STAT406 - Statistical Inference

Ayesha Sultan
Lecturer
Virtual university of Pakistan



STA 406 - Statistical Inference

Lecture No.1

Statistical Inference

It is the procedure by which we reach a conclusion about a population on the basis of information contained in the sample drawn from that population.

Two broad areas of statistical inference

- Estimation
- Hypothesis testing

Estimation

It is the process of calculating some statistics based upon the sample data drawn from a certain population. The statistics is used as an approximation to the population parameter.

Join VU Group: https://chat.whatsapp.com/Cbs4WPb2T3aCASMHRaj6FS

Types of Estimate

For each of parameters, we can have

- **Point Estimate**
- **Interval estimate**

➢ A point estimate is a single numerical value used to estimate the corresponding population parameter.

Example: The mean of the age of men attending a show is 32 years.

➢ An interval estimate consists of two numerical values defining a range of values that, with a specified degree of confidence, includes the parameter being estimated.

Example: The mean of the age of men attending a show is between 28 and 36 years.

Properties of good point estimator

A point estimator is said to be good point estimator if it satisfies the following criteria's :

- Unbiasedness
- Consistency
- Sufficiency
- Efficiency

Unbiasedness

An estimator is defined to be unbiased if statistic used as an estimator has its expected value equal to the true value of the population parameter.

Suppose $\hat{\theta}$ be an estimator of a parameterθ. Then $\hat{\theta}$ will be called an unbiased estimator if $E(\hat{\theta}) = \theta$. If $E(\hat{\theta}) \neq \theta$, the statistic is said to be a biased estimator

Unbiasedness

1. The sample mean $\bar{X}$ is an unbiased estimator of the population mean μ.
2. The sample proportion $\hat{p}$ is also an unbiased estimator of the population P.
3. s^2 is an unbiased estimator of σ^2 but S^2 is a biased estimator of σ^2 .

Example 1

Show that the sample mean is an unbiased estimator of the population mean.

Example 2

If $\bar{X}$ and S^2, defined by $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$, are the sample mean and sample variance of a random sample of size n from a population with mean μ and variance σ^2, then show that S^2 is a biased estimator of σ^2 i.e. $E(S^2) \neq \sigma^2$

Example 3

If $\bar{X}$ and S^2, defined by $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, are the sample mean and sample variance of a random sample of size n from a population with mean μ and variance σ^2, then show that s^2 is an unbiased estimator of σ^2 i.e. $E(s^2) = \sigma^2$

Example 4

A population consists of five numbers 2, 4, 6, 8, 10. Draw all possible samples of size two with replacement from this population. Show that S^2 is an biased estimator of σ^2 and s^2 is and unbiased estimator of σ^2.

Example 5

Show that if the random variable X has the binomial distribution b(x; n, p), then the sample proportion X/n is an unbiased estimator of the parameter P.

Example 6

Show that the sample mean $\bar{x}$ of a random sample of size n from the density function

$$f(x;\theta)=\frac{1}{\theta}e^{-x/\theta},\quad 0<\theta<\infty$$

is an unbiased estimator of θ.

STAT 406 - Statistical Inference

Ayesha Sultan
Lecturer
Virtual university of Pakistan



STAT 406 - Statistical Inference

Lecture No.2

Example 7

Let $X_1, X_2, X_3, \ldots\ldots, X_n$ denote a random sample from a normal distribution $N(0,\theta)$.Show that $\frac{1}{n}\sum_{i=1}^{n} X_i^2$ is an unbiased estimator of θ.

Consistency

An estimator $\hat{\theta}$ is said to be a consistent estimator of the parameter θ if, for any arbitrary small positive quantity e,

$$\lim_{n\to\infty} P\left[\left|\hat{\theta}-\theta\right| \le e\right] = 1.$$

In other words, an estimator $\hat{\theta}$ is called a consistent estimator of θ if the probability that $\hat{\theta}$ is very close to θ, approaches unity with an increase in the sample size.

It should be noted that consistency is a *large sample* property. Another point to be noted is that a consistent estimator *may or may not* be unbiased.

Consistency

- The sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, which is an unbiased estimator of μ, is a consistent estimator of the mean μ.
- The sample proportion $\hat{p}$ is also a consistent estimator of the parameter p of a population that has a binomial distribution.
- The median is *not* a consistent estimator of μ when the population has a skewed distribution.
- The sample variance

$$S^2 = \frac{1}{n}\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2,$$

Though a *biased* estimator but it is a consistent estimator of the population variance σ^2.

To prove that an estimator is consistent we have to follow the following criteria:

"Suppose $\hat{\theta}$ is an estimator of θ based on a sample of size n. Then $\hat{\theta}$ is an consistent estimator of θ, if $\text{Var}(\hat{\theta}) \to 0$ as $n \to \infty$ "

Example1

Prove that :

(i) Asample mean $\bar{X}$ is a consistent estimator of population mean μ.

(ii) Asample proportion $\hat{p}$ is a consistent estimator of population proportion P for a population that has binomialdistribution.

Example2

Show that the sample mean $\bar{x}$ of a random sample of size n from the density function

$$f(x;\theta) = \frac{1}{\theta} e^{-x/\theta} \quad , 0 < x < \infty$$

is a consistent estimator of the parameter θ.

Efficiency

"If $\hat{\theta}_1$ and $\hat{\theta}_2$ denotes two unbiased estimators for the same parameter θ then the estimator $\hat{\theta}_1$ will be said to be *more efficient* than $\hat{\theta}_2$ if Variance($\hat{\theta}_1$)< Variance($\hat{\theta}_2$)".

To compare the relative efficiency of two estimators, we examine the ratio of their variances as"

Efficiency($\hat{\theta}_1$ $\hat{\theta}_2$)=variance($\hat{\theta}_2$)/variance($\hat{\theta}_1$)>1

$$\Rightarrow \hat{\theta}_1 \text{ is an efficient estimator of } \hat{\theta}_2$$

It thus provides a criterion for *comparing* different unbiased estimators of a parameter.

Efficiency

For a normal distribution, both the sample mean and the sample median are unbiased and consistent estimators of μ, but the variance of the sampling distribution of the sample means $\left(\frac{\sigma^2}{n}\right)$ is smaller than the sampling distribution of sample medians $\left(\frac{\pi\sigma^2}{2n}\right)$

i.e. $\frac{Var(median)}{Var(\bar{X})} = \frac{\pi\sigma^2}{2n} \times \frac{n}{\sigma^2} = \frac{\pi}{2} = 1.57 > 1.$ Hence the sample mean is more efficient than the sample median as an estimator of μ,

The sample mean may therefore be preferred as an estimator.

Example 1:

Let X_1, X_2, X_3 *and* X_4 be a random sample of size $n = 4$ from a $N(\mu, \sigma^2)$. A statistician wishes to estimate the mean by using either of the following estimators of the mean μ.

$$T_1 = \frac{X_1 + X_2 + X_3 + X_4}{4} \quad \text{(The sample mean } \overline{X}\text{)}$$

$$T_2 = \frac{X_1 + 2X_2 + 3X_3 + X_4}{7} \quad \text{(A weighted mean, } \overline{X}_w\text{)}$$

Which estimator should be preferred?

To answer this question correctly, we need to compare the expected values and the variance of these statistics.

Let us first see if the two estimators are unbiased.

Now,

$$E(T_1) = E(\frac{X_1 + X_2 + X_3 + X_4}{4}) = \frac{1}{4} E(X_1 + X_2 + X_3 + X_4)$$

$$= \frac{1}{4}[E(X_1) + E(X_2) + E(X_3) + E(X_4)] = \frac{1}{4}(\mu + \mu + \mu + \mu) = \frac{4\mu}{4} = \mu$$

T_1, the sample mean $\overline{X}$, is the unbiased estimator.

Again

$$E(T_2) = E(\frac{X_1 + 2X_2 + 3X_3 + X_4}{7}) = \frac{1}{7}[E(X_1) + 2E(X_2) + 3E(X_3) + E(X_4)]$$

$$E(T_2) = \frac{1}{7}[\mu + 2\mu + 3\mu + \mu] = \frac{7\mu}{7} = \mu$$

T_2, the weighted mean of the observations, is also an unbiased estimator of μ . Thus we see that both means are unbiased estimators of μ .

Next we find the variances of theses estimators.

$$Var(T_1) = Var\left(\frac{X_1 + X2 + X_3 + X_4}{4}\right) = \frac{1}{16}\left[Var(X_1) + Var(X_2) + Var(X_3) + Var(X_4)\right]$$

$$Var(T_1) = \frac{1}{16}\left[\sigma^2 + \sigma^2 + \sigma^2 + \sigma^2\right] = \frac{4}{16}\sigma^2 = \frac{1}{4}\sigma^2$$

$$Var(T_2) = Var\left(\frac{X_1 + 2X_2 + 3X_3 + X_4}{7}\right) = \frac{1}{49}\left(Var(X_1) + Var(2X_2) + Var(3X_3) + Var(X_4)\right)$$

$$Var(T_2) = \frac{1}{49}\left[\sigma^2 + 4\sigma^2 + 9\sigma^2 + \sigma^2\right] = \frac{15}{49}\sigma^2$$

Now,

$$\frac{Var(T_2)}{Var(T_1)} = \frac{15}{49}\sigma^2 / \frac{1}{4}\sigma^2 = 60/49 = 1.4$$

Which is greater than 1, showing that $Var(T_1) < Var(T_2)$. Thus T_1 is more efficient than

T_2 as an estimator for μ.

Hence $T_1 = \overline{X}$ is a better estimator of μ than $T_2 = \overline{X}_w$ and should be preferred.

STA 413 - Statistical Inference

Ayesha Sultan
Lecturer
Virtual university of Pakistan



STA 413 - Statistical Inference

Lecture No.3

Example 2:

If $X_1, X_2 \text{ and } X_3$ is a random sample taken from a population with mean μ and variance σ^2,

(a) Find which of the following estimators of μ are unbiased:

$$T_1 = \frac{1}{4}X_1 + \frac{1}{2}X_2 + \frac{1}{4}X_3$$

$$T_2 = \frac{4}{5}X_1 + \frac{1}{10}X_2 + \frac{1}{10}X_3$$

$$T_3 = \frac{1}{6}X_1 + \frac{2}{3}X_2 + \frac{1}{2}X_3$$

(b) Which of the unbiased estimator is more efficient?

Example 3:

Based on a random sample of 100 observations taken from a normal distribution, consider two estimators of μ : $\bar{X}_1 = \frac{1}{100}\sum_{i=1}^{100} X_i$ and $\bar{X}_2 = \frac{1}{90}\sum_{i=1}^{90} X_i$. Here $\bar{X}_1$ is the mean of all the observations, whereas $\bar{X}_2$ is the mean of first 90 observations. (i) Are $\bar{X}_1$ and $\bar{X}_2$ unbiased? (ii) What is the efficiency of $\bar{X}_2$ relative to $\bar{X}_1$?

Example 3:

Based on a random sample of 100 observations taken from a normal distribution, consider two estimators of μ : $\bar{X}_1 = \frac{1}{100}\sum_{i=1}^{100} X_i$ and $\bar{X}_2 = \frac{1}{90}\sum_{i=1}^{90} X_i$. Here $\bar{X}_1$ is the mean of all the observations, whereas $\bar{X}_2$ is the mean of first 90 observations. (i) Are $\bar{X}_1$ and $\bar{X}_2$ unbiased? (ii) What is the efficiency of $\bar{X}_2$ relative to $\bar{X}_1$?

Example 4:

Based on a random sample of 3 observations, consider 3 possible estimators of μ

$$\bar{X}_1 = \frac{1}{3}X_1 + \frac{1}{3}X_2 + \frac{1}{3}X_3, \quad \bar{X}_2 = \frac{5}{8}X_1 + \frac{1}{8}X_2 + \frac{1}{4}X_3,$$

$$\bar{X}_3 = 0.2X_1 + 0.3X_2 + 0.4X_3$$

Find

(i) Which are unbiased? (ii) The efficiency of each unbiased estimator relative to other.

Sufficiency

An estimator is defined to be sufficient, if the statistic used as estimator uses all the information that is contained in the sample.

A necessary and sufficient condition for a statistic $\hat{\theta}$ to be sufficient estimator for an unknown parameter θ is that the joint distribution of the sample of n values can be factorized into two non-negative functions (factors), one depending on θ $and \hat{\theta}$, other depending on X but not on θ.

Sufficiency

Mathematically, lf $x_1,x_2,..x_n$ be a random sample from a probability distribution $f(x, \theta)$ and let $f(x1,x2,..xn;\theta)$ denote its joint distribution.

Then $\hat{\theta} = h(x_1,x_2,..x_n)$ is a sufficient estimator of θ if and only if:

$f(x_1,x_2,..x_n; \theta) = g(\hat{\theta}, \theta)\ h(x_1,x_2,..x_n)$

Example1:

Let $X_1, X_2, \ldots, X_n$ be a random sample from the Poisson distribution $f(x; \theta) = \frac{e^{-\theta}\theta^{x}}{x!}$, $(x = 0,1,2, \ldots)$.

Then show that $\bar{X}$ is a sufficient estimator for θ.

Given the Poisson distribution

$$f(x;\theta) = \frac{e^{-\theta}\theta^{x}}{x!}, \quad x = 0,1,2, \ldots \qquad (\theta \text{ is used for } \mu)$$

The joint distribution of the sample values is

$$f(x_1,x_2, \ldots, x_n;\theta) = f(x_1; \theta)\, f(x_2; \theta) \ldots f(x_n; \theta)$$

$$= \prod_{i=1}^{n} \frac{e^{-\theta}\theta^{x_i}}{x_i!} = \frac{e^{-n\theta}\theta^{\sum x}}{x_1!x_2!\ldots x_n!}$$

$$= \frac{e^{-n\theta}(n\theta)^{\sum x_i}}{(\sum x_i)!} \cdot \frac{(\sum x_i)!}{x_1!x_2!\ldots x_n!\, n^{\sum x_i}}$$

Now if we write

$g(\hat{\theta};\theta)=\dfrac{e^{-n\theta}(n\theta)^{\sum x_i}}{(\sum x_i)}$, a distribution of $\sum x_i$ that depends on θ, and

$h(x_1,...,x_n)=\dfrac{(\sum x_i)!}{x_1!x_2!...x_n!}\dfrac{1}{n^{\sum x_i}}$, a function not depending on θ; we see that the joint distribution has been split up into two functions, which satisfy the factorization criterion.

This shows that $\sum x_i$ is a sufficient estimator for θ.

$\Rightarrow \bar{X}=\dfrac{\sum_{i=1}^{n} X_i}{n}$ *is also a sufficinet estimator as any function of sufficinet estimator is also a suffient estimator.*

Example 2:

If X_1, X_2 *and* X_3 denote a random sample taken from the Bernoulli distribution

$$f(x;p) = p^x (1-p)^{1-x} \quad , \qquad x = 0,1$$

Show that $\sum X_i$ is a sufficient estimator of P.

STA 413 - Statistical Inference

Ayesha Sultan
Lecturer
Virtual university of Pakistan



STA 413 - Statistical Inference

Lecture No.4

Methods of Point Estimation

- Maximum Likelihood Estimation
- The Method of Moments
- Method of Least Square

Maximum likelihood estimation

Let $X_1, X_2 \ldots X_n$ have a joint pmf or pdf

$$f(x_1,..,x_n;\theta_1,...,\theta_m) = \prod_{i=1}^{n} f(x_i,\theta) \quad \text{----------(1)}$$

Where the parameters $\theta_1,...,\theta_m$ have unknown values. When $x_1,..,x_n$ are observed samples values and (1) is regarded as a function of$\theta_1,...,\theta_m$, it is called the likelihood function.

$$L(\theta_1,...,\theta_m) = f(x_1,..,x_n;\theta_1,...,\theta_m) = \prod_{i=1}^{n} f(x_i,\theta)$$

Then maximum likelihood estimates (MLE's) $\hat{\theta}_1,...,\hat{\theta}_m$ are those values of the θ_i's that maximize the likelihood function.

Example:

Let $X_1, X_2, ..., X_n$ be a random sample from a normal population with parameters μ and σ^2, find MLE for

(i) μ, when population variance σ^2 is known

(ii) σ^2, when population mean μ is known.

Example

Let $X_1, X_2, \ldots, X_n$ be a random sample form the normal distribution.

$$f(x_1,..,x_n;\mu,\sigma^2)=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x_i-\mu)^2/(2\sigma^2)}$$

$$=\left(\frac{1}{2\pi\sigma^2}\right)^{n/2}e^{-\sum_{i=1}^{n}(x_i-\mu)^2/(2\sigma^2)}$$

The natural logarithm of the likelihood function is

$$\ln(f(x_1,..,x_n;\mu,\sigma^2)=\frac{n}{2}\ln(2\pi\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

To find the values of μ and σ^2 that maximize the function, we must take the partial derivatives of ln(f) with respect to μ and σ^2, equate them to zero, and solve the resulting two equations.

Omitting the details, the resulting (MLES) are

$$\hat{\mu}=\bar{X}: \qquad \hat{\sigma}^2=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n}$$

Example 2

Let $X_1, X_2, X_3, ..., X_n$ denote a random sample from the Poisson distribute

$$f(x;\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!} \qquad x = 0,1,2,3,.....$$

Find the maximum likelihood estimator of

Example 3

Suppose that X is a random variable with the density function:

$$f(x;\theta) = \theta e^{-\theta x}, \qquad x > 0 \quad \theta > 0$$

What is the maximum likelihood estimator for θ based on random sample of *n* observations of X?

Example 4

Suppose that X is a Bernoulli random variable with parameter p. Given a random sample of n observations of X, compute the maximum likelihood estimator for p.

Example 5

Let X be a binomial random variable with parameter n (known) and p. Find the maximum likelihood estimator for p based on a random sample of N observation of X.

STA 413 - Statistical Inference

Ayesha Sultan
Lecturer
Virtual university of Pakistan



STA 413 - Statistical Inference

Lecture No.5

Method of moments

The method of moments is the oldest method of deriving point estimators. They are often used because they lead to very simple computations. In this method, we equate sample moments to population moments, thus getting as many equations as are needed to solve for the unknown parameters. The procedure is described below:

Let X_1, X_2… X_n be a random sample of size n from a population. Then the rth sample moment about zero is

$$\mu'_r = \frac{\sum X_i^r}{n}, \quad r = 1, 2, ...$$

The corresponding r^{th} population moment is μ'_r. we then *match* these moments and get *as many* equations as we need to solve for the unknown parameters. The following examples illustrate the method:

Example

The probability density function of the given uniform distribution is

$$f(x) = \frac{1}{\theta}, \quad 0 \le x \le \theta$$

Then, find an estimator of θ by the method of moments.

Since the uniform distribution has only one parameter, (i.e.θ), therefore, in order to find the estimator of θ by the method of moments, we need to consider only *one* equation.

The first sample moment about zero is

$$\mu'_1 = \frac{\sum X_i}{n}.$$

And, the first population moment about zero is

$$\mu'_1 = \int_0^\theta x.f(x)\,dx = \int_0^\theta x.\frac{1}{\theta}\,dx = \frac{1}{\theta}\left[\frac{x^2}{2}\right]_0^\theta = \frac{\theta}{2}$$

Method of moments

Matching these moments, we obtain:

$$\frac{\sum X_i}{n} = \frac{\theta}{2} \text{ or } \theta = 2\overline{X}.$$

Hence, the moment estimator $2\overline{X}$ of θ is given by:

$$\hat{\theta} = 2\overline{X}.$$

Method of Least Square

In least squares (LS) estimation, the unknown values of the parameters, in the regression function are estimated by finding numerical values for the parameters that minimize the sum of the squared deviations between the observed responses and the functional portion of the model.

This method will be discussed in detail in the Regression Analysis course.

Example 2

Let $X_1, X_2, X_3, \ldots\ldots\ldots X_n$ be a random sample from a Poisson distribution with parameter λ .Find the method of moments estimators of λ .

Example 3

Let $X_1, X_2, X_3, \ldots\ldots\ldots X_n$ be a random sample from a normal distribution with mean μ and variance σ^2.Find the method of moments estimator of the parameters μ and σ^2.

Example 4

Let $X_1, X_2, X_3, \ldots\ldots\ldots X_n$ be a random sample from the density function

$$f(x;\theta) = \theta e^{-\theta x}, \qquad 0 < x < \infty$$

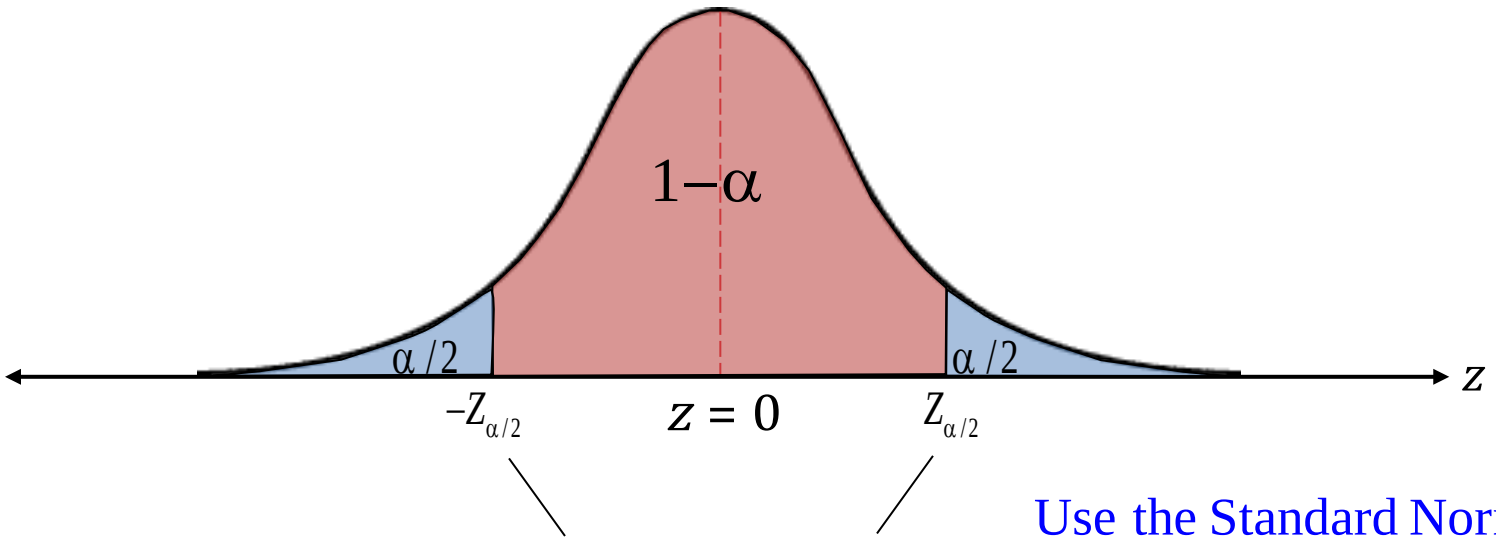
Estimate θ by the method of moments.

Example 5

Let X be a binomial random variable with parameters n and p, both unknown .Given a random sample of N observation s of X, compute the method of moments estimators of n and p.

Level of Confidence

The **level of confidence** $1-\alpha$ is the probability that the interval estimate contains the population parameter.



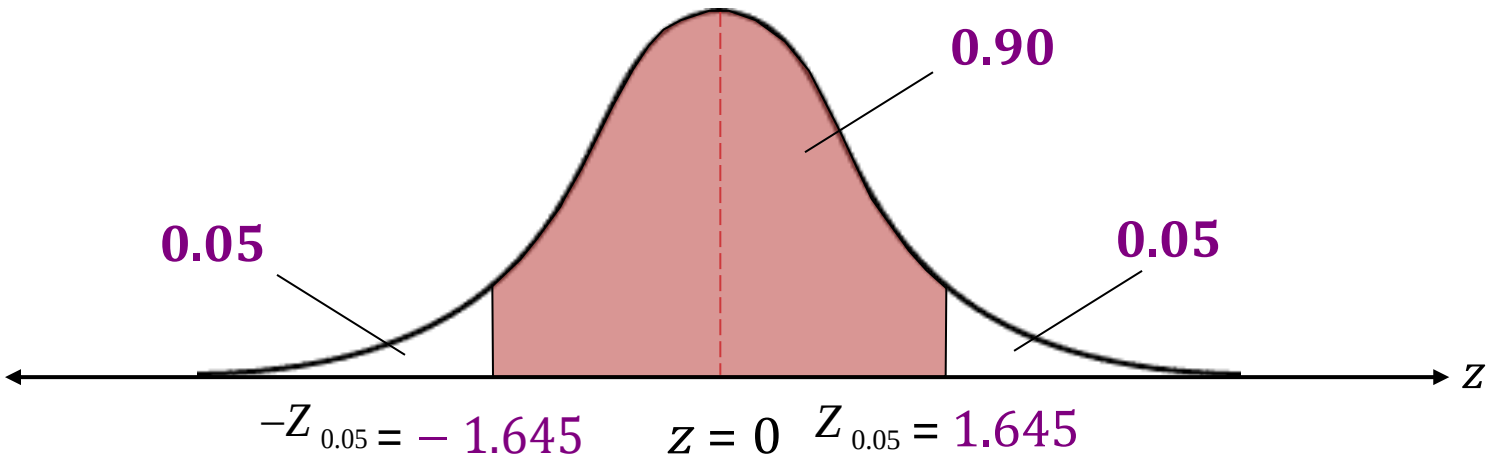
Critical values

Use the Standard Normal Table to find the corresponding *z*-scores.

The remaining area in the tails is $\alpha/2$.

Common Levels of Confidence

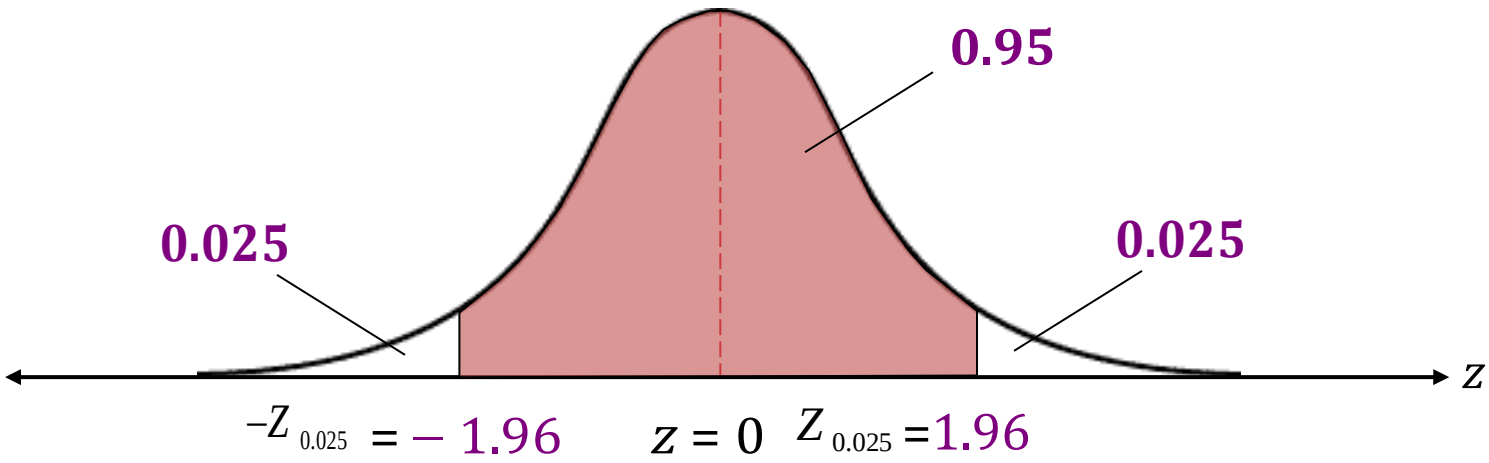
If the level of confidence is 90%, this means that we are 90% confident that the interval contains the population mean, μ.



The corresponding z-scores are ± 1.645.

Common Levels of Confidence

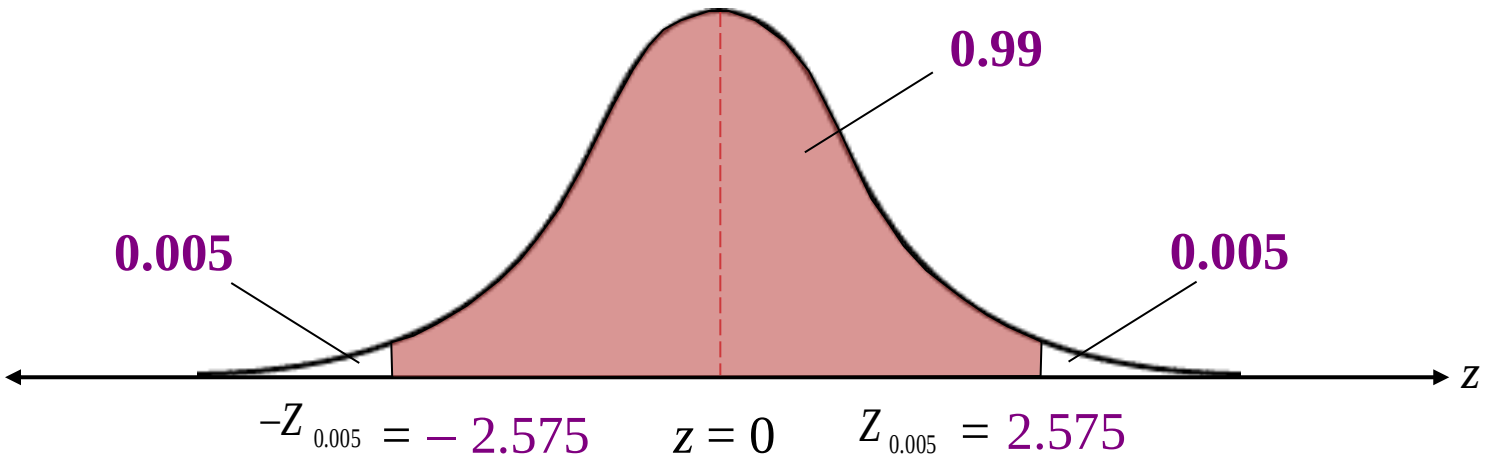
If the level of confidence is 95%, this means that we are 95% confident that the interval contains the population mean, μ.



The corresponding z-scores are $\pm$ 1.96.

Common Levels of Confidence

If the level of confidence is 99%, this means that we are 99% confident that the interval contains the population mean, μ.



The corresponding z-scores are ± 2.575.

Area between 0 and z.

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990

Margin of Error

Given a level of confidence, the **margin of error** (sometimes called the maximum error of estimate or error tolerance) ***E*** is the greatest possible distance between the point estimate and the value of the parameter it is estimating.

$$E = z_c \sigma_{\bar{x}} = z_c \frac{\sigma}{\sqrt{n}},$$

When $n \geq 30$, the sample standard deviation, s, can be used for σ.

$$Where \quad c = \alpha / 2$$

Margin of Error

Example:

A random sample of 32 textbook prices is taken from a local college bookstore. The mean of the sample is $\bar{x} = 74.22$, and the sample standard deviation is $s = 23.44$.

Use a 95% confidence level and find the margin of error for the mean price of all textbooks in the bookstore.

$$E = z_c \frac{\sigma}{\sqrt{n}} = 1.96 \cdot \frac{23.44}{\sqrt{32}}$$

$$\approx 8.12$$

Since $n \geq 30$, s can be substituted for σ.

We are 95% confident that the margin of error for the population mean (all the textbooks in the bookstore) is about $ 8.12.

Case-1: Confidence Intervals for μ (σ Known)

To compute a ***confidence interval*** for the population mean μ , we have to see either the population is normal or not, either the population standard deviation is known or unknown, and either the sample size is large or small.

$$A\ \ 100(1-\alpha)\%\ Confidence\ \text{int}\ erval\ \ for\mu\ is\ givenby:$$

$$P(\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

Confidence Intervals for μ (σ Known)

Example:
A random sample of 25 students had a grade point average with a mean of 2.86. Past studies have shown that the standard deviation is 0.15 and the population is normally distributed.

Construct a 90% confidence interval for the population mean grade point average.

Solution:

$$n = 25 \qquad \bar{x} = 2.86 \qquad \sigma = 0.15$$

$$z_c = 1.645 \qquad E = z_c \frac{\sigma}{\sqrt{n}} = 1.645 \cdot \frac{0.15}{\sqrt{25}} \approx 0.05$$

$$\bar{x} + E = 2.86 \pm 0.05 \longrightarrow 2.81 < \mu < 2.91$$

With 90% confidence we can say that the mean grade point average for all students in the population is between 2.81 and 2.91.

Case-2 Confidence Intervals for μ *(σ unknown and n>30)*

Example:

A random sample of 32 textbook prices is taken from a local college bookstore. The mean of the sample , $\bar{x} = 74.22$
the sample standard deviation is 23.44.

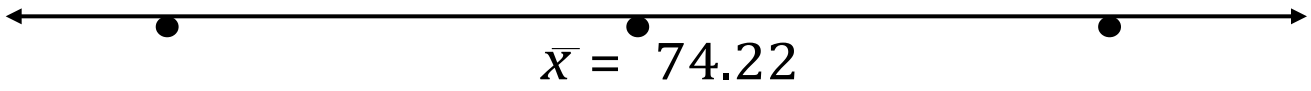
Construct a 95% confidence interval for the mean price of all textbooks in the bookstore.

Solution:

$\bar{x} = 74.22$ $S = 23.44$ $E = 8.12$

Left endpoint = ?

Right endpoint = ?



$\bar{x} - E = 74.22 - 8.12$

$= 66.1$

$\bar{x} + E = 74.22 + 8.12$

$= 82.34$

With 95% confidence we can say that the cost for all textbooks in the bookstore is between \$66.10 and \$82.34.

Case 3: For Non Normal population

If we have a non normal population then:

$$Pr(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) \to 1-\alpha \quad \text{as the}$$

sample size tends to infinity. (We say that the CI has an *asymptotic* level of $1-\alpha$.)

the population is not normal, then we can use the CLT, which says that the distribution of $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ converges to the standard normal distribution as $n \to \infty$. Using the argument above and the Central Limit Theorem, we have

$$Pr(\text{CI Contains } \mu) = Pr(-z_{\alpha/2} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_{\alpha/2}) \to 1-\alpha \ .$$

A $100(1-\alpha)\%$ *Confidence* int *erval for* μ *the mean of the noramal population with* σ *known is given by* :

$$\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

A $100(1-\alpha)\%$ *Confidence* int *erval for* μ *the mean of the noramal population with* σ *unknown is given by* :

$$\bar{X} \pm Z_{\alpha/2} \frac{S}{\sqrt{n}}$$

If we are using without replacement sampling then :

$$\bar{X} \pm Z_{\alpha/2} \frac{S}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

Finding Confidence Intervals for μ

($n \geq 30$ or σ known with a normally distributed population)

Finding a Confidence Interval for a Population Mean($n \geq 30$ or σ known with a normally distributed population)

In Words

1. Find the sample statistics $\bar{x}$.
2. Specify σ, if known. Otherwise, if $n \geq 30$, find the sample standard deviation s and use it as an estimate for σ.
3. Find the critical value z_c that corresponds to the given level of confidence.
4. Find the margin of error E.
5. Find the left and right endpoints and form the confidence interval.

In Symbols

$$\bar{x} = \frac{\sum x}{n}$$

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

Use the Standard Normal Table.

$$E = z_c \frac{\sigma}{\sqrt{n}}$$

Left endpoint: $\bar{x} - E$

Right endpoint: $\bar{x} + E$

Interval: $\bar{x} - E < \mu < \bar{x} + E$

Interpretation of a Confidence Interval

$A\ 100(1-\alpha)\%\ Confidence\ \text{int}\ erval\ \ for\mu\ is:$

$$P(\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

It is to emphasized that in this expression, μ *is cons* tan *t and it is theend ponits of the* int *erval whicharerandom* var *ibales*. Thereforeafter computing theconfidenceintervalfor particular sampleit is erroneous

tosay that :

$$P(\bar{x} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$$

because in this expression *norandom* variable *appears whereas probability statement is made about random* var *iables*.

STA 413 - Statistical Inference

Ayesha Sultan
Lecturer
Virtual university of Pakistan



STA 413 - Statistical Inference

Lecture No.6

To find Sample Size

Determining sample size is a very important issue because samples that are too large may waste time, resources and money, while samples that are too small may lead to inaccurate results. So, it is very important to know that how much sample size of the population should be selected to ensure that the population is properly represented?

When sample mean is calculated from the sample then we can find a sample size by using the given below formula:

$$n = \left(\frac{z\,\sigma}{E} \right)^2$$

Example:

You want to estimate the mean price of all the textbooks in the college bookstore. How many books must be included in your sample if you want to be 99% confident that the sample mean is within $5 of the population mean with a sample standard deviation 23.44?

Sample Size

As, we have from the example,

$E = |\bar{X} - \mu| = 5$ $\quad \sigma \approx s = 23.44$ $\quad z_c = 2.575$

$$n = \left(\frac{z\,\sigma}{E} \right)^2 = \left(\frac{2.575 \cdot 23.44}{5} \right)^2$$

$$= 145.7$$

$$\approx 146$$

Confidence Intervals for the difference of two populations Means, $\mu_1 - \mu_2$.

To construct the confidence interval for the difference between two population means $\mu_1 - \mu_2$,following case will be considered:

Case-1: Confidence Intervals for $\mu_1 - \mu_2$ (σ_1 and σ_2 known)

Case-2: Confidence Intervals for $\mu_1 - \mu_2$ (σ_1 and σ_2 unknown)

Case-1: Confidence Intervals for $\mu_1 - \mu_2$ (σ_1 and σ_2 known)

1. Check Requirements
   - Are σ_1 *and* σ_2 known?
   - Are the samples independent?
   - Assumption of normality?
2. Compute E: $E = z_c\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$
3. Construct the Interval:

$$(\bar{x}_1 - \bar{x}_2) - \mathrm{E} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + \mathrm{E}$$

Example

A research team is interested in the difference between serum uric acid levels in patients with and without mental disorder. In a large hospital for the treatment of the mentally retarded, a sample of 12 individuals with mental disorder have a mean of $\bar{X}_1 = 4.5$ mg/100 ml. In a general hospital a sample of 15 normal individuals of the same age and sex were found to have a mean value of $\bar{X}_2 = 3.4$ mg/100 ml. If it is reasonable to assume that the two populations of values are normally distributed with variances equal to 1 and 1.5 respectively, find the 95 percent confidence interval for $\mu_1 - \mu_2$.

Calculations :

The point estimate for $\mu_1 - \mu_2$ *is* $\bar{X}_1 - \bar{X}_2$

$$\bar{X}_1 - \bar{X}_2 = 4.5 - 3.4 = 1.1$$

The standard error is :

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{1}{12} + \frac{1.5}{15}} = 0.4282$$

The 95% *confidence interval is* :

$$\bar{X}_1 - \bar{X}_2 \pm Z_{(1-\alpha/2)} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

whereas $Z_{(1-\alpha/2)} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = E$ (*Marginof error*)

$= 1.1 \pm 1.96 \ (.4282)$

$= (.26, 1.94)$

We conclude with 95% confidence that *the* average *difference between* the *serum uric acid levels in patients with and withoutmental disorder lies between* 0.26 *and* 1.94.

Case-2: Confidence Intervals for $\mu_1 - \mu_2$ (σ_1 and σ_2 Are Unknown, n>30)

1. Check Requirements
   - σ_1 *and* σ_2 unknown but n>30
   - Are the samples independent?
   - Assumption of normality?
2. Compute E: $E = z_c \sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}$
3. Construct the Interval:

$$(\bar{x}_1 - \bar{x}_2) - \mathrm{E} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + \mathrm{E}$$

Example

Let suppose I surveyed 50 people from a poor area of town and 70 people from an wealthy area of town about their feelings towards minorities. I counted the number of negative comments made. I was interested in comparing their attitudes. The average number of negative comments in the poor area was 14 and in the wealthy area was 12. The standard deviations were 5 and 4 respectively. Find a 95% confidence for the difference $\mu_1 - \mu_2$,where μ_1 is the average number of negative comments made by poor area and μ_2 is the average number made by wealthy area.

$$\bar{X}_1 - \bar{X}_2 = 14 - 12 = 2$$

$$S = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} = \sqrt{\frac{5^2}{50} + \frac{4^2}{70}} = 0.85$$

The margin of error is

$$E = z_c s = (1.96)(0.85) = 1.7$$

The confidence interval is

2 ± 1.7

or

$[0.3, 3.7]$

Hence, we can conclude that the mean difference about the negative comments made by poor and wealthy people is between 0.3 and 3.7.

STA 413 - Statistical Inference

Ayesha Sultan
Lecturer
Virtual university of Pakistan



STA 413 - Statistical Inference

Lecture No.7

Inference about Population Correlation Coefficient

The sample correlation coefficient r is calculated on random sample of n pairs of values (X, Y) from a bivariate population (we assume, is a bivariate normal population) with a theoretical correlation coefficient denoted by ρ .Tests of significance or hypothesis concerning various values of ρ require knowledge of the sampling distribution of r. For $\rho = 0$, this distribution is symmetric and the t distribution can be used to test the significance of r, For $\rho \neq 0$ the t distribution is skewed ,in such case Fisher's Z transformation is used.

Let us discuss different cases for testing the correlation value.

- Case-1: Testing the hypothesis that the Population Correlation Coefficient equal to zero
- Case-2: Testing the hypothesis that the population correlation coefficient equals some value other then zero
- Case-3: Testing the hypothesis that two independent populations have same correlation coefficients

Case-1: Testing the hypothesis that the Population Correlation Coefficient equal to zero

A correlation coefficient based on a sample of size 27 was computed to be 0.40.Can we conclude at significance level of 0.05 that the corresponding population correlation coefficient is significantly greater than zero?

Example 2

Find the least value of r, in a sample of 27 pairs from a normal population that is significance at the 5% level.

Case-2: Testing the hypothesis that the population correlation coefficient equals some value other then zero.

A correlation coefficient based on a sample of size 35 was computed to be 0.5.Can we reject the hypothesis that the population correlation coefficient is (a) as small as $\rho = 0.3$ (b) as large as $\rho = 0.7$, using a 0.05 significance level.

Example 2

In a random sample of 28 pairs of values from a bivariate normal population, the correlation was found to be 0.7.Is this value consistent with the assumption; that the correlation in the population is 0.5? Use a 0.05 level of significance.

Case-3: Testing the hypothesis that two independent populations have same correlation coefficients

Two correlation coefficients obtained from samples of sizes 23 and 28 were computed to be 0.8 and 0.95 respectively. Can we conclude at a significance level of 0.01 that is a significant difference between the two correlation coefficients?

STA 406 - Statistical Inference

Ayesha Sultan
Lecturer
Virtual university of Pakistan



STA 406 - Statistical Inference

Lecture No.3

Confidence Interval Estimates

- **CONFIDENCE INTERVAL for μ**

$$\bar{x} \pm t \frac{s}{\sqrt{n}}$$

- where:
- t = Critical value from t-distribution with n-1 degrees of freedom
- $\bar{X}$ = Sample mean
- s = Sample standard deviation
- n = Sample size

For very small samples (n < 30) and σ is unknown.

Fundamental principles for using the t-distribution for confidence intervals:

1. You cannot use the t-distribution unless you assume that the population distribution of the variable is normally distributed.
2. The t-distribution, like the z-distribution, is bell-shaped and symmetric about a mean of 0.
3. The t-distribution incorporates the fact that for smaller sample sizes the distribution will be more spread out using something called degrees of freedom.
4. For every change in degrees of freedom, the t-distribution changes. The larger the sample size (n), the closer the t-distribution mimics the z-distribution in shape. We construct a confidence interval for a small sample size in the same way as we do for a large sample, except we use the t-distribution instead of the z-distribution.

Degree of Freedom

The degree of freedom for an estimate is equal to the number of values minus the number of parameters estimated en route to the estimate. For example, if there are two values (8 and 5) and we had to estimate one parameter (μ) on the way to estimating the parameter of interest ($\sigma 2$). Therefore, the estimate of variance has 2 - 1 = 1 degree of freedom. Similarly, if there are 12 sampled observations then our estimate of variance would have had 11 degrees of freedom as the degrees of freedom of an estimate of variance is equal to n - 1, where n is the number of observations.

$$s^2 = \frac{\sum_{i=1} (X_i - \bar{X})_2}{n-1}$$

Question 1 :

You know the population mean for a certain test score. You select 10 people from the population to estimate the standard deviation. How many degrees of freedom does your estimation of the standard deviation have?

Answer :

There are 10 independent pieces of information, so there are 10 degrees of freedom.

Question 2:

You do not know the population mean for a different test score. You select 15 people from the population and use this sample to estimate the mean and standard deviation. How many degrees of freedom does your estimation of the standard deviation have?

Answer:

The degree of freedom for an estimate is equal to the number of values minus the number of parameters estimated en route to the estimate in question. You have 15 values in your sample, and you need to estimate one parameter, the mean, in order to find the standard deviation. 15 - 1 = 14.

t distributions

- Very similar to Z~N(0, 1)
- Sometimes called Student's t distribution;

Properties:

i) symmetric around 0 (like z)

ii) degrees of freedom

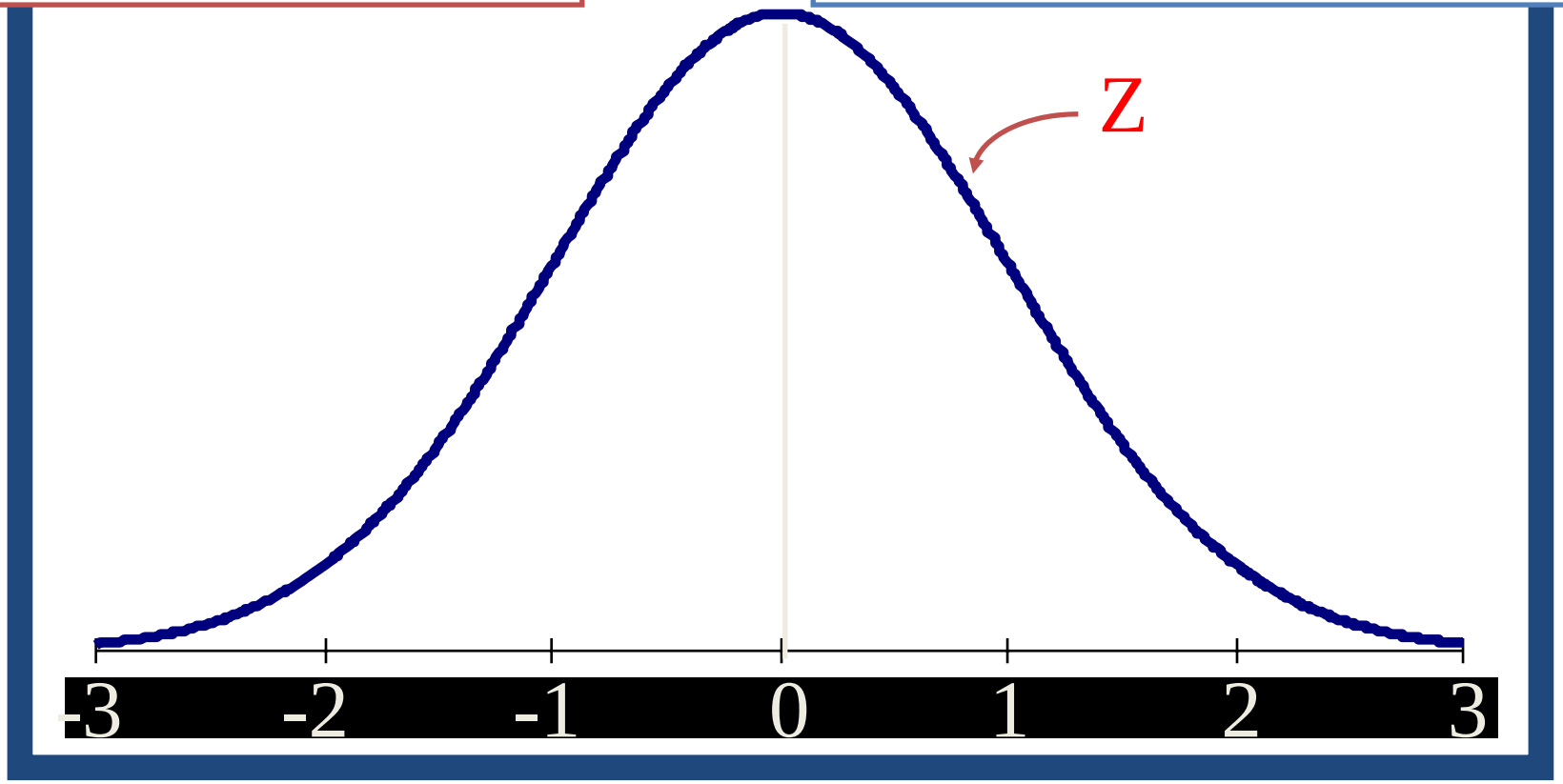
if $\nu > 1$, $E(t) = 0$

if $\nu > 2$, $\sigma = \sqrt{\nu / \nu - 2}$, which is always bigger than 1.

Student's t Distribution

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

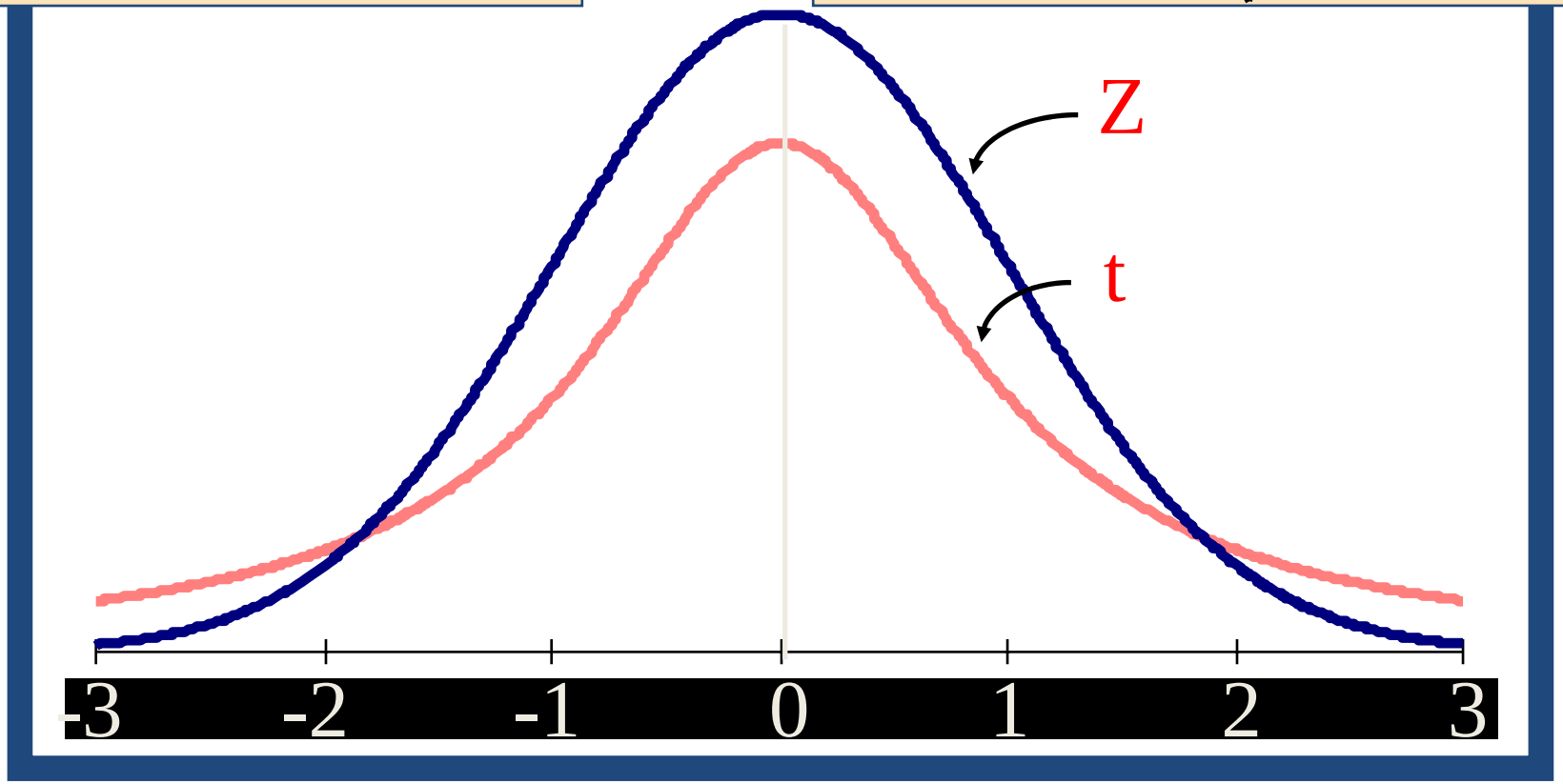
$$t = \frac{\bar{x} - \mu_{\bar{x}}}{s_{\bar{x}}}, \; s_{\bar{x}} = \frac{s}{\sqrt{n}}$$



Student's t Distribution

$$z = \frac{\bar{x} - \mu_x}{\frac{\sigma}{\sqrt{n}}}$$

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



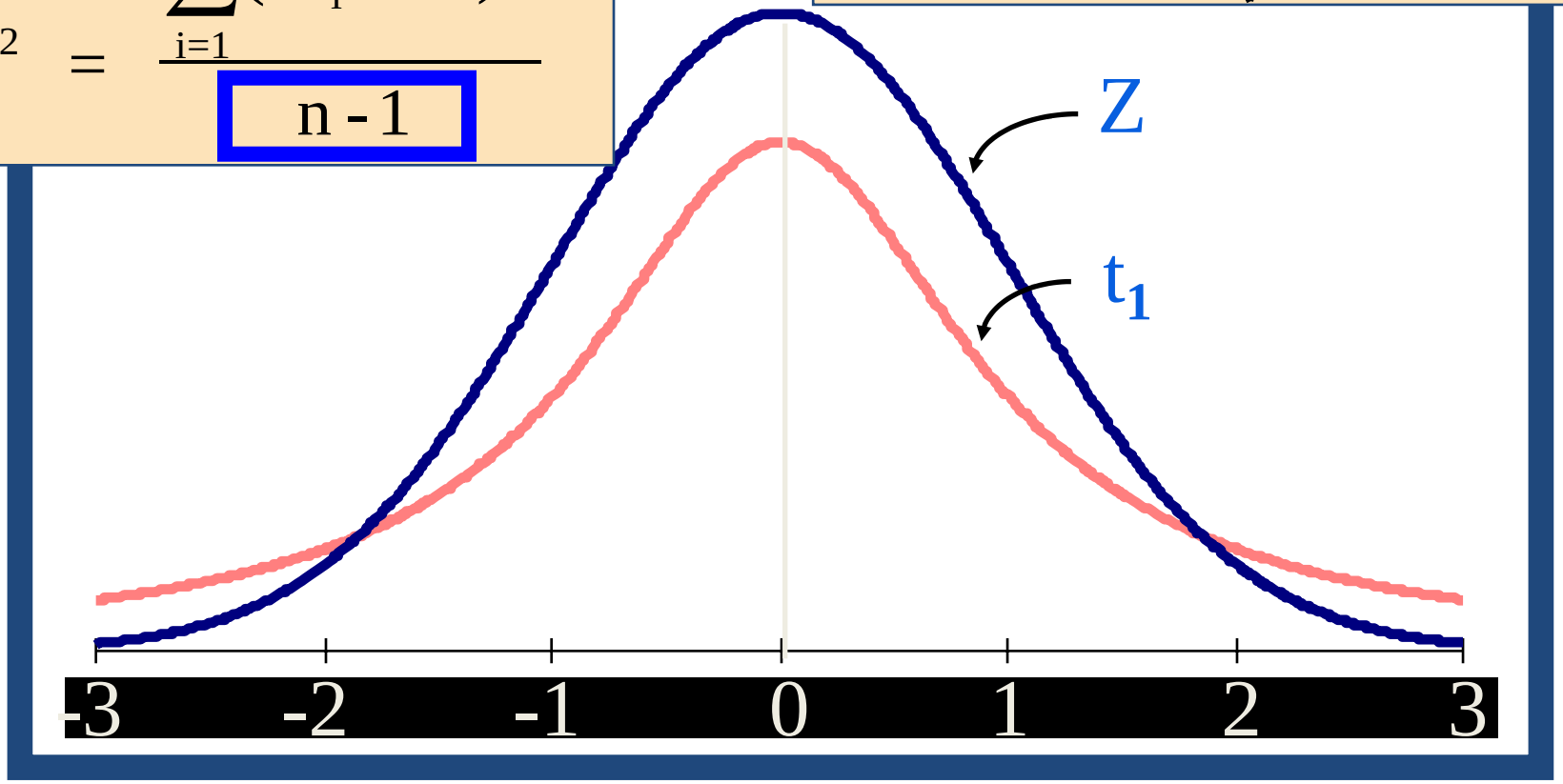
Student's t Distribution

Degrees of Freedom

$$s = \sqrt{s^2}$$

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

$$t = \frac{\overline{x} - \mu_{\overline{x}}}{\frac{s}{\sqrt{n}}}$$



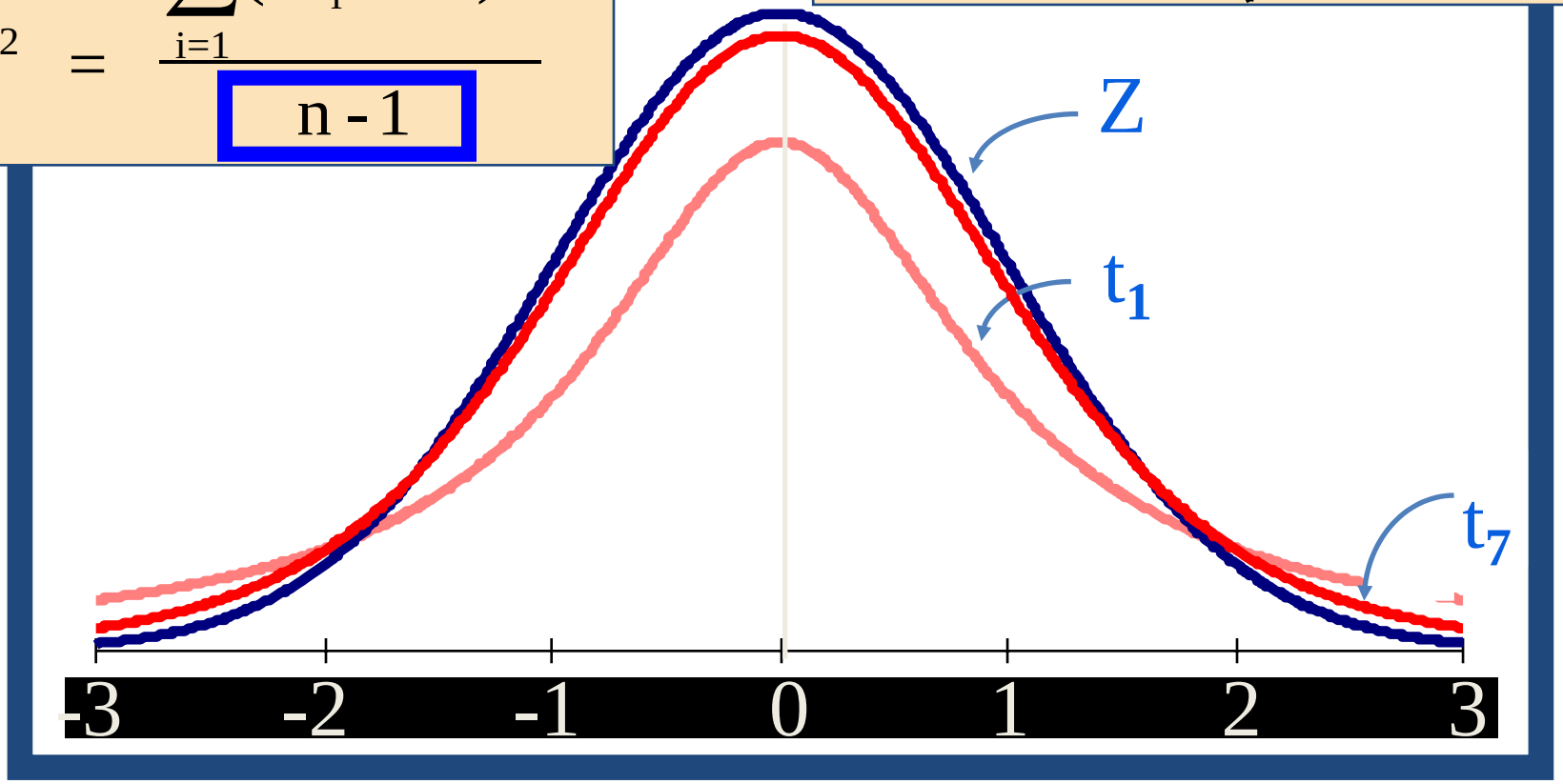
Student's t Distribution

Degrees of Freedom

$$s = \sqrt{s^2}$$

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

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



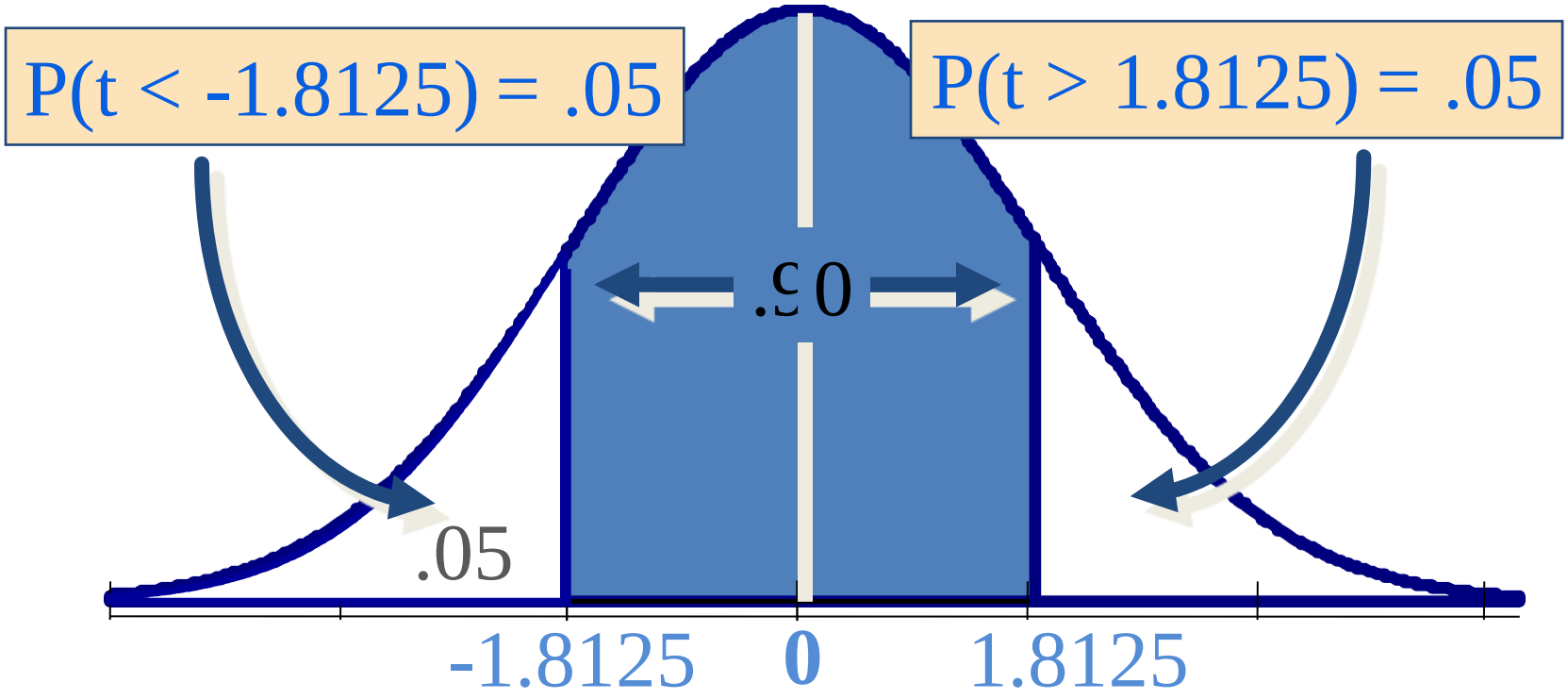
df\p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	4.3178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251

t-Table: text- inside back cover

- 90% confidence interval; df = n-1 = 10

Degrees of Freedom	0.80	0.90	0.95	0.98	0.99
1	3.0777	6.314	12.706	31.821	63.657
2	1.8856	2.9200	4.3027	6.9645	9.9250
.	.	.	.	.	.
.	.	.	.	.	.
10	1.3722	1.8125	2.2281	2.7638	3.1693
.	.	.	.	.	.
.	.	.	.	.	.
100	1.2901	1.6604	1.9840	2.3642	2.6259
∞	1.282	1.6449	1.9600	2.3263	2.5758

$$90\%\,\text{confidence interval}: \bar{x} \pm 1.8125 \frac{s}{\sqrt{11}}$$


P(t < -1.8125) = .05
P(t > 1.8125) = .05
.95
.05
-1.8125
0
1.8125

Comparing t and z Critical Values

	Confidence level	n = 30
z = 1.645	90%	t = 1.6991
z = 1.96	95%	t = 2.0452
z = 2.33	98%	t = 2.4620
z = 2.58	99%	t = 2.7564

Example

An investor is trying to estimate the return on investment in companies that won quality awards last year. A random sample of 25 such companies is selected, and the return on investment is recorded for each company. The data for the 25 companies have $\bar{x} = 14.75 \quad s = 8.18$

Construct a 95% confidence interval for the mean return.

$$\bar{x} \pm t \frac{s}{\sqrt{n}} \qquad d.f. = n - 1$$

$\bar{x} = 14.75 \quad s = 8.18$

degrees of freedom $= 25 - 1 = 24$

from t-table, t $=$ 2.0211

$$\bar{x} \pm t \frac{s}{\sqrt{n}} = 14.75 \pm 2.064 \frac{8.18}{\sqrt{25}}$$

$$= 14.75 \pm 3.376 = [11.37, 18.12]$$

We are 95% confident that the interval (11.37,18.12) contains the population mean return on investment for companies that win quality awards.

Example

Cardiac deaths increase after heavy snowfalls, a study was conducted to measure the cardiac deaths of shoveling snow by hand. The maximum heart rates for 10 adult males were recorded while shoveling snow. The sample mean and sample standard deviation were 175 and 15 respectively.

Find a 90% CI for the population mean maximum heart rate for those who shovel snow.

Solution

$$\bar{x} \pm t\frac{s}{\sqrt{n}} \qquad d.f. = n-1$$

$\bar{x} = 175,\ s = 15\ n = 10$

From the t - table, $t = 1.8331$

$$175 \pm 1.8331\frac{15}{\sqrt{10}} = 175 \pm 8.70$$

$= (166.30, 183.70)$

We are 90% confident that the interval (166.30, 183.70) contains the mean maximum heart rate for snow shovelers

Example

The masses, in grams, of twelve ball bearings taken at random from a batch are 31.4, 33.1, 35.9, 34.7, 33.4, 34.5, 35.0, 32.5, 36.9, 36.4, 35.8 and 33.2. Calculate a 90% confidence interval for the mean mass of the population, supposed normal, from which these masses were drawn.

Confidence Interval Estimates for

$\mu_1 - \mu_2$

STANDARD DEVIATIONS UNKNOWN AND $\sigma_1^2 = \sigma_2^2$

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

= Pooled standard deviation

$t_{\alpha/2}$ = critical value from t-distribution for desired confidence level and degrees of freedom equal to $n_1 + n_2 - 2$

Example

a) Give two random samples of sizes $n_1 = 9 \& n_2 = 16$ from two independent normal populations, with $\bar{x}_1 = 64, \bar{x}_2 = 59, s_1 = 6 \text{ and } s_2 = 5,$ find a 95% confidence interval for $\mu_1 - \mu_2$ assuming that $\sigma_1 = \sigma_2$.

b) A sample from a normal population with unknown variance consists of the observations 34, 25, 43, 37, 45. A sample from a second normal population with the same unknown variance as the first consists of the observations 20, 31, 23, 35, 41, 29, 39. Find a 95% confidence interval for $\mu_1 - \mu_2$

Confidence Interval Estimates for

$\mu_1 - \mu_2$

STANDARD DEVIATIONS UNKNOWN AND $\sigma_1^2 \neq \sigma_2^2$

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

where:

$t_{\alpha/2}$ = critical value from t-distribution for desired confidence level

and degrees of freedom equal to:

$$\frac{(s_1^2/n_1 + s_2^2/n_2)}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$$

Confidence Interval Estimate for paired samples

Paired samples are samples that selected such that each data value from one sample is related (or matched) with a corresponding data value from the second sample. The sample values from one population have the potential to influence the probability that values will be selected from the second population.

Confidence Interval Estimate for paired samples

PAIRED CONFIDENCE INTERVAL ESTIMATE

$$\bar{d} \pm t_{\alpha/2} \frac{s_d}{\sqrt{n}}$$

Confidence Interval Estimate for paired samples

PAIRED DIFFERENCE

$$d = x_1 - x_2$$

where:

d = Paired difference

x_1 and x_2 = Values from sample 1 and 2, respectively

Confidence Interval Estimate for paired samples

MEAN PAIRED DIFFERENCE

$$\bar{d} = \frac{\sum_{i=1}^{n} d_i}{n}$$

where:

d_i = ith paired difference

n = Number of paired differences

STANDARD DEVIATION FOR PAIRED DIFFERENCES

$$s_d = \sqrt{\frac{\sum_{i=1}^{n} (d_i - \bar{d})_2}{n-1}}$$

where:

d_i = i^{th} paired difference

$\bar{d}$ = Mean paired difference

Example

A nutrition scientist is assessing a weight-loss programme to evaluate its effectiveness. Ten people were randomly selected. Both the initial weight and the final weight after 20 weeks on the programme was recorded as:

Subject	Initial Weight	Final Weight
1	180	165
2	142	138
3	126	128
4	138	136
5	175	170
6	205	197
7	116	115
8	142	128
9	157	144
10	136	130

Find the 95% confidence interval for the mean difference between Initial weight and Final weight. Assume that the mean differences are approximately normally distributed.

Example

Twenty-two students were randomly selected from a population of 1000 students. All of the students were given a standardized English test and a standardized Math test. Find the 90% confidence interval for the mean difference between student scores on the Math and English tests. Assume that the mean differences are approximately normally distributed.

Test results are summarized below:

Student	English(x)	Math(Y)
1	95	90
2	89	85
3	76	73
4	92	90
5	91	90
6	53	53
7	67	68
8	88	90
9	75	78
10	85	89
11	90	95
12	85	83
13	87	83
14	85	83
15	85	82
16	68	65
17	81	79
18	84	83
19	71	60
20	46	47
21	75	77
22	80	83

STA 406 - Statistical Inference

Ayesha Sultan

Lecturer

Virtual university of Pakistan



STA 406 - Statistical Inference

Lecture No.4

HYPOTHESIS TESTING

HYPOTHESIS

What do you mean by a Hypothesis?

A supposition or explanation that is provisionally accepted in order to interpret certain events or phenomena, and to provide guidance for further investigation.

OR

A hypothesis is a tentative statement about the relationship between two or more variables.

HYPOTHESIS TESTING

Statistically, an assumption about certain characteristics of a population.

The statistical procedure for testing a hypothesis requires some understanding of the null hypothesis and alternative hypothesis.

TYPES OF HYPOTHESIS

Null Hypothesis

A null hypothesis is "the hypothesis that there is no relationship between two or more variables. In a mathematical formulation of the null hypothesis there will typically be an equal sign. This hypothesis is denoted by H_0.

Alternate Hypothesis

The alternative hypothesis proposes a relationship between two or more variables. In a mathematical formulation of the alternative hypothesis there will typically be an inequality, or not equal to symbol. This hypothesis is denoted by either H_a or by H_1.

EXAMPLE OF HYPOTHESES

For example, you might have come up with a measurable hypothesis that children have a higher IQ if they eat oily fish for a period of time.

Null hypothesis: Children who eat oily fish for six months do not show a higher IQ.

Alternative hypothesis: Children who eat oily fish for six months will show a higher IQ

A common statistical method is to compare a population to the mean.

H_0 : The children will show no increase in mean intelligence.

i.e.

$H_0 : \mu = 100$

H_1 : The children will show an increase in mean intelligence.

i.e.

$H_1 : \mu > 100$

TYPES OF ERRORS

What are errors in Hypothesis Testing?

The purpose of Hypothesis Testing is to reject or not reject the Null Hypothesis based on statistical evidence

Hypothesis Testing is said to have resulted in an error when the decision regarding treatment of the Null Hypothesis is wrong

There are basically **two types of errors** we can make:





TYPES OF ERRORS

Type-I Error (Ho right but rejected)

When Null Hypothesis is rejected despite the test on data showing that the outcome was true

Type-II Error (Ho wrong but not rejected)

When Null Hypothesis is not rejected despite the test on data showing that the outcome was false

Type I Error

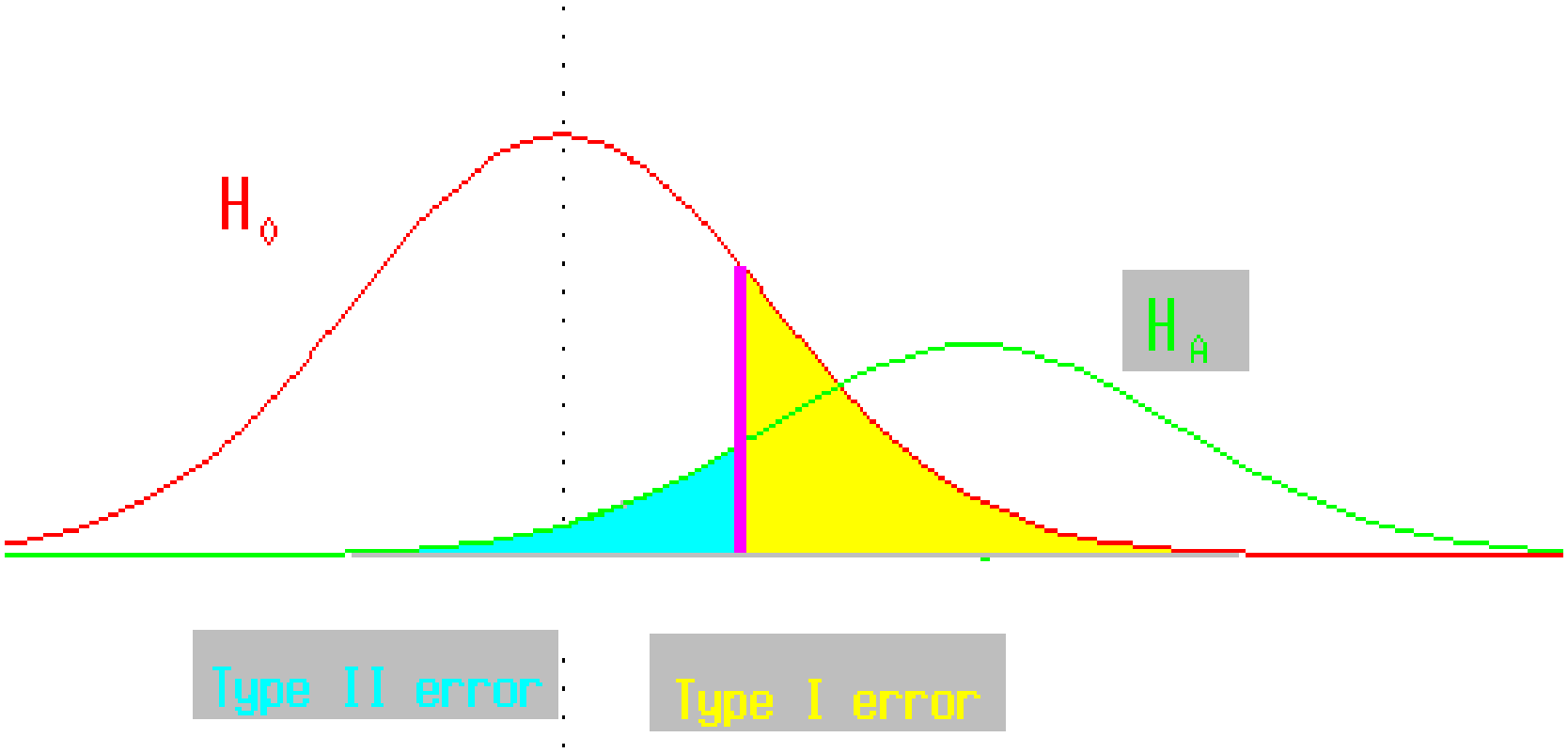
- A **type I error**, also known as an **error of the first kind**, occurs when the null hypothesis (H_0) is true, but is rejected.
- A type I error may be compared with a so called ***false positive.***
- The rate of the type I error is called the *size* of the test and denoted by **the Greek letter α (alpha).**
- It usually equals the **significance level of a test**.
- If type I error is fixed at 5 %, it means that there are about 5 chances in 100 that we will reject H_0 when H_0 is true.

Type II Error

- **Type II error**, also known as an **error of the second kind**, occurs when the null hypothesis is false, but erroneously fails to be rejected.
- Type II error means accepting the hypothesis **which should have been rejected**.
- A type II error may be compared with a so-called ***False Negative.***
- The rate of the type II error is denoted by the **Greek letter β (beta)** and related to the power of a test **(which equals 1-β)**.

In the tabular form two error can be presented as follows:

	Null hypothesis (H_0) is true	Null hypothesis (H_0) is false
Reject null hypothesis	Type I error False decision	Correct True decision
Fail to reject null hypothesis	Correct True decision	Type II error False decision



If we decrease type I error then it will increase type II error (and vice-versa)

Reducing Type I Errors

The chances of making a Type I error is reduced by increasing the level of confidence.

Reducing Type II Errors

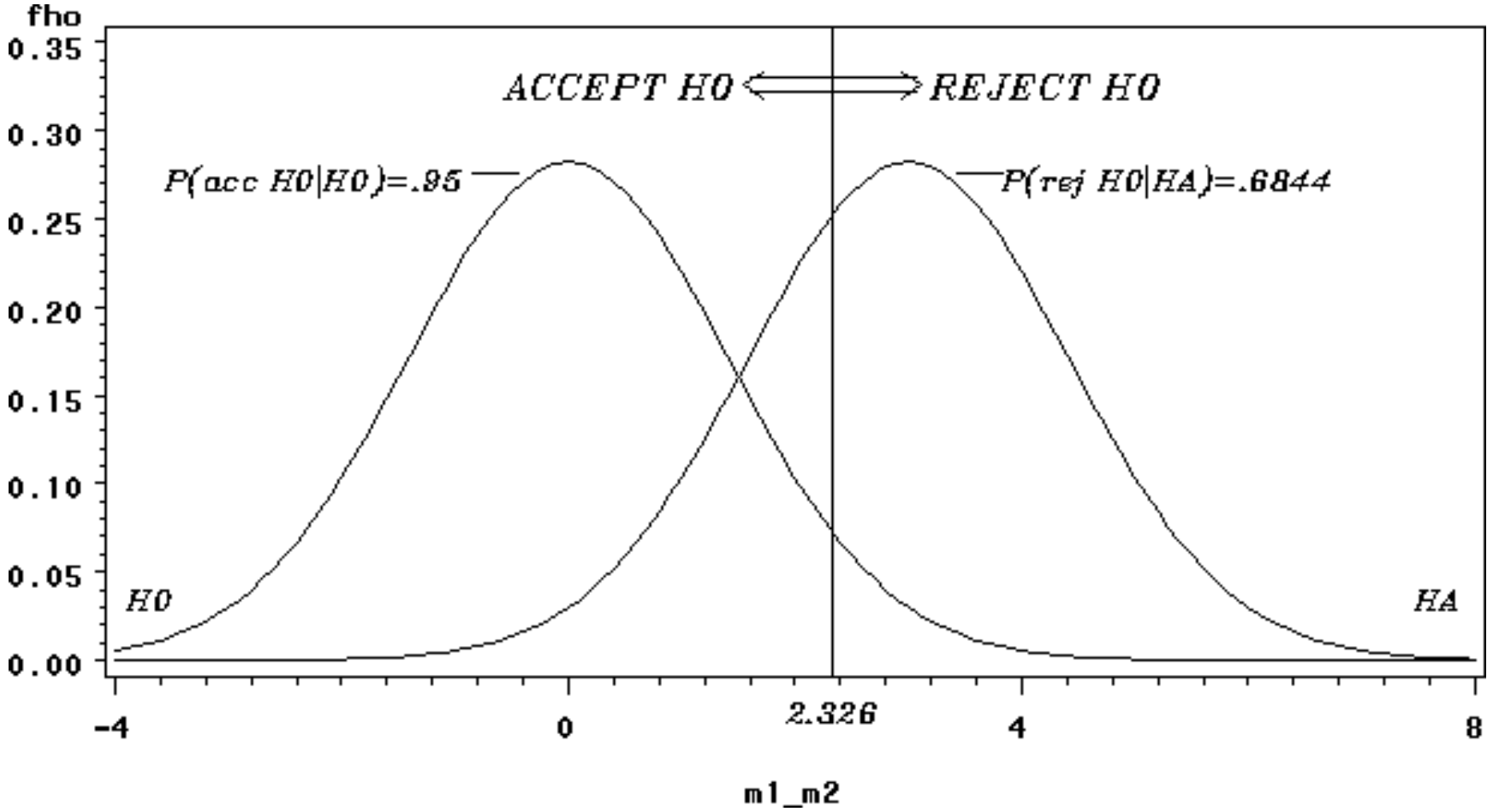
Test condition and acceptance criteria are in turn reduces Type II errors. This increases the number of times we reject the Null hypothesis – **with a resulting increase *in the number of Type I errors*.**

Type II Error and Power

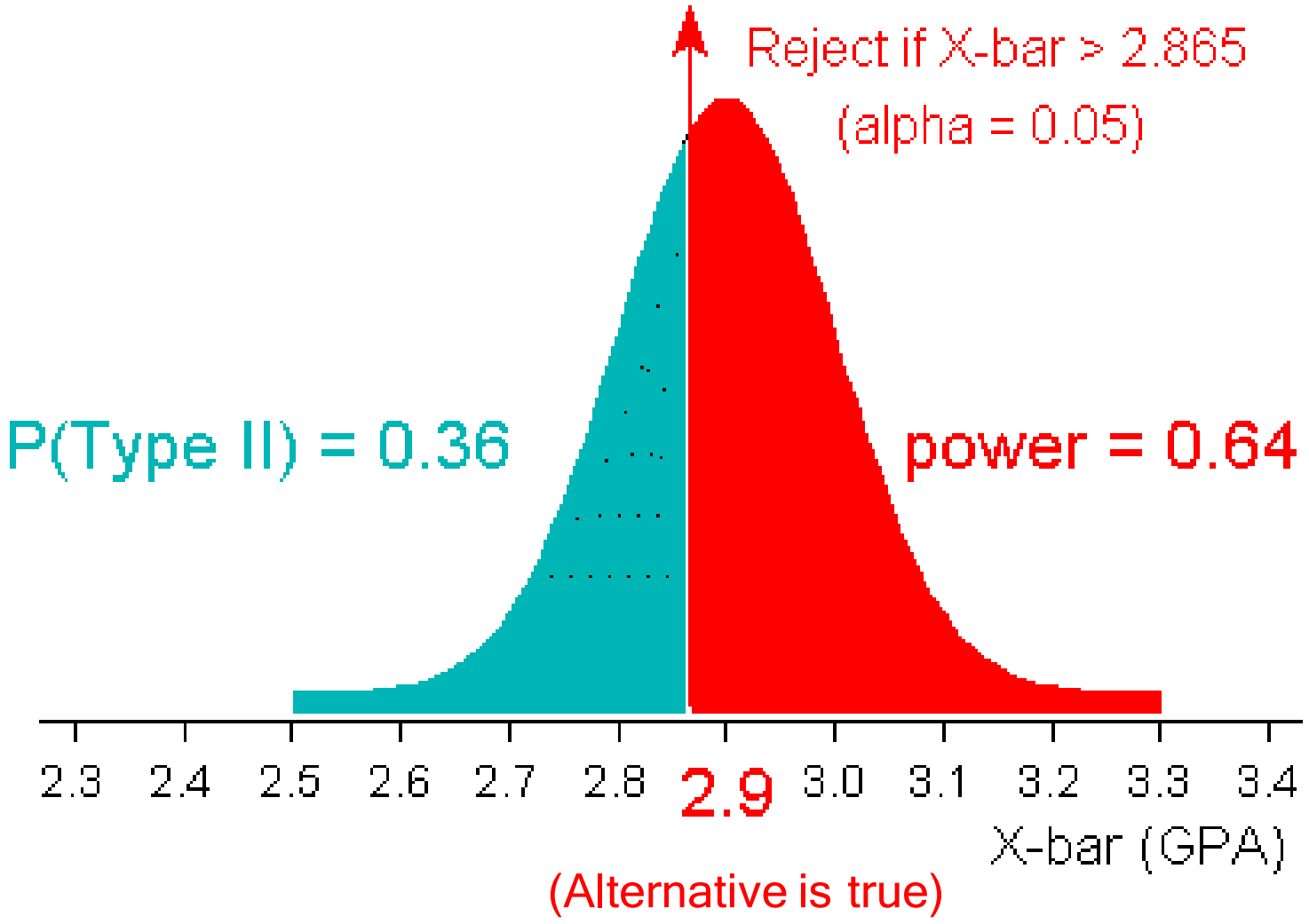
- "**Power**" of a test is the probability of rejecting null when alternative is true.
- **"Power" = 1 - P(Type II error)**
- To minimize the P(Type II error), we equivalently want to maximize power.
- But power depends on the value under the alternative hypothesis .

Power of a Test

Distribution (H_0) Distribution (H_A)



Type II Error and Power

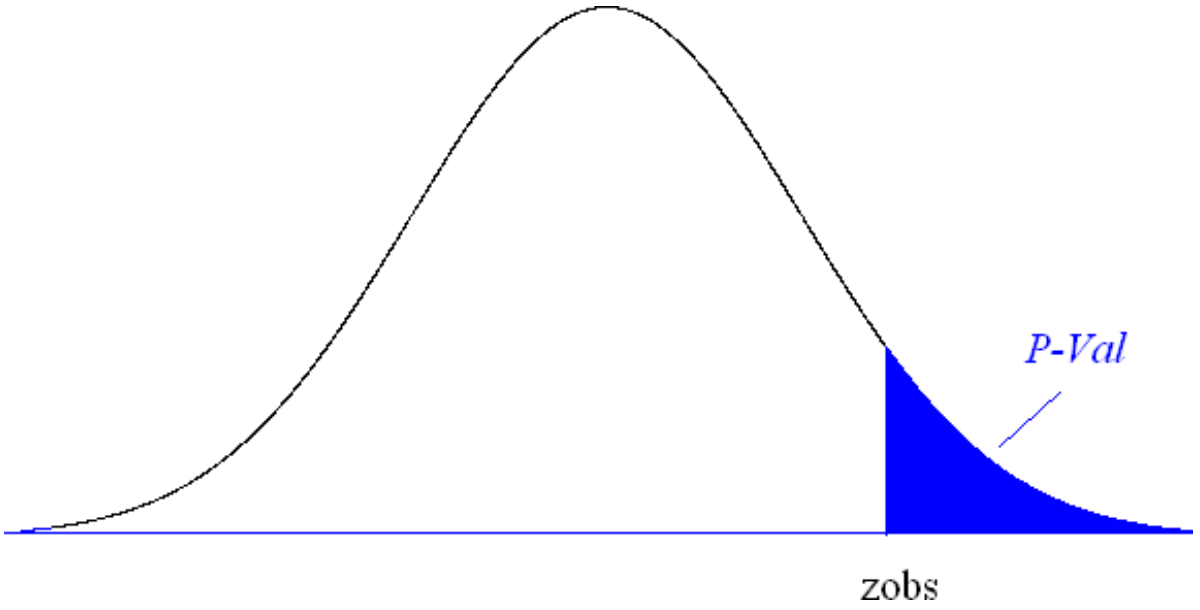


P-value

***P*-value -** Measure of the strength of evidence the sample data provides against the null hypothesis:

P(Evidence This strong or stronger against H_0 | H_0 is true)

$$P-val: p = P(Z \geq z_{obs})$$



INTERPRETING RESULTS

Interpreting the weight of evidence against the Null Hypothesis for rejecting / not rejecting Ho

If the p-value for testing Ho is less than –

- **< 0.10**, we have some evidence that Ho is false
- **< 0.05**, we have strong evidence that Ho is false
- **< 0.01**, we have very strong evidence that Ho is false
- **< 0.001**, we have extremely strong evidence that Ho is false

Simple and Composite Hypothesis

A **simple hypothesis** is one in which all parameters of the distribution are specified. For example, if the heights of college students are normally distributed with , $\sigma^2 = 4$ the hypothesis that its mean μ is, say, $62"$, that is $H : \mu = 62$, we have stated a simple hypothesis, as the mean and variance together specify a normal distribution completely.

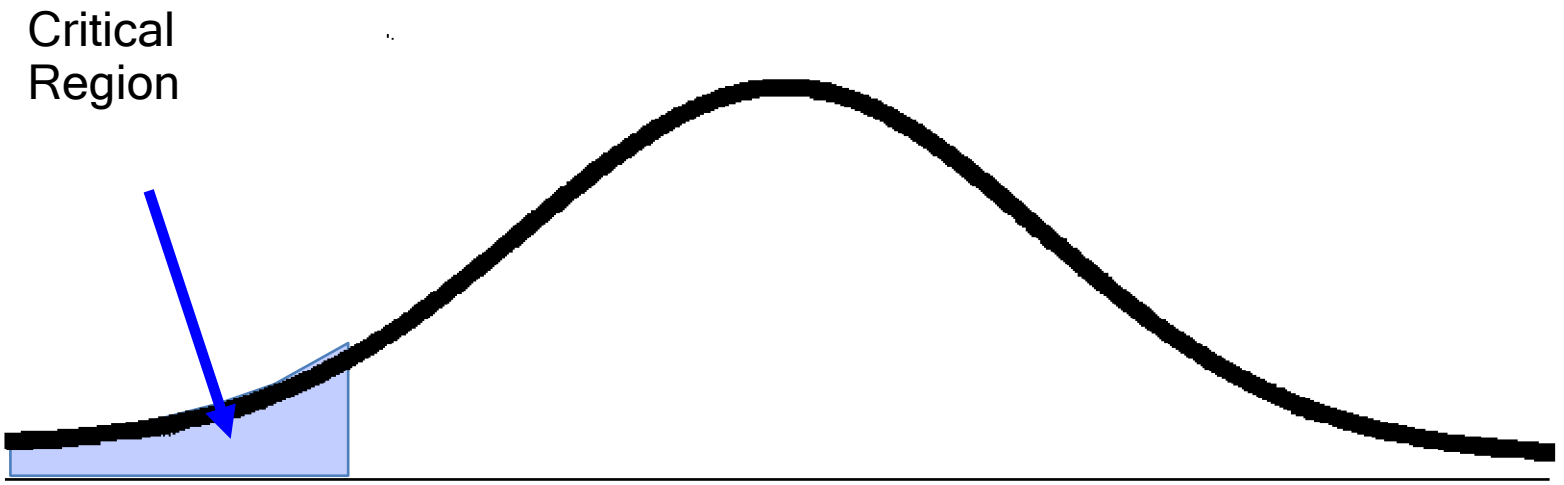
Simple and Composite Hypothesis

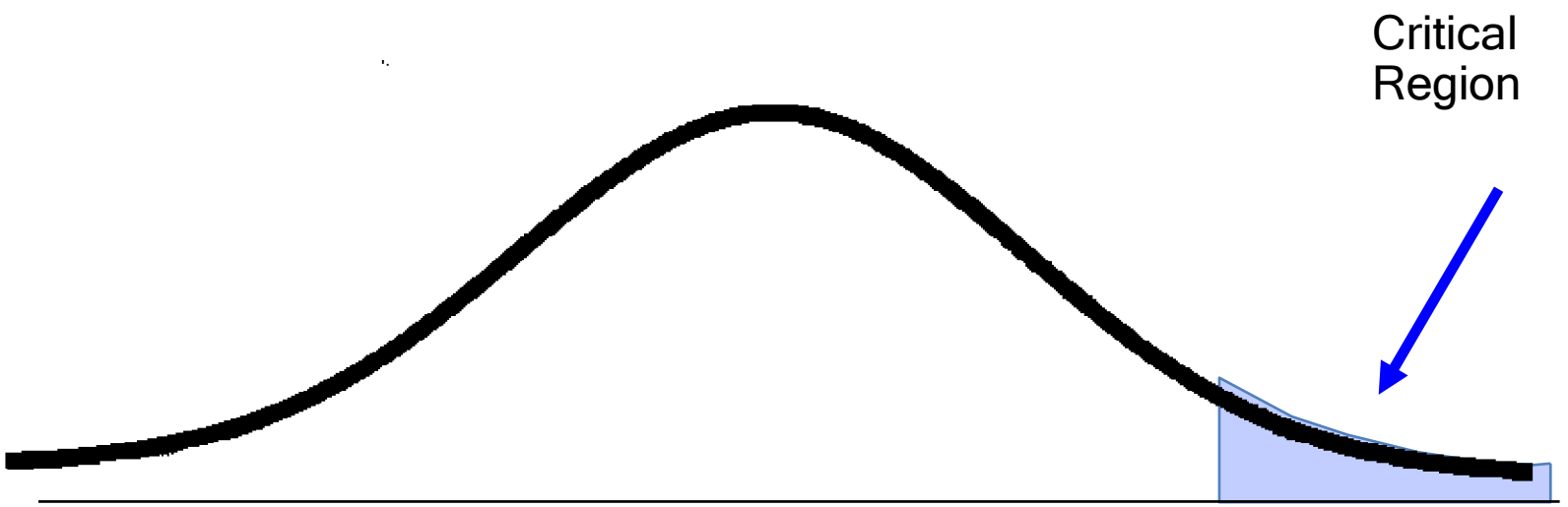
A hypothesis which is not simple (i.e. in which not all of the parameters are specified) is called a **composite hypothesis**. For instance, if we hypothesize that $H : \mu > 62$(and $\sigma^2 = 4$) or $H : \mu = 62$ and $\sigma^2 < 4$, the hypothesis becomes a composite hypothesis because we cannot know the exact distribution of the population in either case. Obviously, the parameters $\mu > 62"$ and $\sigma^2 < 4$ have more than one value and no specified values are being assigned.

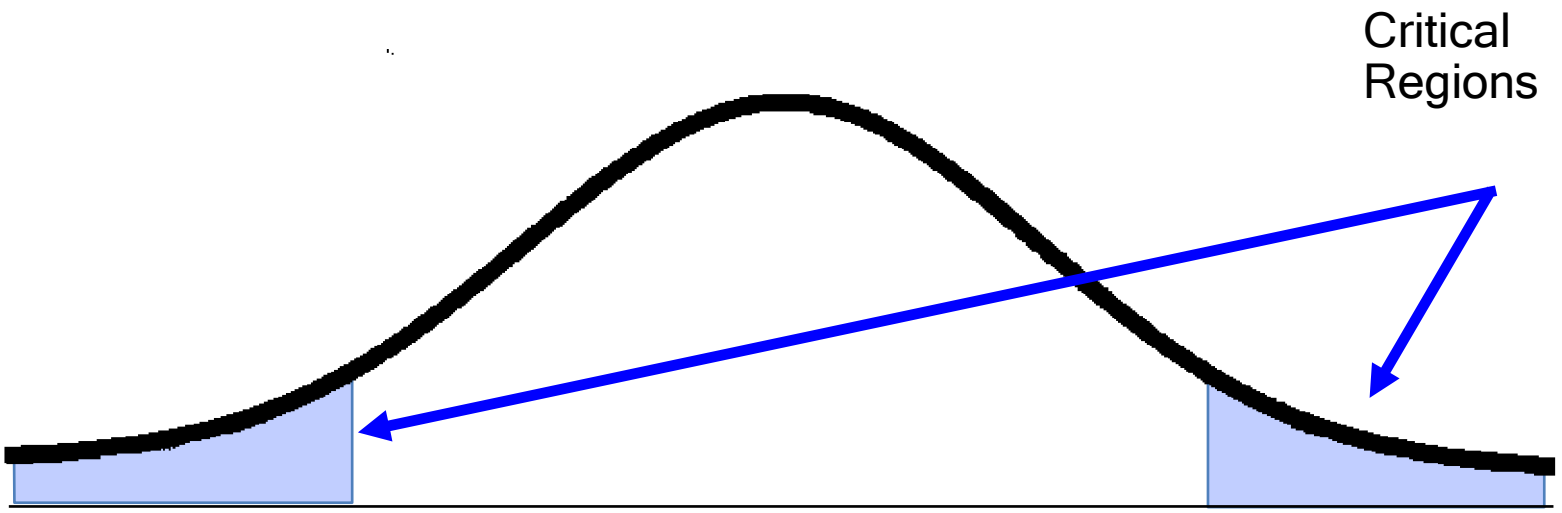
Critical Region (or Rejection Region)

The critical region CR, or rejection region RR, is a set of values of the test statistic for which the null hypothesis is rejected in a hypothesis test. That is, the sample space for the test statistic is partitioned into two regions; one region (the critical region) will lead us to reject the null hypothesis H0, the other will not. So, if the observed value of the test statistic is a member of the critical region, we conclude "Reject H0"; if it is not a member of the critical region then we conclude "Do not reject H0".

Critical Region

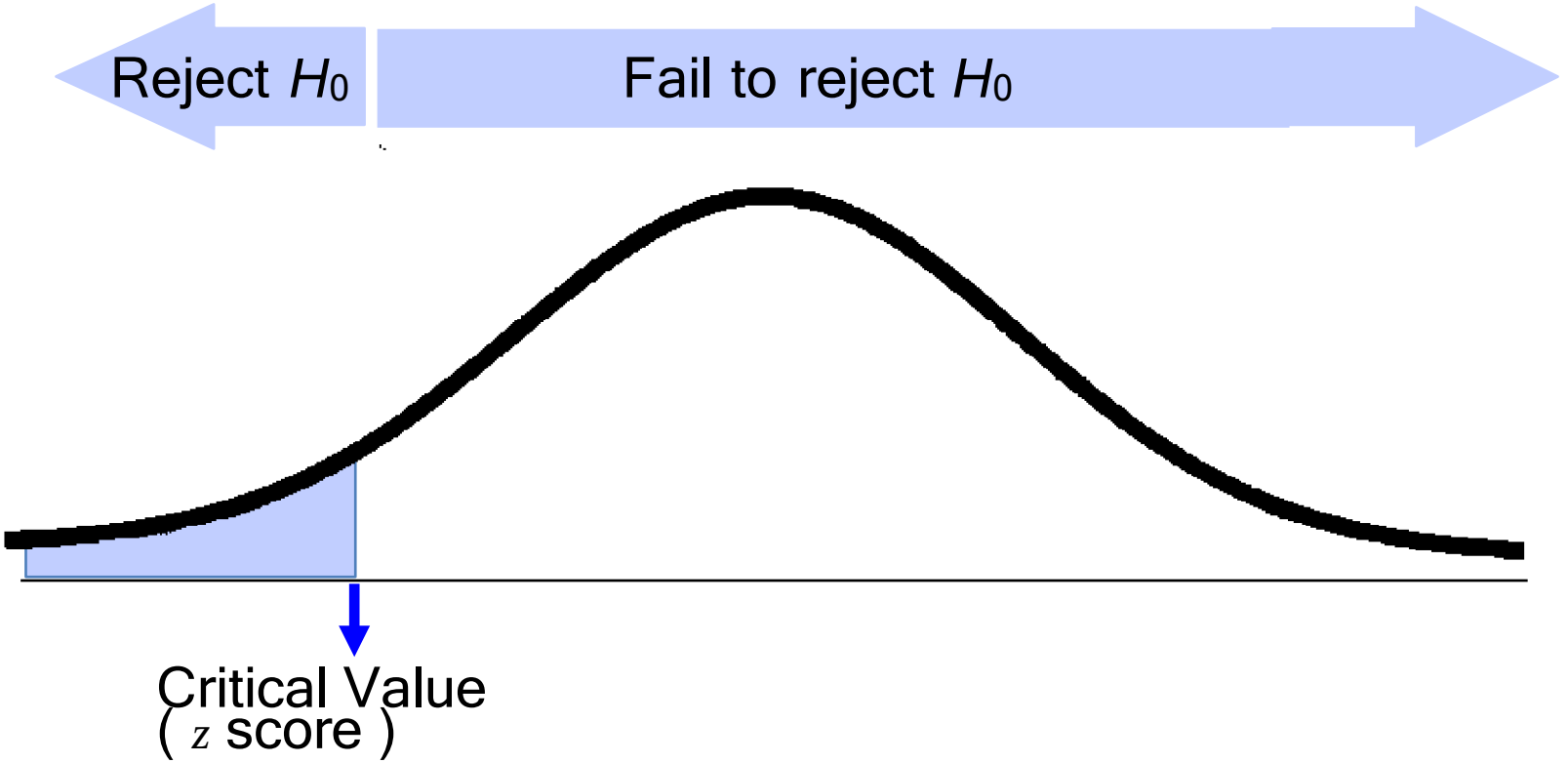



Critical
Region


Critical
Regions

Critical Value

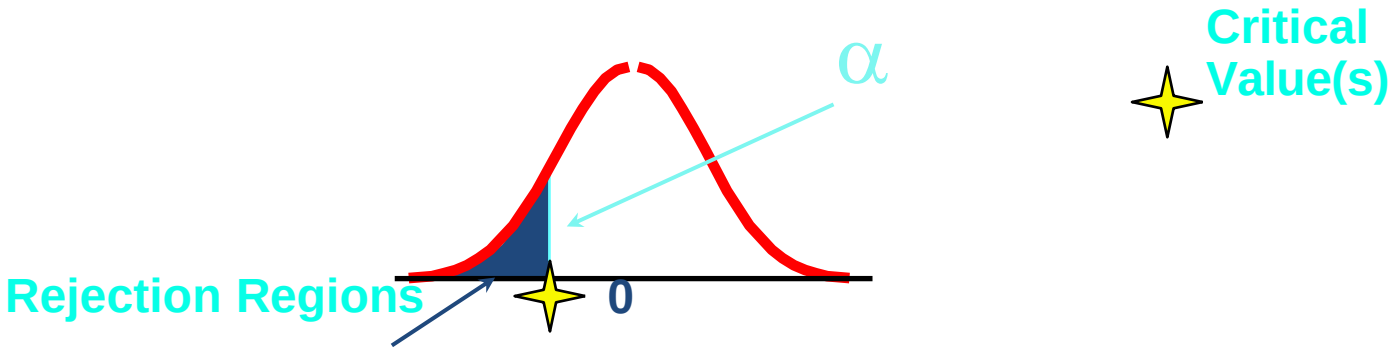
Any value that separates the critical region (where we reject the null hypothesis) from the values of the test statistic that do not lead to a rejection of the null hypothesis



Level of Significance, α and the Rejection Region

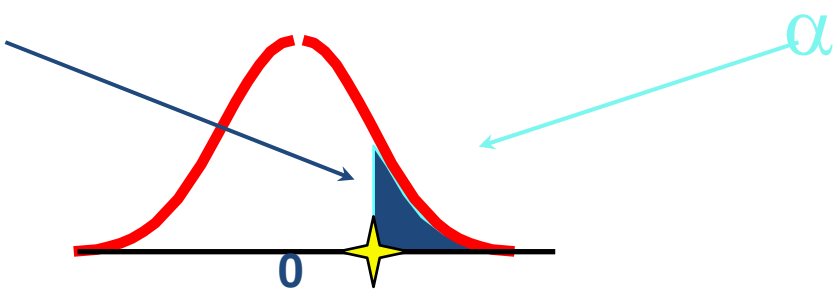
$H_0: \mu \geq 3$

$H_1: \mu < 3$



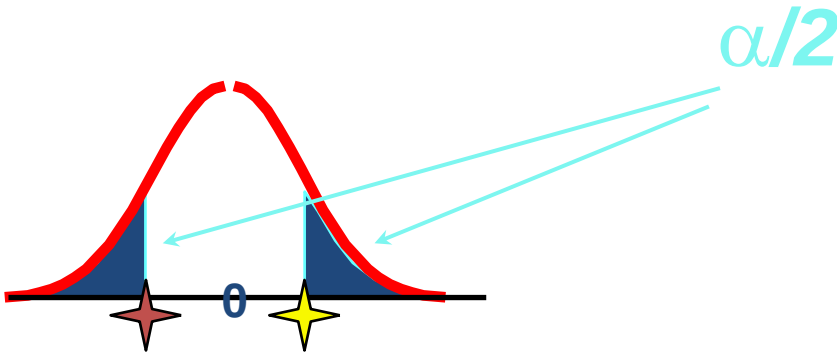
$H_0: \mu \leq 3$

$H_1: \mu > 3$



$H_0: \mu = 3$

$H_1: \mu \neq 3$



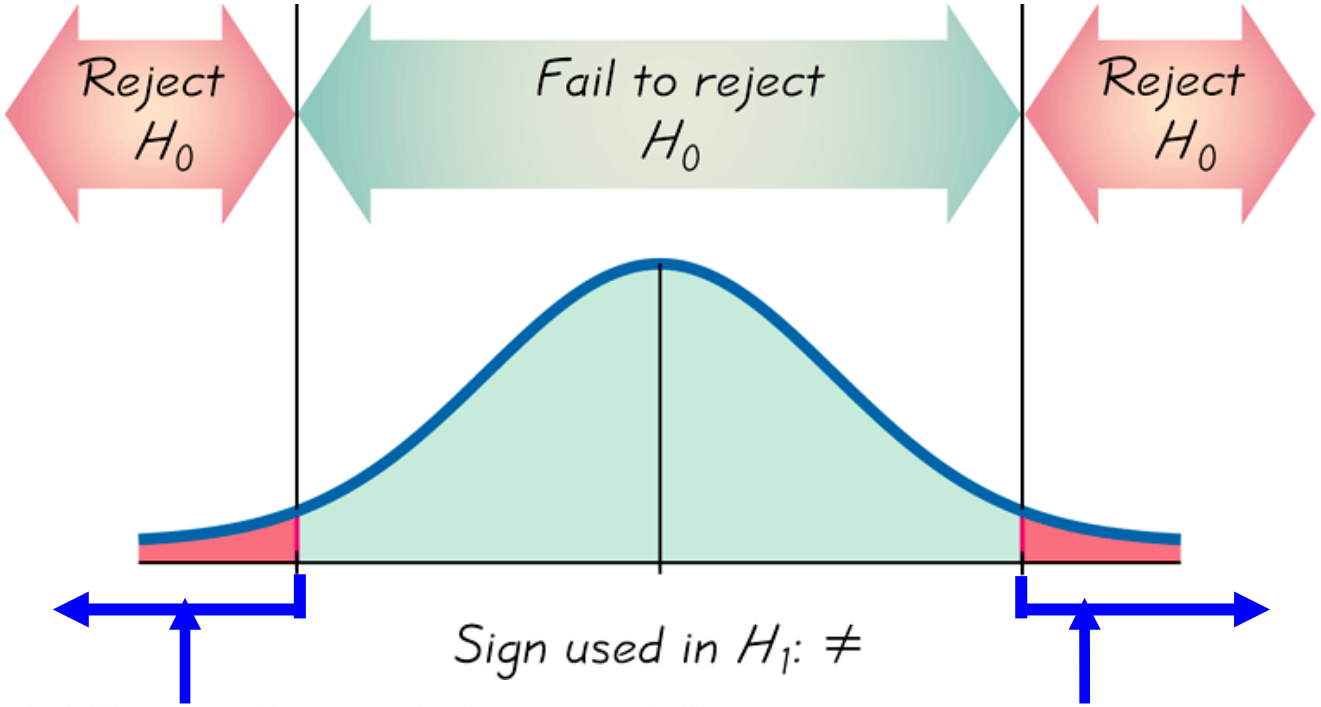
Two-tailed Test

H_0: =

H_1: ≠

α is divided equally between the two tails of the critical region

Means less than or greater than



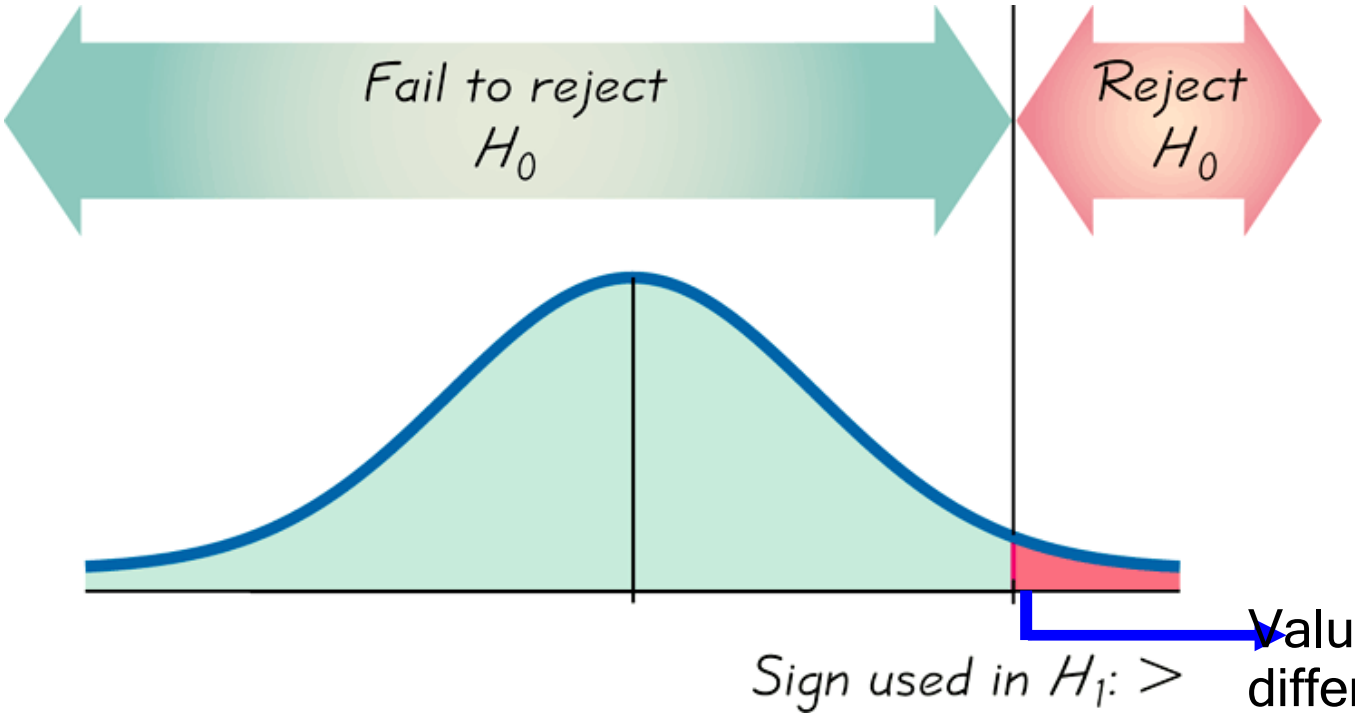
Values that differ significantly from H_0

Right-tailed Test

H_0: =

H_1: >

Points Right

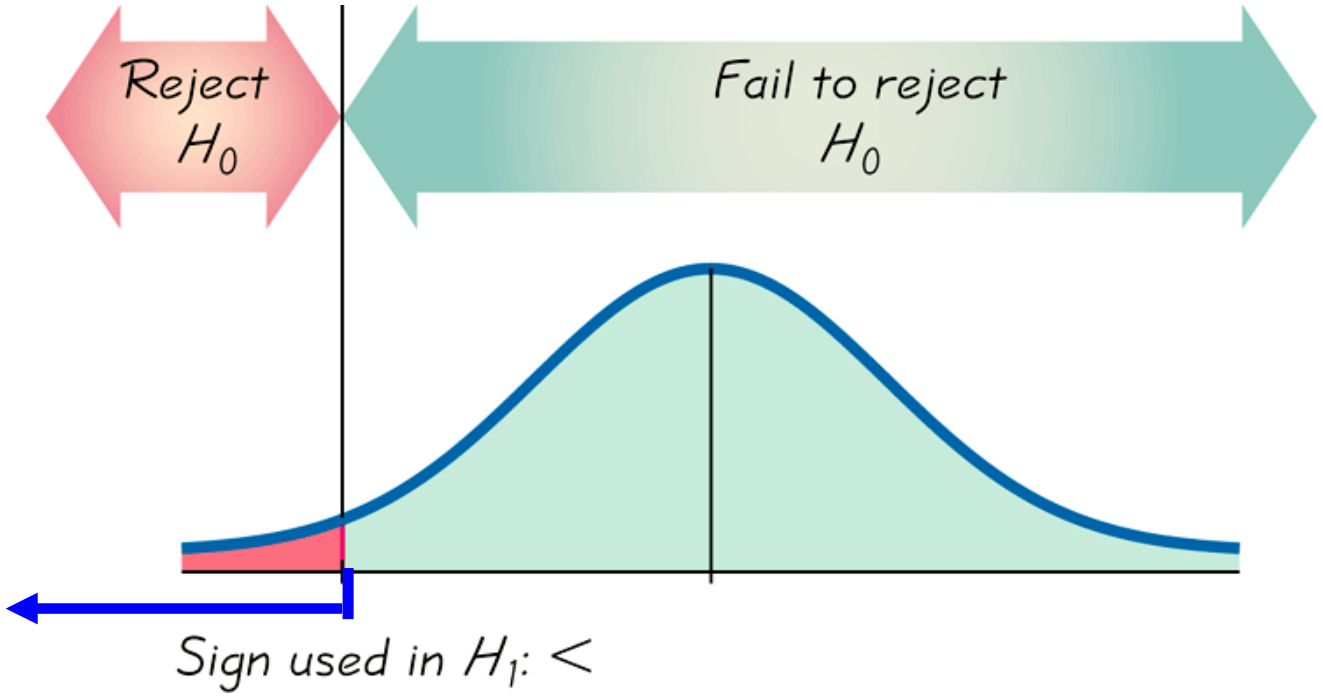


Left-tailed Test

$H_0: =$

$H_1: <$

Points Left



Values that differ significantly from H_o

General Procedure For Testing Hypothesis

Types of Hypothesis Tests Based on Z-distribution

Case-1: To test whether the mean of a normal population is equal to a specified value μ_0 when the population standard deviation σ is known.

Case-2: To test whether the mean of a normal population is equal to a specified value when the population standard deviation is unknown and sample size is large.

Case-3: To test whether the mean of a non-normal population is equal to a specified value when the sample size is large.

Types of Hypothesis Tests Based on Z-distribution

Case-4: To test whether the difference between means of two normal distributions $(\mu_1 - \mu_2)$ is equal to a specified value when σ_1 and σ_2 are known.

Case-5: To test whether the difference between means of two normal distributions is equal to a specified value when σ_1 and are σ_2 unknown.

Case-6: To test whether the difference between means of two non-normal distributions is equal to a specified value when samples sizes are large.

Types of Hypothesis Tests Based on Z-distribution

Case-7: To test whether population proportion is equal to a specified value when the sample size is large.

Case-8: To test whether the difference of population proportion is equal to a specified value when the sample size is large.

Case-1:Testing hypothesis about mean of normal population when σ is known.

(a) Example

Does an average box of cereal contain **368** grams of cereal? A random sample of **25** boxes showed $\bar{X} = 372.5$ and company has specified $\sigma = 15$ grams. Test at the $\alpha = 0.05$ level.

(b) Example:

The heights of college male students are known to be normally distributed with a mean of 67.39 inches and $\sigma = 1.30$ inches. A random sample of 400 students showed a mean height of 67.47 inches. Using a 0.05 significance level, test the hypothesis $H_0 : \mu = 67.39$ against the alternative $H_1 : \mu > 67.39$

(c) Example:

A random sample of size 36 is taken from a normal population with a known variance $\sigma^2 = 25$. If the mean of the sample is 42.6, test the null hypothesis $\mu = 45$ against the alternative hypothesis $\mu < 45$ with $\alpha = 0.05$.

Case-2:Testing hypothesis about mean of normal population when σ is unknown.

Example

The IQ of the college students are known to be normally distributed with a mean of 123. A random sample of 49 students showed an average IQ of $\bar{X} = 120.67$ and S=8.44. Test the hypothesis that $\mu = 123$ against the alternative that it is less. Let $\alpha = 0.05$

Example

Can we reject the claim that the average age of the members of parliament is at least 50, if a random sample of 36 members has a mean of 48.7 with a standard deviation of 3.1 years. Assume all members ages are normally distributed, test at 0.01 level.

STA 406 - Statistical Inference

Ayesha Sultan

Lecturer

Virtual university of Pakistan



STA 406 - Statistical Inference

Lecture No.5

Case-3:Testing Hypothesis about Mean of Non-Normal Population when sample size is large.

Example:

A sample of size 40 from a non-normal population yielded the sample mean $\bar{X} = 71$ and $S^2 = 200$. Test $H_0 : \mu = 72 \, against \, H_1 : \mu \neq 72$, using 1% level of significance.

Case-4:Testing Hypothesis about Difference between Two Population Means $(\mu_1 - \mu_2)$ when σ_1 and σ_2 known .

Example:

A sample of size 6 from a normal population with variance 24 gave $\bar{X}_1 = 15$. A sample of 8 from normal population with variance 80 gave $\bar{X}_2 = 13$. Test $H_0 : \mu_1 - \mu_2 = 0 \; against \; H_1 : \mu_1 - \mu_2 \neq 0$. Let $\alpha = 0.05$

Example:

The two samples A and B detailed below were taken from normal populations of standard deviation 2.5. Decide whether the difference of sample mean is significant at the 0.05 level of significance.

A	16	18	23	26	19	24	25	23	21	22
B	20	21	23	25	25	27	24	26	24	28

Case-5:Testing Hypothesis about Difference between Two Population Means $(\mu_1 - \mu_2)$ when σ_1 and σ_2 are unknown.

Example:

A sample of heights of 6400 Englishmen have a mean of 67.85 inches and a standard deviation of 2.56 inches, while a sample of heights of 1600 Australians has a mean of 68.55 inches and standard deviation of 2.52 inches. Do the data indicate that Australians are on the average taller than Englishmen? Use $\alpha = 0.05$.

Example:

A potential buyer of light bulbs bought 50 bulbs of each of 2 brands. Upon testing the bulbs, he found that brand A had a mean life of 1282 hours with a standard deviation of 80 hours, whereas brand B had a mean life of 1208 hours with a standard deviation of 94 hours. Can the buyer be quite certain that the two brands do differ in quality?

Case-6:Testing Hypothesis about Difference between Two Population Means $(\mu_1 - \mu_2)$ when both populations are non-normal and sample size is large.

Example:

A random sample of 80 light bulbs manufactured by company A had an average lifetime of 1258 hours with standard deviation of 94 hours, while random sample of 60 light bulbs manufactured by company B had an average of 1029 hours with a standard deviation of 68 hours. Because of high cost of bulbs from company A, we are inclined to buy from company B unless the bulbs from company A will last over 200 hours longer on average than those from company B. Run a test using $\alpha = 0.01$ to determine from whom we should buy our bulbs.

Case-7:Testing Hypothesis about Population Proportion when sample size is large.

Example:

A basketball player has hit on 60% of his shots from the floor. If on the next 100 shots he makes 70 baskets, would you say that his shooting has improved? Use a 0.05 level of significance.

Example:

A coin is tossed 900 times and heads appear 490 times. Does the result support the hypothesis that coin is unbiased?

Case-8:Testing Hypothesis about Difference between Two Proportions.

Example:

A random sample of 150 light bulbs manufactured by a firm X showed 12 defective bulbs while a random sample of 100 light bulbs manufactured by another firm Y showed 4 defective bulbs. Is there a significant difference between the proportions of two firms?

Example:

A machine puts out 16 imperfect articles in a sample of 500. After machine is repaired, it puts out 3 imperfect articles in a batch of 100. Has the machine been improved?

STA 406 - Statistical Inference

Ayesha Sultan
Lecturer
Virtual university of Pakistan



STA 406 - Statistical Inference

Lecture No.6

Student’ s t-distribution

Sometimes sample sizes are small and often we do not know the variance of the population. When either of these problems occur, statisticians rely on **t-distribution** (**Student’s t-distribution**) which is a probability distribution that is used to estimate population parameters by using the statistics:

$$t = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$

Where $\bar{X}$ is the sample mean, μ is the population mean, s is the standard deviation of the sample, and n is the sample size.

Degrees of Freedom

Degrees of freedom play an important role when using the Student *t*-distribution. There is different *t*-distribution for every degree of freedom . If our sample size is *n*, then the number of degrees of freedom is *n* - 1. For instance, a sample size of 22 would require us to use the row of the *t*-score table with 21 degrees of freedom.

df\p	0.40	0.25	0.10	0.05	0.025	0.01	0.005	0.0005
1	0.324920	1.000000	3.077684	6.313752	12.70620	31.82052	63.65674	636.6192
2	0.288675	0.816497	1.885618	2.919986	4.30265	6.96456	9.92484	31.5991
3	0.276671	0.764892	1.637744	2.353363	3.18245	4.54070	5.84091	12.9240
4	0.270722	0.740697	1.533206	2.131847	2.77645	3.74695	4.60409	8.6103
5	0.267181	0.726687	1.475884	2.015048	2.57058	3.36493	4.03214	6.8688
6	0.264835	0.717558	1.439756	1.943180	2.44691	3.14267	3.70743	5.9588
7	0.263167	0.711142	1.414924	1.894579	2.36462	2.99795	3.49948	5.4079
8	0.261921	0.706387	1.396815	1.859548	2.30600	2.89646	3.35539	5.0413
9	0.260955	0.702722	1.383029	1.833113	2.26216	2.82144	3.24984	4.7809
10	0.260185	0.699812	1.372184	1.812461	2.22814	2.76377	3.16927	4.5869
11	0.259556	0.697445	1.363430	1.795885	2.20099	2.71808	3.10581	4.4370
12	0.259033	0.695483	1.356217	1.782288	2.17881	2.68100	3.05454	4.3178
13	0.258591	0.693829	1.350171	1.770933	2.16037	2.65031	3.01228	4.2208
14	0.258213	0.692417	1.345030	1.761310	2.14479	2.62449	2.97684	4.1405
15	0.257885	0.691197	1.340606	1.753050	2.13145	2.60248	2.94671	4.0728
16	0.257599	0.690132	1.336757	1.745884	2.11991	2.58349	2.92078	4.0150
17	0.257347	0.689195	1.333379	1.739607	2.10982	2.56693	2.89823	3.9651
18	0.257123	0.688364	1.330391	1.734064	2.10092	2.55238	2.87844	3.9216
19	0.256923	0.687621	1.327728	1.729133	2.09302	2.53948	2.86093	3.8834
20	0.256743	0.686954	1.325341	1.724718	2.08596	2.52798	2.84534	3.8495
21	0.256580	0.686352	1.323188	1.720743	2.07961	2.51765	2.83136	3.8193
22	0.256432	0.685805	1.321237	1.717144	2.07387	2.50832	2.81876	3.7921
23	0.256297	0.685306	1.319460	1.713872	2.06866	2.49987	2.80734	3.7676
24	0.256173	0.684850	1.317836	1.710882	2.06390	2.49216	2.79694	3.7454
25	0.256060	0.684430	1.316345	1.708141	2.05954	2.48511	2.78744	3.7251

Properties of t-distribution

Assumptions

To use the t-distribution, we make the following assumptions.

(i) The sample of n observations is selected randomly.

(ii) The population from which the small sample is drawn, is normal; however, it has shown that slight departures from normality do not seriously affect the tests.

(iii) In case of two small samples, both the samples are selected randomly, both the populations are normal and both the populations have equal variances.

Small Samples Tests About Mean

Case no.1: Testing hypothesis about Mean of a Normal Population when σ is unknown and n< 30.

Case no.2 : Testing hypothesis about difference of Means of two Normal Populations when $\sigma_1 = \sigma_2$

Case no.3: Testing hypothesis about difference of Means of two Normal Populations when $\sigma_1 \neq \sigma_2$

Case no.4: Testing hypothesis about two means with paired observations.

Case no.1: Testing hypothesis about Mean of a Normal Population when σ is unknown and $n < 30$.

Example

A sample of twelve jars of peanut butter was taken from a lot, each jar being labeled “8 ounces net weight.” The individual weights in ounces are: 8.2, 8.0, 7.6, 7.6, 7.7, 7.5, 7.3, 7.4, 7.5, 8.0, 7.4, and 7.5. Test whether these values are consistent with a population mean of 8. Assume that the weights are normally distributed.

Example

A random sample of 16 values from a normal population showed a mean of 41.5 inches and a sum of squares of deviations from this mean equal to 135 $(\text{inches})^2$. Show that the assumption of a mean of 43.5 inches for the population is not reasonable.

Case no.2: Testing hypothesis about difference of Means of two Normal Populations when $\sigma_1 = \sigma_2$

Example

The heights of six randomly selected sailors are in inches: 63, 65, 68, 69, 71 and 72. Those of ten randomly selected soldiers are 61, 62, 65, 66, 69, 70, 71, 72 and 73. Discuss in the light of these data that soldiers are on the average taller than sailors. Assume that the heights are normally distributed.

Example

A group of 12 children are found to have the following intelligence quotients: 112, 109, 125, 113, 116, 131, 112, 123, 108, 113, 132 and 128. A second group of 10 children is tested, resulting in the following IQ's: 117, 110, 106, 109, 116, 119, 107, 106, 105 and 108. Is this group significantly different from the first group?

Case no.3: Testing hypothesis about difference of Means of two Normal Populations when $\sigma_1 \neq \sigma_2$

Example

Given two random samples of size $n_1=9$ and $n_2=16$, from two independent normal populations, with $\bar{X}_2 = 60$ and $\bar{X}_1 = 75$ $s_1=13.61$and $s_2=11.05$, test the hypothesis at the 0.05 level of significance that $\mu_1 = \mu_2$ against the alternative $\mu_1 > \mu_2$. Assume the populations have unequal variances.

Case no.4: Testing hypothesis about two means with paired observations.

Example

The weights of 4 persons before they stopped smoking and 5 weeks after they stopped smoking are as follows:

Person	1	2	3	4
Before	148	176	153	116
After	154	176	151	121

Use the t-test for paired observations to test the hypothesis at the 0.05 level of significance, that giving up smoking has no effect on a person's weight.

Example

To verify whether a course in statistics improved performance, a similar test was given to 12 participants both before and after the course. The original grades recorded in alphabetical order of the participants were 44, 40, 61, 52, 32, 44, 70, 41, 67, 72, 53 and 72. After the course , the grades were in the same order 53, 38, 69, 57, 46, 39, 73, 48, 73, 74, 60 and 78.

Was the course useful, as measured by performance on the test? Consider these 12 participants as a sample from a population.

STA 406 - Statistical Inference

Nabeela Muzaffar

Instructor Statistics

Virtual University of Pakistan



STA 406 - Statistical Inference

Lecture No.7

Contents

- Association of Attributes
- Independence
- Test for independence

Association of Attributes

Variables and Attributes:

Quantitative data usually consist of measurable characteristics called “Variable”. Such as the annual income of family, the weight of a student, the age of a child, etc.

While the marital status of a man, the color of a toy and brand of the cigarettes cannot be measured numerically. Such non-measureable characteristics are qualitative data. A presence or absence can be described is called an attributes.

Differentiate between Variables and Attributes

- Performance of student
- The weight of the person is 68 kg.
- Marital Status
- The marks obtained by student in mathematics is 80.
- Religion
- The age of the person is 54 years.
- A teacher rank students according to their oral language skills
- The people are categorized according to their level of education
- The height of trees at a nursery.

Classification by Attributes

When a population is divided into distinct and mutually exclusive classes according to whether the individuals or objects possess or do not possess a particular characteristics. For example sex is divided into two classes such as male and female or smoking habit is divided as smokers and non-smokers.

- Two-fold division (Dichotomy)

  Gender, Smoking Habit, Marital Status
- Three-fold division (Trichotomy)

  Rain, Performance of Student, Race
- Mani-fold division

  Religion, Language Skills, Level of Education

Notations and Terminology

- We use capital letters A, B, C, … to denote the presence of attributes and Greek letters α, β, γ, … to denote the absence of these attributes.
- The attributes denoted by A, B, C, … are called positive attributes while those denoted by α, β, γ, … are called negative attributes.
- If we are studying two or more attributes, then their combinations can be represented by combining the letters representing the attributes.
- The number of observations assigned to any class is called the class frequency. Class frequencies are denoted by enclosing the class symbols in brackets.

Order and Number of Classes and Class Frequencies

The order of a class and class frequency is determined by the number of attributes specifying a class or class frequency. For example A, AB, ABC, … are classes of first, second, third, … order respectively and (A), (AB), (ABC), … are class frequencies of first, second, third, … order respectively.

The total number of observations denoted by N, is taken as class frequency of zero order because it specifies no attribute.

Example

- If we are considering one attribute such as **A**, then

  Number of class frequencies = 3 (N, (A), (α))

- If the number of attributes is two represented by **A** and **B**, then

  Number of class frequencies $=(3)^2$ = 9 (N, (A), (B), (α), (β), (AB), (Aβ), (αB) and (αβ)

- If the number of attributes is **n**, then

  Number of class frequencies = $(3)^n$ = 27

Ultimate Class Frequencies

The class frequencies of the highest order are called the ultimate class frequencies. For example, if we consider n attributes, then the number of ultimate class frequency is $(2)^n$. We can express every class frequency as the sum of ultimate class frequencies. For example,

$(A) = (AB) + (A\beta) = (ABC) + (AB\gamma) + (A\beta C) + (A\beta\gamma)$

The frequency on the right side are some of the ultimate class-frequencies, if we consider three attributes A, B and C.

In order to express any class-frequency in terms of the known class-frequencies, we can treat the class symbols as operators and can multiply them together like algebraically quantities. Writing A.N for the operation of N according attribute A, we can write

Cont…

$A.N = (A);$

Similarly, $\alpha.N = (\alpha)$

Adding, we get $A.N + \alpha.N = (A) + (\alpha)$

$(A + \alpha).N = N$ $[N = (A) + (\alpha)]$

Omitting N, we get a symbolic relation as $A + \alpha = 1$. Hence we can replace A by $1 – \alpha$ and α by $1 – A$. Similarly, their positive operators are:

$(\alpha B) = \alpha B.N = (1 – A)B.N = (B – AB).N = (B) – (AB)$

$(\alpha\beta\gamma) = \alpha\beta\gamma.N = (1 – A)(1 – B)(1 – C).N$

$= (1 – A – B – C + AB + AC + BC – ABC).N$

$= N – (A) – (B) – (C) + (AB) + (AC) + (BC) – (ABC)$

Consistence

Any class frequencies which have been or might have been observed within one and the same population may be said to be consistent with one another, because they conform with one another and do not conflict in any way. The necessary and sufficient condition for the consistence of a set of independent class frequencies is that no ultimate class frequency be negative. Hence to test the consistence, we calculate the ultimate class-frequencies from the given data. If any ultimate class-frequency turns out to be negative, the given set of class frequency will be inconsistent.

Example

A market investigator returns the following data of 1000 people consulted, 811 liked chocolates, 758 liked toffee and 418 liked boiled sweets, 570 liked chocolates and toffee, 356 liked chocolates and boiled sweets and 348 liked toffee and boiled sweets; 297 liked all three. Show that this information must be incorrect.

Solution:

N = 1000, (A) = 811, (B) = 758, (C) = 418, (AB) = 570, (AC) = 356, (BC) = 348, (ABC) = 297.

The test of consistence is that no ultimate class-frequency should be negative.

Cont…

Now $(\alpha\beta\gamma) = \alpha\beta\gamma.N = (1-A)(1-B)(1-C).N$

$= N – (A) – (B) – (C) + (AB) + (AC) + (BC) – (ABC)$

$= 1000 – 811 – 752 – 418 + 570 + 356 + 348 – 297$

$= -4$, which is negative.

Hence the information is not correct.

Independence

If there is no relationship of any kind between two attributes A and B, we expect to find the same proportion of A's amongst the B's as amongst the β's.

Criterion of Independence

In case of two attributes A and B, the frequencies are grouped into a table with two rows and two columns as in the following table.

Attribute	B	β	Total
A	(AB)	(Aβ)	(A)
α	(αB)	(αβ)	(α)
Total	B	β	N

Cont….

Two attributes A and B are independent if $\frac{(AB)}{(B)} = \frac{(A)}{N}$, i.e. the proportions of A's amongst the B's is the same as in the population at large. This equation can be written $(AB) = \frac{(A)(B)}{N}$. Following the criterion of independence, the above table is replaced by the following table.

Attribute	B	β	Total
A	(AB)/N	(Aβ)/N	(A)
α	(αB)/N	(αβ)/N	(α)
Total	B	β	N

Cont…

The cell frequencies in the table are called expected cell frequencies. In general, the expected cell frequency in the *i*th row and *j*th column (E_{ij}) is given by $(E_{ij}) = R_i \bullet C_j / N$ where R_i and C_j are the corresponding row and column totals and N is the total frequency.

Association

Two words say A and B are associated only if they appear together a large number of times than it is expected, if they are independent.

The association can be measured by:

1. Coefficient of Association
2. Coefficient of Colligation

Coefficient of Association

The degree of association between two attributes A and B can be measured using Yule's coefficient of association defined by the formula.

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

Like the coefficient of association, it lies between -1 and +1. For complete association Q = +1 and for complete disassociation Q = -1. When Q = 0, the attributes are independent.

Coefficient of Colligation

Another coefficient used to measure the degree of association is the coefficient of colligation defined by the formula

$$Y = \frac{1 - \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}{1 + \sqrt{\frac{(A\beta)(\alpha B)}{(AB)(\alpha\beta)}}}$$

The coefficient of association and the coefficient of colligation have the same properties and are related by the equation.

$$Q = \frac{2Y}{1 + Y^2}$$

Example

Find the coefficient of association by yule's method between inoculation against cholera and exemption from attack (not attack) for the following data.

	Not attacked	Attacked
Inoculated	276	3
Not inoculated	473	66

Solution

If A and B denote inoculation and exemption from attack (not attack) respectively, then the data are presented in the following form.

	Not attacked	Attacked
Inoculated	(AB) 276	(Aβ) 3
Not inoculated	(αB) 473	(αβ) 66

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)} = \frac{(276)(66) - (3)(473)}{(276)(66) + (3)(473)} = \frac{16797}{19635} = 0.855.$$

There is high positive association between inoculation and exemption from attack.

Contingency Table

Suppose that the attribute A is divide into s classes and the attribute B into t classes, then there will be $s \times t$ classes of the table. A table consisting of s rows and t columns in which the data are classified according to two attributes A and B is called an $s \times t$ contingency table. The simplest form of a contingency table is the 2×2 table in which the two attributes are dichotomized. The value in the cells of a contingency table are called cell frequencies. The row and column totals are called marginal probabilities.

Attributes	B_1	B_2
A_1	(A_1B_1)	(A_1B_2)
A_2	(A_2B_1)	(A_2B_2)

Test for Independence

The test of the null hypothesis

H_0 : two characteristics or criteria of classification are independent

against the alternate hypothesis

H_1: two characteristics or criteria of classification are not independent

is based on the chi-square (χ^2) statistic. The test statistic compares the expected and the observed cell frequencies.

$$\chi^2 = \sum_i \left(\frac{(O_i - E_i)^2}{E_i} \right)$$

The distribution of the statistic approximates the chi-square distribution with $\nu = (r-1)(c-1)$ degrees of freedom.

Cont…

For a specified level of significance α, if the calculated value of χ^2 is greater than the critical value of χ^2_{α}, then the null hypothesis is rejected otherwise the null hypothesis will be accepted.

Example

The following table shows the results of an experiment to investigate the effect of vaccination of laboratory animals against a particular disease. Using 0.05 significance level, test the hypothesis that there is no difference between vaccinated and unvaccinated groups.

	Got disease	Did not get the disease
Vaccinated	9	42
Not vaccinated	17	28

Solution

Formulation of Hypothesis:

H_0 :*Thevaccinationand thediseaseareindependent*

H_1 :*Thevaccinationand thediseasearenot independent*

Level of Significance:

$\alpha = 0.05$

Test Statistics:

$$\chi^2 = \sum_i \left(\frac{(O_i - E_i)^2}{E_i} \right)$$

Cont….

O_i	E_i	$(O_i - E_i)^2$	$(O_i - E_i)^2 / E_i$
9	13.8125	23.1602	1.6768
42	37.1875	23.1602	0.623
17	12.1875	23.1602	1.90
28	32.8125	23.1602	0.706
			4.91

$$\chi^2 = \sum_i \left(\frac{(O_i - E_i)^2}{E_i} \right) = 4.91$$

Cont…

Critical Region:

Reject Ho if

$$\chi^2 > \chi^2_{\alpha\,(r-1,c-1)}$$

$$4.91 > \chi^2_{0.05(1,1)}$$

$$4.91 > \chi^2_{0.05(1)}$$

$$4.91 > 3.84$$

we reject Ho.

Conclusion:

Since the value fall in the critical region, so we reject Ho or also we conclude that the there is association between vaccination and disease.

STA 406 - Statistical Inference

Amber Asghar

Instructor Statistics

Virtual University of Pakistan



STA 406 - Statistical Inference

Lecture No.8

Contents

- Pearson's Chi square test for Goodness of Fit
- Chi-square test of Homogeneity

Pearson's Chi square test for Goodness of Fit

- An application of the χ^2 distribution is the test of goodness of fit.

- The test is applied when you have one categorical variable from a single population. It is used to determine whether sample data are consistent with a hypothesized distribution.

- Using the χ^2 distribution, it is possible to test the hypothesis that a population has a specified theoretical distribution. The theoretical distribution may be normal, the binomial, Poisson, or any other form.

Conditions for χ^2 testing

- The chi-square goodness of fit test is appropriate when the following conditions are met:
- The sampling method is simple random sampling.
- The population is at least 10 times as large as the sample.
- The variable under study is categorical.

Cont…

- The expected value of the number of sample observations in each level of the variable is at least 5.
- The sampling distribution of χ^2 approaches the chi-square distribution.
- The test involves a comparison of the observed frequency distribution in the sample with the expected frequency distribution.

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

- The symbols Oi and Ei represent the observed and expected frequencies respectively for the ith class and k represents the number of possible outcomes or the number of different classes.

Cont…

- If the observed frequencies are close to the corresponding expected frequencies, the computed χ^2 value will be small, indicating a good fit otherwise poor fit.
- The number of degrees of freedom for this chi-square distribution is $\upsilon = k-1$, when class parameters are known. If expected frequencies can be computed by estimating m population parameters from the sample statistics then the degree of freedom of chi-square distribution is $\upsilon = k-1-m$.

Steps for χ^2 testing

The following steps are involved for χ^2 hypothesis testing:

Formulation of Hypothesis:

$$H_0 : The\ fit\ is\ good$$

$$H_1 : The\ fit\ is\ not\ good$$

Level of Significance:

α = 1%, 5% and 10%

Test Statistics:

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

Cont…

Critical Region:

Reject H_0, if

$$\chi^2_{cal} > \chi^2_{\alpha\,(k-1-m)}, \; If \;\; the \;\; parameter \; is unknown.$$

$$\chi^2_{cal} > \chi^2_{\alpha\,(k-1)}, \; If \;\; the \;\; parameter \; is \; known.$$

Conclusion:

Example

Five pennies were tossed 1,000 times and the number of heads were observed as given below:

Number of Heads	0	1	2	3	4	5
Frequencies	38	144	342	287	164	25

Test whether a binomial distribution gives a satisfactory fit to these data.

Solution

Formulation of Hypothesis:

H_0 : *The fit is binomial*.

H_1 : *The fit is not binomial*.

Level of Significance:

$\alpha = 0.05$

Test Statistics:

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

Cont…

Calculation:

To estimate the value of p, we first compute the mean number of heads, $\bar{x}$. Thus

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{2470}{1000} = 2.47.$$

Theoretically, $\bar{x} = np$, so that $p = \frac{\bar{x}}{n} = \frac{2.47}{5} = 0.494.$

Letting the random variable X represent the number of heads, the above calculations yield the fitted binomial distribution as

$$b(x;5,0.494) = \binom{5}{x}(0.494)^x(0.506)^{n-x}$$

Cont…

Number of Heads	Observed f, (o_i)	Probability P(x)	Expected f, $(e_i) = N \times \mathrm{P(x)}$	$(o_i - e_i)$	$(o_i - e_i)^2 / e_i$
0	38	0.0332	33.2	4.8	0.69
1	144	0.1619	161.9	-17.9	1.98
2	342	0.3162	316.2	25.8	2.15
3	287	0.3087	308.7	-21.7	1.53
4	164	0.1507	150.7	13.3	1.17
5	25	0.0294	29.4	-4.4	0.66
Total	1,000		1,000.1		$\chi^2 = 8.18$

Chi-square Table

The entries in this table are values of $x^2_\alpha(\nu)$, for which the area to their right under the chi-square distribution with ν degrees of freedom is equal to α.

Upper Percentage Points of the Chi-square Distribution									
ν \ α	0.99	0.98	0.975	0.95	0.10	0.05	0.03	0.02	0.01
1	0.0002	0.001	0.001	0.004	2.71	3.84	5.02	5.41	6.64
2	0.020	0.040	0.051	0.103	4.61	5.99	7.38	7.82	9.21
3	0.115	0.185	0.216	0.352	6.25	7.82	9.35	9.84	11.34
4	0.297	0.429	0.484	0.711	7.78	9.49	11.14	11.67	13.28
5	0.554	0.752	0.831	1.145	9.24	11.07	12.83	13.39	15.09
6	0.87	1.13	1.24	1.64	10.64	12.59	14.45	15.03	16.81
7	1.24	1.56	1.69	2.17	12.02	14.07	16.01	16.62	18.48
8	1.65	2.03	2.18	2.73	13.36	15.51	17.54	18.17	20.09
9	2.09	2.53	2.70	3.32	14.68	16.92	19.02	19.68	21.67
10	2.56	3.06	3.25	3.94	15.99	18.31	20.48	21.16	23.21
11	3.05	3.61	3.82	4.58	17.28	19.68	21.92	22.62	24.72
12	3.57	4.18	4.40	5.23	18.55	21.03	23.34	24.05	26.22
13	4.11	4.76	5.01	5.89	19.81	22.36	24.74	25.47	27.69
14	4.66	5.37	5.63	6.57	21.06	23.68	26.12	26.87	29.14
15	5.23	5.98	6.26	7.26	22.31	25.00	27.49	28.26	30.58

Cont…

	α								
v	**0.99**	**0.98**	**0.975**	**0.95**	**0.10**	**0.05**	**0.025**	**0.02**	**0.01**
16	5.81	6.61	6.91	7.96	23.54	26.30	28.84	29.63	32.00
17	6.41	7.26	7.56	8.67	24.77	27.59	30.19	31.00	33.41
18	7.02	7.91	8.23	9.39	25.99	28.87	31.53	32.35	34.81
19	7.63	8.57	8.91	10.12	27.20	30.14	32.85	33.69	36.19
20	8.26	9.24	9.59	10.85	28.41	31.41	34.17	35.02	37.57
21	8.90	9.92	10.28	11.59	29.62	32.67	35.48	36.34	38.93
22	9.54	10.60	10.98	12.34	30.81	33.92	36.78	37.66	40.29
23	10.20	11.29	11.69	13.09	32.01	35.17	38.08	38.97	41.64
24	10.86	11.99	12.40	13.85	33.00	36.42	39.36	40.27	42.92
25	11.52	12.70	13.12	14.61	34.38	37.65	40.65	41.57	44.31
26	12.20	13.41	13.84	15.38	35.56	38.88	41.92	42.86	45.64
27	12.88	14.12	14.57	16.15	36.74	40.11	43.19	44.14	46.96
28	13.56	14.85	15.31	16.93	37.92	41.34	44.46	45.42	48.28
29	14.26	15.57	16.05	17.71	39.09	42.56	45.72	46.69	49.59
30	14.95	16.31	16.79	18.49	40.26	43.77	46.98	47.96	50.89

Cont…

Critical Region:

Reject H_0 , if

$$\chi^2_{cal} > \chi^2_{\alpha\,(k-1-m)}$$

$$8.18 > \chi^2_{0.05(6-1-1)}$$

$$8.18 > \chi^2_{0.05(4)}$$

8.18 < 9.49.

We don’t reject Ho.

Conclusion:

Since the value 8.18 < 9.49 don’t fall in the critical, so we don’t reject Ho or also we conclude that the fit is binomial.

Example

A skilled typist, on routine work, kept a record of mistakes made per day during 300 working days.

Mistakes per day (x)	0	1	2	3	4	5	6
Number of days (f)	143	90	42	12	9	3	1

Test the hypothesis that X has a Poisson distribution by applying the χ^2 goodness-of-fit-test by using 0.01 level of significance.

Solution

Formulation of Hypothesis:

H_0 :*The fit is poisson*.

H_1 :*The fit is not poisson*.

Level of Significance:

$\alpha = 0.01$

Test Statistics:

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

Cont…

Calculation:

To estimate the value of p, we first compute the mean number of mistakes per day, $\bar{x}$. Thus

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{267}{300} = 0.89.$$

Letting the random variable X represent the number of mistakes per day, the above calculations yield the fitted Poisson distribution as

$$\mathrm{p}(x;0.89) = \frac{e^{-0.89}(0.89)^x}{x!}, \quad for\ x = 0,1,2,...$$

Cont…

Number of Mistakes	Observed f, (o_i)	Probability P(x)	Expected f, $(e_i) = N \times \mathrm{P(x)}$	$(o_i - e_i)$	$(o_i - e_i)^2 / e_i$
0	143	0.4106	123.2	19.8	3.18
1	90	0.3655	109.6	-19.6	3.51
2	42	0.1627	48.8	-6.8	0.95
3	12	0.0482	14.5		
4	9 25	0.0107	3.2 18.4	6.6	2.37
5	3	0.002	0.6		
6	1	0.0003	0.1		
Total	1,000		1,000.1		$\chi^2 = 10.01$

Cont…

Critical Region:

Reject H_0, if

$$\chi^2_{cal} > \chi^2_{\alpha(k-1-m)}$$

$$10.01 > \chi^2_{0.01(4-1-1)}$$

$$10.01 > \chi^2_{0.01(2)}$$

$$10.01 > 9.21.$$

We reject Ho.

Conclusion:

Since the value 10.01 > 9.21 fall in the critical region, so we reject Ho or also we conclude that the fit is not Poisson.

Test of Homogeneity

- The hypothesis to be tested is that two or more different random samples come from the same population or that samples are homogeneous as the word homogeneous is statistics is often used to indicate the same or equal. The χ^2 - test applied in such a situation is called a test of homogeneity.
- When there are two random samples, a simpler method proposed by Brandt and Snedecor is used to calculate the value of χ^2.
- Suppose we draw two independent random samples of size n from a population and we wish to test whether the two samples are homogeneous. Let the values of the two samples be presented in the following $2\times n$ table:

Cont…

	1	2	…	i	….	n	Total
Sample I	a_1	a_2	…	a_i	…	a_n	A
Sample II	b_1	b_2	…	b_i	…	b_n	B
Total	c_1	c_2	…	c_i	…	c_n	N

As the samples are independent, therefore the expected values e_i in the first sample are given by $e_i = \frac{Ae_i}{N}$, and the expected values e_i in the second sample are given by $e_i = \frac{Be_i}{N}$.

Now $\chi^2 = \sum_i \frac{(o_i - e_i)^2}{e_i} = \sum \frac{o_i^2 + e_i^2 - 2o_i e_i}{e_i}$

$$= \sum_i \frac{o_i^2}{e_i} + \sum_i e_i - 2\sum_i o_i = \sum_i \frac{o_i^2}{e_i} - N \qquad [\because \sum e_i = \sum o_i = N]$$

Thus $$\chi^2 = \sum_i \left[\frac{a_i^2}{\frac{Ae_i}{N}} + \frac{b_i^2}{\frac{Be_i}{N}} \right] - N$$

$$= N\sum_i \left[\frac{a_i^2}{Ae_i} + \frac{b_i^2}{Be_i} \right] - N$$

Cont…

$$= \sum_i \left[\frac{a_i^2}{Ae_i} + \frac{e_i^2 + a_i^2 - 2e_i a_i}{Be_i} \right] - N, \qquad \left[\because b_i^2 = (c_i - a_i)^2 \right]$$

$$= N\sum_i \frac{a_i^2}{e_i}\left(\frac{1}{A} + \frac{1}{B}\right) + \frac{N}{B}\sum_i e_i - \frac{2N}{B}\sum_i a_i - N$$

$$= N\sum_i \frac{a_i^2}{e_i}\left(\frac{N}{AB}\right) + \frac{N^2}{B} - \frac{2NA}{B} - N \qquad \left[\because \sum_i e_i = N, \sum_i a_i = A \right]$$

$$= N\sum_i \frac{a_i^2}{e_i}\left(\frac{N}{AB}\right) + \frac{N^2 - N(A+B) - NA}{N} \qquad (put\ N = A+B)$$

$$= \frac{N^2}{AB} \cdot \sum_i \frac{a_i^2}{e_i} - \frac{NA}{B}$$

$$= \frac{N^2}{AB}\left[\sum_i \frac{a_i^2}{e_i} - \frac{A^2}{N} \right],$$ which either row can be chosen as the a_i's.

Example

In a certain community, a random sample of 50 men and another sample of 50 women over 21 years of age were asked about their educational background, classified as junior high, senior high or college. The results are:

	Junior High	Senior High	College
Male	13	25	12
Female	23	20	7

Test whether the two samples are homogeneous in respect of educational levels.

Solution

Formulation of Hypothesis:

H_0 : *The two samples are homogeneous*

H_1 : *The two samples are not homogeneous*.

Level of Significance:

$\alpha = 0.05$

Test Statistics:

$$\chi_2 = \frac{N^2}{AB}\left[\sum_i \frac{a_i^2}{e_i} - \frac{A^2}{N}\right]$$

Which has an approximate χ^2 - distribution with (n-1) degrees of freedom.

Calculation:

Cont…

	Junior High	Senior High	College	Total
Male (a_i)	13	25	12	50 = A
Female (b_i)	23	20	7	60 = B
Total (e_i)	36	55	19	100 = N

$$\chi^2 = \frac{N^2}{AB}\left[\sum_i \frac{a_i^2}{e_i} - \frac{A^2}{N}\right]$$

Now

$$= \frac{(100)^2}{(50)(50)}\left[\frac{(13)^2}{36} + \frac{(25)^2}{45} + \frac{(12)^2}{19} - \frac{(50)^2}{100}\right]$$

$$= 4\,(4.69 + 13.89 + 7.58 - 25) = 4 \times 1.16 = 4.64$$

Critical Region:

Reject Ho if,

$$\chi^2_{cal} > \chi^2_{\alpha((r-1)(c-1))}$$

Cont…

$$4.64 > \chi^2_{0.05((1)(2))}$$

$$4.64 > \chi^2_{0.05(2)}$$

$$4.64 < 5.99$$

We don't Reject Ho.

Conclusion:

Since the value 4.64 < 5.99 don't fall in the critical region, so we don't reject Ho. Or also we conclude that the two samples are homogeneous.

STA 406 - Statistical Inference

S. M. Muslim Raza
Instructor Statistics
Virtual university of Pakistan



STA 406 - Statistical Inference

Lecture No.9

Contents

- Introduction
- Properties of Chi-square distribution
- Confidence interval for σ^2 and σ
- Hypothesis Testing (for σ^2)

Introduction

- Chi-square analysis is primarily used to deal with categorical data.

- Chi-square distribution has vast application such as a test of goodness of fit, test of independence and test of homogeneity.

- The chi-square distribution contains only one parameter “n” ,called the number of degree of freedom.

If Z_1, ..., Z_k are independent, standard normal random variables, then the sum of their squares,

$$Q = \sum_{i=1}^{k} Z_i^2,$$

is distributed according to the **chi-squared distribution** with k degrees of freedom. This is usually denoted as:

$$Q \sim \chi^2(k) \text{ or } Q \sim \chi_k^2.$$

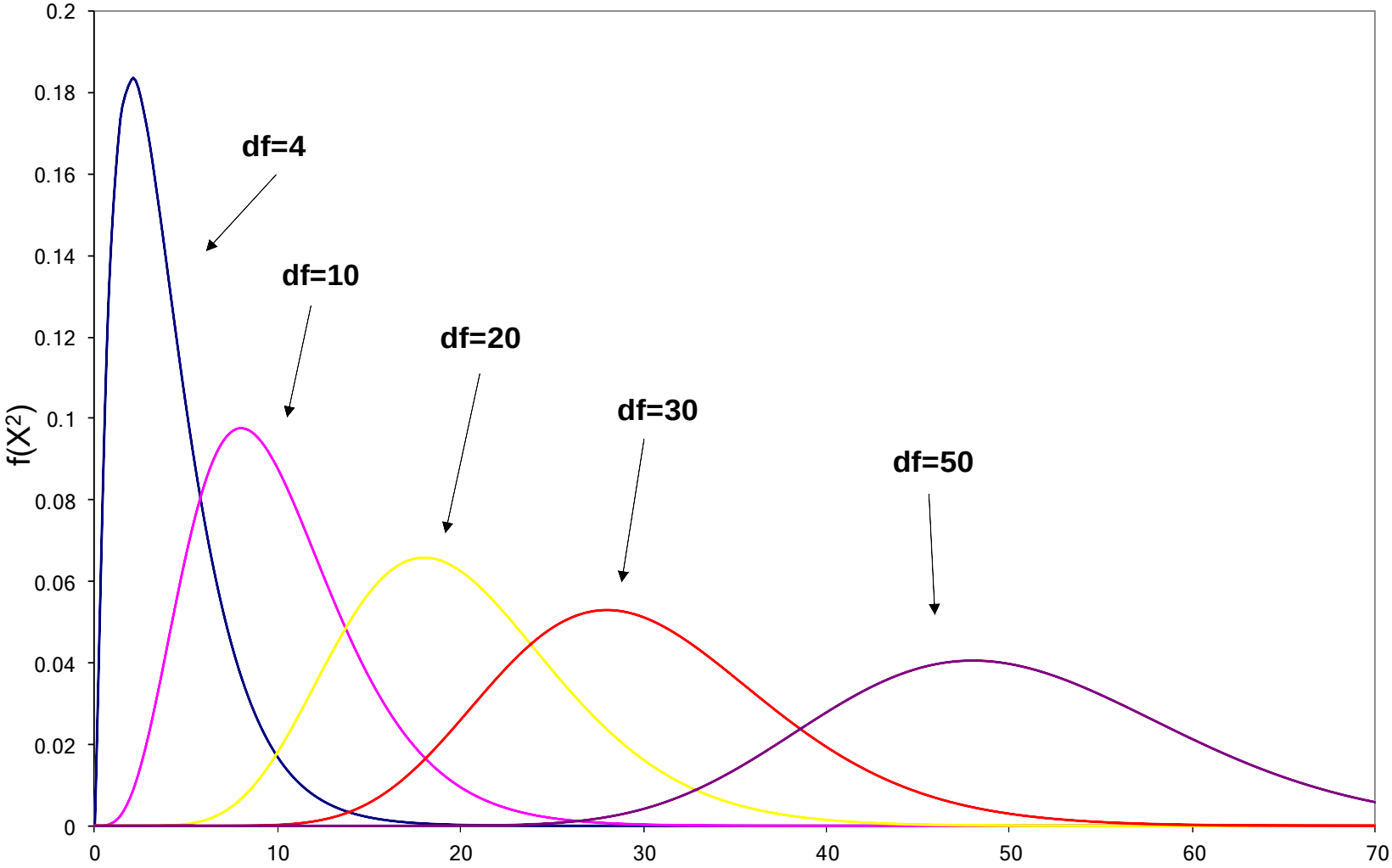
Properties of Chi-square Distribution

- The curve of a χ^2-distribution is positively skewed. The skewness decreases as n increases.
- The χ^2-distribution tends to normal distribution as the number of degrees freedom approaches to infinity.

Chi-square d.f

http://homepage.stat.uiowa.edu/~mbognar/applets/chisq.html

Chi-Square Distributions



- Mean of the Chi-square distribution is =n
- Variance of Chi-square distribution =2n

(Where n is the degree of freedom)

- Additive Property. If X and Y are independent χ^2- random variables with n_1 and n_2 degrees of freedom respectively, then the sum X + Y is a χ^2- random variable with $n_1 + n_2$ degrees of freedom.

- There are many different chi-square distributions, one for each degree of freedom.
- Probability density function (pdf) :

$$\frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)}\, x^{\frac{k}{2}-1} e^{-\frac{x}{2}}$$

- Cumulative density function (cdf) :

$$\frac{1}{\Gamma\left(\frac{k}{2}\right)}\, \gamma\left(\frac{k}{2}, \frac{x}{2}\right)$$

Moment generating function (MGF)

$$M_o(t) = (1 - 2t)^{-k/2} \quad \text{for } t < \tfrac{1}{2}$$

Confidence Interval

(1- α)100% Confidence Interval for σ^2 or σ

Step 1: Obtain a random sample of *n* items from the population, and compute s^2

Step 2: Choose confidence level (1-α)

Step 3: Obtain χ^2_L and χ^2_U from the table of critical values for the chi-square distribution with *n*-1 d.f

Step 4: Compute the confidence interval for σ^2 or σ based on the formula given below:

(1-α) 100% CI for σ^2 : $\left(\frac{(n-1)s^2}{\chi_U^2}, \frac{(n-1)s^2}{\chi_L^2} \right)$

Or

(1-α) 100% CI for σ : $\left(\sqrt{\frac{(n-1)s^2}{\chi_U^2}}, \sqrt{\frac{(n-1)s^2}{\chi_L^2}} \right)$

Step 5: Interpretation

Example

In a study on cholesterol levels a sample of 12 men and women was chosen. The plasma cholesterol levels (mmol/L) of the subjects were as follows: 6.0, 6.4, 7.0, 5.8, 6.0, 5.8, 5.9, 6.7, 6.1, 6.5, 6.3, and 5.8. We assume that these 12 subjects constitute a simple random sample of a population of similar subjects. We wish to estimate the variance of the plasma cholesterol levels with a 95 percent confidence interval.

Step-1: Calculation

6.0 6.4 7.0 5.8 6.0 5.8
5.9 6.7 6.1 6.5 6.3 5.8

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{74.3}{12} = 6.2$$

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

$$s^2(n-1) = \sum (X_i - \bar{X})^2$$

$$\sum (X_i - \bar{X})^2 = (6.0 - 6.2)^2 + (6.4 - 6.2)^2 + ... + (5.8 - 6.2)^2 = 4.31$$

$$s^2(n-1) = 4.31$$

Example

Given that X is normally distributed and from sample values $\bar{x}$ = 42, s = 5 and n = 20. Find the 98 percent confidence interval for σ.

$100(1 - \alpha)\ \% = 98\%$

$1 - \alpha = 0.98$

$\alpha = 1\text{-}0.98 = 0.02$

$$\alpha/2 = 0.02/2 = 0.01$$

$$1 - \alpha/2 = 1 - 0.02/2 = 1 - 0.01 = 0.99$$

Testing of Hypothesis about Variance

Testing Hypothesis about Variance of a Normal Population

Step-1: State Null and alternative hypotheses

$$H_0 : \sigma^2 = \sigma_0^2 \; and \; H_1 : \sigma^2 \neq \sigma_0^2$$
$$H_0 : \sigma^2 \leq \sigma_0^2 \; and \; H_1 : \sigma^2 > \sigma_0^2$$
$$H_0 : \sigma^2 \geq \sigma_0^2 \; and \; H_1 : \sigma^2 < \sigma_0^2$$

Step-2: Decide Significance Level: α

Step-3: Test Statistic to be used

$$\chi^2_{n-1} = \frac{nS^2}{\sigma_0^2}$$

$$or \quad = \frac{(n-1)s^2}{\sigma_0^2}$$

$$or \quad = \frac{\sum (X - \bar{X})^2}{\sigma_0^2}$$

Step-4: Determine the critical region which depends on α and the alternative hypothesis H_1

(a) When H_1 is $\sigma^2 \neq \sigma_0^2$, the critical region is

$$\chi^2 > \chi^2_{\alpha/2,(n-1)} \quad and \quad \chi^2 < \chi^2_{1-\alpha/2,(n-1)}$$

$(TwoTailed\ Test)$

(b) When H_1 is $\sigma^2 > \sigma_0^2$, the critical region falls in the right tail and its value is

$$\chi^2 > \chi^2_{\alpha\,(n-1)}$$

$(One-tailed\ test)$

(c) When H_1 is $\sigma^2 < \sigma_0^2$, the critical region falls will be entirely in the left tail with the critical value $\chi^2 \leq \chi^2_{1-\alpha,(n-1)}$

$(\text{One}-\text{tailedtest})$

Step-5: Compute the value of $\chi^2_{n-1} = \frac{nS^2}{\sigma_0^2}$ from the given data

Step-6: Conclusion/Interpretation

Reject H_0 if the computed value of χ^2 fall in the critical region otherwise accept H_0.

Example

The variability in the tensile strength of a type of steel wire must be controlled carefully. A sample of the wire is subjected to test and it is found that the sample variance is S^2 = 31.5. The sample size was n = 16 observations.

Test the hypothesis that the population variance is 25 against the alternative that the variance is greater than 25. Use a 0.05 level of significance.

STA 406 - Statistical Inference

S. M. Muslim Raza
Instructor Statistics
Virtual university of Pakistan



STA 406 - Statistical Inference

Lecture No.10

Contents

- Introduction
- Properties of F distribution
- Confidence interval
- Hypothesis Testing

The F-Distribution

- It is defined in terms of the ratio of variances of two normally distributed populations. So, it is sometimes also called variance ratio.
- Used for comparing the variances of two populations.
- If $\sigma_1^2 = \sigma_2^2$, then the statistics F = s_1^2 / s_2^2 follows the F distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom.

Properties of F-distribution

- It is positively skewed and always its ranges from 0 to ∞ .
- Mean and variance of F-distribution:

  Mean = $\frac{v_2}{(v_2 - 2)}$ $for\, v_2 > 2.$

  Variance = $\frac{2v_2^2(v_1 + v_2 - 2)}{v_1(v_2 - 2)^2(v_2 - 4)}$, $for\, v_2 > 4.$

- Shape of the F-distribution depends upon the number of degrees of freedom. For different values of v_1 and v_2 we will get different distributions.

- http://homepage.stat.uiowa.edu/~mbognar/applets/f.html

(1-α) 100 percent confidence interval for

$$\sigma_1^2 / \sigma_2^2$$

$$\left[\frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{\alpha/2}(\nu_1,\nu_2)}, \frac{s_1^2}{s_2^2} \cdot F_{\alpha/2}(\nu_2,\nu_1) \right]$$

Example

Given two random samples of size $n_1 = 12$ and $n_2 = 10$ from two independent normal populations, with $s_1 = 2.3$ and $s_2 = 1.5$, find a 90% confidence interval for σ_1^2/σ_2^2 and σ_1/σ_2.

The 90% confidence interval for the ratio σ_1^2 / σ_2^2 is

$$\left[\frac{s_1^2}{s_2^2} \cdot \frac{1}{F_{\alpha/2}(\nu_1,\nu_2)}, \frac{s_1^2}{s_2^2} . F_{\alpha/2}(\nu_2,\nu_1) \right]$$

Here $s_1^2 = (2.3)^2 = 5.29$, $s_2^2 = (1.5)^2 = 2.25$,

$\alpha = 0.10$, $\nu_1 = 12\text{-}1 = 11$ and $\nu_2 = 10\text{-}1 = 9$.

Consulting the F-table, we find that

$F_{0.05}(11,9) = 3.10$ and $F_{0.05}(9,11) = 2.90$.

Substituting these values, we get

$$\left[\frac{5.29}{2.25}\cdot\left(\frac{1}{3.10}\right), \frac{5.29}{2.25}(2.90)\right]$$
$$= (0.76, 6.81)$$

Hence the 90% confidence interval for σ_1^2 / σ_2^2 is (0.76, 6.81).

Taking the square root of the end points

(0.76, 6.81), we get the 90% confidence interval for σ_1/σ_2 as (0.87, 2.61).

Taking the square root of the end points

(0.76, 6.81), we get the 90% confidence interval for σ_1/σ_2 as (0.87, 2.61).

Testing hypothesis about the equality of two Variances

Step-1: Formulate the null and alternative hypothesis

$H_0 : \sigma_1^2/\sigma_2^2 = 1$ and $H_1 : \sigma_1^2/\sigma_2^2 \neq 1$

$H_0 : \sigma_1^2/\sigma_2^2 \leq 1$ and $H_1 : \sigma_1^2/\sigma_2^2 > 1$

$H_0 : \sigma_1^2/\sigma_2^2 \geq 1$ and $H_1 : \sigma_1^2/\sigma_2^2 < 1$

Step-2: Decide on the significance level α.

Step-3: The test-statistic to be used:

$F = \frac{s_1^2}{s_2^2}$, where s_1^2 is larger than s_2^2.

Step-4: Calculate the value of F from the sample data.

Step-5: Determine the critical region which depends on alternative hypothesis :

(a) When H_1 is $\sigma_1^2 / \sigma_2^2 > 1$ (i.e. $H_1 : \sigma_1^2 > \sigma_2^2$), the

critical region will be $F \geq F_{\alpha(\nu_1,\nu_2)}$.

(b) When H_1 is $\sigma_1^2/\sigma_2^2 < 1$ (i.e. $H_1 : \sigma_1^2 < \sigma_2^2$), we interchange the role of two samples and use

$F = \frac{s_1^2}{s_2^2}$, then the critical region will be $F < 1/F_{\alpha(\nu_2,\nu_1)}$.

(c) When H_1 is $\sigma_1^2/\sigma_2^2 \neq 1$ (i.e. $H_1 : \sigma_1^2 \neq \sigma_2^2$), the critical region will be

$F \geq F_{\alpha/2(\nu_1,\nu_2)}$, when $s_1^2 > s_2^2$. or

$F < 1/F_{\alpha/2(\nu_2,\nu_1)}$, when $s_2^2 > s_1^2$.

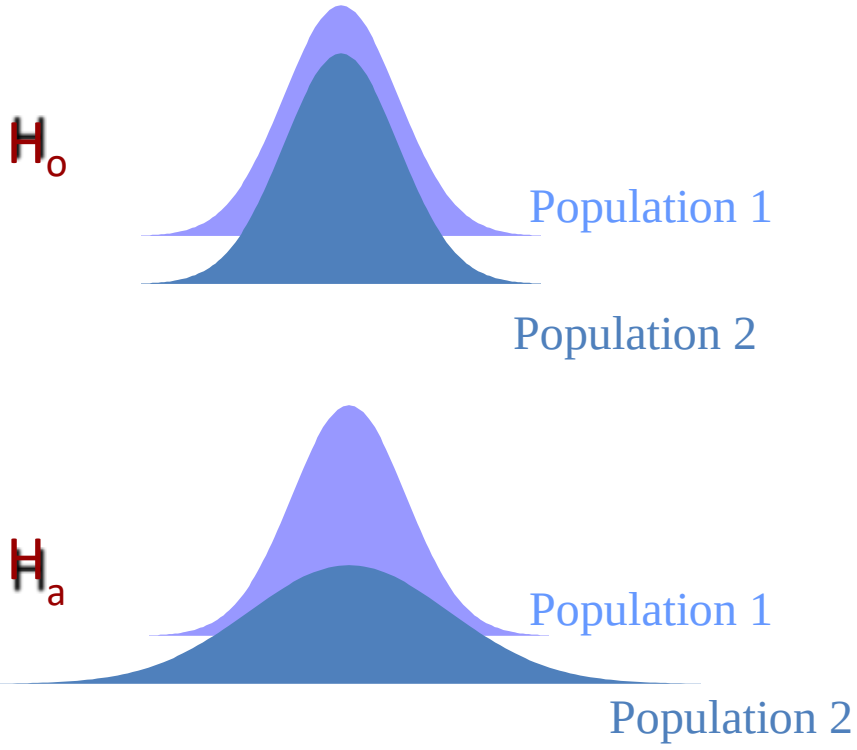
This procedure avoids the use of the left-hand tail. However, if one wishes to use the left hand tail (two-sided test) also, then the critical region will be F ≥ $F_{\alpha/2(\nu_1,\nu_2)}$, and F ≤ 1/ $F_{\alpha/2(\nu_2,\nu_1)}$.

Step-6: Conclusion:

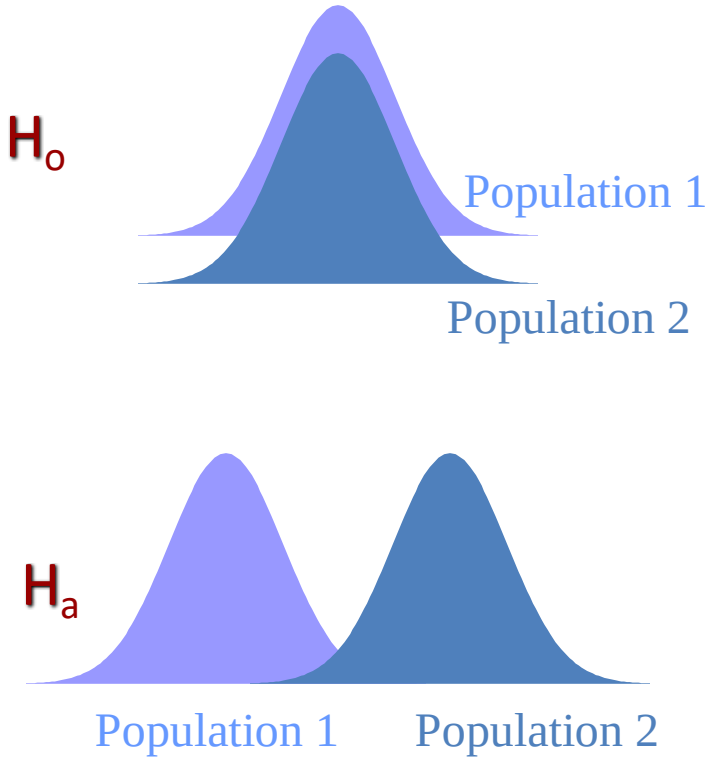
Reject H_0 if the computed value of F fall in the critical region otherwise accept H_0.

Two Sample Tests

TEST FOR EQUAL VARIANCES



TEST FOR EQUAL MEANS



Hypothesis Tests for Two Population Variances

Two-Tailed Test

$$H_0 : \sigma_1^2 = \sigma_2^2$$
$$H_A : \sigma_1^2 \neq \sigma_2^2$$

Upper One-Tailed Test

$$H_0 : \sigma_1^2 \leq \sigma_2^2$$
$$H_A : \sigma_1^2 > \sigma_2^2$$

Lower One-Tailed Test

$$H_0 : \sigma_1^2 \geq \sigma_2^2$$
$$H_A : \sigma_1^2 < \sigma_2^2$$

F-TEST STATISTIC FOR TESTING WHETHER TWO POPULATIONS HAVE EQUAL VARIANCES

$$F = \frac{s_i^2}{s_j^2} \; (df = D_1 = n_{i-1} \; and \; D_2 = n_{j-1})$$

where:

n_i = Sample size from ith population

n_j = Sample size from jth population

s_i^2 = Sample variance from ith population

s_j^2 = Sample variance from jth population

Hypothesis Tests for Two Population Variances

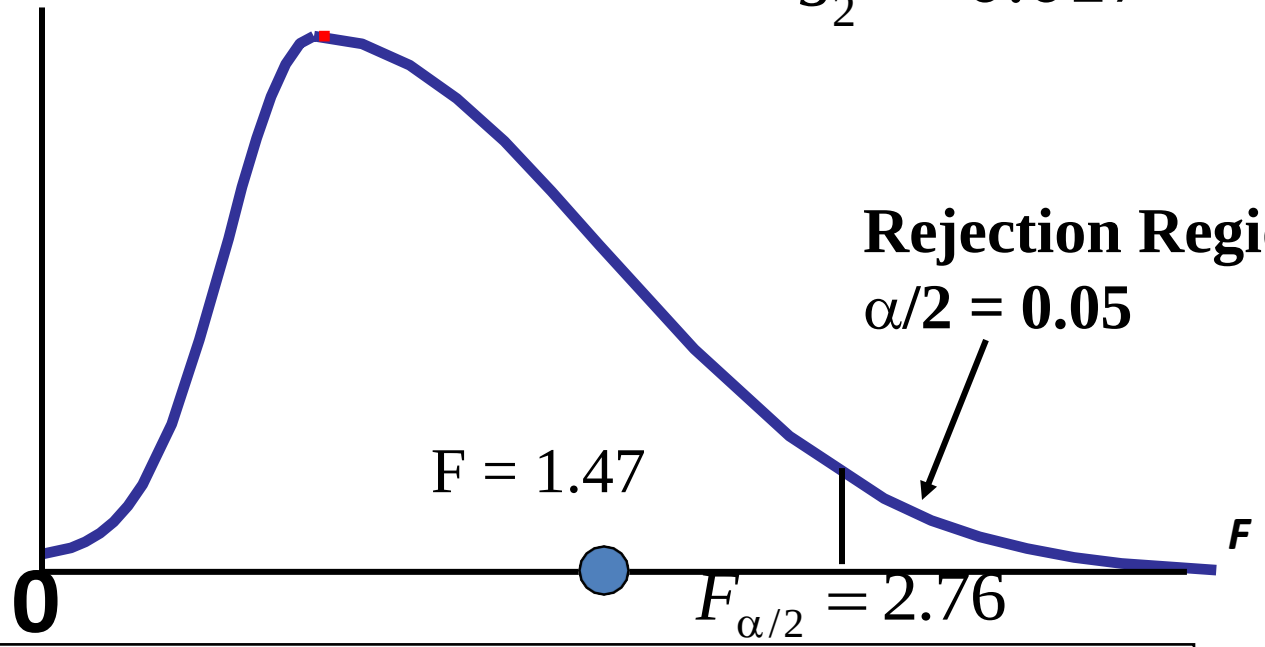
$H_0 : \sigma_1^2 \leq \sigma_2^2$

$H_A : \sigma_1^2 > \sigma_2^2$

$\alpha = .10$

$$F = \frac{s_1^2}{s_2^2} = \frac{0.025}{0.017} = 1.47$$

***df*: $D_i = 10$, $D_j = 12$**



Since $F = 1.47 \leq F_{\alpha/2} = 2.76$, do not reject H_0

Example

The variability in the amount of impurities present in a batch of chemicals used for a particular process depends on the length of time that the process is in operation. Suppose a sample of size 25 is drawn from the normal process which is to be compared to a sample of a new process that has been developed to reduce the variability of impurities.

Example

	Sample 1	Sample 2
n	25	25
s^2	1.04	0.51

	Sample 1	Sample 2
n	25	25
s^2	1.04	0.51

State Null and alternative hypothesis:

$$H_0 : \sigma_1^2 = \sigma_2^2$$

$$H_1 : \sigma_1^2 > \sigma_2^2$$

Level of significance: 0.05

The test-statistic to be used:

$$F = \frac{S_1^2}{S_2^2} = \frac{1.04}{0.51} = 2.04$$

Critical region: $F > F_{\alpha(\nu_1,\nu_2)}$

$$: F > F_{0.05(24,24)} = 1.98$$

Conclusion: As Calculated valued F=2.04 fall in the critical region so we will reject H_0 and conclude that the variability in the new process (Sample 2) is less than the variability in the original process (Sample 1).

Example

In an experiment on reaction times in seconds of two individuals A and B, measured under identical conditions, the following results were obtained:

A	0.41, 0.38, 0.37, 0.42, 0.35, 0.38
B	0.32, 0.36, 0.38, 0.33, 0.38

Example

Test the hypothesis at 0.05 level of significance that

$H_0 : \sigma_1^2 = \sigma_2^2$ against

$H_1 : \sigma_1^2 \neq \sigma_2^2$.

(i) We state our null and alternative hypothesis as

$$H_0 : \sigma_A^2 = \sigma_B^2 \text{ against } H_1 : \sigma_A^2 \neq \sigma_B^2 .$$

(ii) The level of significance is set at $\alpha = 0.05$

(iii) The test statistic to use is

$$F = S_A^2 / S_b^2$$

$$if\ S_A^2 > S_B^2$$

$$Or$$

$$F = S_B^2 / S_A^2$$

$$if\ S_B^2 > S_A^2$$

(iv) Computations:

A	**A^2**	**B**	**B^2**
0.41	0.1681	0.32	0.1024
0.38	0.1444	0.36	0.1296
0.37	0.1369	0.38	0.1444
0.42	0.1764	0.33	0.1089
0.35	0.1225	0.38	0.1444
0.38	0.1444		
2.31	**0.8927**	**1.77**	**0.6297**

$$\bar{A} = \frac{\sum A}{n_1} = \frac{2.31}{6} = 0.385, \quad \bar{B} = \frac{\sum B}{n_2} = \frac{1.77}{5} = 0.354,$$

$$s_A^2 = \frac{1}{n_1 - 1}\left[\sum A^2 - \frac{\left(\sum A\right)^2}{n_1}\right] = \frac{1}{5}\left[0.8927 - \frac{(2.31)^2}{6}\right]$$

$$= 0.00067$$

$$s_B^2 = \frac{1}{n_2 - 1}\left[\sum B^2 - \frac{\left(\sum B\right)^2}{n_2}\right] = \frac{1}{4}\left[0.6297 - \frac{(1.77)^2}{5}\right]$$

$$= 0.00078$$

$$\therefore \quad F = \frac{s_A^2}{s_B^2} = \frac{0.00067}{0.00078} = 0.86$$

(v) The critical region is F ≥ $F_{0.025(5,4)} = 9.36,$, and

$$F \leq \frac{1}{F_{0.025}(4,5)} = \frac{1}{7.39} = 0.14$$

(vi) **Conclusion:** Since the calculated value

F = 0.86 does not fall in the critical region, so we accept H_0, and conclude that the two population variances are equal.

Critical Values of *F*-Distributions

- Notation: $F_{a,\ df1,\ df2}$ is the value with upper tail area of a above it for the *F*-distribution with degrees' of freedom df_1 and df_2, respectively

- $F_{1-a,\ df1,\ df2} = 1/\ F_{a,\ df2,\ df1}$ (Lower tail critical values can be obtained from upper tail critical values with "reversed" degrees of freedom)