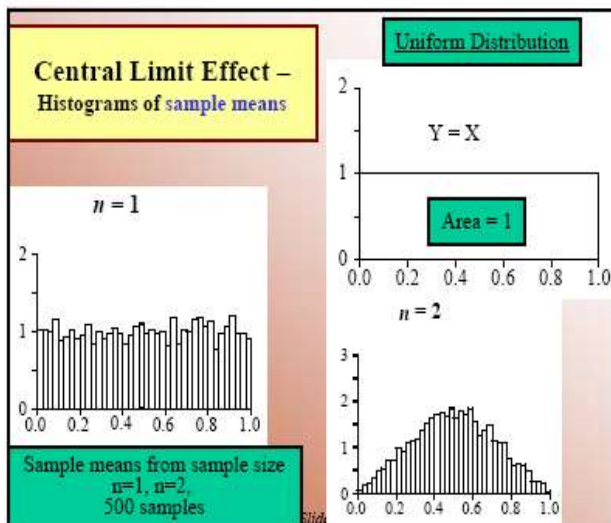
$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

approaches to “standard normal distribution $N(0,1)$ as $n \rightarrow \infty$.

The amazing and counter-intuitive thing about the central limit theorem is that **no matter what the shape of the original distribution, the sampling distribution of the mean approaches a normal distribution**. Furthermore, for most distributions, a normal distribution is approached very quickly as n increases. **Keep in mind that n is the sample size for each mean and not the number of such samples.**

Remember in a sampling distribution the number of samples is assumed to be infinite. The sample size is the number of observations in each sample; it is the number of scores that goes into the computation of each mean.

Below are shown the resulting frequency distributions each based on 500 means. For $n = 2$, 2 scores were sampled from a uniform distribution 500 times and the mean computed each time. The same method was followed with means of 4 scores for $n = 4$ and 10 scores for $n = 10$. Below are shown the resulting frequency distributions each based on 500 means.



Two things should be noted about the effect of increasing n :

1. The distributions becomes more and more normal.
2. The spread of the distributions decreases.

Determination of sample Size

When estimating the population mean using an independent and identically distributed (*iid*) sample of size n , where data values has variance σ^2 , the standard error of the sample mean is σ/\sqrt{n} .

This expression describes quantitatively how the estimate becomes more precise as the sample size increases. Using the central limit theorem to justify approximating the sample mean with a normal distribution yields an approximate 95% confidence interval of the form $(\bar{x} - 2\sigma/\sqrt{n}, \bar{x} + 2\sigma/\sqrt{n})$. For 99% confidence interval, the z or t value will be 2.58, i.e. interval would be $(\bar{x} - 2.58\sigma/\sqrt{n}, \bar{x} + 2.58\sigma/\sqrt{n})$.

If we wish to have a 95% confidence interval that is W units in width, known as Margin of Error, we would solve

$$4\sigma/\sqrt{n} = W$$

for n , resulting the sample size $n = \left(\frac{4\sigma}{W}\right)^2$.

For example, suppose we would like to estimate the average amount of drug that lowers a subject's blood pressure, with 95% confidence and a margin of error of six units. Further, it is known that the standard deviation of blood pressure in the population is 15, then the required sample size is 100.

Finding the boundary values of the acceptance region of a test

Two-sided Test

- Upper boundary = $\mu_0 + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ and lower boundary = $\mu_0 - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
- Upper boundary = $\mu_0 + t_{\alpha/2} \frac{s}{\sqrt{n}}$ and lower boundary = $\mu_0 - t_{\alpha/2} \frac{s}{\sqrt{n}}$

One-sided test

- For $H_1: \mu > \mu_0$, upper boundary = $\mu_0 + Z_{\alpha} \frac{\sigma}{\sqrt{n}}$ or $\mu_0 + t_{\alpha} \frac{s}{\sqrt{n}}$
- For $H_1: \mu < \mu_0$, lower boundary = $\mu_0 - Z_{\alpha} \frac{\sigma}{\sqrt{n}}$ or $\mu_0 - t_{\alpha} \frac{s}{\sqrt{n}}$

Finding the probability of type II error, known variance case [β]

Two-sided Test

We can find the probability of type II error as:

$$\beta = P(\bar{C}_{LB} \leq \bar{x} \leq \bar{C}_{UB} | \mu \neq \mu_0, \text{ i.e. under } H_1),$$

where \bar{C}_{LB} and \bar{C}_{UB} respectively are the lower and upper boundary of the acceptance region for the test corresponding to the null hypothesis, H_0 .

One-sided Test

- For $H_1: \mu > \mu_0$, $\beta = P(\bar{x} \leq \bar{C}_{UB} | \mu > \mu_0, \text{ i.e. under } H_1)$
- For $H_1: \mu < \mu_0$, $\beta = P(\bar{C}_{LB} \leq \bar{x} | \mu < \mu_0, \text{ i.e. under } H_1)$

\bar{C}_{LB} and \bar{C}_{UB} are as stated earlier.

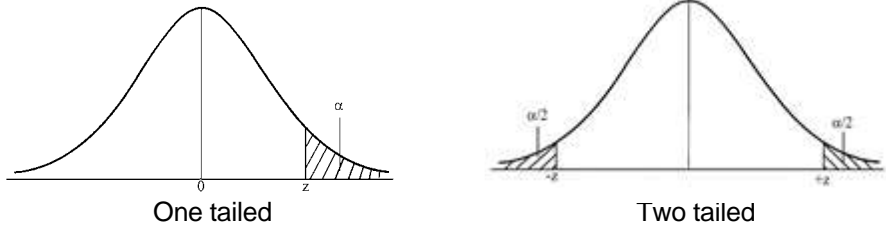
Testing for the Mean of a Population

The methods described in this section require the population to be measured on the interval or ratio scales and to be normally distributed.

Before deciding which test to use we have to ask whether the sample size is large or small and whether the population standard deviation is known or has to be estimated from the sample. The methods used are basically the same but different formulae and tables need to be used.

Choice of statistical test:

Sample size	Population standard deviation	
	Known	Unknown
	Standard error = $\frac{\sigma}{\sqrt{n}}$	Standard error = $\frac{s}{\sqrt{n}}$
Large (≥ 40)	z-test	z -test
Small	z -test	t-test

PERCENTAGE POINTS OF THE STANDARD NORMAL CURVE						
						
One tail	5%	2.5%	1%	0.5%	0.1%	0.05%
Two tails	10%	5%	2%	1%	0.2%	0.1%
Z	1.645	1.96	2.33	2.576	3.09	3.29

Above table gives, for the same Z-value what would be the level of significance (α) for one tail and two tail tests.

PERCENTAGE POINTS OF THE t-DISTRIBUTION

df	One-Tail = .4 Two-Tail = .8	.25 .5	.1 .2	.05 .1	.025 .05	.01 .02	.005 .01	.0025 .005	.001 .002	.0005 .001
1	0.325	1.000	3.078	6.314	12.706	31.821	63.657	127.32	318.31	636.62
2	0.289	0.816	1.886	2.920	4.303	6.965	9.925	14.089	22.327	31.598
3	0.277	0.765	1.638	2.353	3.182	4.541	5.841	7.453	10.214	12.924
4	0.271	0.741	1.533	2.132	2.776	3.747	4.604	5.598	7.173	8.610
5	0.267	0.727	1.476	2.015	2.571	3.365	4.032	4.773	5.893	6.869
6	0.265	0.718	1.440	1.943	2.447	3.143	3.707	4.317	5.208	5.959
7	0.263	0.711	1.415	1.895	2.365	2.998	3.499	4.029	4.785	5.408
8	0.262	0.706	1.397	1.860	2.306	2.896	3.355	3.833	4.501	5.041
9	0.261	0.703	1.383	1.833	2.262	2.821	3.250	3.690	4.297	4.781
10	0.260	0.700	1.372	1.812	2.228	2.764	3.169	3.581	4.144	4.587
11	0.260	0.697	1.363	1.796	2.201	2.718	3.106	3.497	4.025	4.437
12	0.259	0.695	1.356	1.782	2.179	2.681	3.055	3.428	3.930	4.318
13	0.259	0.694	1.350	1.771	2.160	2.650	3.012	3.372	3.852	4.221
14	0.258	0.692	1.345	1.761	2.145	2.624	2.977	3.326	3.787	4.140
15	0.258	0.691	1.341	1.753	2.131	2.602	2.947	3.286	3.733	4.073
16	0.258	0.690	1.337	1.746	2.120	2.583	2.921	3.252	3.686	4.015
17	0.257	0.689	1.333	1.740	2.110	2.567	2.898	3.222	3.646	3.965
18	0.257	0.688	1.330	1.734	2.101	2.552	2.878	3.197	3.610	3.922
19	0.257	0.688	1.328	1.729	2.093	2.539	2.861	3.174	3.579	3.883
20	0.257	0.687	1.325	1.725	2.086	2.528	2.845	3.153	3.552	3.850
21	0.257	0.686	1.323	1.721	2.080	2.518	2.831	3.135	3.527	3.819
22	0.256	0.686	1.321	1.717	2.074	2.508	2.819	3.119	3.505	3.792
23	0.256	0.685	1.319	1.714	2.069	2.500	2.807	3.104	3.485	3.767
24	0.256	0.685	1.318	1.711	2.064	2.492	2.797	3.091	3.467	3.745
25	0.256	0.684	1.316	1.708	2.060	2.485	2.787	3.078	3.450	3.725
26	0.256	0.684	1.315	1.706	2.056	2.479	2.779	3.067	3.435	3.707
27	0.256	0.684	1.314	1.703	2.052	2.473	2.771	3.057	3.421	3.690
28	0.256	0.683	1.313	1.701	2.048	2.467	2.763	3.047	3.408	3.674
29	0.256	0.683	1.311	1.699	2.045	2.462	2.756	3.038	3.396	3.659
30	0.256	0.683	1.310	1.697	2.042	2.457	2.750	3.030	3.385	3.646
40	0.255	0.681	1.303	1.684	2.021	2.423	2.704	2.971	3.307	3.551
60	0.254	0.679	1.296	1.671	2.000	2.390	2.660	2.915	3.232	3.460
120	0.254	0.677	1.289	1.658	1.980	2.358	2.617	2.860	3.160	3.373
∞	0.253	0.674	1.282	1.645	1.960	2.326	2.576	2.807	3.090	3.291

Example 3

A packaging device is set to fill detergent packets with a mean weight of 150g. The standard deviation is **known** to be 5.0g. It is important to check the machine periodically because if it is overfilling it increases the cost of the materials, whereas if it is underfilling the firm is liable to prosecution. A random sample of 25 filled boxes is weighed and shows a mean net weight of 152.5g.

- Can we conclude that the machine is no longer filling packs with 150g. detergents? Use a 5% significance level.
- Find the probability of type II error, when mean weight shifts to 152 g.

Solution

a) Given $\mu = 150\text{g}$, $\sigma = 5\text{g}$, $n = 25$, $\bar{x} = 152.5\text{g}$

The machine can 'no longer fill 150g packs' so it could be either over-filling or under-filling, therefore the appropriate test is two tailed.

$$H_0: \mu = 150\text{g} \quad H_1: \mu \neq 150\text{g}$$

Significance level (α): 5% (0.05)

Critical value: σ known therefore z test, 5%, two tailed $\rightarrow 1.96$.

$$\text{Test statistic: } z = \frac{|\bar{x} - \mu|}{\sigma/\sqrt{n}} \Rightarrow \frac{152.5 - 150}{5/\sqrt{25}} = \frac{2.5}{1} = 2.5$$

Conclusion: The test statistic exceeds the critical value so reject H_0 and conclude that the mean weight packed is no longer 150g.

b) Upper boundary of acceptance region = $150 + 1.96 \times 1 = 151.96$.

Lower boundary of acceptance region = $150 - 1.96 \times 1 = 148.04$.

$$\begin{aligned} \text{Prob(type II error)} &= \text{Prob}(148.04 \leq \bar{x} \leq 151.96, \mu = 152 \text{ \& } \sigma_{\text{Mean}} = 1) \\ &= P(-3.96 \leq Z \leq -0.04) \\ &= P(Z \leq -0.04) - P(Z \leq -3.96) \\ &= 0.484047 - 0.000037 = \mathbf{0.484010}. \end{aligned}$$

Example 4

The mean and standard deviation of the weights produced by this same packaging device set to fill detergent packets with a mean weight of 150g, are known to drift upwards over time due to the normal wearing of some bearings. Obviously it cannot be allowed to drift too far, so a large random sample of 100 boxes is taken and the contents weighed. This sample observations has a mean weight of 151.0 g and a standard deviation of 6.5g. Can we conclude that the mean weight produced by the machine has increased? Use a 5% significance level.

Solution

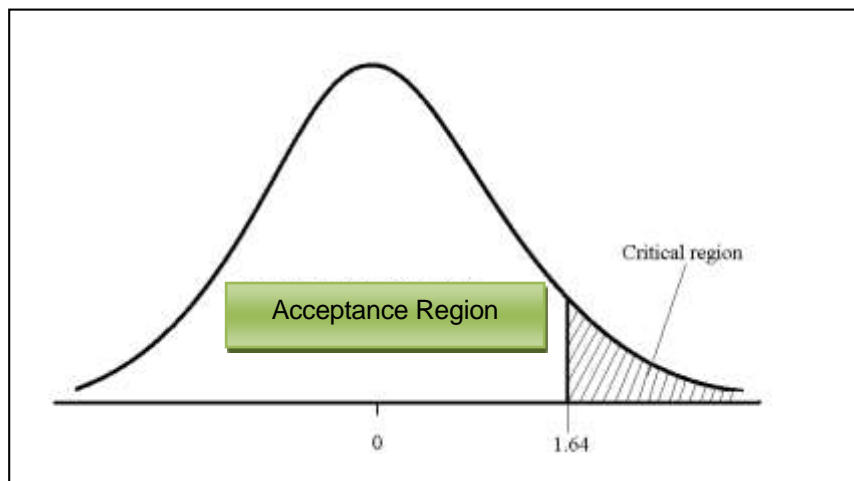
Given: $\mu = 150\text{g}$, $n = 100$, $\bar{x} = 151.0\text{g}$, $s = 6.5\text{g}$,

We are only interested in whether the mean weight has increased or not so a one-tailed test is appropriate.

$$H_0: \mu = 150 \qquad H_1: \mu > 150$$

Significance level (α): 0.05

Critical value: σ is unknown but the sample size is large, therefore use z test, 5%, one tailed $\rightarrow 1.64$.



Test statistic: $z = \frac{|\bar{x} - \mu|}{s/\sqrt{n}} = \frac{151 - 150}{6.5/\sqrt{100}} = \frac{10}{6.5} = 1.538.$

Conclusion: Value of the test statistic is less than the critical value so we *fail to reject* H_0 and conclude that the mean weight produced by the machine has remained unchanged.

[Find the probability of type II error assuming mean is shifted to 152 gm]

Example 5

The personnel department of a company developed an aptitude test for screening potential employees. The person who devised the test asserted that the mean mark attained would be 100. The following results were obtained with a random sample of applicants:

$$\bar{x} = 96, \quad s = 5.2, \quad n = 13.$$

Based on this observation, can it be assumed that the mean mark is less than 100 at the 1% significance level.

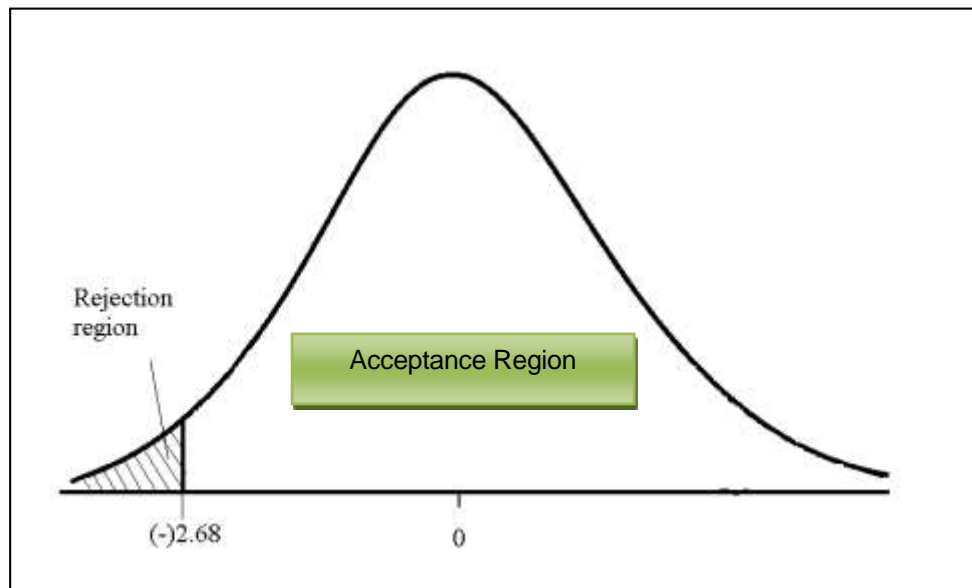
Solution

Given: $\mu = 100$, $\bar{x} = 96$, $s = 5.2$, $n = 13$.

$$H_0: \mu = 100 \qquad H_1: \mu < 100$$

Significance level (α): 1% (0.01)

Critical Value: σ unknown and small sample, so we have to use t-test. Therefore, the tabulated value at 1% significant, 1 tail, $n - 1 = 12$ df is 2.68 and the critical value corresponding to the alternate hypothesis is -2.68.



Test statistic: $t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{96 - 100}{5.2/\sqrt{13}} = \frac{-4 \times \sqrt{13}}{5.2} = -2.72$.

Conclusion: The test statistic falls in the critical region, we reject the null hypothesis H_0 and conclude that the mean mark is less than 100.

Testing for Proportions, Percentages

In this case the test statistic for testing proportion is $= \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$, where \hat{p} is the sample proportion, p_0 is hypothesized population proportion and n is the sample size. Since sample sizes are generally large, above statistic can be assumed to follow standard normal distribution.

So, the critical values can be obtained from Standard Normal Distribution table.

Example 6

A company manufacturing a certain brand of breakfast cereal claims that 60% of all housewives prefer its brand to any other. A random sample of 300 housewives include 165 who do prefer the brand. Is the true percentage is lower than the claims made by the company at the 5% significance level?

Solution

Given Information: $p_0 = 60\%$, $\hat{p} = \frac{165}{300} \times 100 = 55\%$, $n = 300$
 $H_0: p = 60\%$ $H_1: p < 60\%$

Critical Value: Large sample size, z test, 1 tail, 5% significance

→ - 1.64 (based upon H_1)

Test Statistic:
$$\frac{\hat{p} - p_0}{\sqrt{\frac{p_0(100 - p_0)}{n}}} = \frac{55 - 60}{\sqrt{\frac{60 \times 40}{300}}} = \frac{-5}{\sqrt{8}} = -1.77$$

Conclusion: Test statistic < critical value, so reject H_0 . Therefore, the % of housewives using the breakfast cereal brand is < 60%.

Example 7

Previous study shows that 10% of invoices for a certain company are incorrect. To test this claim a random sample of 200 invoices are checked and 24 are found to be incorrect. Test at 1% significant if the claim is supported by the sample evidence.

Solution

Given Information: $p_0 = 10\%$, $\hat{p} = \frac{24}{200} \times 100 = 12\%$, $n = 200$.

$$H_0 : p = 10\% \quad H_1 : p \neq 10\%$$

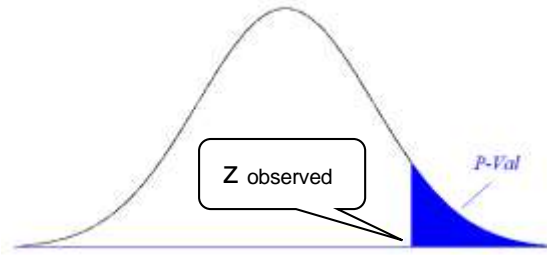
Critical Value: z test as sample size is large, 2 tail, 1% level of significance → 2.58.

Test Statistic

$$\frac{|\hat{p} - p_0|}{\sqrt{\frac{p_0(100 - p_0)}{n}}} = \frac{|12 - 10|}{\sqrt{\frac{10 \times 90}{200}}} = \frac{2}{\sqrt{9/2}} = 0.943$$

Conclusion: Test statistic is less than critical value so H_0 cannot be rejected. Hence, the percentage of incorrect invoices is consistent with the study findings of 10%.

The P-value Method of Hypothesis Testing



The *P*-value (or *p*-value or probability value) is the probability of getting a value of the test statistic that is at least as extreme as the one obtained from sample data, assuming that the null hypothesis is true. By extreme we mean: far from what we would expect to observe if the null hypothesis is true. Clearly, it is the smallest level of significance, i.e. smallest value of α , at which the null hypothesis, H_0 can possibly be rejected.

Obviously, a small *P*-value indicates that observation of the test statistic would be unlikely if the null hypothesis is true. Lower the *P*-value, the more evidence there is in favor of rejecting the null hypothesis.

Decision Rule Based on *P*-value

To use a *P*-value to make a conclusion in a hypothesis test, compare the *P*-value with α and

- a) if $P \leq \alpha$, then reject H_0 ,
- b) if $P > \alpha$, then fail to reject H_0 .

Determination of P-value

Null Hypothesis	Alternate Hypothesis	Test Statistic	<i>P</i> -value
$H_0: \mu = \mu_0$ [σ^2 known]	$H_1: \mu > \mu_0$	$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	$1 - P(Z \leq z_0)$
	$H_1: \mu \neq \mu_0$		$2[1 - P(Z \leq z_0)]$
	$H_1: \mu < \mu_0$		$P(Z \leq -z_0)$

Similarly, *P*-values can also be determined for the tests that based on *t*-distribution (i.e. σ unknown) and for **two sample tests** also.

Connection between Hypothesis Testing and Confidence Interval

A close relationship exists between the test of hypothesis about a parameter, say θ , and the confidence interval for θ .

If $[l, u]$ is a $100(1 - \alpha)\%$ confidence interval for the parameter μ with confidence coefficient α , then the hypothesis

$$H_0: \mu = \mu_0 \qquad H_1: \mu \neq \mu_0$$

will lead to rejection of the null hypothesis at significant level α , if and only if μ_0 is not in the $100(1 - \alpha)\%$ confidence interval $[l, u]$.

We know that null hypothesis can not be rejected, if $\left| \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \right| \leq Z_{\alpha/2}$. This leads to confidence interval of μ as $\bar{x} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ and consequently the null hypothesis will be rejected if μ_0 falls outside this confidence interval.

For example, let us consider the detergent packet example (Example 3) with

$$\bar{x} = 152.5\text{g}, \sigma = 5\text{g}, \text{ and } n = 25$$

The null hypothesis $H_0: \mu = 150\text{g}$ was rejected at $\alpha = 0.05$. The 95% confidence interval on μ can be calculated as $152.5 \pm 1.96(5/\sqrt{25}) = 152.5 \pm 1.96$. Above leads to the 95% confidence interval as $[150.54, 154.46]$.

Since the value $\mu_0 = 150$ is not included in above interval, the null hypothesis $H_0: \mu = 150\text{g}$ is rejected.

Summary of Hypothesis Testing

I. Normal Population $N(\mu, \sigma^2)$: σ^2 known - $H_0: \mu = \mu_0$

Hypothesis	Test	Test Statistic	Reject H_0 at level of significance α , if	$100(1-\alpha)\%$ confidence interval for μ
$H_1: \mu \neq \mu_0$	Two-tailed test	$z_0 = (\bar{x} - \mu_0) / (\sigma / \sqrt{n})$	$ z_0 > Z_{\alpha/2}$	$\bar{x} - \frac{\sigma}{\sqrt{n}} Z_{\alpha/2} \leq \mu \leq \bar{x} + \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}$
$H_1: \mu > \mu_0$	Right-tailed test	$z_0 = (\bar{x} - \mu_0) / (\sigma / \sqrt{n})$	$z_0 > Z_\alpha$	$\mu \geq \bar{x} - \frac{\sigma}{\sqrt{n}} Z_\alpha$
$H_1: \mu < \mu_0$	Left-tailed test	$z_0 = (\bar{x} - \mu_0) / (\sigma / \sqrt{n})$	$z_0 < -Z_\alpha$	$\mu \leq \bar{x} + \frac{\sigma}{\sqrt{n}} Z_\alpha$

II. Normal Population $N(\mu, \sigma^2)$: σ^2 unknown - $H_0: \mu = \mu_0$

Hypothesis	Test	Test Statistic	Reject H_0 at level of significance α , if	$100(1-\alpha)\%$ confidence interval for μ
$H_1: \mu \neq \mu_0$	Two-tailed test	$t_0 = (\bar{x} - \mu_0) / (s / \sqrt{n})$	$ t_0 > t_{\alpha/2, n-1}$	$\bar{x} - \frac{s}{\sqrt{n}} t_{\alpha/2, n-1} \leq \mu \leq \bar{x} + \frac{s}{\sqrt{n}} t_{\alpha/2, n-1}$
$H_1: \mu > \mu_0$	Right-tailed test	$t_0 = (\bar{x} - \mu_0) / (s / \sqrt{n})$	$t_0 > t_{\alpha, n-1}$	$\mu \geq \bar{x} - \frac{s}{\sqrt{n}} t_{\alpha, n-1}$
$H_1: \mu < \mu_0$	Left-tailed test	$t_0 = (\bar{x} - \mu_0) / (s / \sqrt{n})$	$t_0 < -t_{\alpha, n-1}$	$\mu \leq \bar{x} + \frac{s}{\sqrt{n}} t_{\alpha, n-1}$

III. Normal Populations: $\sigma_1^2 \neq \sigma_2^2$ known - $H_0: \mu_1 = \mu_2$

Hypothesis	Test	Test Statistic	Reject H_0 at level of significance α , if
$H_0: \mu_1 \neq \mu_2$	Two-tailed test	$z_0 = (\bar{x}_1 - \bar{x}_2) / \sqrt{[\sigma_1^2/n_1 + \sigma_2^2/n_2]}$	$ z_0 > Z_{\alpha/2}$
$H_1: \mu_1 > \mu_2$	Right-tailed test	$z_0 = (\bar{x}_1 - \bar{x}_2) / \sqrt{[\sigma_1^2/n_1 + \sigma_2^2/n_2]}$	$z_0 > Z_\alpha$
$H_1: \mu_1 < \mu_2$	Left-tailed test	$z_0 = (\bar{x}_1 - \bar{x}_2) / \sqrt{[\sigma_1^2/n_1 + \sigma_2^2/n_2]}$	$z_0 < -Z_\alpha$

IV. Normal Populations: $\sigma_1^2 = \sigma_2^2 = \sigma^2$ unknown - $H_0: \mu_1 = \mu_2$

Hypothesis	Test	Test Statistic	Reject H_0 at level of significance α , if
$H_1: \mu_1 \neq \mu_2$	Two-tailed test	$t_0 = (\bar{x}_1 - \bar{x}_2) / S_p \sqrt{[1/n_1 + 1/n_2]}$	$ t_0 > t_{\alpha/2, n_1+n_2-2}$
$H_1: \mu_1 > \mu_2$	Right-tailed test	$t_0 = (\bar{x}_1 - \bar{x}_2) / S_p \sqrt{[1/n_1 + 1/n_2]}$	$t_0 > t_{\alpha, n_1+n_2-2}$
$H_1: \mu_1 < \mu_2$	Left-tailed test	$t_0 = (\bar{x}_1 - \bar{x}_2) / S_p \sqrt{[1/n_1 + 1/n_2]}$	$t_0 < -t_{\alpha, n_1+n_2-2}$

Note: $S_p = \sqrt{\{(n_1-1)s_1^2 + (n_2-1)s_2^2\} / (n_1+n_2-2)}$

V. Normal Populations: $\sigma_1^2 \neq \sigma_2^2$ unknown - $H_0: \mu_1 = \mu_2$

Hypothesis	Test	Test Statistic	Reject H_0 at level of significance α , if
$H_1: \mu_1 \neq \mu_2$	Two-tailed test	$t_0 = (\bar{x}_1 - \bar{x}_2) / \sqrt{[s_1^2/n_1 + s_2^2/n_2]}$	$ t_0 > t_{\alpha/2, \vartheta}$
$H_1: \mu_1 > \mu_2$	Right-tailed test	$t_0 = (\bar{x}_1 - \bar{x}_2) / \sqrt{[s_1^2/n_1 + s_2^2/n_2]}$	$t_0 > t_{\alpha, \vartheta}$
$H_1: \mu_1 < \mu_2$	Left-tailed test	$t_0 = (\bar{x}_1 - \bar{x}_2) / \sqrt{[s_1^2/n_1 + s_2^2/n_2]}$	$t_0 < -t_{\alpha, \vartheta}$

Note: $\vartheta \cong \left\{ \frac{s_1^2}{n_1} + \frac{s_2^2}{n_2} \right\}^2 / \left\{ \frac{(s_1^2/n_1)^2}{n_1-1} + \frac{(s_2^2/n_2)^2}{n_2-1} \right\}$ is the degrees of freedom for this test (Ref. Welch-Satterthwaite Equation to find approximate DF of linear combination independent sample variances)

VI. Normal Population: μ and σ^2 unknown - $H_0: \sigma^2 = \sigma_0^2$

Hypothesis	Test	Test Statistic	Reject H_0 at level of significance α , if
$H_1: \sigma^2 \neq \sigma_0^2$	Two-tailed test	$\chi_0^2 = (n-1)s^2 / \sigma_0^2$	$\chi_0^2 > \chi_{\alpha/2, n-1}^2$ or $\chi_0^2 < \chi_{1-\alpha/2, n-1}^2$
$H_1: \sigma^2 > \sigma_0^2$	Right-tailed test	$\chi_0^2 = (n-1)s^2 / \sigma_0^2$	$\chi_0^2 > \chi_{\alpha, n-1}^2$
$H_1: \sigma^2 < \sigma_0^2$	Left-tailed test	$\chi_0^2 = (n-1)s^2 / \sigma_0^2$	$\chi_0^2 < \chi_{1-\alpha, n-1}^2$

VII. Normal Populations: μ 's and σ^2 's unknown - $H_0: \sigma_1^2 = \sigma_2^2$

Hypothesis	Test	Test Statistic	Reject H_0 at level of significance α , if
$H_1: \sigma_1^2 \neq \sigma_2^2$	Two-tailed test	$F_0 = s_1^2 / s_2^2$	$F_0 > F_{\alpha/2, n_1-1, n_2-1}$ or $F_0 < F_{1-\alpha/2, n_1-1, n_2-1}$
$H_1: \sigma_1^2 > \sigma_2^2$	Right-tailed test	$F_0 = s_1^2 / s_2^2$	$F_0 > F_{\alpha, n_1-1, n_2-1}$

Exact Test for Binomial Parameter

Suppose a random sample of size n is drawn from an infinite population for which the proportion of entities having a specific property, say p , is unknown. In order to test the hypothesis

$$H_0 : p = p_0$$

we make use of the statistic x , the number of members in the sample having that specific property, which is a sufficient statistic for p .

Now, under the null hypothesis, x is distributed as binomial with parameters n and p_0 .

Let the observed value of x be x_0 .

Case I: $H_1 : p > p_0$

In this situation we will compute the probability:

$$P[x \geq x_0 | p = p_0] = \sum_{x \geq x_0} \binom{n}{x} p_0^x (1 - p_0)^{n-x}$$

If the calculated probability is small, then null hypothesis is unlikely to be true. Hence, if the calculated probability found to be less than specified level of significance, we shall consider x_0 to be an unlikely value under the null hypothesis and consequently reject the null hypothesis. Otherwise if the calculated probability is found to be greater than the specified level of significance, the decision will be failure to reject the null hypothesis.

Case II: $H_1 : p < p_0$

Here we will compute

$$P[x \leq x_0 | p = p_0] = \sum_{x \leq x_0} \binom{n}{x} p_0^x (1 - p_0)^{n-x}$$

and shall reject the null hypothesis if this probability become less than the specified level of significance, else we fail to reject the null hypothesis.

Example 8

A researcher **believes that more than 10%** of all cricket helmets have manufacturing flaws (that could potentially cause injury to the user). A sample of 20 helmets revealed that 2 helmets contained such defects. Does the finding support the researcher's claim? Use $\alpha = 0.05$.

Solution

$$H_0 : p \leq 0.1$$

$$H_1 : p > 0.1$$

Here, $n = 20$, and $x = 2$. So, we can calculate the probability of getting atleast 2 defects under null hypothesis as

$$\begin{aligned} P[x \geq 2 | p = 0.1] &= \sum_{x \geq 2} \binom{20}{x} (0.1)^x (0.9)^{20-x} \\ &= 1 - \sum_{x=0}^1 \binom{20}{x} (0.1)^x (0.9)^{20-x} \\ &= 1 - 0.391747 \\ &= 0.608253 \end{aligned}$$

Thus, probability of getting at least 2 defects under null hypothesis is quite high.

Since, this probability exceeds 0.05, the null hypothesis can not be rejected at 5% level of significance. Thus the data fails to support the researcher's claim.

Pearson's Chi-squared Test

Pearson's chi-squared test (χ^2) is the statistical procedures whose results are evaluated by reference to the chi-squared distribution. Its properties were first investigated by Karl Pearson in 1900. **It tests a null hypothesis that the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution.** **The events considered must be mutually exclusive and have total probability 1.** A common case for this is, where the events are outcome of a **categorical variable**. A simple example is the hypothesis that an ordinary six-sided die is "fair", i.e. all six outcomes are equally likely to occur.

Pearson's chi-squared is used to assess two types of comparison: *tests of goodness of fit* and *tests of independence*.

- **A test of goodness of fit** establishes whether or not an observed frequency distribution of **a single attribute fit a particular theoretical distribution**. For example, whether number of arrivals per minute in a bank follows Poisson Distribution?
- **A test of independence** assesses whether paired observations on **two attributes**, expressed in a contingency table, **are independent of each other or not**. For example, whether people from different genders differ in their income levels.

Pearson's Goodness-of-Fit Test

In Pearson's chi square test for goodness of fit, in each response category we compare the **observed frequencies** to the **expected frequencies under the null hypothesis** and uses the following statistic

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

In this formula, O_i is a count of the number of observed items in category i , and E_i is the number of expected items in category i **under null hypothesis**.

Now, if under null hypothesis p_i be the probability that an observation falls into category i , then clearly $E_i = np_i$, where n is the total number of observations (also known as **sample size**) that were classified into k mutually exclusive classes.

When chi square statistic is used to test several proportions simultaneously, then the number of degrees of freedom will be **$k - 1$** , where k is the number of categories.

Pearson's Goodness-of-Fit Test is always an upper tail test. A value of $\chi^2 = 0$, would represent a perfect match between observed and expected frequencies.

The formula given above is the formula traditionally quoted, yet a slightly easier formula exists for computational purposes. The easier formula will produce the exactly same value for the test statistic, and it is derived as follows:

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^k \left(\frac{O_i^2}{E_i} - 2O_i + E_i \right) = \sum_{i=1}^k \frac{O_i^2}{E_i} - n$$

If you are having to do the computations "by hand", this result is easier. The **benefit of the traditional formula is to be able to identify which categorie(s) are furthest "out of line" and contribute largely to the value of the test statistic.**

Mathematical Justification

Consider the Binomial case, where there are just two categories. For this situation $p_1 + p_2 = 1$, and $x_1 + x_2 = n$,

where p_i is the probability of getting an observation in the i th category, and x_i represents the number of observations in the i th category.

Now, expected number of observations in the first and second categories will be np_1 and np_2 respectively. From, central limit theorem, we know that $z = \frac{x_1 - np_1}{\sqrt{np_1(1-p_1)}}$ can be approximated by a $N(0,1)$ distribution, when n is large.

So, z^2 will follow χ^2 distribution with one degree of freedom.

$$\begin{aligned} \text{Now, } z^2 &= \frac{(x_1 - np_1)^2}{np_1(1-p_1)} \\ &= \frac{(x_1 - np_1)^2(1-p_1) + (x_1 - np_1)^2 p_1}{np_1(1-p_1)} \\ &= \frac{(x_1 - np_1)^2}{np_1} + \frac{(x_1 - np_1)^2}{n(1-p_1)} \end{aligned}$$

Again, since $(x_1 - np_1)^2 = (n - x_2 - n + np_2)^2 = (x_2 - np_2)^2$, so the we have

$$z^2 = \frac{(x_1 - np_1)^2}{np_1} + \frac{(x_2 - np_2)^2}{np_2},$$

which has a χ^2 distribution with one degree of freedom. Under null hypothesis, observed values x_1 and x_2 should be close to corresponding expected values np_1 and np_2 , and consequently calculated value z^2 will be small. Otherwise, z^2 will be large. So, **larger value of chi square will lead to the possible rejection of the null hypothesis**. Exactly how large the χ^2 value must be in order to be considered large enough to reject the null hypothesis, can be determined from the level of significance and the chi square table.

In general, for k random variable with corresponding expected values np_i , a statistic measuring the **closeness** of the observations to their expectations is the sum

$$\frac{(x_1 - np_1)^2}{np_1} + \frac{(x_2 - np_2)^2}{np_2} + \dots + \frac{(x_k - np_k)^2}{np_k} = \sum_{i=1}^k \frac{(x_i - np_i)^2}{np_i}$$

which has a χ^2 distribution with $k - 1$ degrees of freedom. This is because, we know that sum of all the probabilities p_1, p_2, \dots, p_k must equal 1 and thus we can derive p_k by subtracting sum of the first $k - 1$ probabilities from 1.

It may be noted that above chi-squared statistic has an approximate chi-squared distribution for large n . **The chi-squared approximation improves as expected values increase and expected values ≥ 5 is usually sufficient for a decent approximation.**

Should an expected value of a cell be less than 5 that will generally be on tails of the distribution, it has to be cumulated with the expected values of adjacent cell/cells, till the expected value of the combined class become ≥ 5 . The corresponding observed frequencies would also then be cumulated and the degree of freedom of the χ^2 statistic will be (# of classes after combining) -1. **This rule that the expected frequency of any class should be ≥ 5 , is to be used in all χ^2 tests for goodness of fit.** Note that binomial distributions are approximately normal when $np \geq 5$ and $nq \geq 5$, and that a normally distributed random variable X implies that Z^2 has a chi-square distribution with one degree of freedom.

In general, large values of z^2 i. e. χ^2 will imply rejection of null hypothesis. So, if the calculated value of chi-square is more than χ^2_{k-1} for the given significance level, we will reject the null hypothesis, otherwise the null hypothesis will be accepted (loosely speaking).

Since, χ^2 is a right-tailed test, so in RMMR table **upper tail value corresponding to one-sided test will give us the critical value.**

Note: Some statistician suggest that chi-square approximation could even be reasonable if : (Ref. Yates, et al (1999))

1. All expected frequencies are greater than or equal to 1, and
2. No more than 20% of expected frequencies are less than 5.

χ^2 distribution critical values

df	$\alpha = .10$	$\alpha = .05$	$\alpha = .01$
1	2.706	3.841	6.635
2	4.605	5.991	9.210
3	6.251	7.815	11.345
4	7.779	9.488	13.277
5	9.236	11.070	15.086
6	10.645	12.592	16.812
7	12.017	14.067	18.475
8	13.362	15.507	20.090
9	14.684	16.919	21.666
10	15.987	18.307	23.209
11	17.275	19.675	24.725
12	18.549	21.026	26.217
13	19.812	22.362	27.688
14	21.064	23.685	29.141
15	22.307	24.996	30.578
16	23.542	26.296	32.000
17	24.769	27.587	33.409
18	25.989	28.869	34.805
19	27.204	30.144	36.191
20	28.412	31.410	37.566
21	29.615	32.671	38.932
22	30.813	33.924	40.289
23	32.007	35.172	41.638
24	33.196	36.415	42.980
25	34.382	37.652	44.314
26	35.563	38.885	45.642
27	36.741	40.113	46.963
28	37.916	41.337	48.278
29	39.087	42.557	49.558
30	40.256	43.773	50.892

Example 9

A new casino game involves rolling 3 dice. The winnings are directly proportional to the total number of sixes rolled. Suppose a gambler plays the game 100 times, with the following observed counts:

Number of Sixes	Number of Rolls
0	47
1	35
2	15
3	3

The casino becomes suspicious of the gambler and wishes to determine whether the dice are fair. What do they conclude? Assume $\alpha = 0.05$.

Solution

If a die is fair, we would expect the probability of rolling a 6 on any given toss to be $1/6$. So, the hypothesis under question is

$$H_0: p = \frac{1}{6}$$

$$H_1: p \neq \frac{1}{6}$$

Assuming the 3 dice are independent (the roll of one die should not affect the roll of the others), we might assume that the number of sixes in three rolls of dice is distributed as **Binomial**(3,1/6). To determine whether the gambler's dice are fair, we may compare his results with the results expected under this distribution. The expected values for 0, 1, 2, and 3 sixes under the **Binomial**(3,1/6) distribution are the following:

Under null hypothesis:

$$\begin{aligned}
 p_1 &= P(\text{roll 0 sixes}) &= P(x=0 | n=3) &= 0.578703 \\
 p_2 &= P(\text{roll 1 six}) &= P(x=1 | n=3) &= 0.347223 \\
 p_3 &= P(\text{roll 2 sixes}) &= P(x=2 | n=3) &= 0.069445 \\
 p_4 &= P(\text{roll 3 sixes}) &= P(x=3 | n=3) &= 0.00463
 \end{aligned}$$

Since the gambler plays 100 times, the expected counts are the following:

Number of Sixes	Expected Counts	Observed Counts
0	57.87	47
1	34.72	35
2	6.94	15
3	0.46	3
	$\left. \begin{array}{l} 6.94 \\ 0.46 \end{array} \right\} \boxed{7.4}$	
	$\left. \begin{array}{l} 15 \\ 3 \end{array} \right\} \boxed{18}$	

The chi-squared statistic for the above example is

$$\frac{(47 - 57.87)^2}{57.87} + \frac{(35 - 34.72)^2}{34.72} + \frac{(18 - 7.4)^2}{7.4}$$

$$= 2.04176 + 0.00226 + 15.18378 = 17.2278$$

Since $k = 3$ in this case (#of possibilities -1), the test statistic χ^2 is associated with the chi-square distribution with 2 degrees of freedom. If we are interested in a significance level of 0.05 we may reject the null hypothesis (that the dice are fair) if $\chi^2 \geq 5.991$, the value corresponding to the 0.05 significance level for the $\chi^2(2)$ distribution. Since 17.2278 is clearly greater than 5.991, we may reject the null hypothesis that the dice are fair at the 0.05 significance level and conclude that the dice are not fair.

Example 10

Following table gives height (H) of 66 women in inches. Can it be assumed the heights follow Gaussian distribution? Assume $\alpha = 0.05$.

H	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72	73
$n(H)$	1	0	1	4	6	7	13	8	11	2	7	4	1	0	0	1

Solution

First we calculate the **mean** and **sd** of data on height. The values obtained are

$$\bar{H} = 64.9 \text{ inches and } s = 2.7 \text{ inches}$$

H	Range of H	Probability [P]	Expected [66*P]		Observed	χ^2
58	$57.5 \leq h \leq 58.5$	0.005820	0.38413	6.66	1	0.06541
59	$58.5 \leq h \leq 59.5$	0.013865	0.91509		0	
60	$59.5 \leq h \leq 60.5$	0.028840	1.90342		1	
61	$60.5 \leq h \leq 61.5$	0.052378	3.45698		4	
62	$61.5 \leq h \leq 62.5$	0.083063	5.48216	5.48	6	0.04934
63	$62.5 \leq h \leq 63.5$	0.115017	7.59111	7.60	7	0.04737
64	$63.5 \leq h \leq 64.5$	0.139065	9.17827	9.18	13	1.58959
65	$64.5 \leq h \leq 65.5$	0.146817	9.68990	9.69	8	0.29475
66	$65.5 \leq h \leq 66.5$	0.135344	8.93267	8.93	11	0.47983
67	$66.5 \leq h \leq 67.5$	0.108944	7.19030	7.19	2	3.74633
68	$67.5 \leq h \leq 68.5$	0.076572	5.05374	5.05	7	0.75297
69	$68.5 \leq h \leq 69.5$	0.046993	3.10154	5.97	4	0.00015
70	$69.5 \leq h \leq 70.5$	0.025182	1.66201		1	
71	$70.5 \leq h \leq 71.5$	0.011782	0.77764		0	
72	$71.5 \leq h \leq 72.5$	0.004814	0.31769		0	
73	$72.5 \leq h \leq 73.5$	0.001717	0.11332		1	

$$\text{Chi Square} = 7.02573$$

Since we have grouped our data into 9 classes, and since we have used up additional two degrees of freedom by demanding (a) that the mean of the hypothesized distribution be equal to the sample mean, and (b) that the variance of the hypothesized distribution be equal to the sample variance, there are 6, i.e. (9-1-2) degrees of freedom left.

Therefore, the critical value is $\chi^2_{0.05,6} = 12.592$. Since the calculated value, i.e. 7.026 is less than the critical value, we **fail to reject the null hypothesis** that our data are drawn from a Gaussian distribution function.

Contingency Table

A **contingency table** (also referred to as **cross tabulation** or **cross tab**) is a type of table in a matrix format that displays the (multivariate) frequency distribution of two or more variables. It is often **used to record and analyze the relationship between two or more categorical variables**.

Suppose that we have two variables, Gender (male or female) and smoking habit (smoker or non-smoker). Further suppose that 100 individuals are randomly sampled from a very large population as part of a **study of gender differences in smoking habit**. A contingency table can be created to display the numbers of individuals who are male and smokers, male and non-smokers, female and smokers, and female and non-smokers. Such a contingency table is shown below.

	Male	Female	Totals
Smokers	41	5	46
Non-Smokers	11	43	54
Totals	52	48	100

The numbers of the males, females, and smokers and non-smokers are called **marginal totals**. The **grand total**, i.e., the total number of individuals represented in the contingency table, is the number in the bottom right corner. Clearly, marginal distribution of smoking habit is: 46% smokers and 54% non-smokers. Conditional distribution of smoking habit for male is **78.8%** smokers and **21.2%** non-smokers, while the same for female is 10.4% smokers and 89.6% non-smokers.

Thus **percentage of male who are smokers is much more as compared to the percentage of females who are smokers**. The significance of this difference between these two proportions can be assessed with Pearson's chi-squared test, provided the

entries in the table represent individuals **randomly sampled** from the population about which we want to draw a conclusion. If the proportions of individuals in the different columns vary significantly between rows (or vice versa), we say that there is a **contingency** between the two variables, i.e. **there is a definite relationship between the two variables**, or in other words, the two variables are **not independent**. If there is no contingency, we say that the two variables are **independent**.

Sometimes the experimental units in a population can be cross-classified by two factors A (r classes/levels) and B (c classes/levels). Such a table is known as $r \times c$ contingency table.

$r \times c$ Contingency Table

$A \downarrow B \rightarrow$	B_1	B_2	\dots	B_c	Total
A_1	f_{11} p_{11}	f_{12} p_{12}	\dots	f_{1c} p_{1c}	$f_{1\bullet}$ $p_{1\bullet}$
A_2	f_{21} p_{21}	f_{22} p_{22}	\dots	f_{2c} p_{2c}	$f_{2\bullet}$ $p_{2\bullet}$
\vdots	\vdots	\vdots	\dots	\vdots	\vdots
A_r	f_{r1} p_{r1}	f_{r2} p_{r2}	\dots	f_{rc} p_{rc}	$f_{r\bullet}$ $p_{r\bullet}$
Total	$f_{\bullet 1}$ $p_{\bullet 1}$	$f_{\bullet 2}$ $p_{\bullet 2}$	\dots	$f_{\bullet c}$ $p_{\bullet c}$	$f_{\bullet\bullet} (= n)$

Note: $p_{ij} = f_{ij}/n$, $p_{i\cdot} = f_{i\cdot}/n$ and $p_{\cdot j} = f_{\cdot j}/n$. **Clearly, they can be thought of as the corresponding probability of occurrence also.**

Such a classification of data can be used for the following two purposes:

- i) **test for independence**, and
- ii) **test for homogeneity**.

In the chi squared test of Homogeneity, the data consists of **two/more independent random samples** from two/more groups, whereas in the test of Independence, the data consists of **a single random sample in which the groupings are retrospectively determined**. The hypotheses are often written differently for the 2 tests. **For the**

test of homogeneity the null hypothesis is that there is no difference in group proportions. In the test of independence, the null hypothesis is that the outcome is independent of group. For example, suppose that you are interested to investigate the association between income level (high vs. low) and gender (male vs female).

1. **In a chi squared test of homogeneity** you would collect a sample of income data for females and an independent sample of income data for males and categorize each sample according to the corresponding income level.
2. **In a chi squared test of independence** you would collect a single sample of data from the mixed population, after that you would figure out how many of them are males and how many of them are females, and finally you would also figure out how many high and low income people are there within each group.

Test for Independence

Let p_{ij} be the probability that a randomly selected entity falls in the i -th class of A and j -th class of B . This probability p_{ij} define the joint probability distribution of A and B . Also the marginal total $p_{i\cdot}$ gives the marginal probability distribution of A , whereas $p_{\cdot j}$ gives the marginal probability distribution of B .

In the $r \times c$ contingency table, there are in total rc number of cells and so the Pearsonian chi-square is

$$\sum_i \sum_j \frac{(f_{ij} - np_{ij})^2}{np_{ij}} \cap \chi_{rc-1}^2.$$

[The table have rc cells and total of all cell frequencies is known, so clearly amongst the rc cells only $rc-1$ cells are independent. So *degrees of freedom for the chi-square test will be $rc-1$*]

Here we define the hypothesis as

$$\begin{aligned} H_0: A \text{ and } B \text{ are independent} \\ H_1: A \text{ and } B \text{ are related} \end{aligned}$$

So, under the null hypothesis

$$P(A = A_i, B = B_j) = P(A = A_i) P(B = B_j)$$

So, the null hypothesis becomes

$$H_0 : p_{ij} = p_{i\cdot} \times p_{\cdot j} \forall i \text{ and } j, \text{ i.e. } A \text{ and } B \text{ are independent.}$$

Thus, under null hypothesis

$$\sum_i \sum_j \frac{(f_{ij} - np_{i\cdot} p_{\cdot j})^2}{np_{i\cdot} p_{\cdot j}} \cap \chi_{rc-1}^2$$

However, since $p_{i\cdot}$ and $p_{\cdot j}$ are, in general, not known we use their estimates, which are given by $(f_{i\cdot}/n)$ and $(f_{\cdot j}/n)$ respectively. Thus, we need to fix the row totals and the column totals, thereby imposing $(r + c)$ constraints on the cell frequencies.

But since

$$\sum_{i=1}^r f_{i\cdot} = \sum_{j=1}^c f_{\cdot j} = n,$$

i.e. row totals and column totals add to grand total, the total number of independent constraints imposed is $(r-1) + (c-1)$. Therefore, the degrees of freedom for the chi-square statistic is

$$(rc-1) - [(r-1) + (c-1)] = (r-1)(c-1).$$

Hence, finally we have under H_0 ,

$$\sum_i \sum_j \frac{\left(f_{ij} - \frac{f_{i\cdot} \times f_{\cdot j}}{n}\right)^2}{\frac{f_{i\cdot} \times f_{\cdot j}}{n}} \cap \chi_{(r-1)(c-1)}^2$$

$$\text{or, } n \left[\sum_i \sum_j \frac{f_{ij}^2}{f_{i\bullet} \times f_{\bullet j}} - 1 \right] \cap \chi_{(r-1)(c-1)}^2.$$

If $\chi_{\text{Calculated}}^2 < \chi_{\alpha, (r-1)(c-1)}^2$, then we **fail to reject the null hypothesis**, which implies that the two attributes are independent. Otherwise, we **reject the null hypothesis** implying that the two attributes are associated/related.

Example 11

A car manufacturer wants to know if in the preference of customers in respect of vehicle style and color are independent or not. They randomly sample their sales in the past year, and observe the following.

	Silver	Black	White	Red	Total
Sedan	21	28	16	23	88
Minivan	17	15	19	18	69
Truck	13	22	18	20	73
Total	51	65	53	61	230

Solution

Here the hypotheses are

H_0 : Vehicle style and color are independent

H_1 : Vehicle style and color are related

Vehicle Style	Color	Observed	Expected	(O-E) ²	(O-E) ² /E
Sedan	Silver	21	19.51	2.2110	0.1133
	Black	28	24.87	9.7996	0.3940
	White	16	20.28	18.3035	0.9026
	Red	23	23.34	0.1150	0.0049
Minivan	Silver	17	15.30	2.8900	0.1889
	Black	15	19.50	20.2500	1.0385
	White	19	15.90	9.6100	0.6044
	Red	18	18.30	0.0900	0.0049
Truck	Silver	13	16.19	10.1567	0.6275
	Black	22	20.63	1.8757	0.0909
	White	18	16.82	1.3883	0.0825
	Red	20	19.36	0.4085	0.0211
Total					4.0736

So, the value of chi-squared statistic is 4.0736 and the corresponding degrees of freedom is $(3-1)*(4-1) = 6$.

Now, critical value at 5% level of significance is $\chi^2_{0.05,6} = 12.592$. Since the calculated value is less than the critical value, we **fail to reject the null hypothesis**. Hence, there is insufficient evidence to conclude that vehicle style and color are related.

Note: When there is no association between two variables, observed counts and expected counts are likely to be close to each other giving a value of chi square statistic that is quite small.

Test for Homogeneity of Proportions

Here there are r populations of interest and each population is divided into the same c categories. A random sample is then taken from the i -th population, and counts are entered in the appropriate columns (i.e. category) of the i -th row resulting in a $r \times c$ contingency table.

For the same $r \times c$ contingency table, we can test the following null hypothesis:

$$H_0 : p_{1j} = p_{2j} = \cdots = p_{rj} = p_j, \quad \forall j = 1, 2, \dots, c.$$

i.e. **category-wise proportions are identical across different populations.**

Now, let us consider the first row (i.e. population) only, i.e. $i = 1$.

Row 1	f_{11} p_{11}	f_{12} p_{12}	...	f_{1c} p_{1c}	$f_{1\cdot}$ $p_{1\cdot}$
-------	----------------------	----------------------	-----	----------------------	------------------------------

Then

$$\sum_{j=1}^c \frac{(f_{1j} - f_{1\cdot} \times p_{1j})^2}{f_{1\cdot} \times p_{1j}} \cap \chi_{c-1}^2$$

Similarly, for each $1 \leq i \leq r$, we have

$$\sum_{j=1}^c \frac{(f_{ij} - f_{i\cdot} \times p_{ij})^2}{f_{i\cdot} \times p_{ij}} \cap \chi_{c-1}^2$$

So, we get

$$\sum_{i=1}^r \sum_{j=1}^c \frac{(f_{ij} - f_{i\cdot} \times p_{ij})^2}{f_{i\cdot} \times p_{ij}} \cap \chi_{r(c-1)}^2$$

Therefore, under the null hypothesis

$$\sum_{i=1}^r \sum_{j=1}^c \frac{(f_{ij} - f_{i\bullet} \times p_j)^2}{f_{i\bullet} \times p_j} \cap \chi_{r(c-1)}^2$$
, where p_j is the common proportion of j -th category.

The value of p_j , in general, is unknown and need to be estimated, which is given by $\hat{p}_j = f_{\bullet j}/n$. By this we will be imposing additional $c-1$ constraints and consequently degrees of freedom for the chi-square test will be

$$r(c-1) - (c-1) = (r-1)(c-1).$$

Thus, we have the final expression for the test statistic as

$$\sum_i \sum_j \frac{\left(f_{ij} - \frac{f_{i\bullet} \times f_{\bullet j}}{n} \right)^2}{\frac{f_{i\bullet} \times f_{\bullet j}}{n}} \cap \chi_{(r-1)(c-1)}^2.$$

So, as earlier we can simplify above to

$$n \left[\sum_i \sum_j \frac{f_{ij}^2}{f_{i\bullet} \times f_{\bullet j}} - 1 \right] \cap \chi_{(r-1)(c-1)}^2$$

Note: *One can use either of the above two expressions for χ^2 .*

If $\chi_{\text{Calculated}}^2 < \chi_{\alpha, (r-1)(c-1)}^2$, we fail to reject the null hypothesis, i.e. we conclude that the similarly classified populations are homogeneous. On the other hand, if $\chi_{\text{Calculated}}^2 > \chi_{\alpha, (r-1)(c-1)}^2$, we conclude that the similarly classified populations are non-homogeneous.

Example 12

A researcher selected a sample of 150 seniors from three high schools of an area and asked each senior, “Do you drive to school in a car owned by either you or your parents?” The data are shown in the table. At $\alpha = 0.05$, test the claim that the proportion of students who drives their own or their parents’ cars is the same at all three schools.

	School 1	School 2	School 3	Total
Yes	18	22	16	56
No	27	28	39	94
Total	45	50	55	150

Solution

Here, the “Schools” represent three different populations and the hypotheses to be tested is

$$H_0: p_1 = p_2 = p_3$$

H_1 : At least one proportion is different from others

Table of expected values, i.e. $(f_{i\cdot} \times f_{\cdot j})/n$

	School 1	School 2	School 3	Total
Yes	16.80	18.67	20.53	56
No	28.20	31.33	34.47	94
Total	45	50	55	150

So, the value of chi-squared for homogeneity is

$$0.085714 + 0.593942 + 0.999557 + 0.051064 + 0.353939 + 0.595326 = 2.679542,$$

and corresponding degrees of freedom is $(2-1) \times (3-1) = 2$.

Critical value at 5% level of significance is 5.991.

Since, calculated value of the chi-squared for homogeneity is less than the corresponding critical value, we **fail to reject** the null hypothesis.

There is not enough evidence to reject the null hypothesis that the proportions of high school students who drive their own or their parents' cars to school are equal for each school.

Example 13

Suppose that over a 2-year period, 120 patients with heart disease were treated with one of two drugs (A or B). After a period of time, each patient's condition was rated as no change, improved, or greatly improved. The following contingency table gives the distribution of frequency counts. Determine whether the patients' conditions are similar with respect to the two drugs?

Drug Type	Patient's Condition			TOTAL
	No Change	Improved	Greatly Improved	
A	15	22	33	70
B	20	18	12	50
TOTAL	35	40	45	120

Solution

In this problem **patients treated under two Drug Type represents two populations**, and the hypotheses here are

H_0 : The proportions of patients falling into the three categories are the same for drug A and drug B.

H_1 : The proportions of patients falling into the three categories are **not** the same for drug A and drug B.

Or in other words

$$H_0: p_{A_i} = p_{B_i}, \quad i = 1, 2, 3$$

$$H_1: p_{A_i} \neq p_{B_i}, \quad \text{for atleast one } i$$

Now, we calculate the expected frequency **under the null hypothesis**. Following table gives the observed and corresponding **expected frequency** (in braces).

Drug Type	Patient's Condition			TOTAL
	No Change	Improved	Greatly Improved	
A	15 (20.42)	22 (23.33)	33 (26.25)	70
B	20 (14.58)	18 (16.67)	12 (18.75)	50
TOTAL	35	40	45	120

Next we calculate the observed value of the test statistic

$$\chi^2_{obs} = \frac{(15 - 20.42)^2}{20.42} + \frac{(22 - 23.33)^2}{23.33} + \frac{(33 - 26.25)^2}{26.25} + \frac{(20 - 14.58)^2}{14.58}$$

$$+ \frac{(18 - 16.67)^2}{16.67} + \frac{(12 - 18.75)^2}{18.75}$$

$$= 7.801$$

Again, the critical value at 5% level of significance and 2 degrees freedom is 5.991. So, we reject the null hypothesis as the observed value is more than the critical value.

Therefore, based on data, there is significant evident that the patients' conditions are not similar with respect to the two drugs.

Note: Table $O_i - E_i$: an observation

Drug Type	Patient's Condition			TOTAL
	No Change	Improved	Greatly Improved	
A	- 5.42	- 1.33	6.75	0
B	5.42	1.33	- 6.75	0
TOTAL	0	0	0	

Yates's Correction for Continuity

Yates's correction for continuity (or **Yates's chi-squared test**) is used in situations when testing for **independence in a 2×2 contingency table**. In such cases, Frank Yates suggested a correction to the Pearson's χ^2 based on the notion that a test of discrete variables (**dichotomous**) should follow a discrete distribution, but are tested using a normal approximation, the chi-squared distribution.

Yates argued that ordinary (uncorrected) Pearson's χ^2 statistic might produce a χ^2 *value* that are too large and probabilities that are too small leading reserachers to incorrectly reject the null hypothesis, i.e. increase of type I error.

To overcome this Yates suggested a correction for continuity that adjusts the formula for Pearson's chi-squared test by **subtracting 0.5 from the absolute difference between each observed value and corresponding expected value in a 2×2 contingency table**. This reduces the chi-squared value obtained and thus increases its p -value. Note that only in case of a **2×2 contingency table, i.e. when the degrees of freedom is equal to 1, one should always use Yates' correction for continuity**.

The effect of Yates's correction is to prevent over estimation of statistical significance, i.e. over estimation of the *true* discrepancy between observed and expected frequencies, for small data. *This formula is mainly used when at least one cell of the table has an expected frequency smaller than 5*. Many researchers recommended that Yates' correction be applied to every 2×2 contingency table, even if no theoretical cell frequency is less than 5, whereas a large body of reserchers has found that the correction is too strict.

The following is Yates's corrected version of Pearson's chi-squared statistic:

$$\chi_{Yates}^2 = \sum_{i=1}^N \frac{(|O_i - E_i| - 0.5)^2}{E_i}$$

where:

Note: By the suggested transformation as advocated by Yates

a) If $O_i > E_i$, then O_i is reduced by 0.5.

b) If $O_i < E_i$, then O_i is increased by 0.5.

In other words,

a) If $O_i - E_i > 0$, subtract 0.5 from the difference,

b) If $O_i - E_i < 0$, add 0.5 to the difference.

So that differences between observed and expected outcome become less compared to the original.

O_i = an observed frequency

E_i = an expected (theoretical) frequency, under the null hypothesis

N = number of distinct events

Yates points out that this adjusted test statistic produces a p -value sometimes greater and sometimes less than the exact p -value, but this p -value is typically much closer to the exact value than the p -value corresponding to the uncorrected χ^2 test statistic.

Note: To ensure continuity, the general correction is to either add or subtract 0.5 to each discrete x -value. This fills in the gaps to make it continuous. Here in case of χ^2_{Yates} , under null hypothesis we want the value of the test statistic to be smaller, so that the corresponding p -value becomes larger. Hence, 0.5 is subtracted from the absolute value (irrespective of the algebraic sign) of the numerator contribution of each cell.

Moreover, addition of 0.5 (or, subtraction of 0.5) to (or, from) cell frequencies, depending upon whether the expected frequency is more or less than the theoretical frequency keeps the marginal frequencies same.

Example 14

To test the efficacy of a vaccine, two batches of 70 and 60 animals were selected. First batch was vaccinated, whereas the second batch was given placebo. The number of infected and non-infected animals are given in the following table for both cases. Can it be assumed that proportion of non-infection under vaccination is more than that under placebo. Test at 5% level of significance?

	Infected	Not Infected	TOTAL
Vaccine	19	51	70
Placebo	28	32	60
TOTAL	47	83	130

Solution> Let p_1 and p_2 respectively be the proportions of non-infection under vaccine and placebo.

$$H_0: p_1 = p_2 \quad \text{vs} \quad H_1: p_1 > p_2$$

Table of Observed and Expected counts:

	Infected	Not Infected	TOTAL
Vaccine	19 (25.308)	51 (44.692)	70
Placebo	28 (21.692)	32 (38.308)	60
TOTAL	47	83	130

Calculation of χ^2 considering Yates's Correction

	O_i	E_i	$ O_i - E_i - 0.5$	$(O_i - E_i - 0.5)^2 / E_i$
Vaccine, Infected	19	25.308	5.808	1.3328
Vaccine, Not Infected	51	44.692	5.808	0.7547
Placebo, Infected	28	21.692	5.808	1.5549
Placebo, Not Infected	32	38.308	5.808	0.8805

ChiSquare calculated = 4.5229 and ChiSquare Critical = 3.84

We reject the null hypothesis and conclude that the proportion of non-infection under vaccine is more than that under placebo. This proves that the vaccine is effective against the disease.

2×2 Contingency χ^2

Row	Column		Total
	1	2	
1	a	b	$a+b$
2	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d (=n)$

For such contingency tables, expected value for a cell is determined by the following formula

expected value of a cell = (row total x column total) / total number (n)

e. g. $\hat{a} = \frac{(a+c)(a+b)}{n}$

Similarly, $\hat{b} = \frac{(b+d)(a+b)}{n}$, $\hat{c} = \frac{(a+c)(c+d)}{n}$ and $\hat{d} = \frac{(b+d)(c+d)}{n}$.

After simple mathematics, we arrive at the following simplified forms of the chi-squared statistic

$$\chi_1^2 (\text{uncorrected}) = \frac{(ad - bc)^2 \cdot n}{(a+b)(b+d)(d+c)(c+a)} \quad [\text{using } \sum \frac{O_i^2}{E_i} - n]$$

$$\chi_1^2 (\text{corrected}) = \frac{\left[|ad - bc| - \frac{n}{2} \right]^2 \cdot n}{(a+b)(b+d)(d+c)(c+a)}$$

[According to Yates' correction, we subtract (or add) $1/2$ from a and d and add (or subtract) to b and c , so that marginal totals are not disturbed. Thus, corrected value of chi-square is

$$\chi_1^2 = \frac{n \left[\left(a \mp \frac{1}{2} \right) \left(d \mp \frac{1}{2} \right) - \left(b \pm \frac{1}{2} \right) \left(c \pm \frac{1}{2} \right) \right]^2}{(a+b)(b+d)(d+c)(c+a)}$$

$$\text{Numerator} = n \left[(ad - bc) \mp \frac{1}{2} (a + d + b + c) \right]^2 = n \left[|ad - bc| - \frac{n}{2} \right]^2.$$

Since, if $(ad - bc) > 0$, subtract $\frac{n}{2}$, else add $\frac{n}{2}$.

So, we get the expression for χ_1^2 (corrected)]

Example 15

A chemotherapeutic agent used for bone marrow destruction in preparation for transplantation comes in either generic or non-generic form. Patients in a **particular hospital** received either the generic or non-generic formulation. The study outcome is cerebellar toxicity, a well known side-effect of the chemotherapy procedure. Cross-tabulation revealed the following

Agent type	Cerebellar Toxicity		Total
	Presence	Absence	
Generic	11	14	25
Non-Generic	3	31	34
Total	14	45	59

Can we conclude that occurrence of cerebellar toxicity is independent of agent type?
Use $\alpha = 0.05$.

Solution

Here the hypotheses are

H_0 : Cerebellar toxicity and agent type are independent

H_1 : Cerebellar toxicity and agent type are related

We calculate the chi-squared statistic incorporating Yates's correction for continuity and the value of the test statistic is

$$\chi^2 = \frac{[|11 \times 31 - 3 \times 14| - 59/2]^2 \times 59}{25 \times 34 \times 45 \times 14} = \frac{4285184.75}{535500} = 8.0022$$

Degrees of freedom for the test is 1. So the critical value at 5% level of significance is 3.841.

Since calculated value of chi-squared is more than the critical value, we reject the null hypothesis and we conclude that occurrence of cerebellar toxicity is dependent on the agent type.

Fisher's Exact Test

Fisher's exact test is a statistical significance test used in the analysis of contingency tables where **sample sizes are small**. It is named after its inventor, R. A. Fisher, and is one of a class of exact tests, so called because the **significance of the deviation from a null hypothesis can be calculated exactly**.

With large samples, a chi-squared test can be used. However, the significance value it provides is only an approximation, because the sampling distribution of the test statistic that is calculated is only approximately equal to the theoretical chi-squared distribution. **The approximation is inadequate when sample sizes are small, or the data are very unequally distributed among the cells of the table, resulting in the cell counts predicted on the null hypothesis (the "expected values") being much different compared to the observed value.** The chi-squared test is not suitable when the expected values in any of the cells of a contingency table **are below 5**, (or below 10) when there is only one degree of freedom, i.e for 2×2 contingency tables.

Let us consider the following 2×2 contingency table.

Characteristic of interest	Sample		Total
	1	2	
present	a	b	$a+b$
absent	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d (= n)$

Fisher's test determines whether the presence of the characteristic of interest differ in the proportion with which they fall into the two populations. For the table above, the test would determine whether proportion of characteristic of importance in Population 1 (p_1) **differ** significantly from the proportion of characteristic of importance in Population 2 (p_2).

So, here the hypotheses is

$$H_0: p_1 = p_2$$

$$H_1: p_1 > p_2$$

Fisher showed that the probability of obtaining any such set of values was given by the hypergeometric distribution:

$$p = \frac{\binom{a+c}{a} \binom{b+d}{b}}{\binom{n}{a+b}} = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!}$$

This formula gives the exact probability of observing this particular arrangement of the data under the null hypothesis that the proportion of characteristic of importance in both the populations are equal, assuming the given marginal totals. This hypergeometric formula gives the conditional probability of observing the values a , b , c , d in the four cells, subject to the observed marginals. Thus using above expression, one can calculate the exact probability of any such arrangement of the n objects into the four cells of the table given the marginals.

In Fisher's Exact test, a **directional hypothesis is generally assumed**. The directional hypothesis assumed in the Fisher Exact test is nothing but the hypothesis based on the one tailed test. In order to calculate the significance of the observed data, i.e. the total probability of observing data as extreme or more extreme if the null hypothesis is true, we need to consider only those cases where the arrangements are at least as extreme as the observed arrangement keeping the marginal totals same.

So, we need to compute, in addition to the observed case, the **probability of more extreme outcomes, with the same marginal totals**. By "more extreme", we mean relative to the null hypothesis of equal proportions and finally "the significance probability" is obtained by adding probabilities of observed outcome and the extreme outcomes.

This value can be interpreted as the sum of evidence provided by the observed data or any more extreme table—under the null hypothesis (that there is no difference in the proportions p_1 and p_2). The smaller the value of p , the greater the evidence for rejecting the null hypothesis. So, if the total probability, thus obtained, is less than the significance level (α) we reject the null hypothesis.

But for a **two-tailed test** we must also consider tables that are equally extreme but in the either direction. Unfortunately, classification of the tables according to whether or not they are 'as extreme' is problematic.

Example 16

A study was carried out with 20 patients – 9 female and 11 male. Following table gives outcome of specific treatment on each of them. Can we say that success rate of female is more than that of male? Assume $\alpha = 0.05$.

Patient	Effect of Treatment		Total
	Success	No success	
Female	8	1	9
Male	4	7	11
Total	12	8	20

Solution

These data would not be suitable for analysis by **Pearson's chi-squared test**, because the expected values in some of the cells in the table will be below 10 (rather below 5), and in a 2×2 contingency table, the number of degrees of freedom is also 1.

Let p_1 be the success rate for the female and p_2 be the success rate for the male. So, the hypotheses will be

$$H_0: p_1 = p_2$$

$$H_1: p_1 > p_2$$

Using hypergeometric distribution, we calculate the probability for the observed arrangement.

$$p = \frac{9!11!12!8!}{8!1!4!7!20!} = \frac{9 \times 11}{19 \times 17 \times 13} = 0.02358.$$

Now, there is only one 'more extreme' case, which is given by

9	0	9
3	8	11
12	8	20

Clearly, in this configuration success rate for the female, i.e. p_1 will be more than that for the original configuration.
[9/9 vs. 8/9]

The probability corresponding to this arrangement is

$$p = \frac{9!11!12!8!}{9!3!8!20!} = \frac{11}{2 \times 19 \times 17 \times 13} = 0.00131.$$

So, the total probability is $0.02358 + 0.00131 = 0.02489$, which is less than $\alpha = 0.05$. So, we can reject the null hypothesis and conclude that success rate of female is more than that of male.

Note: We will use **Example 16** to show that statistic based on Yates' correction for continuity gives a p -value that is closer to the exact value of **0.0249** compared to the p -value given by the uncorrected χ^2 .

Test Statistic	χ^2 value	p -value (one tail)
Corrected χ^2	3.712	0.0270
Uncorrected χ^2	5.690	0.0085

Possible arrangements leading to two-tailed tests for success rate.

For example, consider the following table relating success rate of female and male

Patient	Effect of Treatment		Total
	Success	No success	
Female	1	3	4
Male	4	3	7
Total	5	6	11

Fisher's exact test for two tailed test would look at the set of all tables that have same row totals of (4,7) and column totals of (5,6) with different success rates. They are

	0 4	1 3	2 2	3 1	4 0
	5 2	4 3	3 4	2 5	1 6
	$H_1: p_1 < p_2$			$H_1: p_1 > p_2$	
Probability	0.04545	0.30303	0.45454	0.18182	0.01515

The calculation of a one-tailed p value begins by ordering the set of all tables with the same marginals (according to the value of the cell in the upper right hand corner, say). The probability of observing each table is calculated by using the hypergeometric distribution. Then the probabilities are summed from the observed table to the extreme arrangement in either direction to get the one tailed test probabilities. In this example, the two sums are $0.04545 + 0.30303 = 0.34848$ and $0.30303 + 0.45454 + 0.18182 + 0.01515 = 0.95454$, and the one-tailed p value is the smaller of these two, i.e. case in which minimum rearrangements are possible and its value is 0.34848.

Yates (1984) argues that a two-tailed p value should be obtained by doubling the one-tailed p value, but most statisticians would compute the two tailed p value as the sum of the probabilities, under the null hypothesis, of all tables having a probability of occurrence **less or equal** to the observed table. In this case it is $0.04545 + 0.30303 + 0.18182 + 0.01515 = 0.54545$.

Measures of Association

A problem with the chi square test for independence is that the size of the chi square statistic may not provide a reliable guide to the strength of the statistical relationship between the two variables. When two different cross classification tables have the **same sample size**, the variables in the table with the **larger chi square value are more strongly related** than are the two variables in the table with the smaller chi square value (note: *null hypothesis assumes that the variables are not related*) . But when the **sample sizes for two tables differ**, the size of the **chi square statistic** is a **misleading indicator** of the extent of relationship between two variables.

Effect of Sample Size on the Chi Square Statistic

Let us consider two hypothetical examples to show the misleading nature of chi-square statistic when sample sizes are different.

Opinion	Male	Female	Total		Opinion	Male	Female	Total
Agree	65 (60)	25 (30)	90		Agree	260 (240)	100 (120)	360
Disagree	35 (40)	25 (20)	60		Disagree	140 (160)	100 (80)	240
Total	100	50	150		Total	400	200	600
$\chi^2 = 0.417 + 0.833 + 0.625 + 1.250 = 3.125$					$\chi^2 = 1.667 + 3.333 + 2.5 + 5.0 = 12.5$			
$p = 0.0771$					$p = 0.000407$			

It may be noted that in right Table sample size is increased from $n = 150$ to $n = 600$. In order to preserve the nature of the relationship, each of the observed numbers of cases in the cells of left Table are multiplied by 4 to get the right Table. But now the chi square statistic is dramatically increased, and becomes 4 times the original value. Thus the larger sample size in the right Table has increased the value of the chi square statistic so that even the relatively weak relationship between gender and opinion becomes very significant statistically.

This example shows that value of the chi square statistic is sensitive to the sample size. The degrees of freedom stay unchanged, so that the larger chi square value appears to imply a much stronger statistical relationship between gender and opinion.

Considerable caution should be exercised when comparing the chi square statistic, and its significance, for two tables. If the sample size for the two tables is the same, and the dimensions of the table are also identical, then the table with the larger chi square value generally provides stronger evidence for a relationship between the two variables. But when the sample sizes, or the dimensions of the table differ, the chi square statistic and its statistical significance may not provide an accurate idea of the extent of the relationship between the two variables.

One way to solve some of the problems associated with the chi square statistic is to adjust the chi square statistic for the sample size. Following are measures of association that carry out this adjustment, using the chi square statistic.

Phi (ϕ)

The measure of association, phi, is a measure which adjusts the chi square statistic by the sample size. Phi is most easily defined as

$$\phi = \sqrt{\frac{\chi^2}{n}}$$

Phi is usually less than one and very easy to compute when chi square statistic for the contingency table is determined.

It may be noted that in the previous example of relationship between gender and opinion, value of ϕ for both the Tables is 0.144 though the chi square statistic showed that one has a strong relationship compared to the other.

Coefficient of Contingency

A slightly different measure of association is the contingency coefficient. This is another chi square based measure of association, and one that also adjusts for different sample sizes. The contingency coefficient can be defined as

$$C_{\text{obs}} = \sqrt{\frac{\chi^2}{\chi^2 + n}}$$

The larger the value of C , the greater is the degree of association, whereas the minimum value it can attain is 0. Clearly, C will be zero for $\chi^2 = 0$, whereas for large value of χ^2 ; C will be large. Note that coefficient of contingency may be less than 1 even when the two variables are perfectly related to each other. This is not desirable, so C is adjusted so that it reaches 1 when there is complete association in a table of any number of rows and columns. So, the coefficient of contingency is adjusted by dividing C by the maximum value it can attain.

For a $k \times k$ contingency table, maximum value of the contingency coefficient is calculated as $C_{\max} = \sqrt{\frac{k-1}{k}}$. In case of $r \times c$ contingency tables, $k = \min(r, c)$.

Then the standardized contingency coefficient is calculated as $C_{\text{Standard}} = \frac{C_{\text{obs}}}{C_{\max}}$, which varies between 0 and 1 such that 0 indicating independence and 1 perfect dependence.

Cramer's V

Cramer's V is the preferred measure among the χ^2 based measures. When it is computed, Cramer's V takes into account, in addition to sample size, the dimensions of the table too, and attempts to adjust the effect of both. So that Cramer's V can be used to compare the strength of association between any two cross classification tables with different sample size and/or dimensions.

For $k \times k$ tables the correlation coefficient between attributes (or classifications) is defined as

$$r = \sqrt{\frac{\chi^2}{n(k-1)}}$$

Cramer's V equals 0 when there is no relationship between the two variables, and generally has a maximum value of 1, regardless of the dimension of the table or the sample size. Thus, tables with larger values for Cramer's V can be considered to have a strong relationship between the variables, and smaller value for V indicates a weaker relationship.

For $r \times c$ contingency tables Cramer's V is defined as

$$r = \sqrt{\frac{\chi^2}{n[\min\{r-1, c-1\}]}}$$

Exercises

1. An automatic filling machine is used to fill bottles with liquid detergents. A random sample of 20 bottles results in a sample variance of fill volume of 0.0153. If the variance of fill volume exceeds 0.01, an unacceptable portion of bottles will be underfilled or overfilled. Is there evidence in the sample data to suggest that the manufacturer has problem with underfilled or overfilled bottles.? Use $\alpha = 0.05$, and assume that fill volume has a normal distribution.

Hint. $H_0: \sigma^2 = 0.01$ $H_1: \sigma^2 > 0.01$

2. A semiconductor manufacturer produces controllers used in automobile engine applications. The customer requires that the process fallout or fraction defective at a critical manufacturing step should not exceed 0.05 and the manufacturer must demonstrate process capability at this level of quality using $\alpha = 0.05$. The manufacturer takes a random sample of 200 devices and find that 4 of them are defective. Will the manufacturer be able to demonstrate process capability to the customer.

Ans. $H_0: p = 0.05$ $H_1: p < 0.05$

This alternative will allow the manufacture to make a strong claim about the process capability if null hypothesis is rejected.

Here test statistic to be used is

$$z_0 = \frac{x - np_0}{\sqrt{np_0(1 - p_0)}}, \text{ where } x = 4, n = 200, \text{ and } p_0 = 0.05.$$

On computation we get, $z_0 = -1.95$. Critical value is $-z_{0.05} = -1.645$. Since $-1.95 < -1.645$, we reject the null hypothesis and conclude that the process fraction defective p is less than 0.05.

3. Consider the following frequency table of observations on the random variable X

Values	0	1	2	3	4
Observed frequency	24	30	31	11	4

Based on these 100 observations, is a Poisson distribution with mean of 1.2 an appropriate model? Perform a goodness-of-fit procedure with $\alpha = 0.05$.

4. A company has to choose among three pension plans. Management wishes to know whether the preference for plans is independent of job classification and wants to use $\alpha = 0.05$. The opinions of a random sample 500 employees are as under

Job Classification	Pension Plan			Totals
	1	2	3	
Regular Workers	160	140	40	340
Casual Workers	40	60	60	160
Totals	200	200	100	500

5. Patients in a hospital are classified as surgical or medical. A record is kept of the times patients require nursing service during the night and whether or not these patients are on Medicare. The data are presented below:

Under Medicare	Patient Category	
	Surgical	Medical
Yes	46	52
No	36	43

Test the hypothesis, using $\alpha = 0.01$, that calls by patients are independent of whether patients are under Medicare or not.

6. Two machines are used for filling plastic bottles with net volume of 16.0 ounces. The fill volume can be assumed normal, with standard deviation of $\sigma_1 = 0.02$ and $\sigma_2 = 0.025$ ounces. A member of quality engineering staff suspects that both machines do not fill the same mean net volume, whether or not this volume is 16.0 ounces. A random sample of 10 bottles is taken from output of each machine.

Machine 1		Machine 2	
16.03	16.01	16.02	16.03
16.04	15.96	15.97	16.04
16.05	15.98	15.96	16.02
16.05	16.02	16.01	16.01
16.02	15.99	15.99	16.00

Do you think the engineer is correct? What is the p-value of the test?

[Hint: $H_0 : \mu_1 = \mu_2$ $H_1 : \mu_1 \neq \mu_2$ and test statistic will be

$$z_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{10} + \frac{\sigma_2^2}{10}}}]$$

7. The viscosity of a liquid detergent is supposed to have an average of 800 centistokes at $25^\circ C$. A random sample of 16 batches of detergent is collected, and the average viscosity is 812. Suppose we know that the standard deviation of viscosity is $\sigma = 25$ centistokes.
- State the hypothesis to be tested?
 - Test these hypotheses using $\alpha = 0.05$. What is your conclusion?
 - What is the P-value of the test?
 - Find a 95% confidence interval about the mean.

8. Two types of plastic are suitable for use by an electronic calculator manufacturer. The breaking strength of this plastic is important. It is known that $\sigma_1 = \sigma_2 = 1.0$ psi. From random samples of $n_1 = 10$ and $n_2 = 12$ we obtain $\bar{y}_1 = 162.5$ and $\bar{y}_2 = 155.0$. The company **will not adopt plastic 1 unless its breaking strength exceeds that of plastic 2 by at least 10 psi**. Based on the sample information, should they use plastic 1? In answering this questions, set up and test appropriate hypotheses using $\alpha = 0.01$. Construct a 99 percent confidence interval on the true mean difference in breaking strength.

$$H_0 : \mu_1 - \mu_2 = 10$$

$$H_1 : \mu_1 - \mu_2 > 10$$

$$z_0 = \frac{\bar{y}_1 - \bar{y}_2 - 10}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}, \text{ since } \sigma^2 \text{ is known.}$$

9. The following are the burning times of chemical flares of two different formulations. The design engineers are interested in both the means and variance of the burning times.

Type 1	Type 2
65 82	64 56
81 67	71 69
57 59	83 74
66 75	59 82
82 70	65 79

- (a) Test the hypotheses that the two variances are equal. Use $\alpha = 0.05$.
 (b) Using the results of (a), test the hypotheses that the mean burning times are equal. Use $\alpha = 0.05$.

Soln.

$$\begin{aligned} \text{a)} \quad H_0 : \sigma_1^2 &= \sigma_2^2 & s_1 &= 9.264 \\ H_1 : \sigma_1^2 &\neq \sigma_2^2 & s_2 &= 9.364 \\ F_0 &= s_2^2 / s_1^2 = 1.011 \end{aligned}$$

$$F_{0.025,9,9} = 4.03 \quad F_{0.975,9,9} = \frac{1}{F_{0.025,9,9}} = \frac{1}{4.03} = 0.248.$$

So, F_0 is neither $> F_{0.025,9,9}$ nor $< F_{0.975,9,9}$ \rightarrow fail to reject H_0 .

$$\text{b) } s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2} = 86.775 \Rightarrow s_p = 9.32.$$

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{70.4 - 70.2}{9.32 \sqrt{\frac{1}{10} + \frac{1}{10}}} = 0.048.$$

Critical value $t_{0.025,18} = 2.101$.

Since, $t_0 = 0.048 \ll t_{0.025,18} = 2.101$, we **fail to reject** the null hypothesis and conclude that mean burning times are equal.