# Part 1: BASIC STATISTICAL CONCEPTS (continued)

## 1. Probability

## 2. Random variables

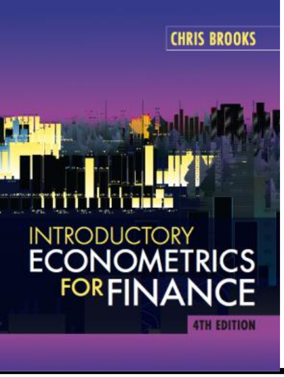
## 3. Some important theoretical probability distributions

- Descriptive Statistics
- Normal distribution, Student's t-distribution,
- Chi-square distribution, F distribution

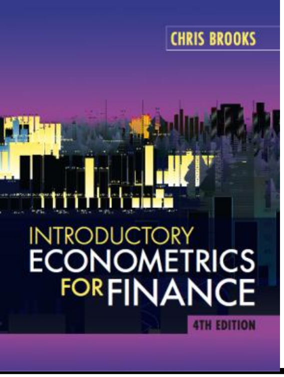
## 4. Statistical Inference: Estimation

- Point Estimation
- Interval Estimation (Confidence Interval)

## 5. Statistical Inference: Hypothesis Testing

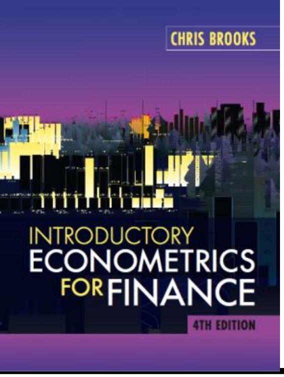



# 0. DESCRIPTIVE STATISTICS




# Measures of central tendency

- The average value of a series is its *measure of location* or *measure of central tendency*, capturing its ‘typical’ behaviour
- There are three broad method to calculate the average value of a series: the *mean*, *median* and *mode*
- The **mean** is the very familiar sum of all $N$ observations divided by $N$
- More strictly, this is known as the **arithmetic mean**
- The **mode** is the most frequently occurring value in a set of observations
- The **median** is the middle value in a series when the observations are arranged in ascending order
- Each of the three methods of calculating an average has advantages and disadvantages

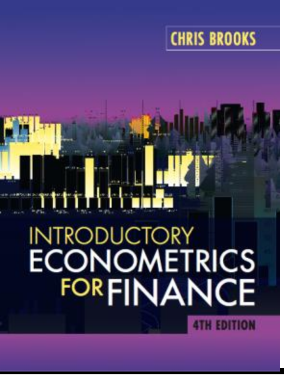



# The geometric mean

- The **geometric mean** involves calculating the *N*th root of the product of the *N* observations, more relevant for growth
- So the geometric mean of six numbers in a series would be obtained by multiplying them together and taking the sixth root
- **In finance**, when the numbers in the series can be negative or 0 (like returns), we can use a slightly different method to calculate the **geometric mean**
- Here we add one to each data point, then multiply together, take the *N*th root and then subtract one at the end

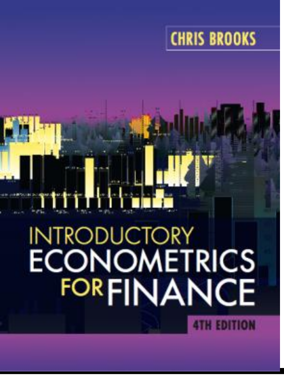
$$\overline{R}_G = [(1+r_1)(1+r_2)\ldots(1+r_N)]^{1/N} - 1$$

- where $r_1$, $r_2$ etc. are the data points that we wish to take the geometric mean of
- The geometric mean will always be smaller than the arithmetic mean unless all of the data points are the same.



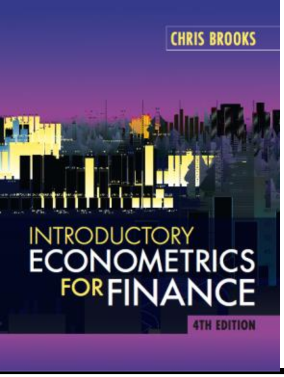

# Measures of spread

- The spread of a series about its mean value can be measured using the ***variance* or *standard deviation*** (which is the square root of the variance)
- This quantity is an important measure of risk in finance
- The standard deviation scales with the data, whereas the variance scales with the square of the data. So, for example, if the units of the data points are US dollars, the standard deviation will also be measured in dollars whereas the variance will be in dollars squared
- Other measures of spread include the ***range*** (the difference between the largest and smallest of the data points) and the ***interquartile range*** (the difference between the third and first quartile points in the series)
- The ***coefficient of variation*** divides the standard deviation by the sample mean to obtain a unit-free measure of spread that can be compared across series with different scales.



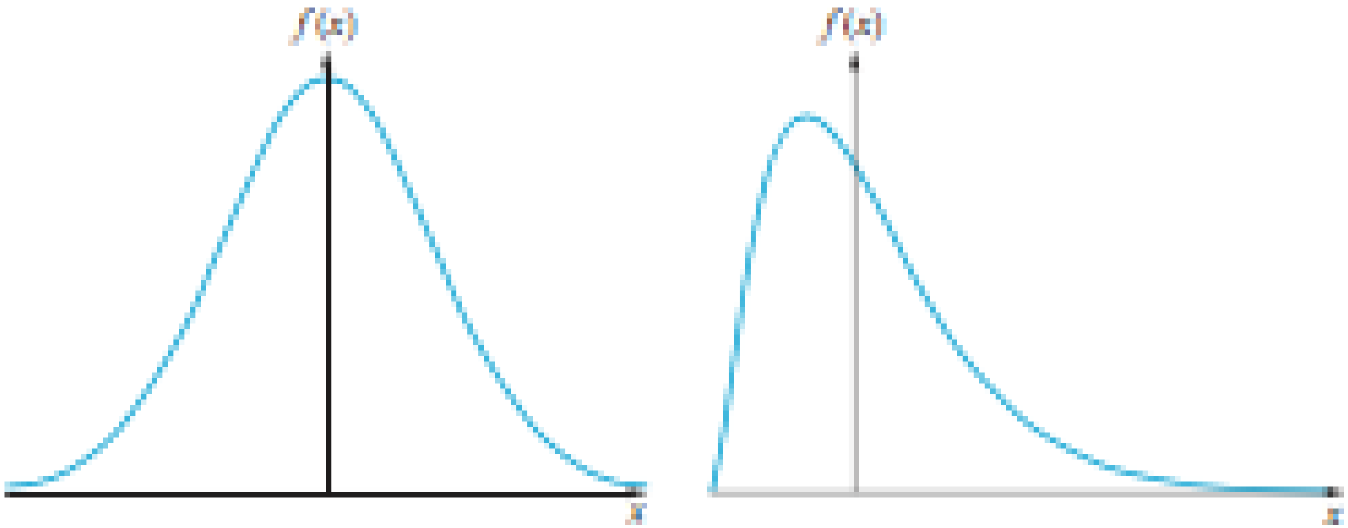

# Higher moments

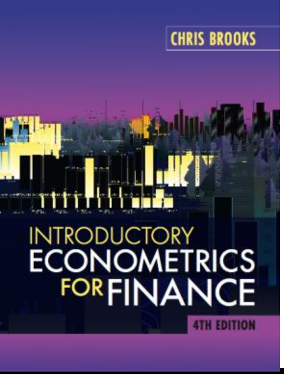
- The higher moments of a data sample give further indications of its features and shape.
- **Skewness** is the standardised third moment of a distribution, defines the shape of the distribution, and indicates the extent to which it is not symmetric about the mean
- **Kurtosis** is the standardised fourth moment which measures the fatness of the tail and how peaked is the distribution at the mean
- Skewness can be positive or negative while kurtosis can only be positive
- Skewness is equal to 0, kurtosis is equal to 3 for a normal distribution



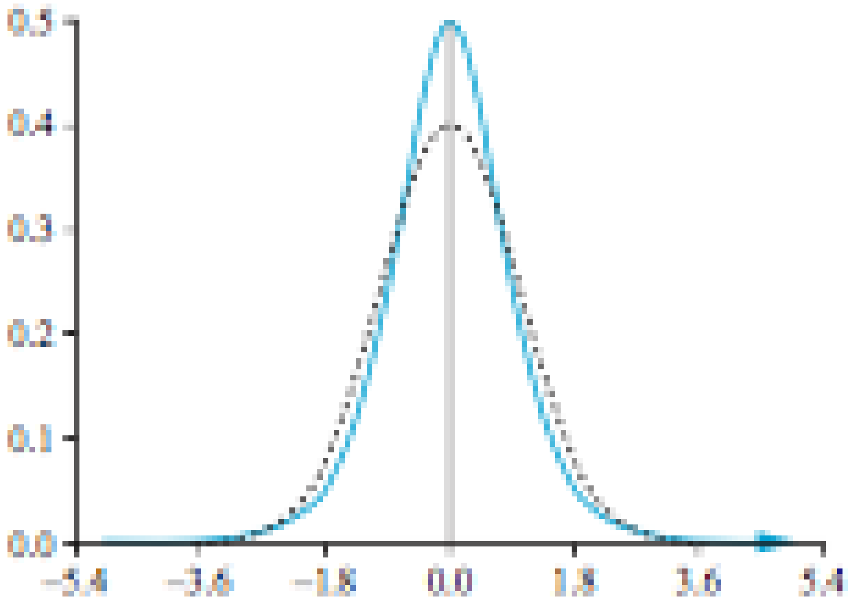

$$\text{Kurtosis (X)} = E\left[\left(\frac{X-\mu}{\sigma}\right)^4\right]$$

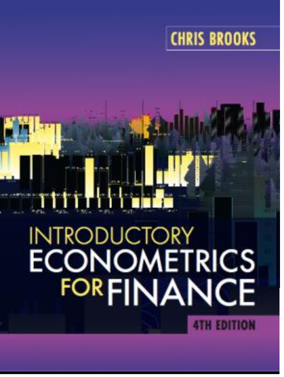
# Plot of a skewed series versus a normal distribution



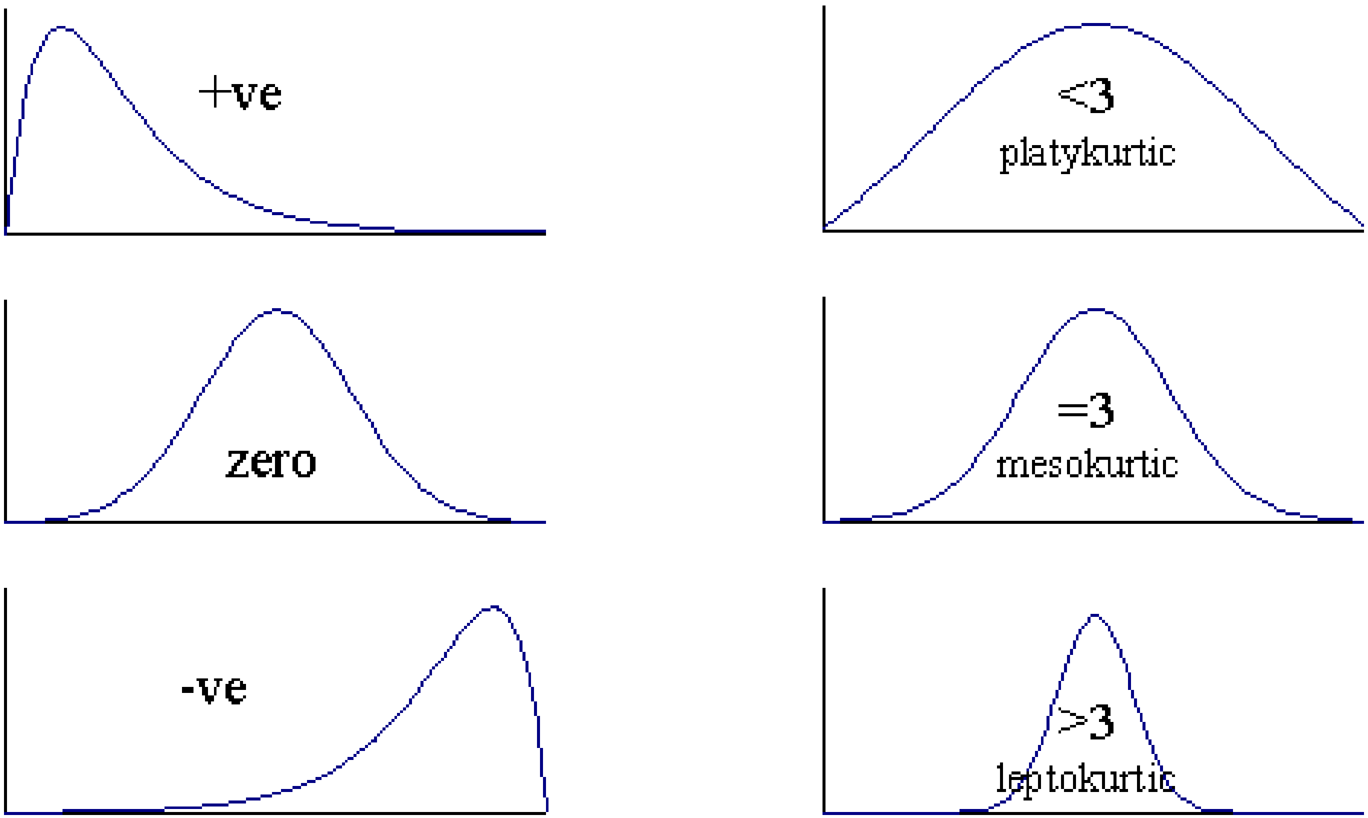


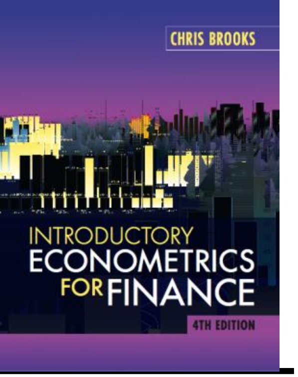

# Plot of a leptokurtic series versus a normal distribution

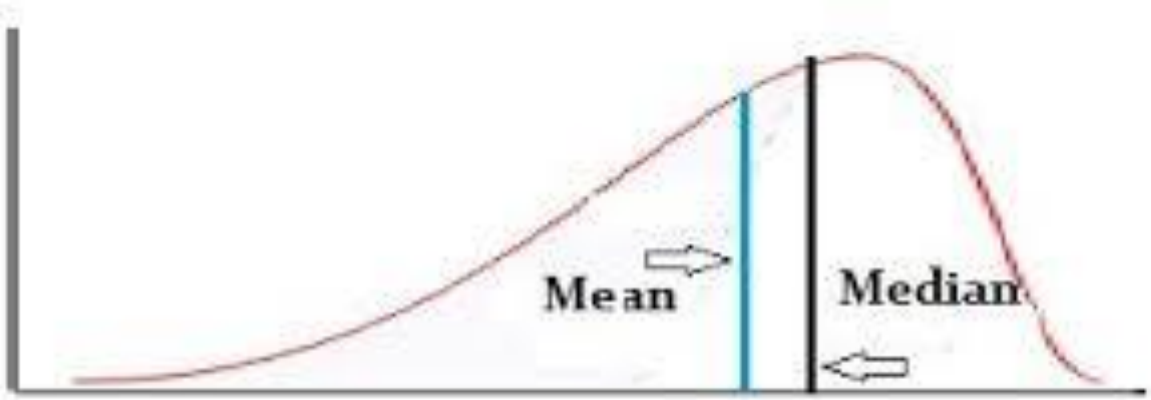




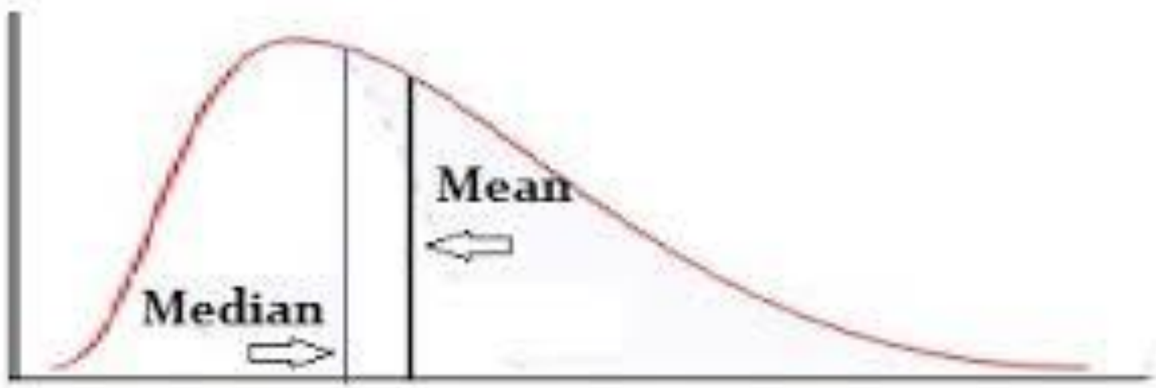

## Skewness Kurtosis



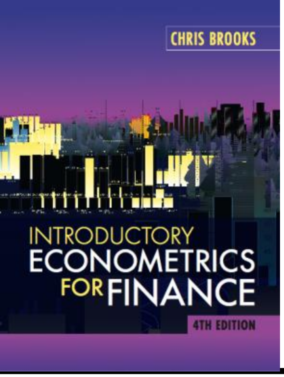



Left skewed: Mean is to the left

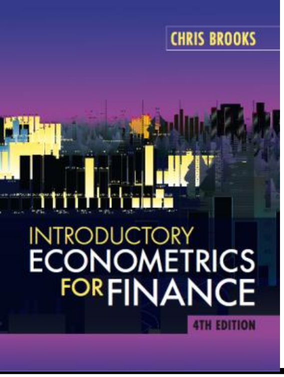



Right skewed distribution: Mean is to the right




# Measures of association

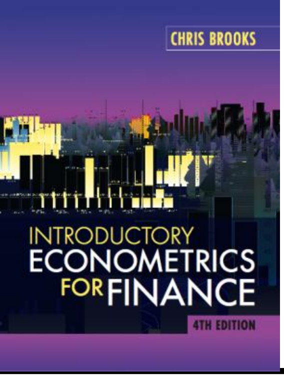
- **Covariance** is a linear measure of association between two series.
- It is simple to calculate but scales with the standard deviations of the two series
- The **correlation coefficient** is another measure of association that is calculated by dividing the covariance between two series by the product of their standard deviations
- Correlations are unit-free and must lie between (-1,+1)
- The correlation calculated in this way is more specifically known as Pearson's correlation measure between continuous variables.
- An alternative measure is known as Spearman's rank correlation measure, involving ordinal variables.




# Some algebra useful for working with means, variances and covariances

## Means

- The mean of a random variable $y$ is also known as its expected value, written $E(y)$.
- The expected value of a constant is the constant, e.g. $E(c) = c$
- The expected value of a constant multiplied by a random variable is equal to the constant multiplied by the expected value of the variable: $E(cy)=c\,E(y)$. It can also be stated that $E(cy+d)= cE(y)+d$, where d is also a constant.
- For two independent random variables, $y_1$ and $y_2$, $E(y_1y_2) = E(y_1)\,E(y_2)$




# Some algebra useful for working with means, variances and covariances 2

## Variances

- The variance of a random variable $y$ is usually written var($y$).
- The variance of a random variable y is given by var($y$) = E[$y$ − E($y$)]$^2$. The variance of a constant is zero: var($c$) = 0
- For c and d constants, var($cy + d$) = $c^2$var($y$)
- For two independent random variables, $y_1$ and $y_2$, var($cy_1 + dy_2$) = $c^2$var($y_1$) + $d^2$var($y_2$)

## Covariances

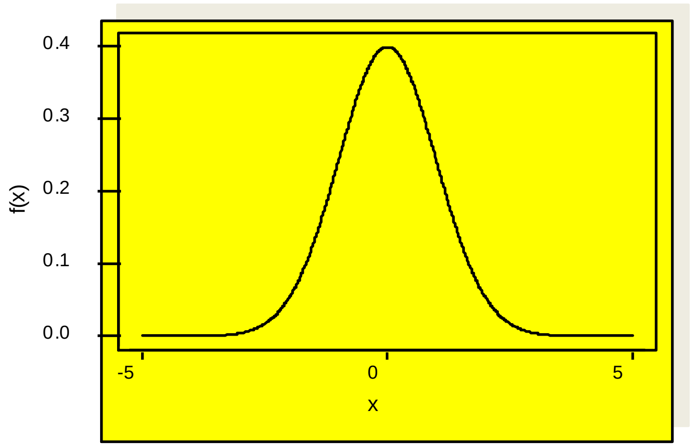
- The covariance between two random variables, $y_1$ and $y_2$ may be expressed as cov($y_1$, $y_2$)
- cov($y_1$, $y_2$) = E[($y_1$ − E($y_1$))($y_2$ − E($y_2$))]
- For two independent random variables, $y_1$ and $y_2$, cov($y_1$, $y_2$) = 0
- For four constants, $c$, $d$, $e$, and $f$, cov($c+dy_1$, $e+fy_2$)=$df$cov($y_1$, $y_2$).

# 1. Normal Distribution

**Normal Probability Density Function:**

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

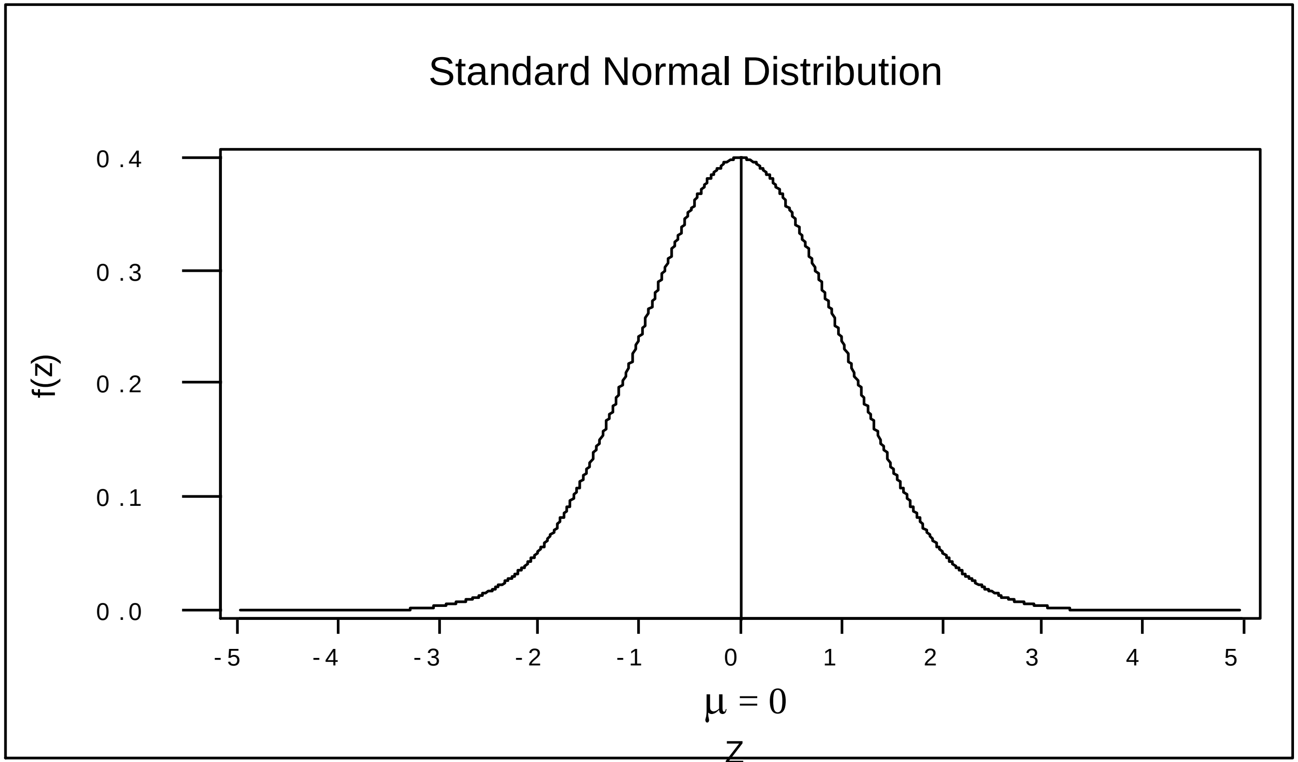
Normal Distribution: $\mu = 0$, $\sigma^2 = 1$



# The Standard Normal Distribution

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

The **standard normal random variable**, Z, is the normal random variable with mean $\mu = 0$ and standard deviation $\sigma = 1$: $Z \sim N(0,1^2)$.



# Properties of normal distribution

- **Symmetric** about the mean
- $P(a<x<b)$= **area of the region** between the density function, horizontal axis and vertical lines $x=a$, $x=b$
- **Sum of independent normal R.V**. is normally distributed
- **Central Limit Theorem**: sample mean is normally distributed as sample size increases to $\infty$
- Skewness=0, kurtosis=3

# The Central Limit Theorem

Let $S_n = X_1 + \ldots + X_n$ be the sum of independent random variables with the same distribution. Then for large n (n>30), the distribution of $S_n$ is approximately normal with mean

$$E(S_n) = n\mu \text{ and } SD(S_n) = \sigma\sqrt{n},$$
$$\text{where } \mu = E(X_i) \text{ and } \sigma = SD(X_i)$$

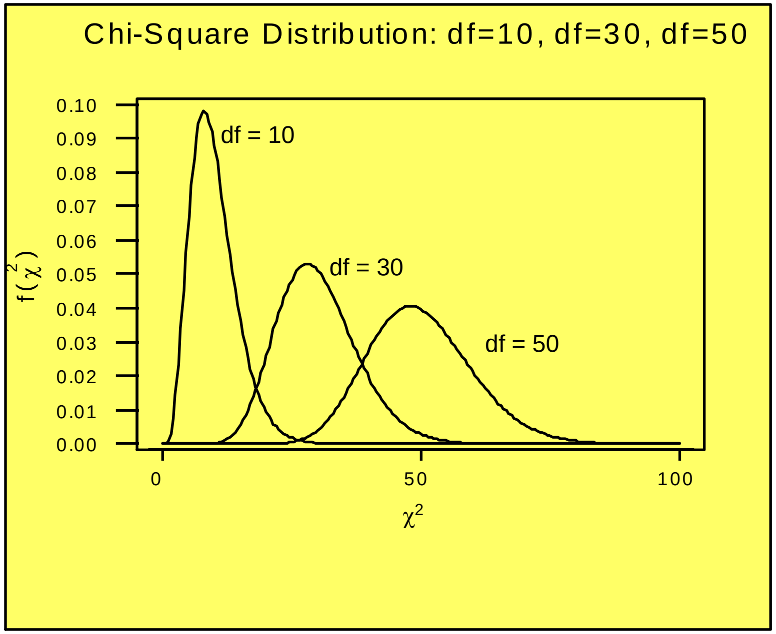
# 2. The Chi-Square ($\chi^2$) Distribution

$$Z = \sum_{i=1}^{k} Z_i^2 \sim \chi_k^2$$

- ✓The **chi-square distribution** is the probability distribution of the sum of **k** independent, squared standard normal random variables.

- ✓The mean of the chi-square distribution is equal to the degree of freedom parameter, **E[$\chi^2$] = *k*.** The variance of a chi-square is equal to twice the degree of freedom, **Var[$\chi^2$] = *2k***.

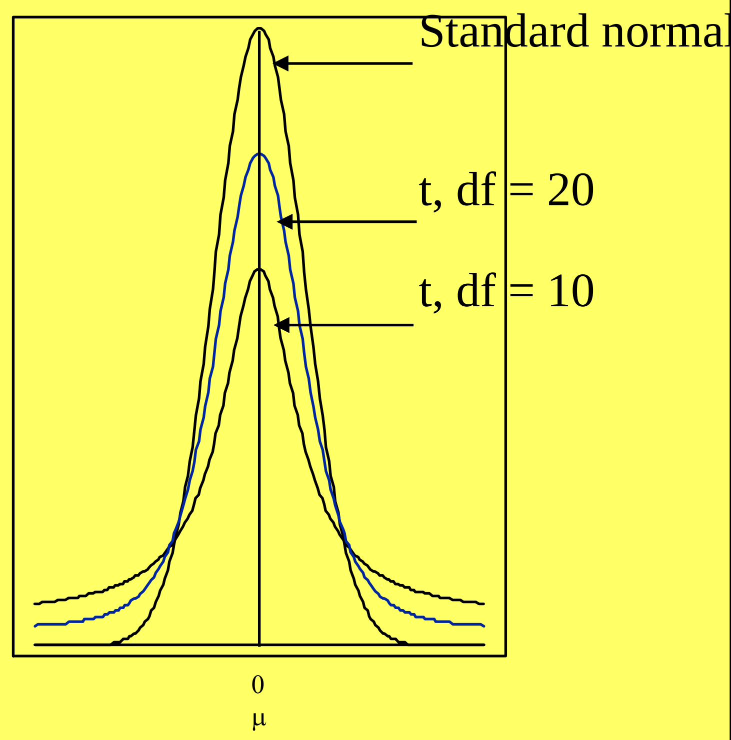
- The **chi-square** random variable cannot be negative, so it is bounded by zero on the left.
- The **chi-square** distribution is skewed to the right.
- The **chi-square** distribution approaches a **normal distribution** as the degree of freedom increases.
- Sum of independent **chi-square** RV is also a chi-square RV.



# 3. The t Distribution (Student')

$$t = \frac{Z_1}{\sqrt{\chi_k^2 / k}} \sim t_k$$

- The ***t* distribution** is a family of bell-shaped and symmetric distributions.
- The **expected value** of *t* distribution is **0**.
- For *df > 2*, the **variance** of *t* distribution is **df/(df-2)**.
- The *t* distribution is flatter and has fatter tails than standard normal.
- The *t* distribution **approaches a standard normal** as the number of degree of freedom increases



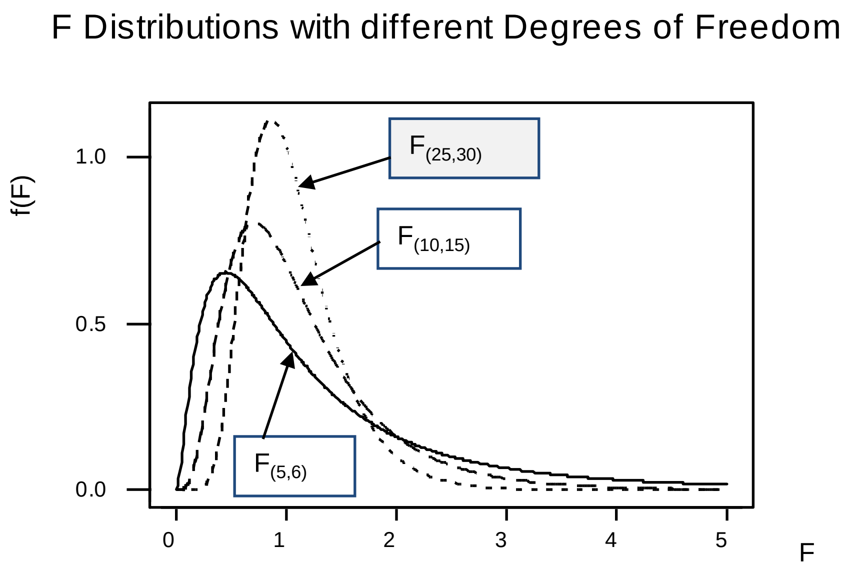
# 4. The *F* Distribution

The **F distribution** is the distribution of the **ratio of two chi-square** random variables that are **independent** of each other, each of which is divided by its own degree of freedom.

An ***F*** random variable with $k_1$ and $k_2$ degrees of freedom:

$$F_{(k_1,\ k_2)} = \frac{\chi_1^2 / k_1}{\chi_2^2 / k_2}$$

# The *F* Distribution




- The ***F* random variable** cannot be negative, so it is bounded by zero on the left.
- The ***F* distribution** is skewed to the right.
- The ***F* distribution** is identified by the **degree of freedom in the numerator**, $\mathbf{k_1}$, and the **degree of freedom in the denominator**, $\mathbf{k_2}$.

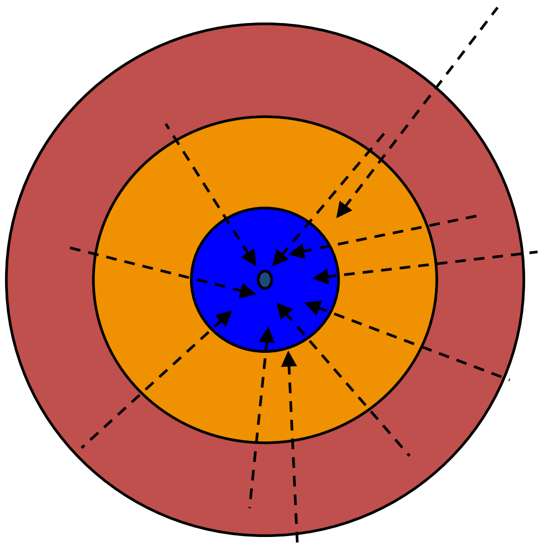
# 5. Estimators and Properties

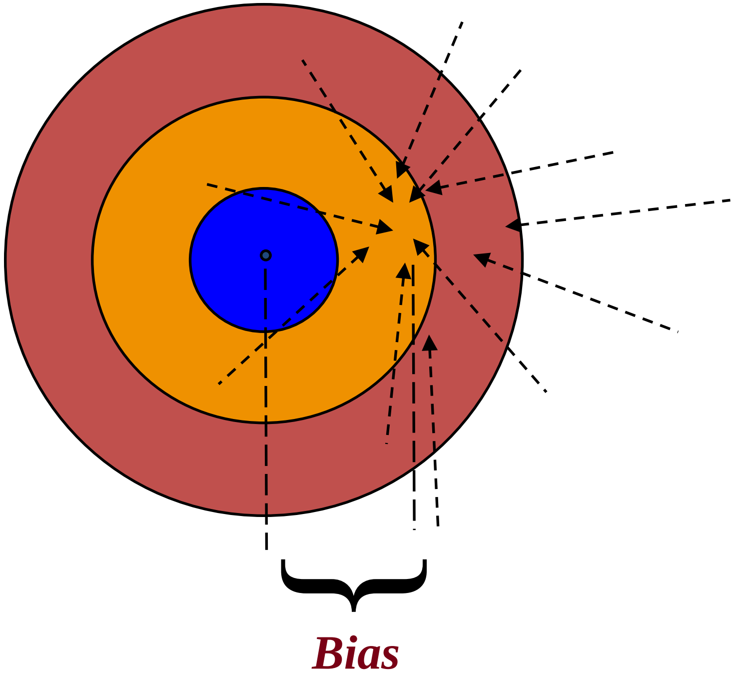
An **estimator** of a population parameter is a sample statistic used to estimate the parameter. The most commonly-used estimator of the:

| Population Parameter | | Sample Statistic |
|---|---|---|
| Mean ($\mu$) | is the | Mean ($\bar{X}$) |
| Variance ($\sigma^2$) | is the | Variance ($s^2$) |
| Standard Deviation ($\sigma$) | is the | Standard Deviation ($s$) |
| Proportion (p) | is the | Proportion $(\hat{p})$ |

- Desirable properties of estimators include:
  - ✓ Unbiasedness
  - ✓ Efficiency
  - ✓ Consistency

# Unbiased and Biased Estimators



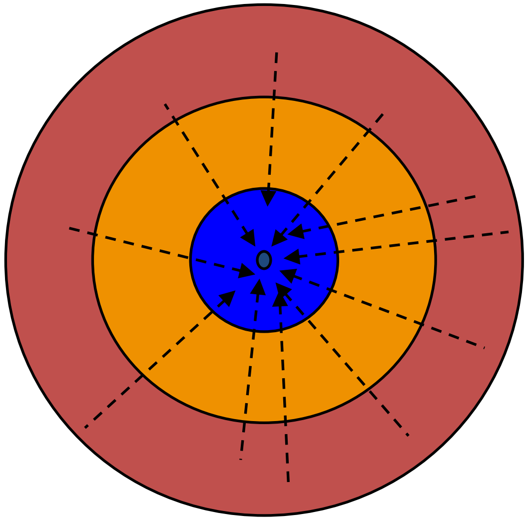


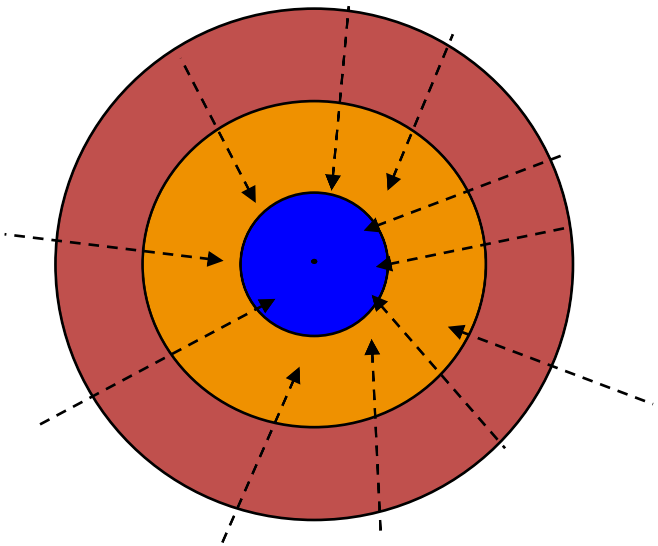

An **unbiased** estimator is on target on average.

A **biased** estimator is off target on average.

# Efficiency

An estimator is **efficient** if it has a relatively small variance (and standard deviation).





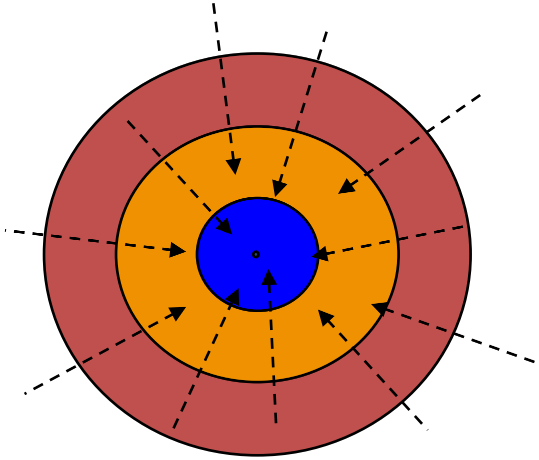
An **efficient** estimator is, on average, closer to the parameter being estimated.

An **inefficient** estimator is, on average, farther from the parameter being estimated.

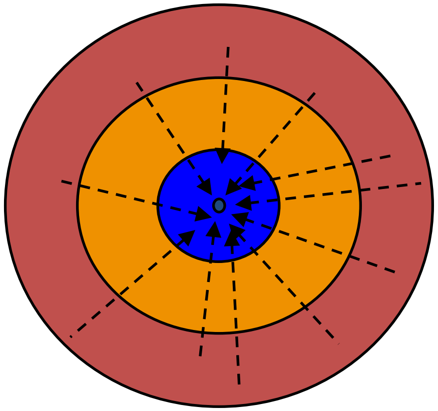
# Consistency

An estimator is said to be **consistent** if its probability of being close to the parameter it estimates increases as the sample size increases.

***Consistency***



n = 10



n = 100

# 6. Confidence Interval (Interval Estimate)

**2 types of estimators:**

- **Point Estimate**
  - ✓A single-valued estimate.
  - ✓Conveys little information about the actual value of the population parameter, about the accuracy of the estimate.
- **Interval Estimate** (or **Confidence Interval**)
  - ✓An ***interval*** or range of values believed to include the unknown population parameter.
  - ✓Associated with the interval is a measure of the ***confidence*** we have that the interval does indeed contain the parameter of interest.

# Confidence Interval for the mean μ by sample mean

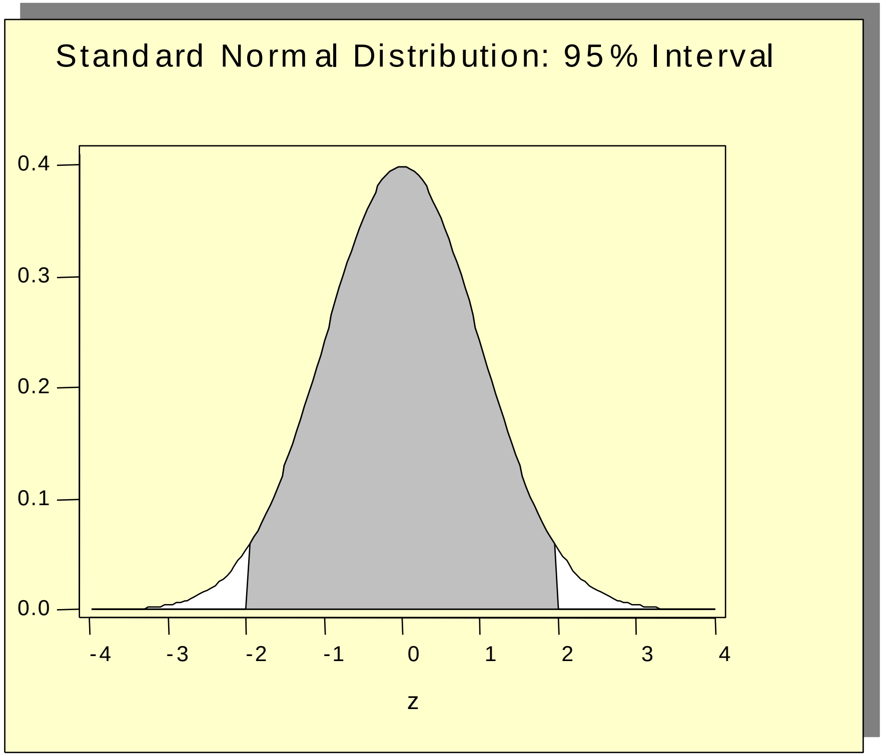
According to the Central Limit Theorem, $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

Then:

$$P\left(-1.96 < \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < 1.96\right) = 0.95$$

$$\Updownarrow$$

$$\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$$



# 7. Statistical Hypothesis Testing

- A **null hypothesis**, denoted by $H_0$, is an assertion about one or more population parameters. This is the assertion we hold to be true until we have sufficient statistical evidence to conclude otherwise.

  ✓ $H_0: \mu = 100$

- The **alternative hypothesis**, denoted by $H_1$, is the assertion of all situations *not* covered by the null hypothesis.

  ✓ $H_1: \mu \neq 100$

# The Null Hypothesis, $H_0$

- The **null hypothesis**:
  - ✓ Often represents an existing belief.
  - ✓ Is maintained to be true, until a **test** leads to its rejection in favor of the alternative hypothesis.
  - ✓ Is accepted or rejected on the basis of a consideration of a ***test statistic***.

# Concepts of Hypothesis Testing

- A **test statistic** is a sample statistic computed from sample data. The value of the test statistic is used in determining whether or not we may reject the null hypothesis.
- The **decision rule** of a statistical hypothesis test is a rule that specifies the conditions under which the null hypothesis may be rejected.

Consider $H_0$: $\mu = 100$. We may have a decision rule that says: "Reject $H_0$ if the sample mean is less than 95 or more than 105."

In a courtroom we may say: "The accused is innocent until proven guilty beyond a reasonable doubt."

# Decision Making

- There are two possible states of nature:
  - ✓ $H_0$ is *true*
  - ✓ $H_0$ is *false*
- There are two possible decisions:
  - ✓ *Do not reject $H_0$*
  - ✓ *Reject $H_0$*

# Type I and Type II Errors

A **contingency table** illustrates the possible outcomes of a statistical hypothesis test.

| | State of Nature | |
|---|---|---|
| **Decision** | $H_0$ True | $H_0$ False |
| Do not Reject $H_0$ | Correct | Type II Error ($\beta$) |
| Reject $H_0$ | Type I Error ($\alpha$) | Correct |

# 1-Tailed and 2-Tailed Tests

The tails of a statistical test are determined by the need for an action. If action is to be taken if a parameter is greater than some value *a*, then the alternative hypothesis is that the parameter is greater than *a*, and the test is a **right-tailed** test.

$H_0: \mu \leq 50$

$H_1: \mu > 50$

If action is to be taken if a parameter is less than some value *a*, then the alternative hypothesis is that the parameter is less than *a*, and the test is a **left-tailed** test.

$H_0: \mu \geq 50$

$H_1: \mu < 50$

If action is to be taken if a parameter is either greater than or less than some value *a*, then the alternative hypothesis is that the parameter is not equal to *a*, and the test is a **two-tailed** test.

$H_0: \mu = 50$

$H_1: \mu \neq 50$

**FIGURE 7–6** A Right-Tailed Test: The Rejection Region for $H_0$: $\mu \leq 1,000$; $\alpha = 5\%$

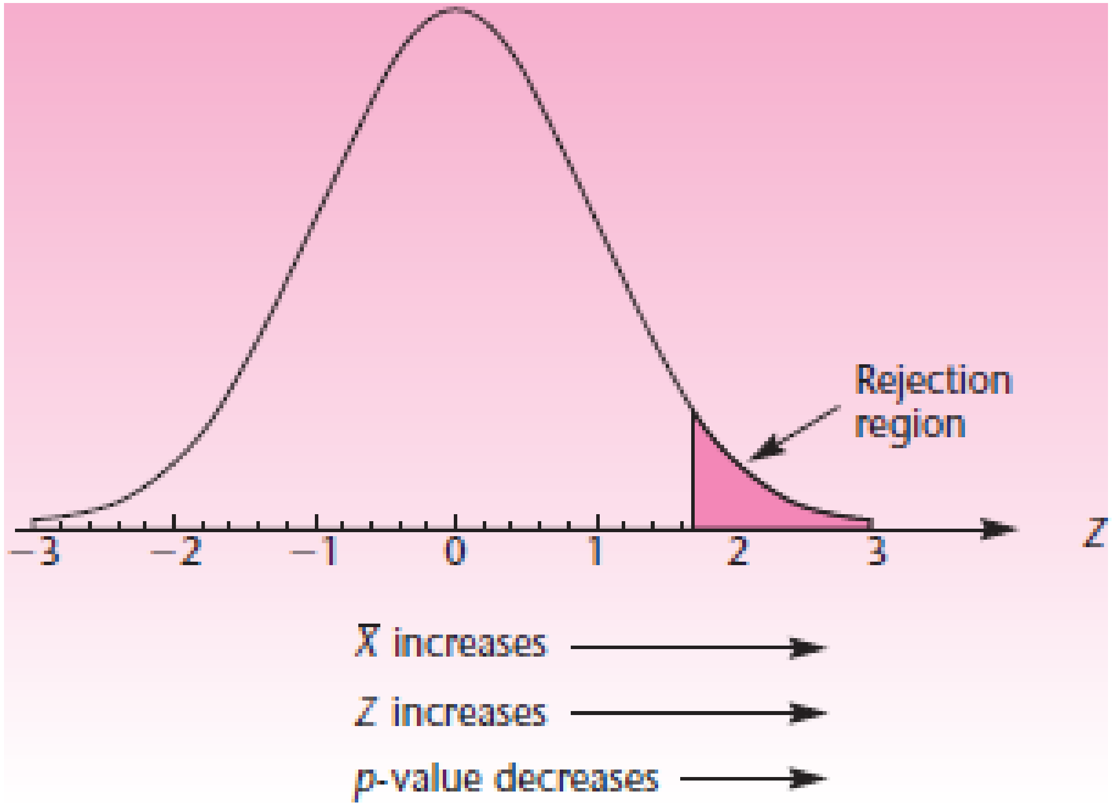


FIGURE 7–5 A Left-Tailed Test: The Rejection Region for $H_0$: $\mu \geq 1{,}000$; $\alpha = 5\%$

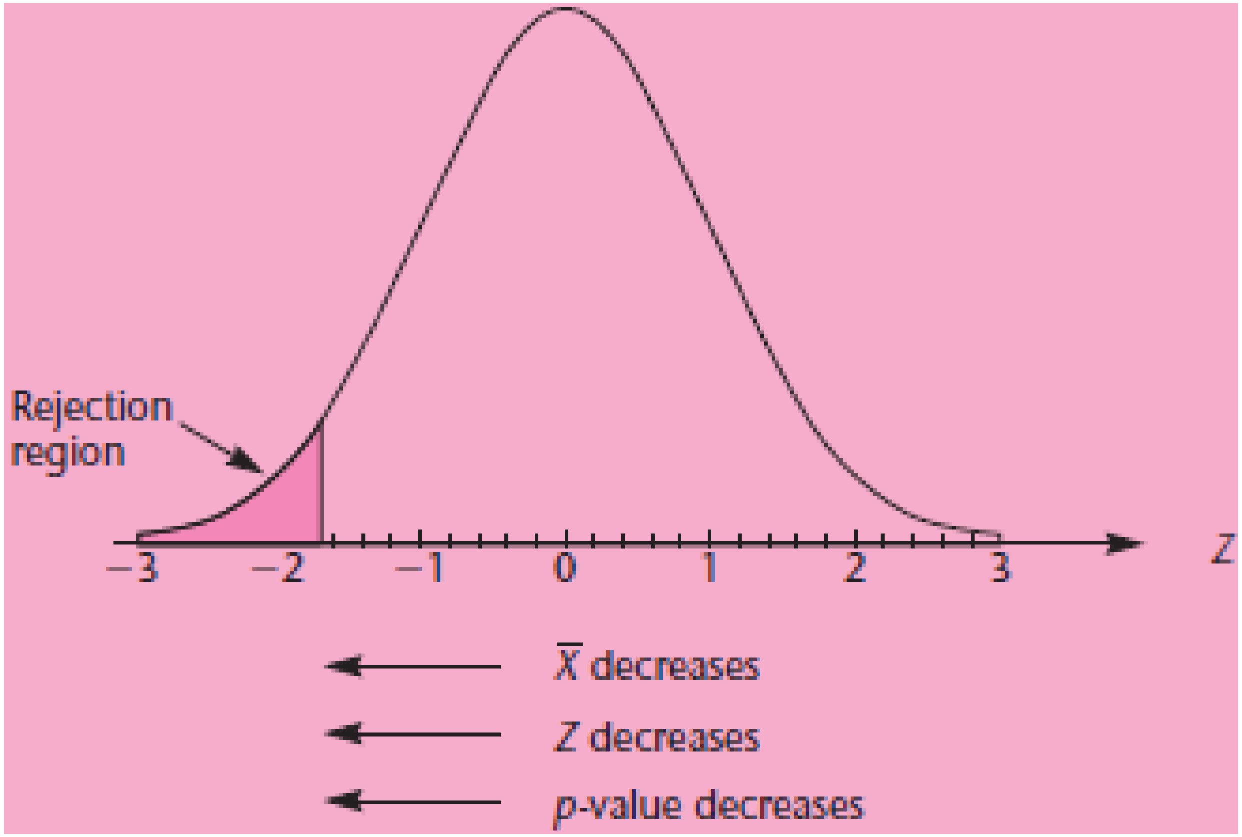
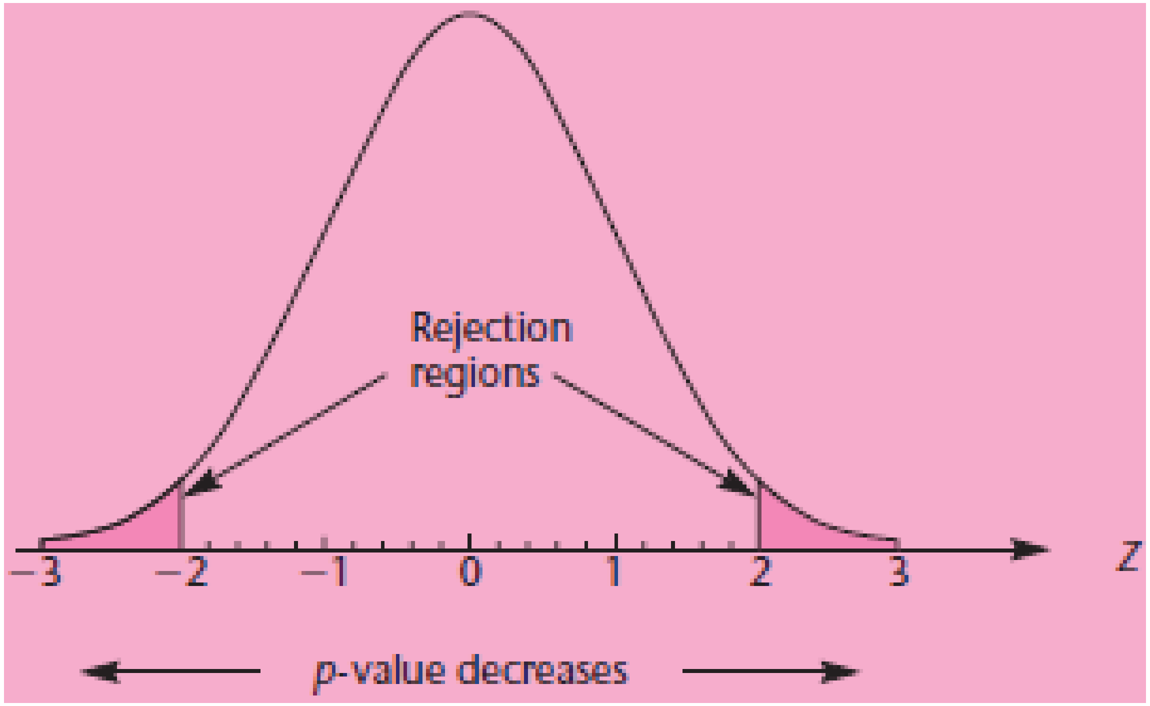


FIGURE 7–7 A Two-Tailed Test: Rejection Region for $H_0$: $\mu = 1{,}000$, $\alpha = 5\%$



# Rejection Region

- The **rejection region** of a statistical hypothesis test is the range of numbers that will lead us to reject the null hypothesis in case the test statistic falls within this range.
- The rejection region, also called the **critical region**, is defined by the **critical points**.
- The rejection region is defined so that, before the sampling takes place, our test statistic will have a probability $\alpha$ of falling within the rejection region if the null hypothesis is true.
- So, the rejection region has the area equal to $\alpha$

# Five-Step Procedure for Hypothesis Testing: 1st approach

- **Step 1:** State the null hypothesis $H_0$.
- **Step 2:** State the alternative hypothesis $H_1$.
- **Step 3:** Compute the test statistic (T.S.) value.
- **Step 4:** Determine the rejection region for a given level of significance $\alpha$.
- **Step 5:** Conclude based on the rule:
    - If T.S. belongs to the rejection region, we reject $H_0$.
    - Else we don't reject $H_0$.

# Five-Step Procedure for Hypothesis Testing: 2nd approach

- **Using confidence interval (for 2 tailed-tests)**

# The *p*-Value

**The p-value** is the area of the rejection region when the Test Statistics is equal to the Critical Value.

**The p-value** is the smallest level of significance $\alpha$, at which the null hypothesis may be rejected using the obtained value of the test statistic.

**RULE: When the *p*-value is less than or equal to the significance level $\alpha$ , reject $H_0$.**

# Five-Step procedure for Hypothesis Testing: 3rd approach using p-value

- **Step 1:** State the null hypothesis $H_0$.
- **Step 2:** State the alternative hypothesis $H_1$.
- **Step 3:** Compute the test statistic (T.S.) value.
- **Step 4:** Get the p-value corresponding to T.S. (often with software).
- **Step 5:** Compare the p-value and the level of significance $\alpha$.
  - If p-value $\leq \alpha$ we reject $H_0$.
  - Else we don't reject $H_0$.

**FIGURE 7–6** A Right-Tailed Test: The Rejection Region for $H_0$: $\mu \le 1{,}000$; $\alpha = 5\%$

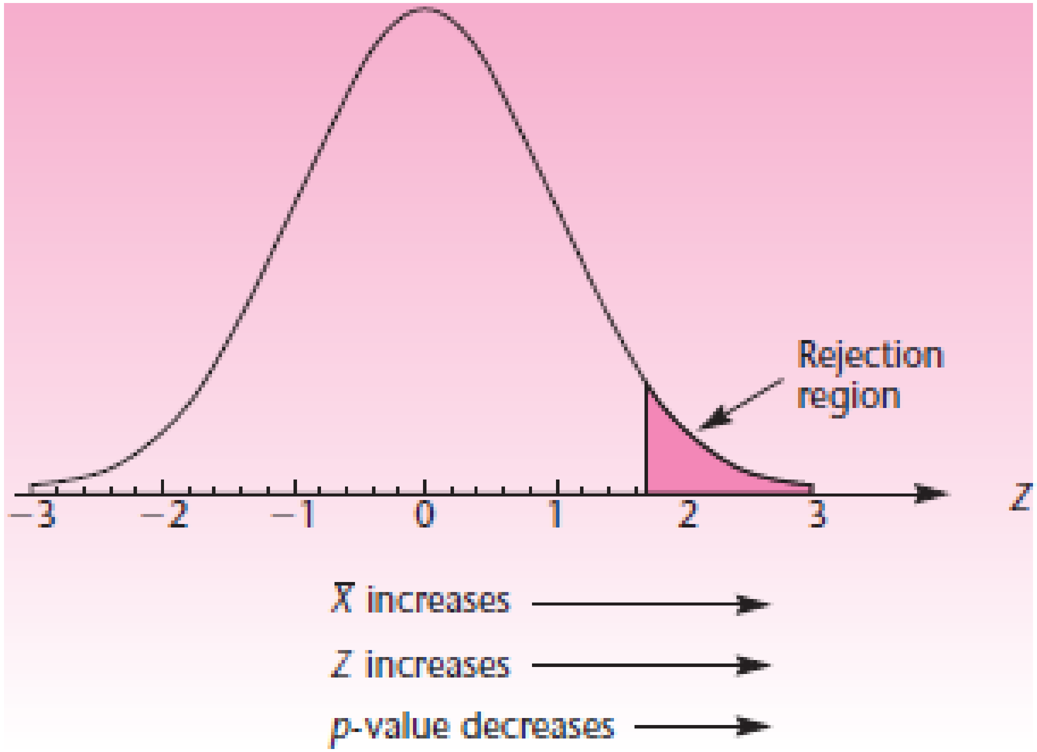


FIGURE 7–5 A Left-Tailed Test: The Rejection Region for $H_0$: $\mu \geq 1{,}000$; $\alpha = 5\%$

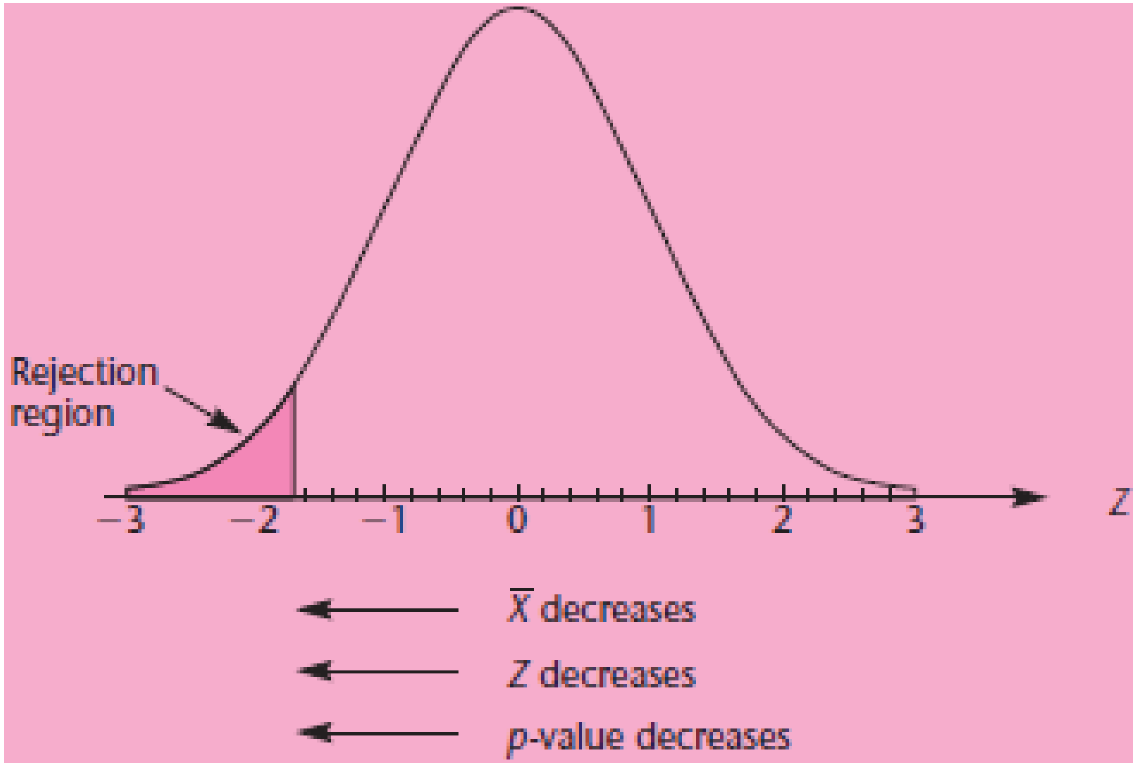
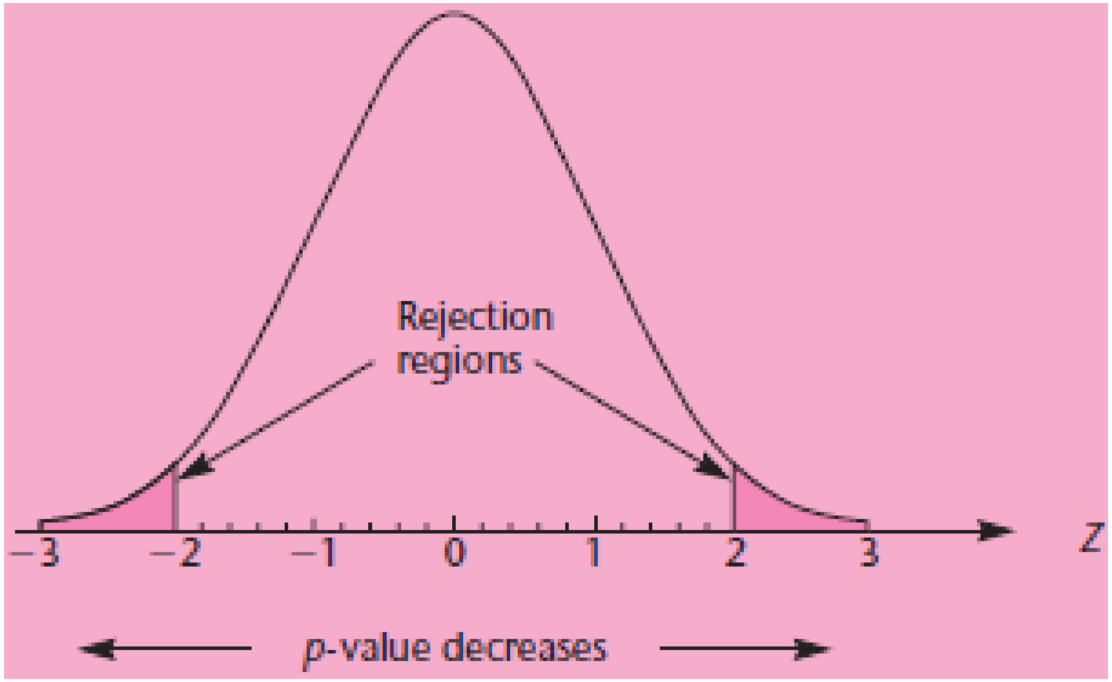


**FIGURE 7–7** **A Two-Tailed Test: Rejection Region for $H_0$: $\mu = 1{,}000$, $\alpha = 5\%$**



# Testing Population Means

- Cases in which the **test statistic** is ***Z*** (when using σ)

*The formula for calculating Z is:*

$$z = \frac{\bar{x} - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

# Testing Population Means

- Cases in which the **test statistic** is ***t*** (when using s)

*The formula for calculating t is:*

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