



ACADGILD

Mastering Data
Science



Statistics



Session 6 - Introduction to Statistics



Agenda

1 Statistics

2 Introduction to Basic Terms

3 Variables

4 Mean, Median and Mode

5 Measure of Dispersion

6 Range

7 Sample Variance

8 Standard Deviation

9 Population Vs. Sample

10 Chebysheff's Theorem

11 Law of Expected Values and Variance

12 Probability Density Function

- Statistics is the science of collecting, organizing, presenting, analyzing, and interpreting data to help in making more effective decisions.
- Statistical Analysis is implemented to manipulate, summarize and investigate data, so that useful decision-making information results are obtained.

- Descriptive Statistics is a method of organizing, summarizing, and presenting data in an informative way.
- Inferential Statistics is a method which is used in determining something about a population on the basis of a sample.
 - Population - The entire set of individuals or objects of interest or the measurements obtained from all individuals or objects of interest.
 - Sample - A portion, or part, of the population of interest.

- Population - A collection/set of individuals/objects/events whose properties are to be analyzed. There are two kinds:
 - Finite
 - Infinite
- Sample - A population subset.

- Variable - A **characteristic** about each individual element of a population/sample.
- Data (singular) - A **value** of the associated variable with one element of a population/sample. This value may be a number, a word, or a symbol.
- Data (plural) - A **set of values** collected for the variable from each of the elements belonging to the sample.
- Experiment - A **planned activity** whose results yield a set of data.
- Parameter - A **numerical value** which summarizes the entire population data.
- Statistics - A **numerical value** which summarizes the sample data.

Qualitative, or Attribute, or Categorical, Variable

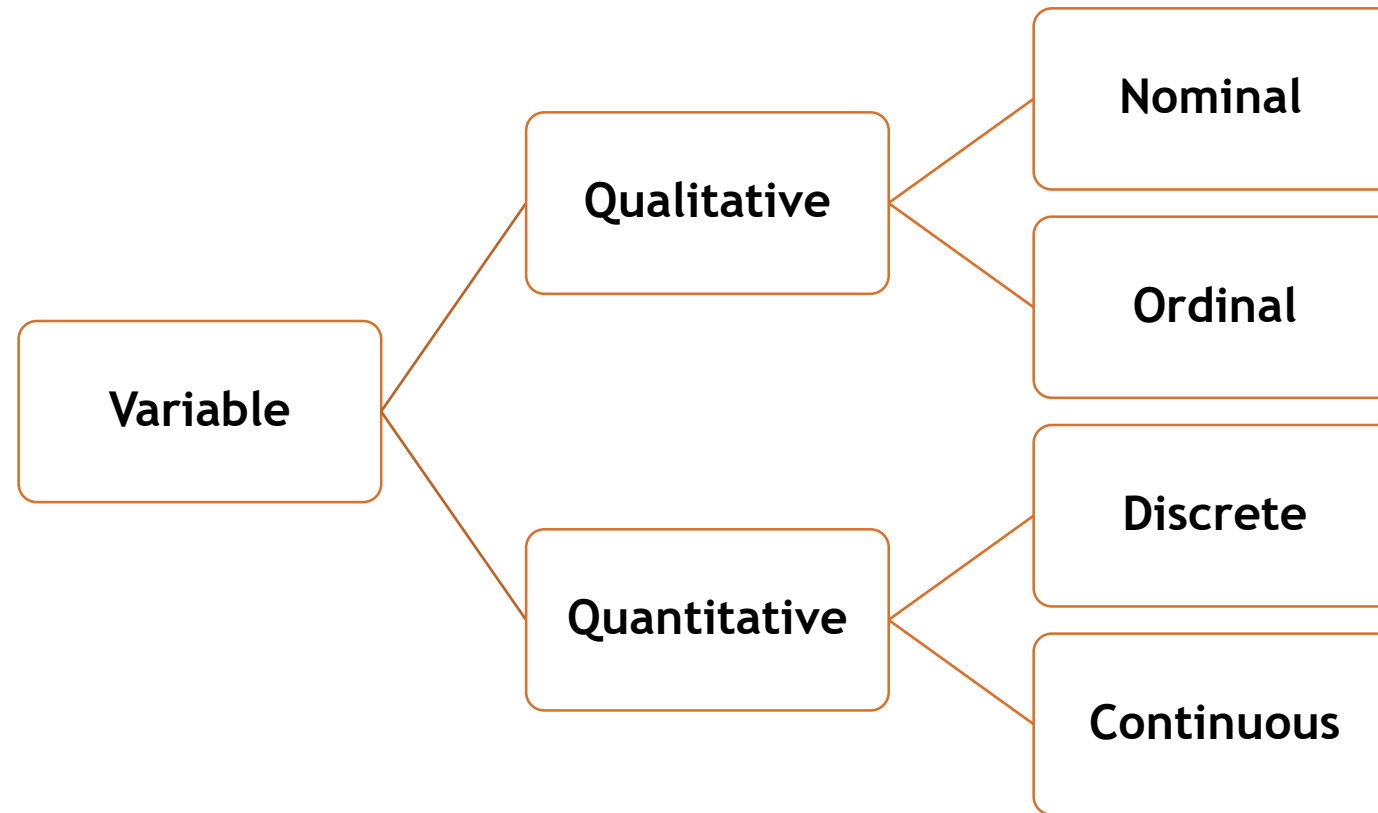
- A variable that categorizes or describes a population element.

Note: Arithmetic operations such as addition and averaging, are not meaningful for data resulting from a qualitative variable.

Quantitative, or Numerical, Variable

- A variable that quantifies a population element.

Note: Arithmetic operations such as addition and averaging, are meaningful for data resulting from a quantitative variable.



Two Kinds of Variables



- Nominal Variable - A **qualitative variable** that categorizes (or describes, or names) a population element.
- Ordinal Variable - A **qualitative variable** that incorporates an ordered position or ranking.
- Discrete Variable - A **quantitative variable** that can assume a countable number of values.
 - This can assume values corresponding to the isolated points along a line interval.
 - There is a gap between any two values
- Continuous Variable - A **quantitative variable** that can assume an uncountable number of values.
 - This can assume any value along a line interval
 - Including every possible value between any two values

- Let $x_1, x_2, x_3, \dots, x_n$ be the realized values of a random variable 'X', from a sample of size 'n'.

The **sample arithmetic mean** is defined as:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

Example

➤ The systolic blood pressure of seven middle aged men were as follows:

151, 124, 132, 170, 146, 124 and 113.

$$\begin{aligned}\text{The Mean is } \bar{x} &= \frac{(151 + 124 + 132 + 170 + 146 + 124 + 113)}{7} \\ &= 137.14\end{aligned}$$

- The median for the sample data arranged in an increasing order is defined as:
 - i. If “n” is an odd number - Middle value
 - ii. If “n” is an even number - Midway between the two middle values
- The mode is the most commonly occurring value.

Example - n is odd

The re-ordered systolic blood pressure data seen earlier are:

113, 124, 124, 132, 146, 151, and 170.

- The **Median** is the middle value of the ordered data, i.e. 132.
- Two individuals have systolic blood pressure = 124 mm Hg, so the **Mode** is 124.

Example - n is even

Six men with high cholesterol participated in a study to investigate the effects of diet on cholesterol level. At the beginning of the study, their cholesterol levels (mg/dL) were as follows:

366, 327, 274, 292, 274 and 230

Rearrange the data in numerical order as follows:

230, 274, 274, 292, 327 and 366.

- The **Median** is **half way between the middle two readings**, i.e. $(274+292) / 2 = 283$.
- The **mode** between the **two men having the same cholesterol level** = 274.

- If the histogram of the data is **right-skewed** then large sample values tend to inflate the mean.
- If the distribution is **skewed** then the median is not influenced by large sample values and is a better measure of centrality.

Note - If **mean = median = mode** then the data are said to be symmetrical.

For example,

- In the CK measurement study, the sample mean = 98.28.
- The median = 94.5, i.e. mean is larger than median indicating that mean is inflated by two large data values 201 and 203.

- The concept **Measures of Dispersion** characterize how to spread out the distribution, i.e., how variable the data are.
- The commonly used dispersion measures include:
 - Range
 - Variance and Standard Deviation

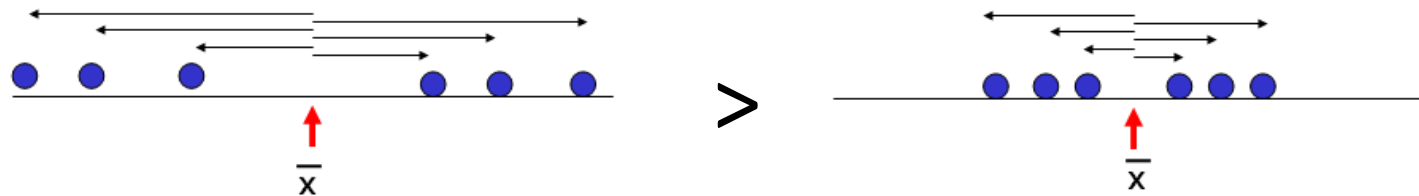
- The **Range** is the difference between the largest and the smallest observations in the sample.
- For example, the minimum and maximum blood pressure is 113 and 170 respectively. Hence the **range** is 57 mmHg
 - Easy to calculate;
 - Implemented for both “best” or “worst” case scenarios
 - Too sensitive for extreme values

Sample Variance



- The sample variance, s^2 , is the **arithmetic mean** of the squared deviations from the sample mean:

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



- The sample standard deviation (s) is the **square-root of the variance**.

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

- The sample standard deviation has an advantage of being in the same units as the **original variable (x)**.

$$\mu = \frac{\sum_{i=1}^N x_i}{N}$$

Population Mean

Vs.

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

Sample Mean



	Population	Sample
Size	N	n
Mean		

Population Vs. Sample



	Population	Sample
Size	N	n
Mean		
Variance		

Population Vs. Sample



- The variance of a population is:

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N}$$

Population Mean

Population Size

The diagram shows the formula for population variance. A blue arrow points from the text 'Population Mean' to the Greek letter mu in the numerator. Another blue arrow points from the text 'Population Size' to the letter N in the denominator.

- The variance of a sample is:

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

Sample Mean

The diagram shows the formula for sample variance. A blue arrow points from the text 'Sample Mean' to the x-bar symbol in the numerator. Another blue arrow points from the text 'Note! the denominator is sample size (n) minus one !' to the 'n-1' in the denominator.

Note! the denominator is sample size (n) minus one !

➤ The **square root of the variance** is termed as the Standard Deviation, thus:

- The population Standard Deviation = $\sigma = \sqrt{\sigma^2}$

- The Sample Standard deviation = $s = \sqrt{s^2}$

- A more general interpretation of the standard deviation is derived from **Chebysheff's Theorem**, which applies to all shapes of histograms (except bell shaped).
- The proportion of observations in any sample that lie within k standard deviations of the mean is at least:

$$1 - \frac{1}{k^2} \text{ for } k > 1$$

For k=2 (say), the theorem states that at least 3/4 of all observations lie within 2 standard deviations of the mean. This is a “lower bound” compared to Empirical Rule's approximation (95%).

Two Types of Random Variables



Discrete Random Variable

- Takes on a **countable number** of values
- For example, values on the roll of dice: 2, 3, 4, ..., 12

Continuous Random Variable

- Values are not **discrete**, not **countable**
- For example, time (30.1 minutes? 30.10000001 minutes?)

Analogy

- Integers are **discrete**, while Real Numbers are **Continuous**

➤ $E(C) = C$

- The expected Value of a Constant is just the value of the constant.

➤ $E(X + C) = E(X) + C$

➤ $E(CX) = cE(X)$

- We can “pull” a constant out of the expected value expression (either as part of a sum with a random variable X or as a coefficient of random variable X).

➤ $V(c) = 0$

- The Variance of constant (c) is zero.

➤ $V(X + c) = V(X)$

- The Variance of random variable and a constant is just the variance of the random variable (per 1 above).

➤ $V(cX) = c^2 V(X)$

- The Variance of a random variable and a constant co-efficient is the co-efficient squared times in the variance of the random variable.

Probability Density Functions



Unlike a discrete random variable, a continuous random variable is one that can assume an uncountable number of values.

- We cannot list the possible values because there is an infinite number of them.
- The probability of each individual value is virtually 0 as there is an infinite number of values

Point Probabilities are Zero



If the probability of each individual value is virtually 0 then there is an infinite number of values.

Thus, we can determine the probability of a **range of values** only.

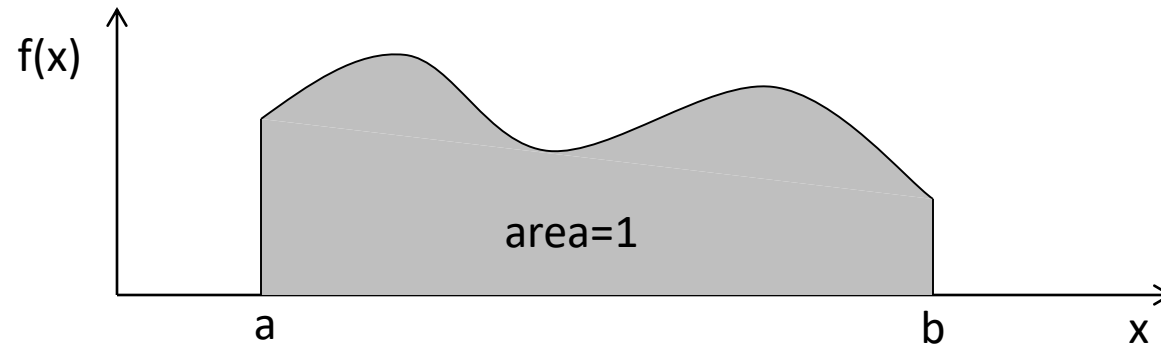
- For example, with a **discrete** random variable like tossing a die, it is meaningful to talk about $P(X=5)$
- In a **continuous** setting (e.g. with time as a random variable), the probability the random variable of interest say task length, takes exactly 5 minutes is infinitely small, hence $P(X=5) = 0$.

Probability Density Function



A function $f(x)$ is called a Probability Density Function over the range $a \leq x \leq b$ if it meets the following requirements:

1. $f(x) \geq 0$ for all x between a and b , and



2. The total area under the curve between a and b is 1.0



Email us - support@acadgild.com



ACADGILD

Mastering Data
Science



Statistics



Session 7 - Distributions and CLT



Agenda

- 1 Probability Function
- 2 Binomial Distribution
- 3 Binomial Random Variable
- 4 Poisson Distribution
- 5 Poisson Probability Distribution
- 6 The Normal Distribution
- 7 Standard Normal Distribution
- 8 Calculating Normal Probabilities
- 9 Using the Normal Table
- 10 Finding and Using the Values of Z
- 11 Central Limit Theorem
- 12 Sampling Distribution of the Sample

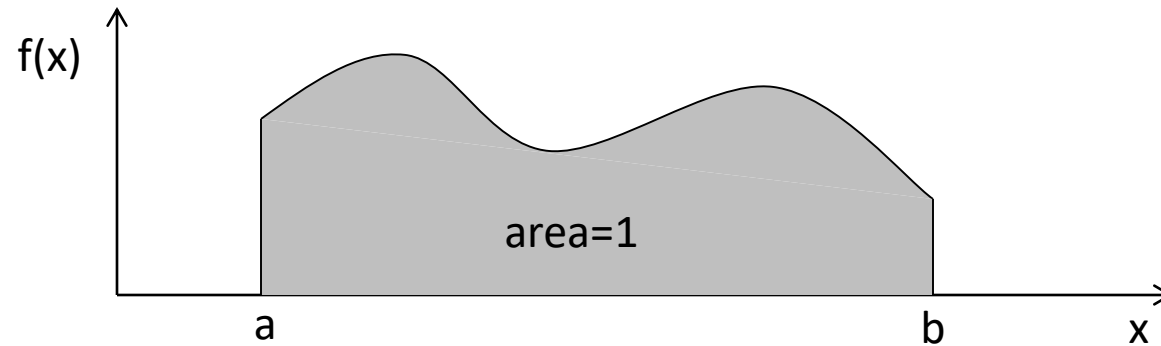
Mean

Probability Density Function



A function $f(x)$ is called a Probability Density Function over the range $a \leq x \leq b$ if it meets the following requirements:

1. $f(x) \geq 0$ for all x between a and b , and



2. The total area under the curve between a and b is 1.0

- The binomial distribution is the probability distribution that results from doing a “binomial experiment”. Binomial experiments have the following properties:
 1. Fixed number of trials, represented as n .
 2. Each trial has two possible outcomes, a “success” and a “failure”.
 3. $P(\text{success})=p$ (and thus: $P(\text{failure})=1-p$), for all trials.
 4. The trials are independent, which means that the outcome of one trial does not affect the outcomes of any other trials.

- The binomial random variable counts the number of successes in n trials of the binomial experiment. It can take on values from 0, 1, 2, ..., n . Thus, it's a discrete random variable.
- To calculate the probability associated with each value we use combinatorics:

$$P(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

for $x=0, 1, 2, \dots, n$

- As you expect, Statisticians have developed general formulas for the mean, variance, and standard deviation of a binomial random variable. They are:

$$\mu = np$$

$$\sigma^2 = np(1 - p)$$

$$\sigma = \sqrt{np(1 - p)}$$

- Named after **Simeon Poisson**, the Poisson distribution is a **discrete probability distribution** and refers to the number of events (a.k.a. successes) within a specific time period or region of space.
- For example,
 - The number of cars arriving at a service station in 1 hour. (The interval of time is 1 hour)
 - The number of flaws in a bolt of cloth. (The specific region is a bolt of cloth)
 - The number of accidents in 1 day on a particular stretch of highway. (The interval is defined by both time, 1 day, and space and the particular stretch of highway.)

The Poisson Experiment



Similar to binomial experiment, a Poisson experiment has four defining characteristic properties:

1. The number of successes that occur in any interval is independent of the number of successes that occur in any other interval.
2. The probability of a success in an interval is the same for all equal-size intervals
3. The probability of a success is proportional to the size of the interval.
4. The probability of more than one success in an interval approaches 0 as the interval becomes smaller.

- The Poisson random variable is the number of successes that occur in a period of time or an interval of space in a Poisson experiment.
- For example, on average, 96 trucks arrive at a border crossing every hour.
- For example, the number of typographic errors in a new textbook edition averages 1.5 per 100 pages.

 Success
 Time Period
 Interval

Poisson Probability Distribution



- The probability that a Poisson random variable assumes a value of x is given by:

$$P(x) = \frac{e^{-\mu} \mu^x}{x!} \quad \text{for } x = 0, 1, 2, \dots$$

where μ is the mean number of successes in the interval

e – natural logarithm base

$$E(X) = V(X) = \mu$$

Example

- The number of typographical errors in new editions of textbooks varies considerably from book to book. After some analysis it concludes that the number of errors is Poisson distributed with a mean of 1.5 per 100 pages. The instructor randomly selects 100 pages of a new book. What is the probability that there are no typos?

That is, what is $P(X=0)$ given that $\mu = 1.5$?

$$P(0) = \frac{e^{-\mu} \mu^x}{x!} = \frac{e^{-1.5} 1.5^0}{0!} = .2231$$

“There is about a 22% chance of finding zero errors”

- The probability of success is **proportional** to the size of the interval.
- Thus knowing an error rate of 1.5 typos per 100 pages, we can determine a mean value for a 400 page book as:

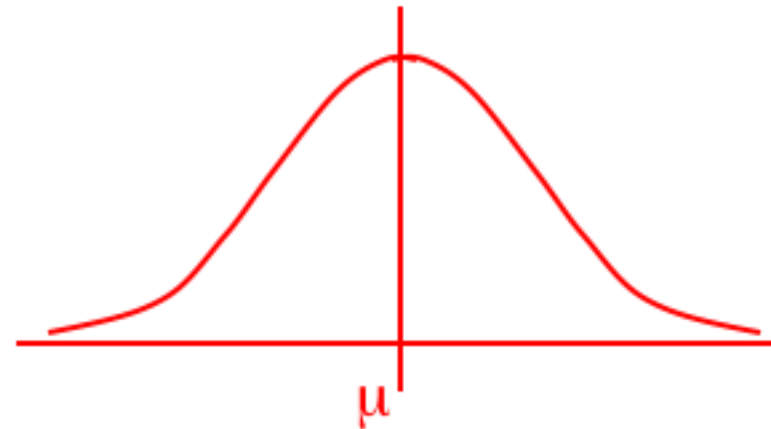
$$\mu = 1.5 (4) = 6 \text{ typos/400 pages}$$

$$\mu = 1.5 (4) = 6 \text{ typos / 400 pages}$$

- The normal distribution is the most important of all probability distributions. The probability density function of a **normal random variable** is given by:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad -\infty < x < \infty$$

- It looks like **bell shaped**, **symmetrical** around the **mean, μ**



- The normal distribution is completely defined by two parameters **Standard Deviation** and **Mean**.

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad -\infty < x < \infty$$

- The normal distribution is bell shaped and symmetrical about the **mean**.

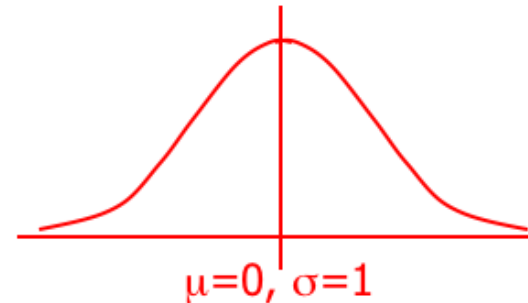
Standard Normal Distribution



- A normal distribution whose mean is zero and standard deviation is 1 is called the standard normal distribution.

$$f(x) = \frac{1}{\textcolor{blue}{1}\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\textcolor{red}{0}}{\textcolor{blue}{1}}\right)^2} \quad -\infty < x < \infty$$

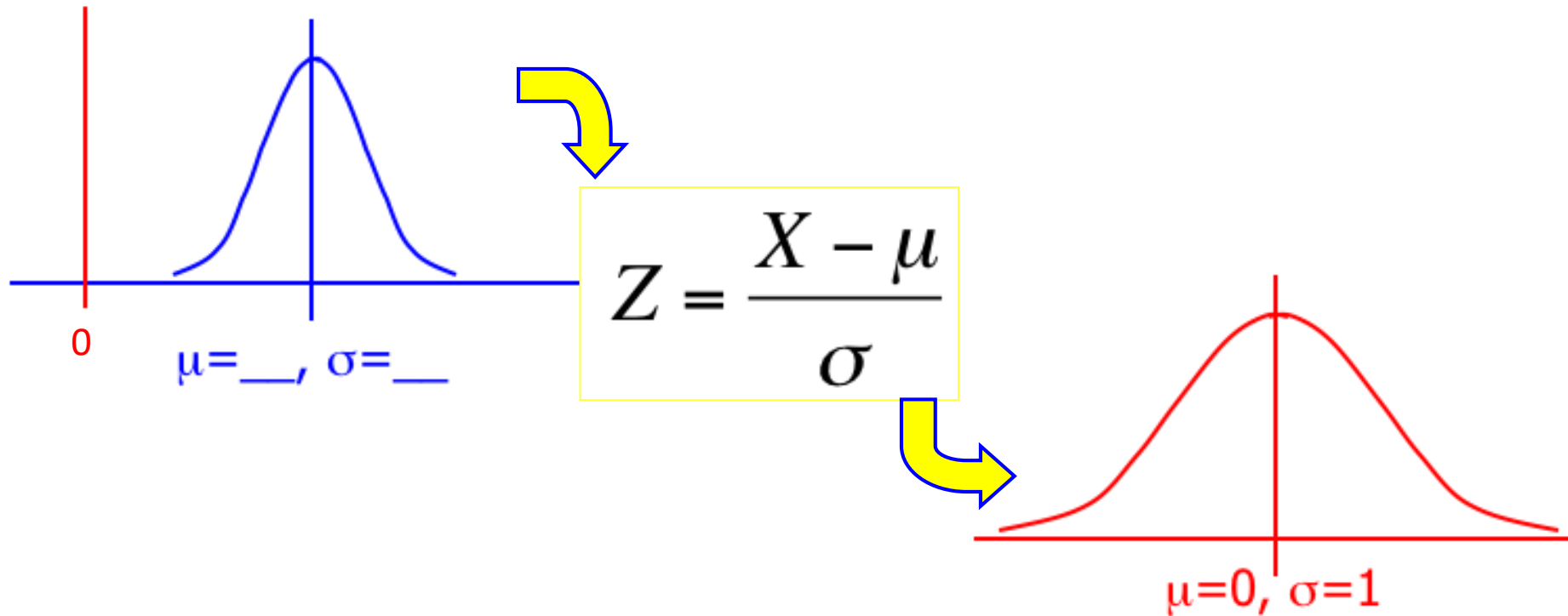
- As we shall see shortly, any normal distribution can be *converted* to a standard normal distribution with simple algebra.
This makes calculations much easier.



Calculating Normal Probabilities



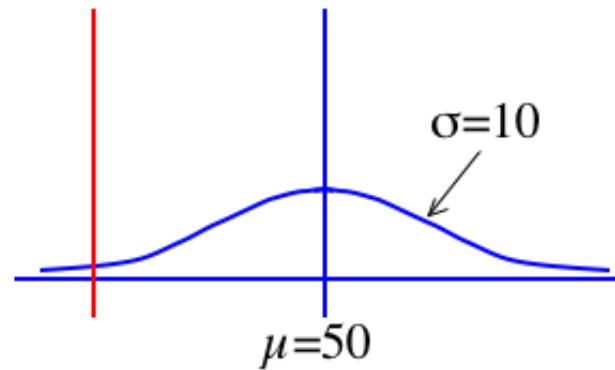
- We can use the following function to convert any normal random variable to a standard normal random variable



Calculating Normal Probabilities



Example: The time required to build a computer is normally distributed with a mean of 50 minutes and a standard deviation of 10 minutes.

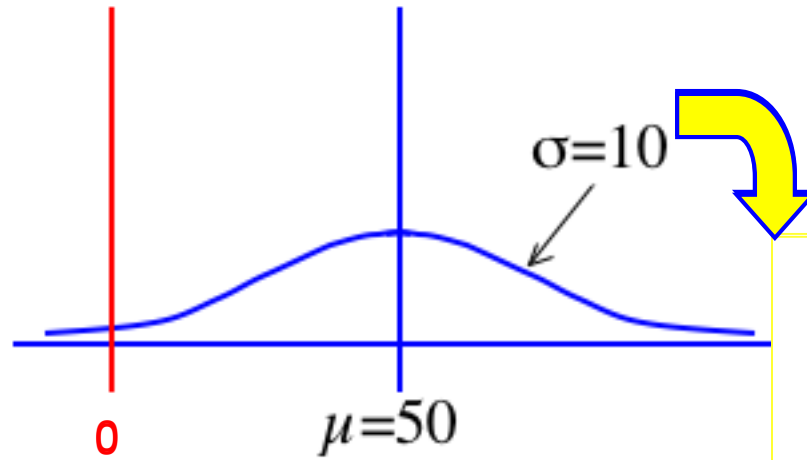


- What is the probability that a computer is assembled in a time between 45 and 60 minutes?
- Algebraically speaking, what is $P(45 < X < 60)$?

Calculating Normal Probabilities



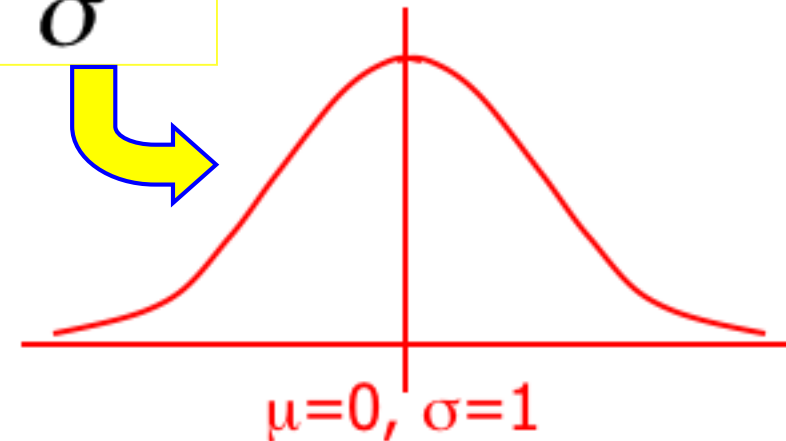
$P(45 < X < 60)$?



Mean of 50 minutes and a standard deviation of 10 minutes...

$$Z = \frac{X - \mu}{\sigma}$$

$$\begin{aligned} P(45 < X < 60) &= \\ P\left(\frac{45 - 50}{10} < \frac{X - \mu}{\sigma} < \frac{60 - 50}{10}\right) &= \\ P(-.5 < Z < 1) \end{aligned}$$



Calculating Normal Probabilities



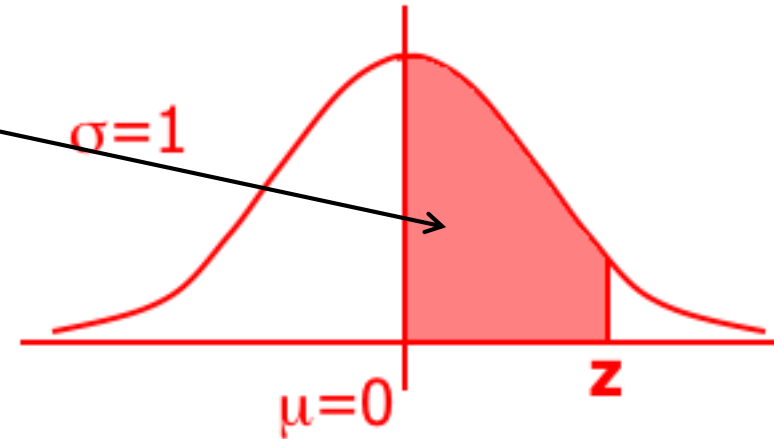
We can use z-table probabilities $P(0 < Z < z)$

We can break up $P(-.5 < Z < 1)$ into:
 $P(-.5 < Z < 0) + P(0 < Z < 1)$

The distribution is *symmetric* around zero, so we have:

$$P(-.5 < Z < 0) = P(0 < Z < .5)$$

$$\text{Hence: } P(-.5 < Z < 1) = P(0 < Z < .5) + P(0 < Z < 1)$$



Calculating Normal Probabilities



How to use **z-table**?

This table gives probabilities $P(0 < Z < z)$

First column = integer + first decimal

Top row = second decimal place

$P(0 < Z < 0.5)$

$P(0 < Z < 1)$

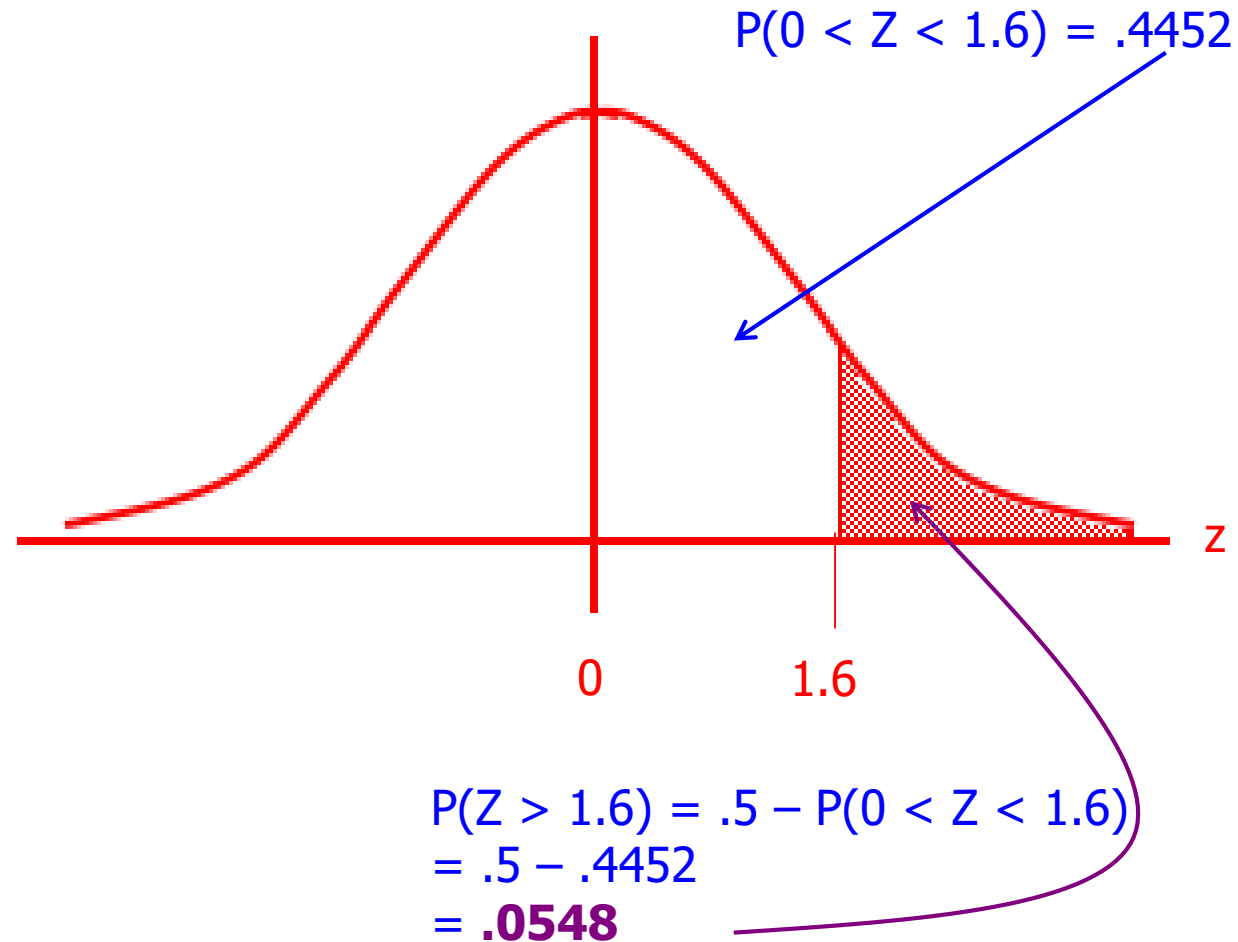
$P(-.5 < Z < 1) = .1915 + .3414 = .5328$

z	.00	.01	.02	.03
0.0	.0000	.0040	.0080	.0120
0.1	.0398	.0438	.0478	.0517
0.2	.0793	.0832	.0871	.0910
0.3	.1179	.1217	.1255	.1293
0.4	.1554	.1591	.1628	.1664
0.5	.1915	.1950	.1985	.2019
0.6	.2257	.2291	.2324	.2357
0.7	.2580	.2611	.2642	.2673
0.8	.2881	.2910	.2939	.2967
0.9	.3159	.3186	.3212	.3238
1.0	.3413	.3438	.3461	.3485
1.1	.3643	.3665	.3686	.3708
1.2	.3849	.3869	.3888	.3906

Using the Normal Table



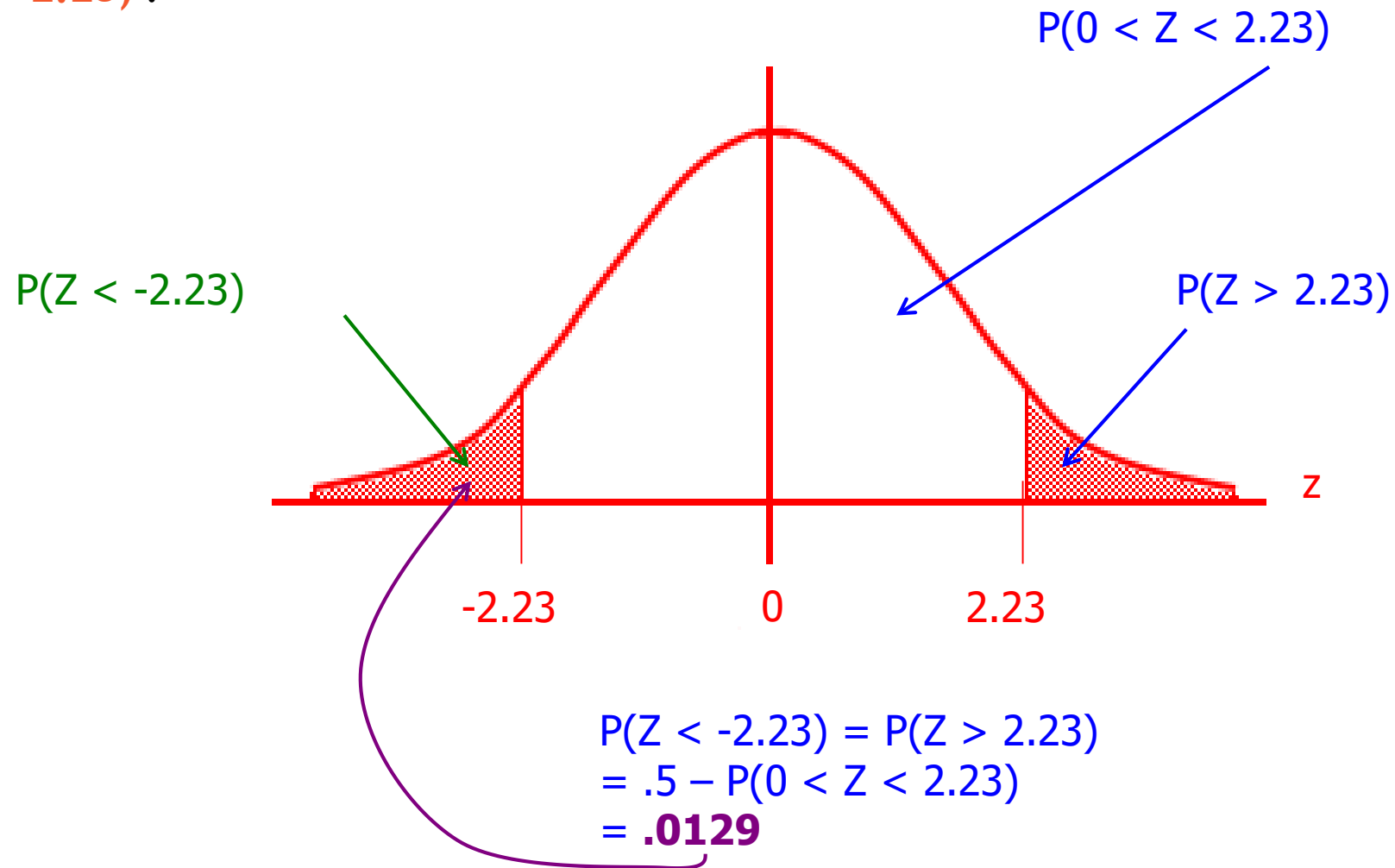
What is $P(Z > 1.6)$?



Using the Normal Table



What is $P(Z < -2.23)$?

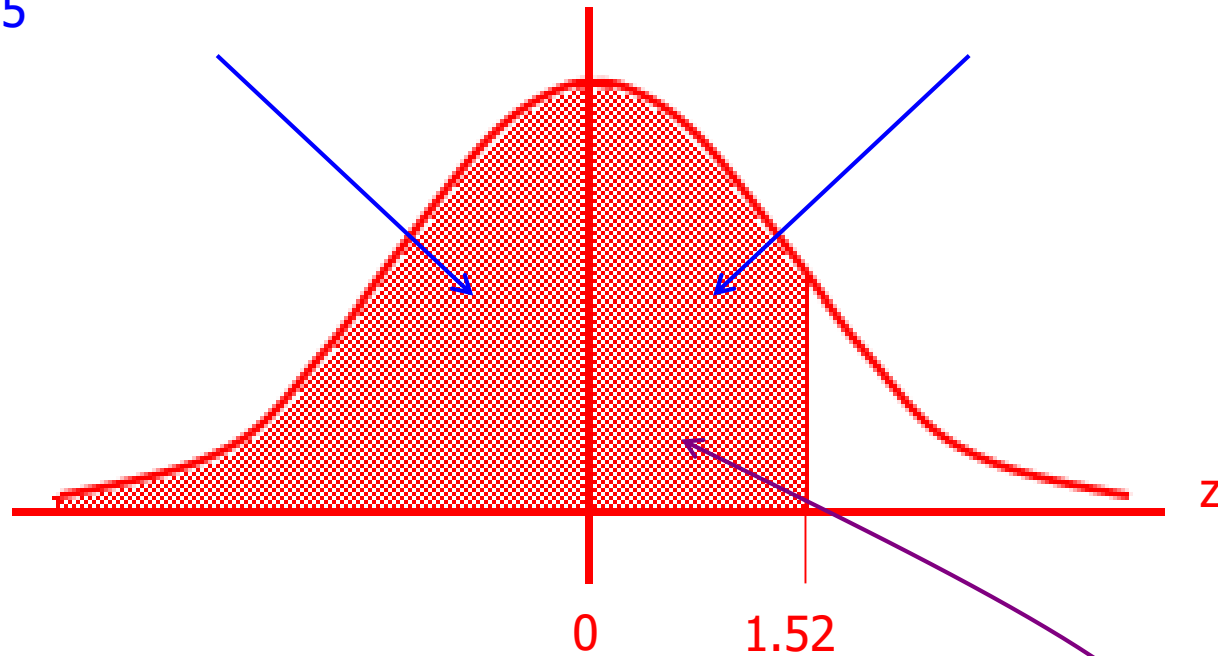


Using the Normal Table



What is $P(Z < 1.52)$?

$$P(Z < 0) = .5$$

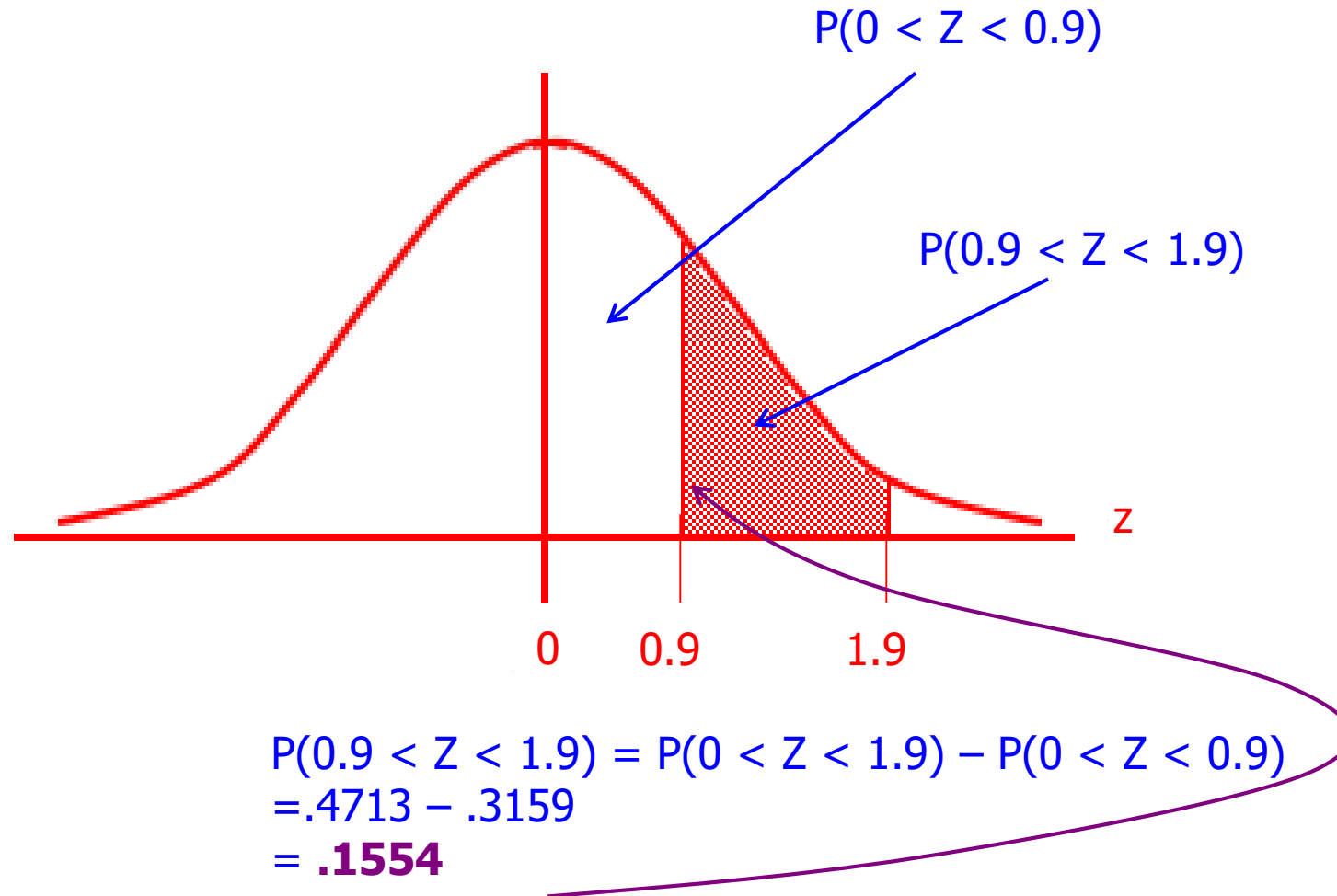


$$\begin{aligned} P(Z < 1.52) &= .5 + P(0 < Z < 1.52) \\ &= .5 + .4357 \\ &= \mathbf{.9357} \end{aligned}$$

Using the Normal Table



What is $P(0.9 < Z < 1.9)$?



➤ The other values of Z are:

- $Z.05 = 1.645$
- $Z.01 = 2.33$

Using the Values of Z



Because $z_{.025} = 1.96$ and $-z_{.025} = -1.96$, it follows that we can state:

$$P(-1.96 < Z < 1.96) = .95$$

Similarly

$$P(-1.645 < Z < 1.645) = .90$$

Central Limit Theorem



- The sampling distribution of the mean of a random sample drawn from any population is approximately normal for a sufficiently large sample size.
- The larger the sample size, the more closely the sampling distribution of 'X' will resemble a normal distribution.

- If the population is **normal**, then 'X' is **normally distributed** for all values of n .
- If the population is **not-normal**, then 'X' is **approximately normal** only for larger values of n .
- In many practical situations, a sample size of 30 may be sufficiently large to allow us to use the normal distribution as an approximation for the sampling distribution of X .

1. $\mu_{\bar{x}} = \mu$

2. $\sigma_{\bar{x}}^2 = \sigma^2 / n$ and $\sigma_{\bar{x}} = \sigma / \sqrt{n}$

3. If X is normal, \bar{X} is normal. If X is not normal then \bar{X} is approximately normal for sufficiently large sample sizes.

Note:

- The definition of “sufficiently large” depends on the extent of non-normality of X .
- For example, heavily skewed; multimodal

Sampling Distributions of the Sample Mean



For Example,

The foreman of a bottling plant has observed that the amount of soda in each “32-ounce” bottle is actually a normally distributed random variable, with a mean of 32.2 ounces and a standard deviation of 0.3 ounce.

If a customer buys one bottle, what is the probability that the bottle will contain more than 32 ounces?



Email us - support@acadgild.com

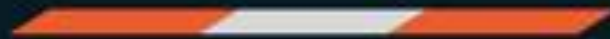


ACADGILD

Mastering Data
Science



Statistics



Session 8 - Sampling Distribution and Hypothesis Testing



Agenda

- 1 Sampling Distribution of the Mean
- 2 Sampling Distribution of Two Dice
- 3 Difference between 2 Mean
- 4 Estimation
- 5 Point and Interval Estimation
- 6 Confidence Levels
- 7 Interval Width
- 8 Selecting Sample Size
- 9 Sample Size to Estimate Mean Size
- 10 Hypothesis Testing
- 11 Concepts of Hypothesis Testing
- 12 Types of Errors

Sampling Distribution of the Mean



A fair die is thrown **infinitely** many times, with the random variable $X = \#$ of spots on any throw.

The probability distribution of X is:

x	1	2	3	4	5	6
$P(x)$	1/6	1/6	1/6	1/6	1/6	1/6

and the mean and variance are calculated as well:

$$\mu = \sum xP(x) = 1(\frac{1}{6}) + 2(\frac{1}{6}) + \dots + 6(\frac{1}{6}) = 3.5$$

$$\sigma^2 = \sum (x - \mu)^2 P(x) = (1 - 3.5)^2 (\frac{1}{6}) + \dots + (6 - 3.5)^2 (\frac{1}{6}) = 2.92$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{2.92} = 1.71$$

Sampling Distribution of Two Dice



A sampling distribution is created by looking at all samples of size $n=2$ (i.e. two dice) and their means

Sample		Sample		Sample	
1, 1	1.0	3, 1	2.0	5, 1	3.0
1, 2	1.5	3, 2	2.5	5, 2	3.5
1, 3	2.0	3, 3	3.0	5, 3	4.0
1, 4	2.5	3, 4	3.5	5, 4	4.5
1, 5	3.0	3, 5	4.0	5, 5	5.0
1, 6	3.5	3, 6	4.5	5, 6	5.5
2, 1	1.5	4, 1	2.5	6, 1	3.5
2, 2	2.0	4, 2	3.0	6, 2	4.0
2, 3	2.5	4, 3	3.5	6, 3	4.5
2, 4	3.0	4, 4	4.0	6, 4	5.0
2, 5	3.5	4, 5	4.5	6, 5	5.5
2, 6	4.0	4, 6	5.0	6, 6	6.0

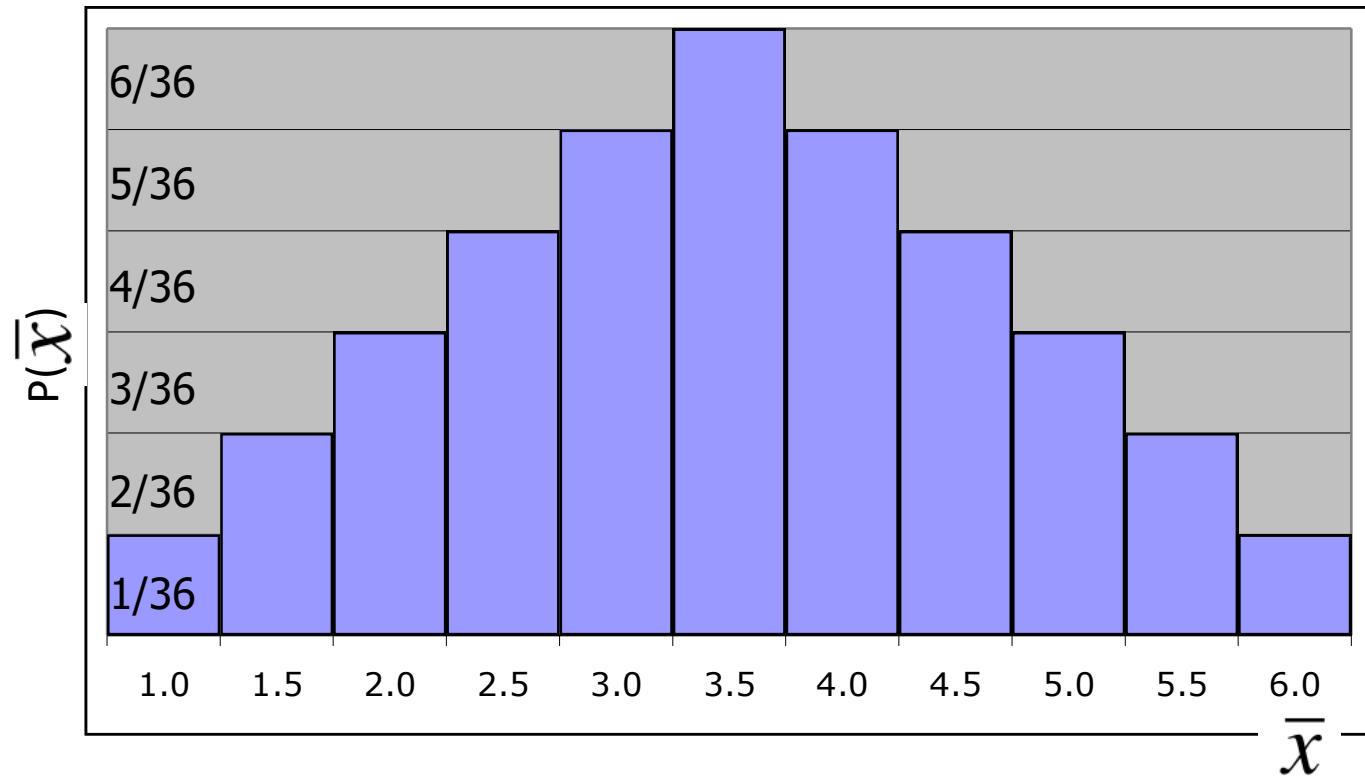
While there are 36 possible samples of size 2, there are only 11 values for \bar{x} , and some (e.g. $\bar{x} = 3.5$) occur more frequently than others. (for example, e.g. $\bar{x} = 1$).

Sampling Distribution of Two Dice



The Sampling Distribution of \bar{x} is shown below:

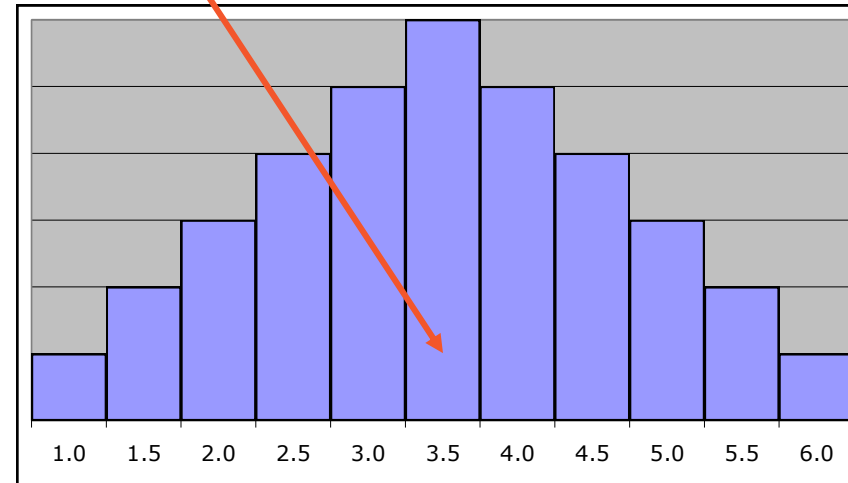
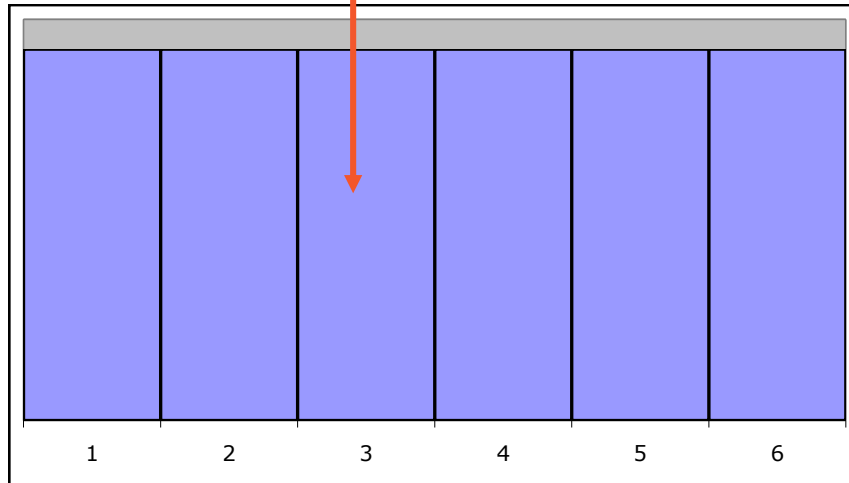
\bar{x}	$P(\bar{x})$
1.0	1/36
1.5	2/36
2.0	3/36
2.5	4/36
3.0	5/36
3.5	6/36
4.0	5/36
4.5	4/36
5.0	3/36
5.5	2/36
6.0	1/36



Compare



Compare the **distribution of X** with the sampling distribution of \bar{x}



As well as note that: $\mu_{\bar{x}} = \mu$

$$\sigma_{\bar{x}}^2 = \sigma^2 / 2$$

Sampling Distribution - Difference between Two Means



- The final sampling distribution introduced is that of the difference between **two sample means**. This requires:
 - Independent random samples be drawn from each of two normal populations
- If this condition is met, then the sampling distribution of the difference between the two sample means, i.e. $\bar{X}_1 - \bar{X}_2$ will be **normally distributed**.

Note: If the two populations are not both normally distributed, but the sample sizes are “large” (>30), the distribution of $\bar{X}_1 - \bar{X}_2$ is approximately normal)

Sampling Distribution - Difference between Two Means



➤ The expected value and variance of the sampling distribution of $\bar{X}_1 - \bar{X}_2$ are given by:

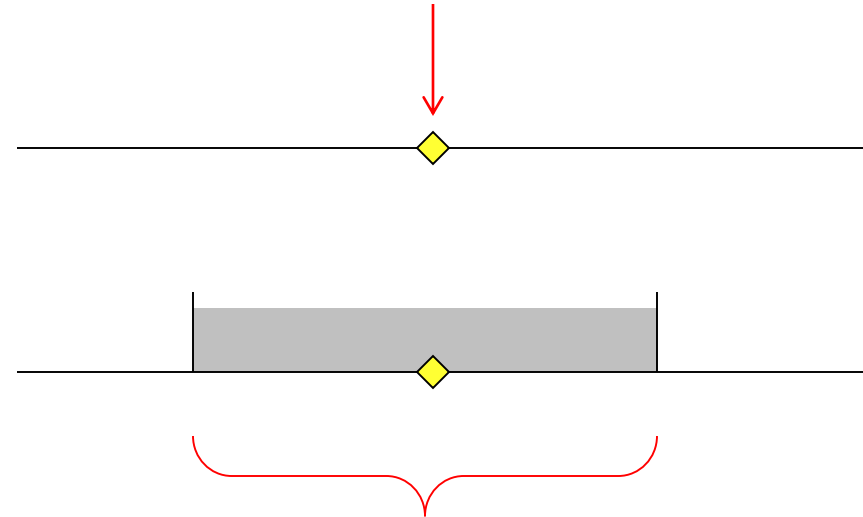
$$\text{Mean, } \mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$$

$$\text{Standard deviation} = \sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Also called the standard error of the difference between two means)

- Estimation and Hypothesis Testing are the two types of Inferences. Whereas Estimation is introduced first.
- The objective of Estimation is to determine the approximate value of a population parameter on the basis of a sample statistic.
- E.g., the sample mean (\bar{x}) is employed to estimate the population mean (μ).

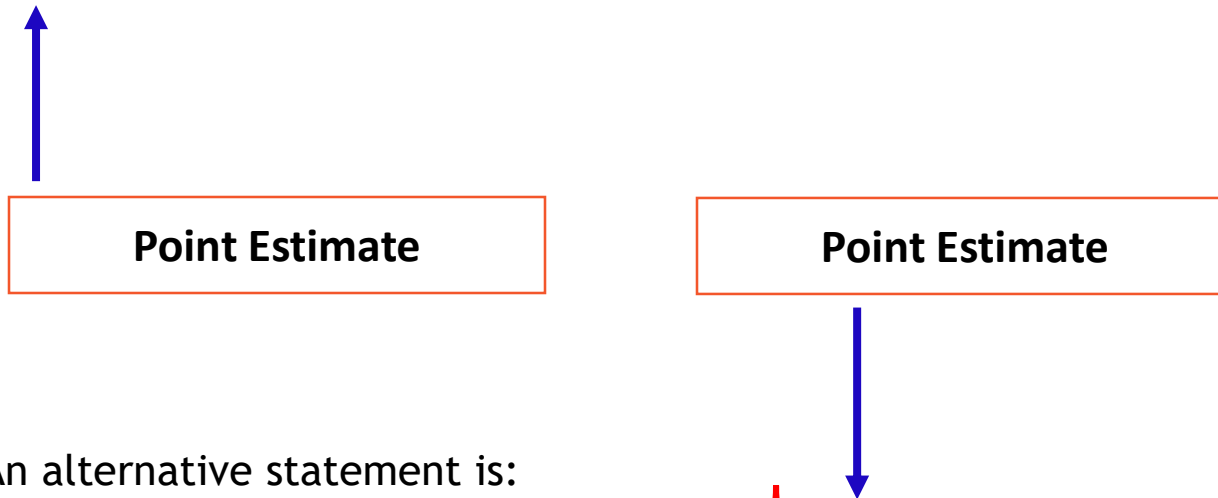
- The objective of estimation is to determine the approximate value of a population parameter on the sample statistic basis.
- There are two types of estimators:
 - Point Estimator
 - Interval Estimator



Point and Interval Estimation



- For example, suppose we want to estimate the mean summer income of a class of business students. For $n=25$ students, \bar{x} is calculated to be \$400/week.



An alternative statement is:

The mean income is *between* 380 and 420 \$/week

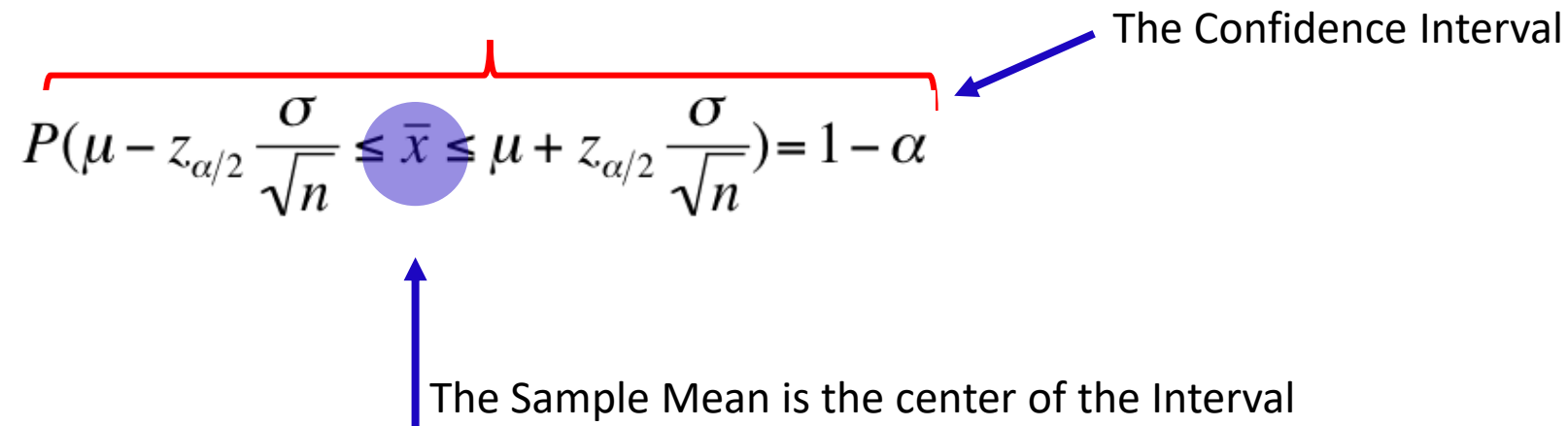
Estimating μ when σ is known

➤ We established

$$P\left(\mu - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \bar{x} \leq \mu + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

The Confidence Interval

The Sample Mean is the center of the Interval



➤ Thus the Probability that the interval is, $\bar{X} \pm Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = \left\{ \bar{X} - Z_{\alpha/2} \frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \right\}$

contains the population mean μ is $1 - \alpha$. This is a *confidence interval estimator for μ* .

Commonly Used Confidence Levels



There are four commonly used Confidence Levels.

Confidence Level



$1 - \alpha$	α	$\alpha / 2$	$z_{\alpha/2}$
.90	.10	.05	$z_{.05} = 1.645$
.95	.05	.025	$z_{.025} = 1.96$
.98	.02	.01	$z_{.01} = 2.33$
.99	.01	.005	$z_{.005} = 2.575$

Example



A computer company samples demand during lead time over 25 time periods:

235	374	309	499	253
421	361	514	462	369
394	439	348	344	330
261	374	302	466	535
386	316	296	332	334

It is known that the standard deviation of demand over lead time is 75 computers. We want to estimate the mean demand over lead time with 95% confidence in order to set inventory levels

Example



In order to use our confidence interval estimator, we need the following pieces of data:

\bar{x}	370.16	} Calculated from the data...
$z_{\alpha/2}$	1.96	
σ	75	} Given
n	25	

$$1 - \alpha = .95, \therefore \alpha/2 = .025$$
$$\text{so } z_{\alpha/2} = z_{.025} = 1.96$$

Therefore, the **lower** and **upper** confidence limits are 340.76 and 399.56.

Example - Interpretation



- The estimation for the mean demand during lead time lies between 340.76 and 399.56 – we can use this as input in developing an inventory policy.
- That is, we estimated that the mean demand during lead time falls between 340.76 and 399.56, and this type of estimator is 95% accurate of the time. That also means that 5% of the time the estimator will be incorrect.
- Incidentally, the media often refer to the 95% figure as “19 times out of 20,” which emphasizes the **long-run** aspect of the confidence level.

A wide interval provides little information.

For example, suppose we estimate with 95% confidence that an accountant's average starting salary is between \$15,000 and \$100,000.

In **Contrast** with this: a 95% confidence interval estimate of starting salaries between \$42,000 and \$45,000.

The second estimate is much narrower, providing accounting students more precise information about starting salaries.

Interval Width



The width of the confidence interval estimate is a function of the **confidence level**, the **population standard deviation**, and the **sample size**...

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

A diagram illustrating the components of the confidence interval formula. A blue arrow points from the text 'confidence level' to the $z_{\alpha/2}$ term. A red arrow points from the text 'population standard deviation' to the σ term. A green arrow points from the text 'sample size' to the \sqrt{n} term. The arrows originate from the text above and point to the corresponding parts of the formula.

Selecting the Sample Size



We can control the width of the interval by determining the sample size necessary to produce narrow intervals.

Suppose we want to estimate the mean demand “to within 5 units”; i.e. we want the interval estimate to be: $\bar{x} \pm 5$

Since: $\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$


It follows that $z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 5$

Sample Size to Estimate Mean Size



The general formula for the sample size needed to estimate a population mean with an interval estimate of: $\bar{x} \pm W$

Requires a sample size of at least this large: $n = \left(\frac{z_{\alpha/2} \sigma}{W} \right)^2$

A blue line starts from the 'W' in the interval estimate formula above, goes down, then left, and finally ends with an arrow pointing to the 'W' in the denominator of the sample size formula below.

Example



A lumber company must estimate the mean diameter of trees to determine whether or not there is sufficient lumber to harvest an area of forest. They need to estimate this to within 1 inch at a confidence level of 99%. The tree diameters are normally distributed with a standard deviation of 6 inches.

How many trees need to be sampled?

Example



Things we know:

Confidence level = 99%, therefore $\alpha = .01$

$1 - \alpha$	α	$\alpha / 2$	$z_{\alpha/2}$
.90	.10	.05	$z_{.05} = 1.645$
.95	.05	.025	$z_{.025} = 1.96$
.98	.02	.01	$z_{.01} = 2.33$
.99	.01	.005	$z_{.005} = 2.575$

$$z_{\alpha/2} = z_{.005} = 2.575$$

We want $\bar{x} \pm 1$, Hence $W = 1$.

We are given that $\sigma = 6$.

We Compute,

$$n = \left(\frac{z_{\alpha/2} \sigma}{W} \right)^2 = \left(\frac{(2.575)(6)}{1} \right)^2 = 239$$

That is, we will need to sample at least 239 trees to have a 99% confidence interval of $\bar{x} \pm 1$

- A **criminal trial** is an example of hypothesis testing without the statistics.
- In a trial a jury must decide between two hypotheses. The null hypothesis is
 - H_0 : The defendant is innocent
- The alternative hypothesis or research hypothesis is
 - H_1 : The defendant is guilty

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

There are two possible errors:

➤ Type I Error

- A Type I error occurs when we **reject a true null hypothesis**. That is, a Type I error occurs when the jury convicts an innocent person.
- The probability of Type I error is denoted as α (Greek Letter Alpha)

➤ Type II Error

- A Type II error occurs when we **don't reject a false null hypothesis**. That occurs when a guilty defendant is acquitted.
- The probability of Type II error is denoted as β (Greek Letter Beta)

The two probabilities are **inversely related**. Decreasing one increases the other.

There are two hypotheses:

- Null hypothesis (H_0)
- Alternative or research hypothesis (H_1)

The null hypothesis (H_0) will always states that the parameter equals the value specified in the alternative hypothesis (H_1).

Concepts of Hypothesis Testing



Consider example, Mean demand for computers during assembly lead time again. Rather than estimate the mean demand, our operations manager wants to know whether the mean is different from 350 units. We can rephrase this request into a test of the Hypothesis:

$$H_0: \mu = 350$$

Thus our research hypothesis becomes:

$$H_1: \mu \neq 350$$

Concepts of Hypothesis Testing



There are two possible decisions that can be made:

1. Conclude that there is enough evidence to support the alternative hypothesis
 - Also stated as, rejecting the null hypothesis in favor of the alternative hypothesis
2. Conclude that there is not enough evidence to support the alternative hypothesis
 - Also stated as, not rejecting the null hypothesis in favor of the alternative hypothesis

Note: We do not say that we accept **Null Hypothesis**

Concepts of Hypothesis Testing



Once the null and alternative hypotheses are stated, the next step is to randomly sample the population and calculate a test statistic (In this example, sample mean)

If the test statistic's value is inconsistent with the null hypothesis we reject the null hypothesis and infer the alternate hypothesis is true.


For example, if we're trying to decide whether the mean is not equal to 350, a large value of \bar{x} (say, 600) would provide enough evidence. If \bar{x} is close to 350 (say, 355) we could not say that this provides a great deal of evidence to infer that the population mean is different than 350.

Concepts of Hypothesis Testing



A Type I error occurs when we **reject a true null hypothesis** (i.e. Reject H_0 When it is True)

A Type II error occurs when we **don't reject a false null hypothesis** (i.e. Do not Reject H_0 When it is False)

H_0	T	F
Reject	I	
 Reject		II



Email us - support@acadgild.com