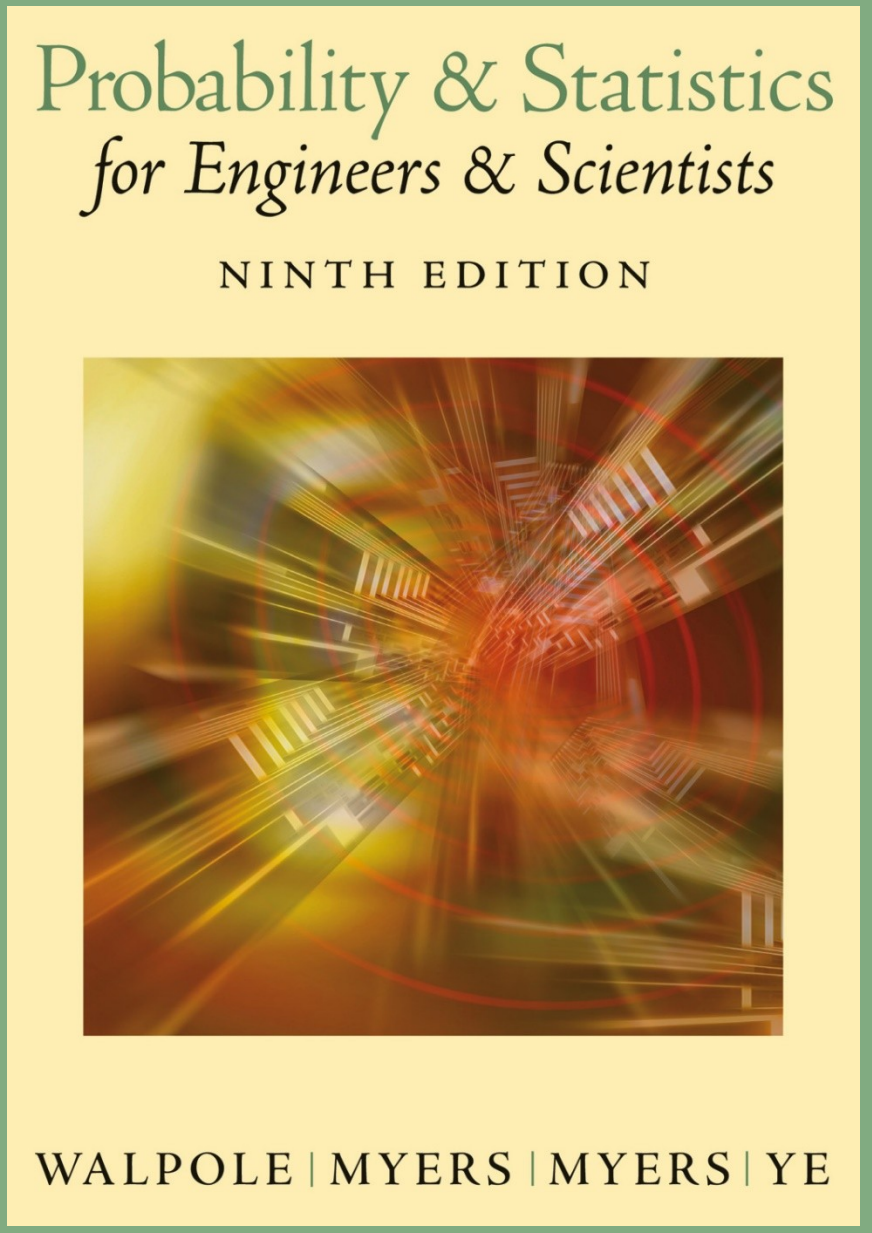
Chapter 8

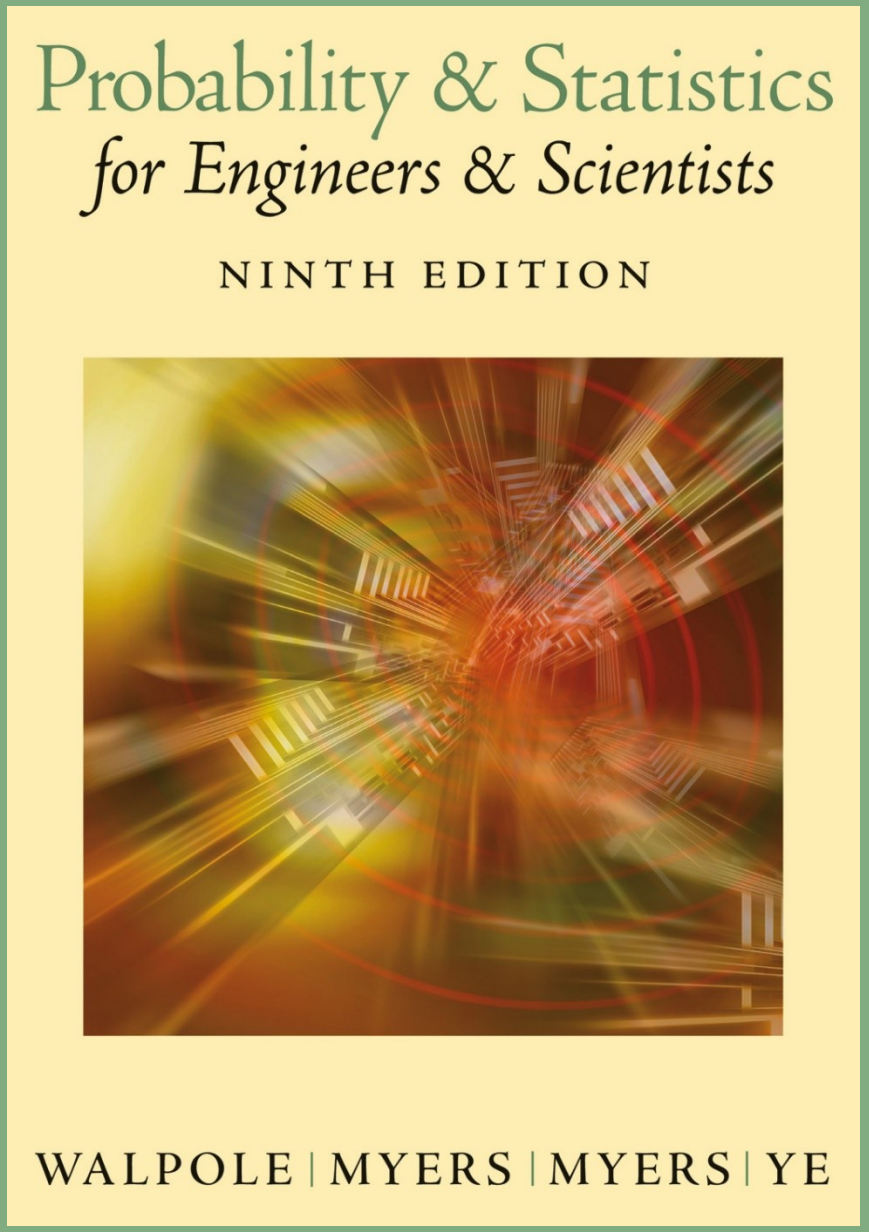
Fundamental Sampling Distributions and Data Descriptions



PEARSON Addison Wesley

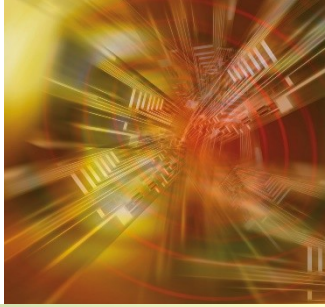
Section 8.1

Random Sampling



PEARSON
Addison Wesley

8.1 Random Sampling:



Definition 8.1:

A population consists of the totality of the observations with which we are concerned. (Population=Probability Distribution)

A sample is a subset of a population.

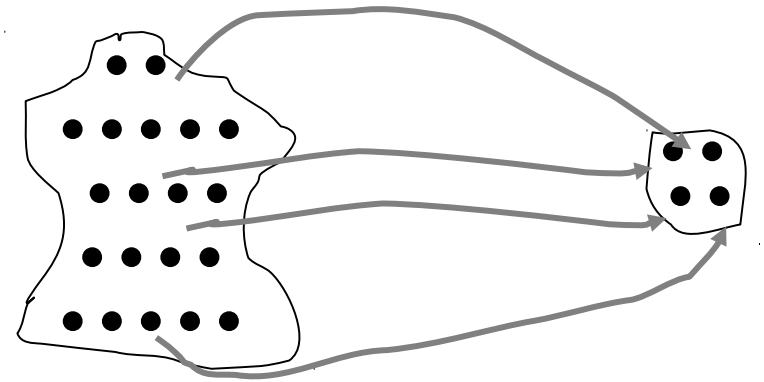
Note:

- **Each observation in a population is a value of a random variable X having some probability distribution f(x).**
- **To eliminate bias in the sampling procedure, we select a random sample in the sense that the observations are made independently and at random.**
- **The random sample of size *n* is: $X_1, X_2, \ldots, X_n$**

It consists of *n* observations selected independently and randomly from the population.

Statistics： Sample

Population: **Sample:** **independent identically distributed (iid)**

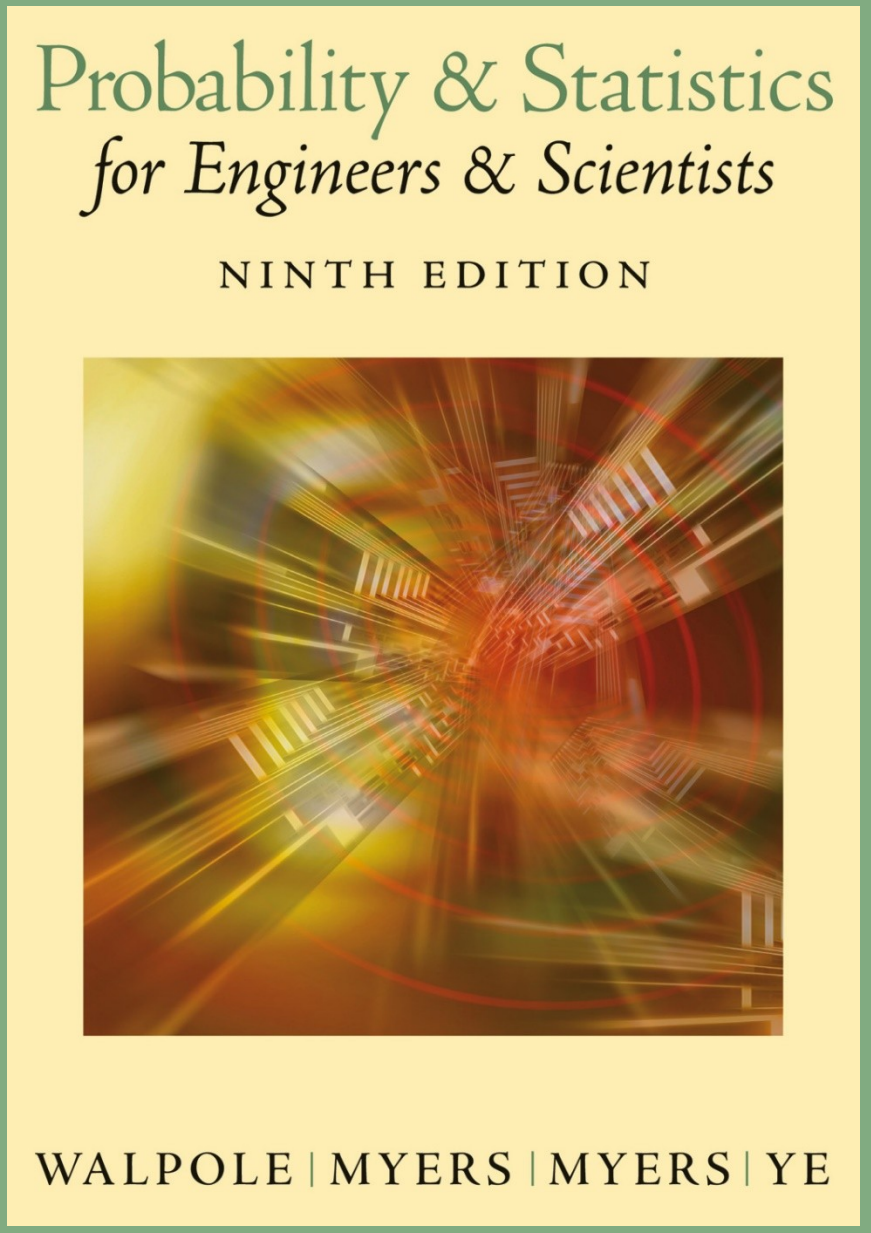


Examples:
- **production**
- **marketing research**

Sample function: **a function of the observed values in the sample used for making general conclusion about the entire population.**

Section 8.2

Some Important Statistics



PEARSON
Addison Wesley

8.2 Some Important Statistics:

Definition 8.4:

Any function of the random sample $X_1, X_2, \ldots, X_n$ is called a statistic.

Central Tendency in the Sample:

Definition 8.5:

If $X_1, X_2, \ldots, X_n$ represents a random sample of size *n*, then the sample mean is defined to be the statistic:

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{\sum_{i=1}^{n} X_i}{n} \qquad \text{(unit)}$$

Note:

- $\overline{X}$ **is a statistic because it is a function of the random sample $X_1, X_2, \ldots, X_n$.**
- $\overline{X}$ **has same unit of $X_1, X_2, \ldots, X_n$.**
- $\overline{X}$ **measures the central tendency in the sample (location).**

Variability in the Sample:

Definition 8.9:

If $X_1, X_2, \ldots, X_n$ represents a random sample of size *n*, then the sample variance is defined to be the statistic:

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} = \frac{(X_1 - \overline{X})^2 + (X_2 - \overline{X})^2 + \cdots + (X_n - \overline{X})^2}{n-1} \quad \text{(unit)}^2$$

Theorem 8.1: **(Computational Formulas for S^2)**

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

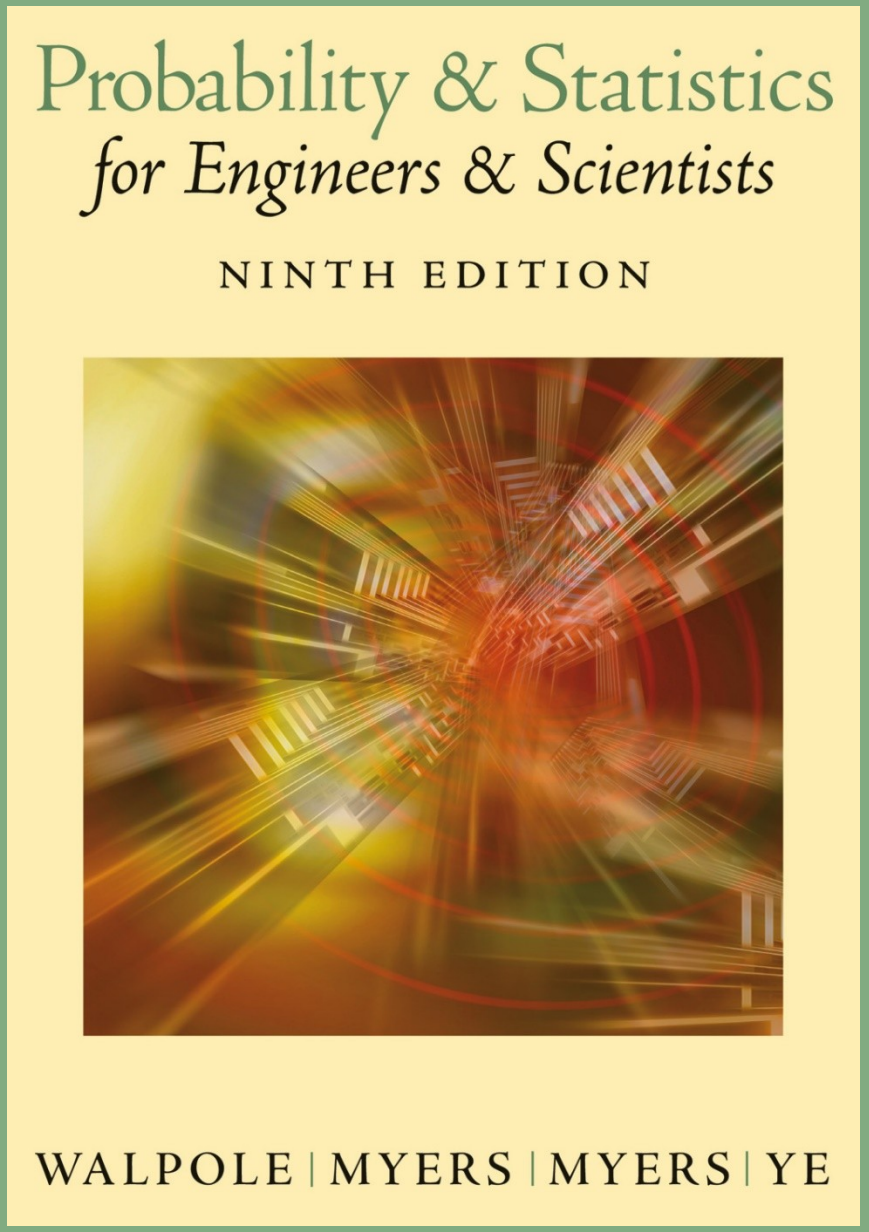
Note:

- **S^2 is a statistic because it is a function of the random sample $X_1, X_2, \ldots, X_n$.**
- **S^2 measures the variability in the sample.**

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}} \quad \text{(unit)}$$

Section 8.3

Sampling Distributions



PEARSON
Addison Wesley

8.4 Sampling distribution:

Definition 8.13:

The probability distribution of a statistic is called a sampling distribution.

- **Example: If $X_1, X_2, \ldots, X_n$ represents a random sample of size *n*, then the probability distribution of $\overline{X}$ is called the sampling distribution of the sample mean $\overline{X}$.**

8.5 Sampling Distributions of Means:

Result:

If $X_1, X_2, \ldots, X_n$ is a random sample of size *n* taken from a normal distribution with mean μ and variance σ^2, i.e. $N(\mu,\sigma)$, then the sample mean $\overline{X}$ has a normal distribution with mean

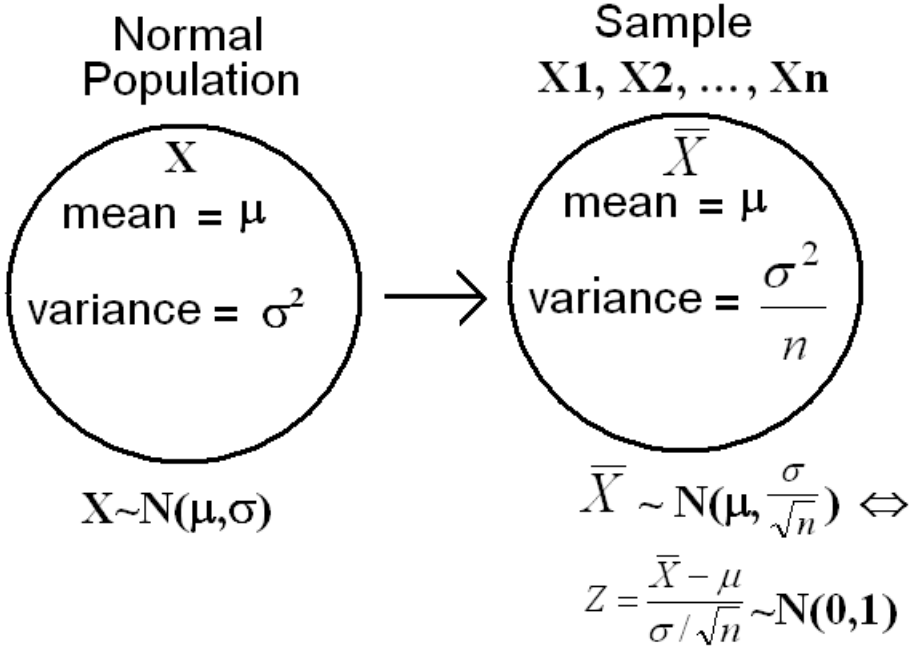
$$E(\overline{X}) = \mu_{\overline{X}} = \mu$$

and variance

$$Var(\bar{X}) = \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$$

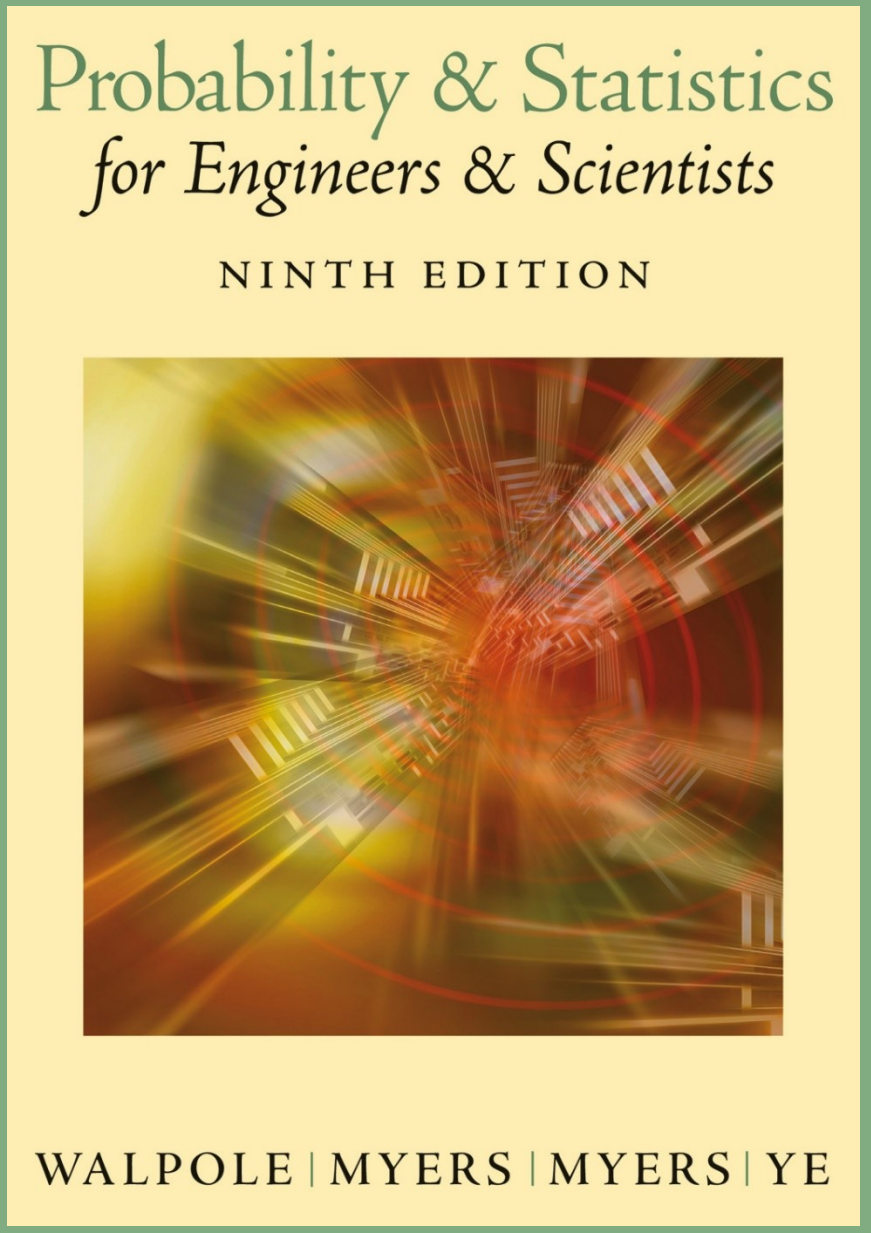
- **If $X_1, X_2, \ldots, X_n$ is a random sample of size *n* from N(μ,σ), then $\bar{X}$ ~N($\mu_{\bar{X}}$, $\sigma_{\bar{X}}$) or $\bar{X}$ ~N(μ, $\frac{\sigma}{\sqrt{n}}$).**

- $$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}}) \Leftrightarrow Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$



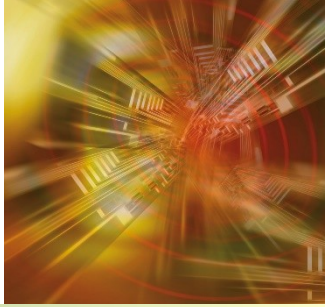
Section 8.4

Sampling Distribution of Means and the Central Limit Theorem



PEARSON Addison Wesley

Sample mean Distribution



The Central Limit Theorem (CLT):

Let $X_1, X_2, \ldots X_n$ be independent identically distributed random variables with same mean μ and same finite variance σ^2. Then the distribution of

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

will tend towards the standard normal distribution as $n \rightarrow \infty$.

How large should n be before the approximation is good?

- **Most distributions:** $n \geq 30$
- **Normal distribution :** **for all n**

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$

is approximately standard normal random variable, i.e.,

$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$ **approximately.**

- $Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1) \Leftrightarrow \overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$
  - **We consider *n* large when *n*≥30.**
- **For large sample size *n*, $\overline{X}$ has approximately a normal distribution with mean μ and variance $\frac{\sigma^2}{n}$, i.e.,**

$\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$ **approximately.**

- The sampling distribution of $\overline{X}$ is used for inferences about the population mean μ.

Example 8.13:

An electric firm manufactures light bulbs that have a length of life that is approximately normally distributed with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours.

Solution:

X= the length of life

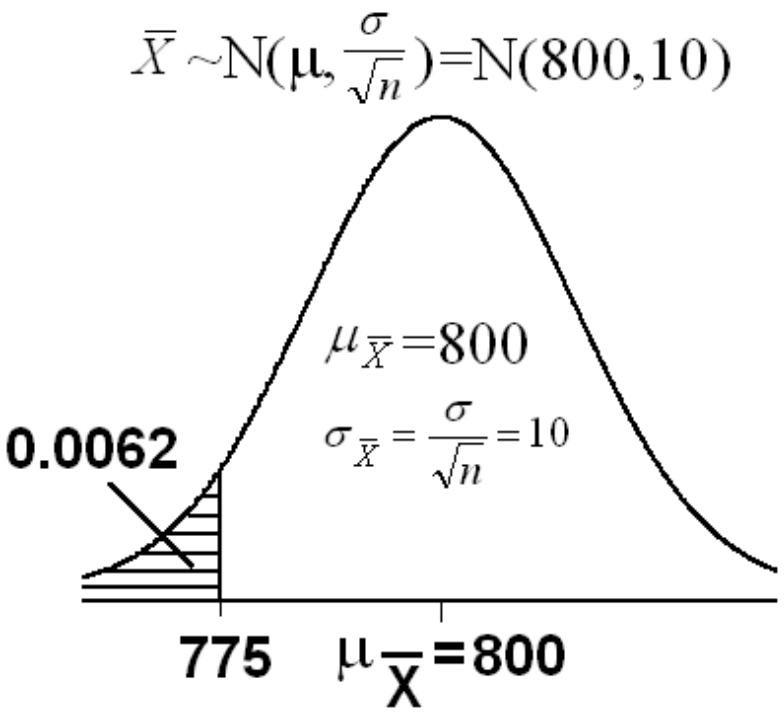
μ=800 , σ=40

X~N(800, 40)

***n*=16**

$$\mu_{\bar{X}} = \mu = 800$$

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{40}{\sqrt{16}} = 10$$



$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}}) = N(800,10)$$

$$\Leftrightarrow Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = Z = \frac{\bar{X} - 800}{10} \sim N(0,1)$$

$$= P\left[\frac{\bar{X} - 800}{10} < \frac{775 - 800}{10}\right]$$

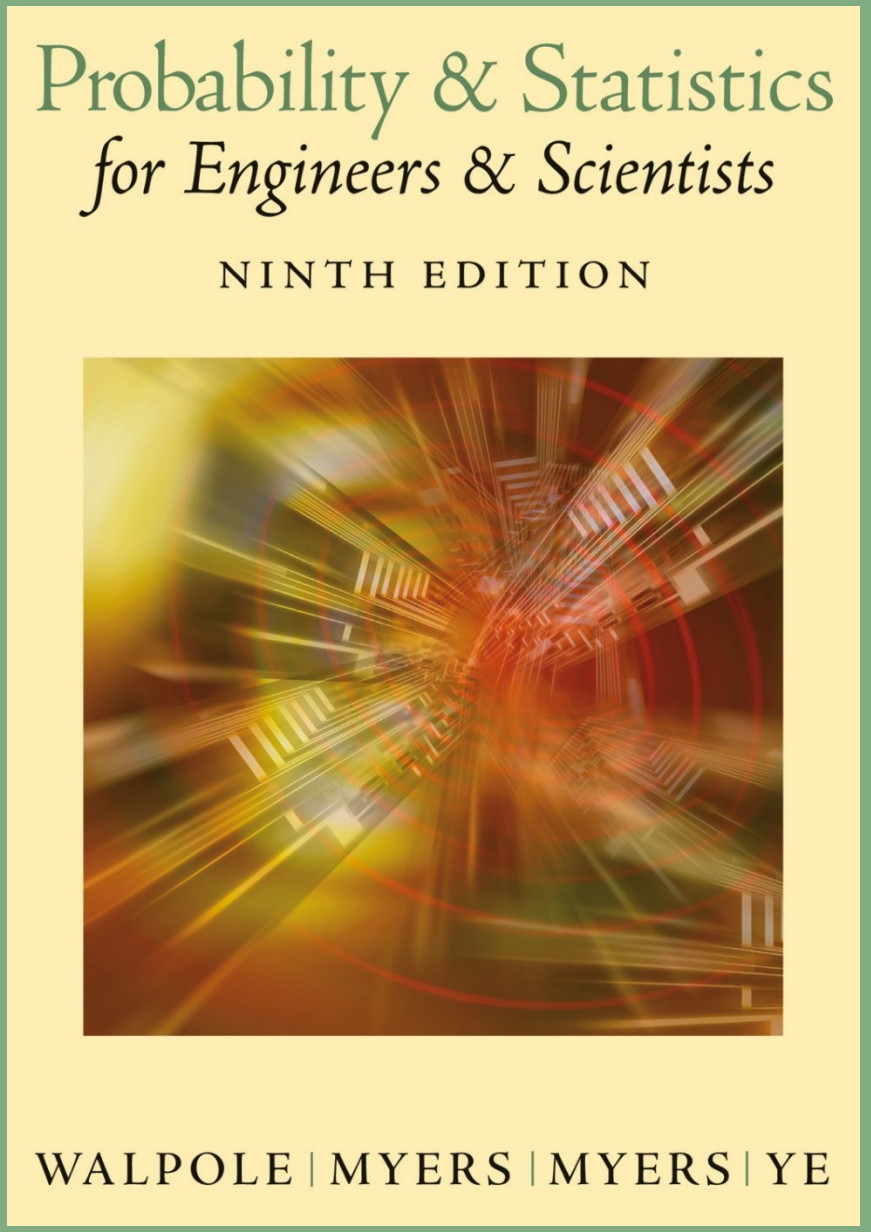
$$= P\left[Z < \frac{775 - 800}{10}\right]$$

$$= P(Z < -2.50)$$

$$= 0.0062$$

Section

Introduction to Inference

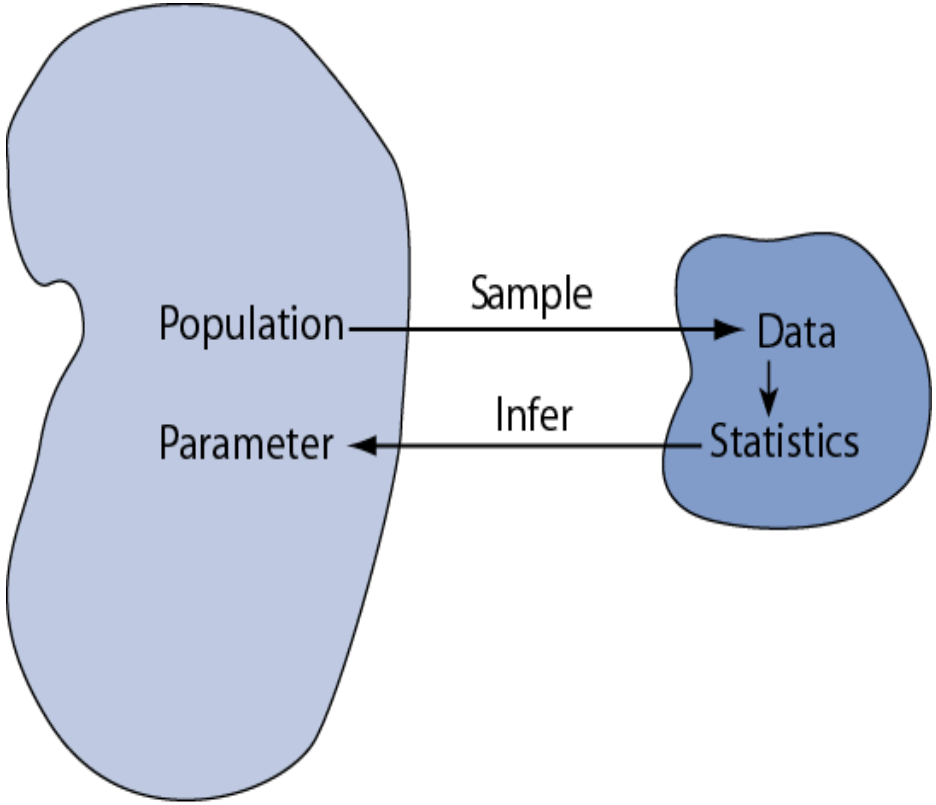


PEARSON
Addison Wesley

Statistical inference is the act of generalizing from a sample to a population with calculated degree of certainty.

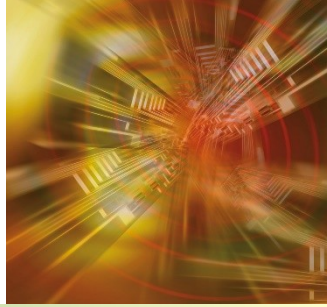
We want to learn about population

parameters…



…but we can only calculate *sample statistics*

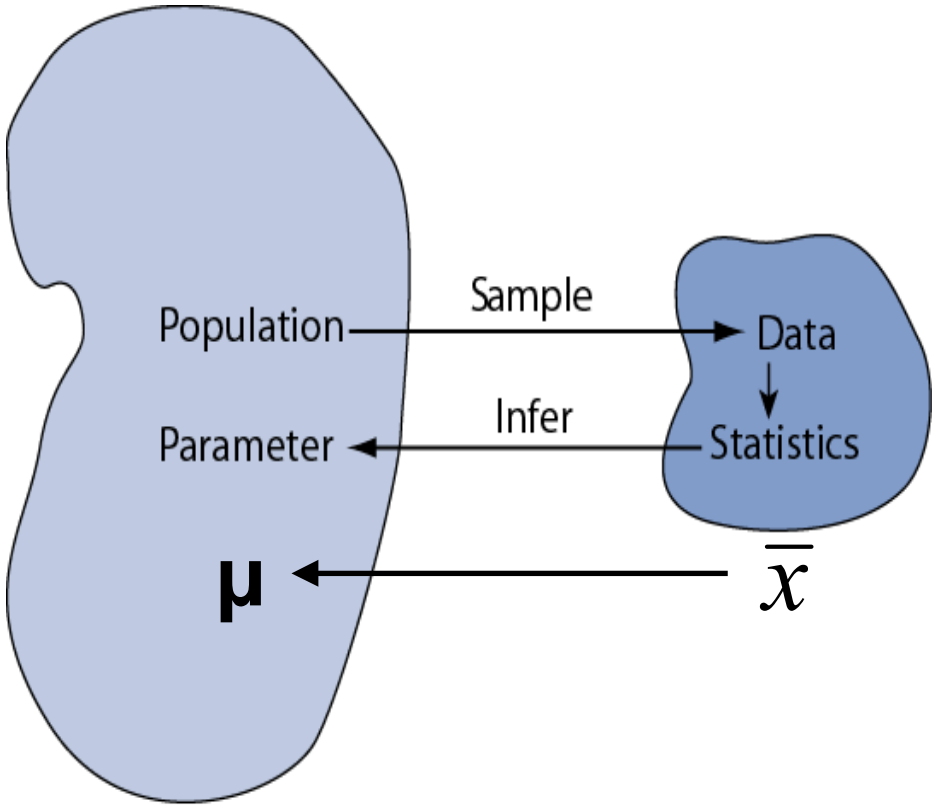
Parameters and Statistics



It is essential that we draw distinctions between parameters and statistics

	Parameters	**Statistics**
Source	Population	Sample
Calculated?	No	Yes
Constants?	Yes	No
Examples	μ, σ, p	$\bar{x}, s, \hat{p}$

We are going to illustrate inferential concept by considering how well a given sample mean "x-bar" reflects an underling population mean μ



Precision and reliability

- How **precisely** does a given sample mean (x-bar) reflect underlying population mean (μ)? How reliable are our inferences?
- To answer these questions, we consider a **simulation experiment** in which we take all possible samples of size *n* taken from the population

Simulation Experiment

- Population (Figure A)

  $N = 10{,}000$

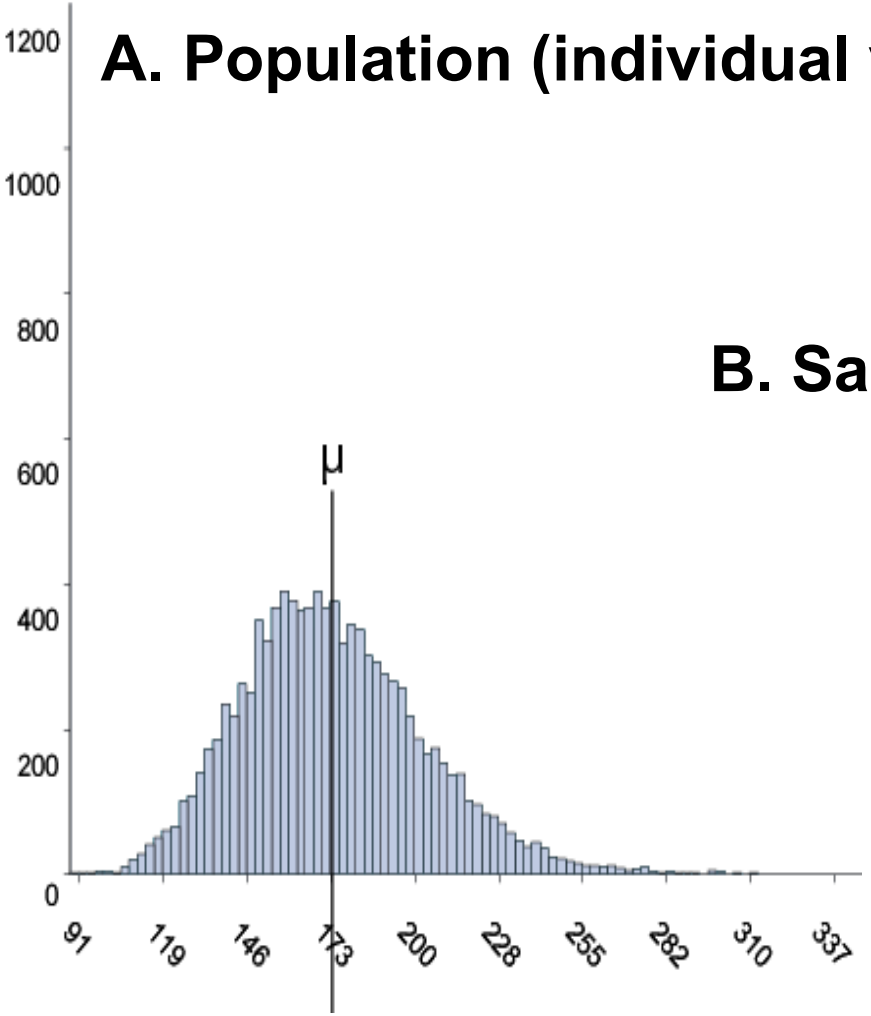
  normal shape (positive skew)

  $\mu = 173$

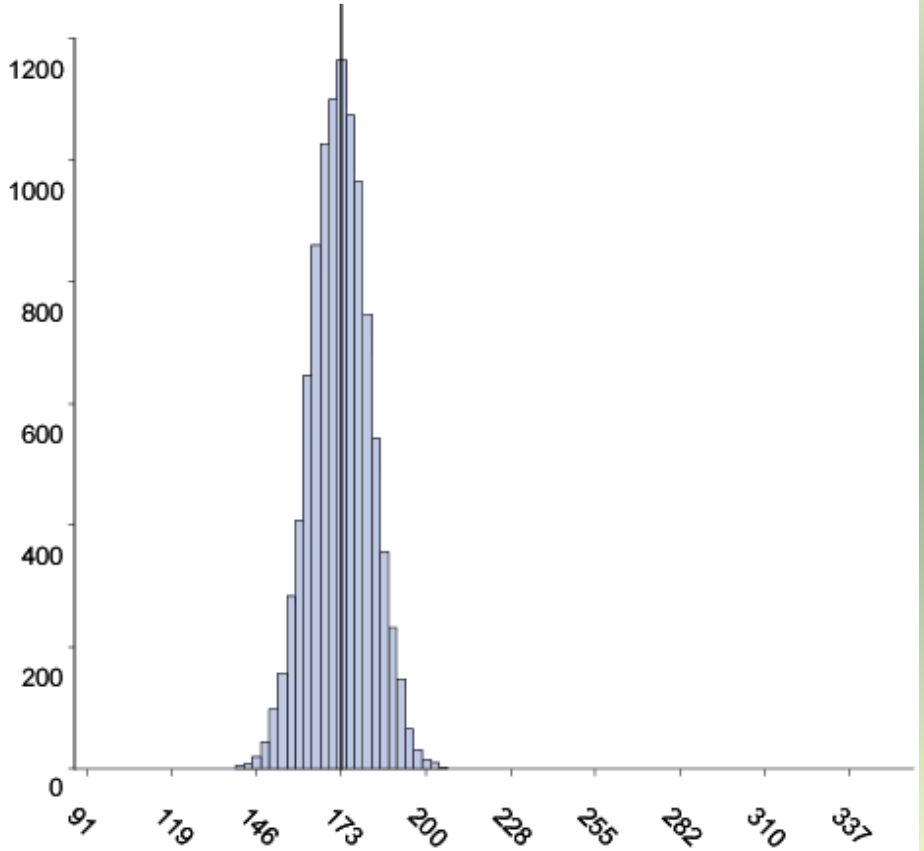
  $\sigma = 30$

- Take repeated SRSs (simple random samplings), each of $n = 10$
- Calculate x-bar in each sample
- Plot x-bars (Figure B , next slide)

A. Population (individual values)

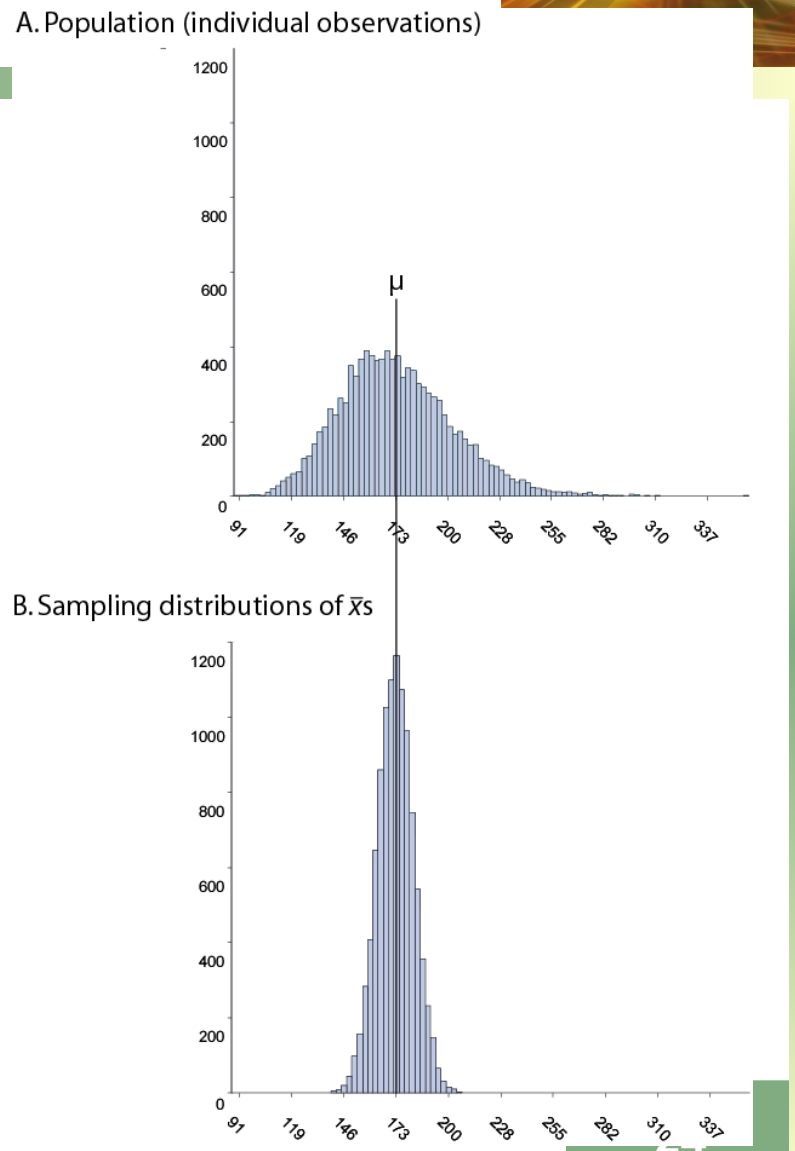


B. Sampling distribution of x-bars



Simulation Experiment Results

1. Distribution B is more Normal than distribution A ⇔ **Central Limit Theorem**
2. Both distributions centered on μ ⇔ **x-bar is unbiased estimator of μ**
3. Distribution B is skinnier than distribution A ⇔ related to **"square root law"**



Reiteration of Key Findings

- **Finding 1 (central limit theorem):** the sampling distribution of x-bar tends toward Normality even when the population is not Normal (esp. strong in large samples).
- **Finding 2 (unbiasedness):** the expected value of x-bar is μ
- **Finding 3** is related to the **square root law,** which says:

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

Standard Deviation of the Mean

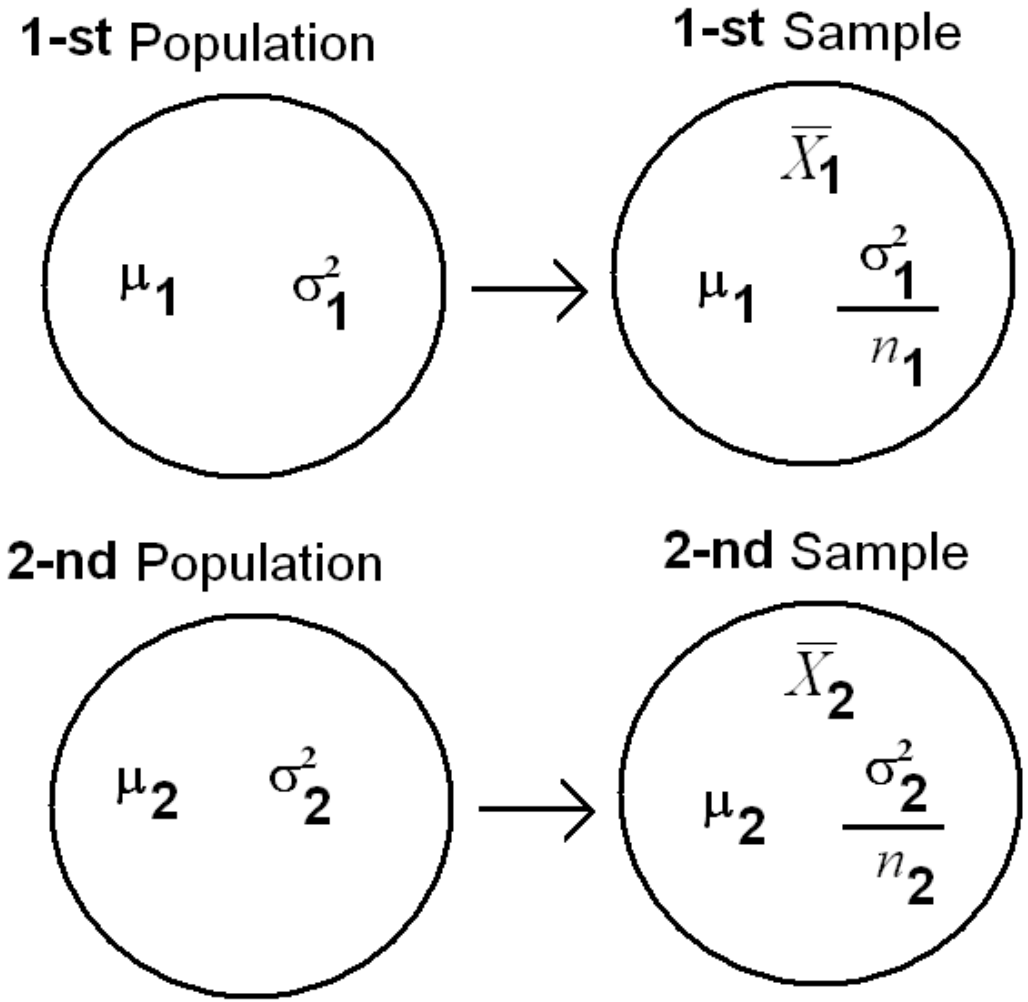
- The standard deviation of the sampling distribution of the mean has a *special name*: **standard error of the mean** (denoted σ_{xbar} or SE_{xbar})
- The square root law says:

$$\sigma_{\bar{x}} \equiv SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Sampling Distribution of the Difference between Two Means:

Suppose that we have two populations:

- **1-st population with mean μ_1 and variance σ_1^2**
- **2-nd population with mean μ_2 and variance σ_2^2**
- **We are interested in comparing μ_1 and μ_2, or equivalently, making inferences about μ_1-μ_2.**
- **We <u>independently</u> select a random sample of size n_1 from the 1-st population and another random sample of size n_2 from the 2-nd population:**
- **Let $\bar{X}_1$ be the sample mean of the 1-st sample.**
- **Let $\bar{X}_2$ be the sample mean of the 2-nd sample.**
- **The sampling distribution of $\bar{X}_1 - \bar{X}_2$ is used to make inferences about μ_1-μ_2.**


1-st Population
μ_1 σ_1^2
1-st Sample
$\overline{X}_1$
μ_1 $\frac{\sigma_1^2}{n_1}$
2-nd Population
μ_2 σ_2^2
2-nd Sample
$\overline{X}_2$
μ_2 $\frac{\sigma_2^2}{n_2}$

Theorem 8.3:

If n_1 and n_2 are large, then the sampling distribution of $\bar{X}_1 - \bar{X}_2$ approximately normal with mean

$$E(\bar{X}_1 - \bar{X}_2) = \mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$$

and variance

$$Var(\bar{X}_1 - \bar{X}_2) = \sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

that is:

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Note:

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\sigma^2_{\bar{X}_1 - \bar{X}_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \neq \sqrt{\frac{\sigma_1^2}{n_1}} + \sqrt{\frac{\sigma_2^2}{n_2}} = \frac{\sigma_1}{\sqrt{n_1}} + \frac{\sigma_2}{\sqrt{n_2}}$$

Example 8.16:

The television picture tubes of manufacturer *A* have a mean lifetime of 6.5 years and standard deviation of 0.9 year, while those of manufacturer *B* have a mean lifetime of 6 years and standard deviation of 0.8 year. What is the probability that a random sample of 36 tubes from manufacturer *A* will have a mean lifetime that is at least 1 year more than the mean lifetime of a random sample of 49 tubes from manufacturer *B*?

Solution:

Population A	Population B
μ_1=6.5	μ_2=6.0
σ_1=0.9	σ_2=0.8
n_1=36	n_2=49

- We need to find the probability that the mean lifetime of manufacturer *A* is at least 1 year more than the mean lifetime of manufacturer *B* which is P($\bar{X}_1 \geq \bar{X}_2 + 1$).
- The sampling distribution of $\bar{X}_1 - \bar{X}_2$ is

- We need to find the probability that the mean lifetime of manufacturer *A* is at least 1 year more than the mean lifetime of manufacturer *B* which is P($\bar{X}_1 \geq \bar{X}_2 + 1$).
- The sampling distribution of $\bar{X}_1 - \bar{X}_2$ is

$$\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}})$$

$$E(\bar{X}_1 - \bar{X}_2) = \mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2 = 6.5 - 6.0 = 0.5$$

$$Var(\bar{X}_1 - \bar{X}_2) = \sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} = \frac{(0.9)^2}{36} + \frac{(0.8)^2}{49} = 0.03556$$

$$\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{0.03556} = 0.189$$

$$\bar{X}_1 - \bar{X}_2 \sim N(0.5, 0.189)$$

Recall

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

$$= P(\bar{X}_1 \geq \bar{X}_2 + 1) = P(\bar{X}_1 - \bar{X}_2 \geq 1)$$

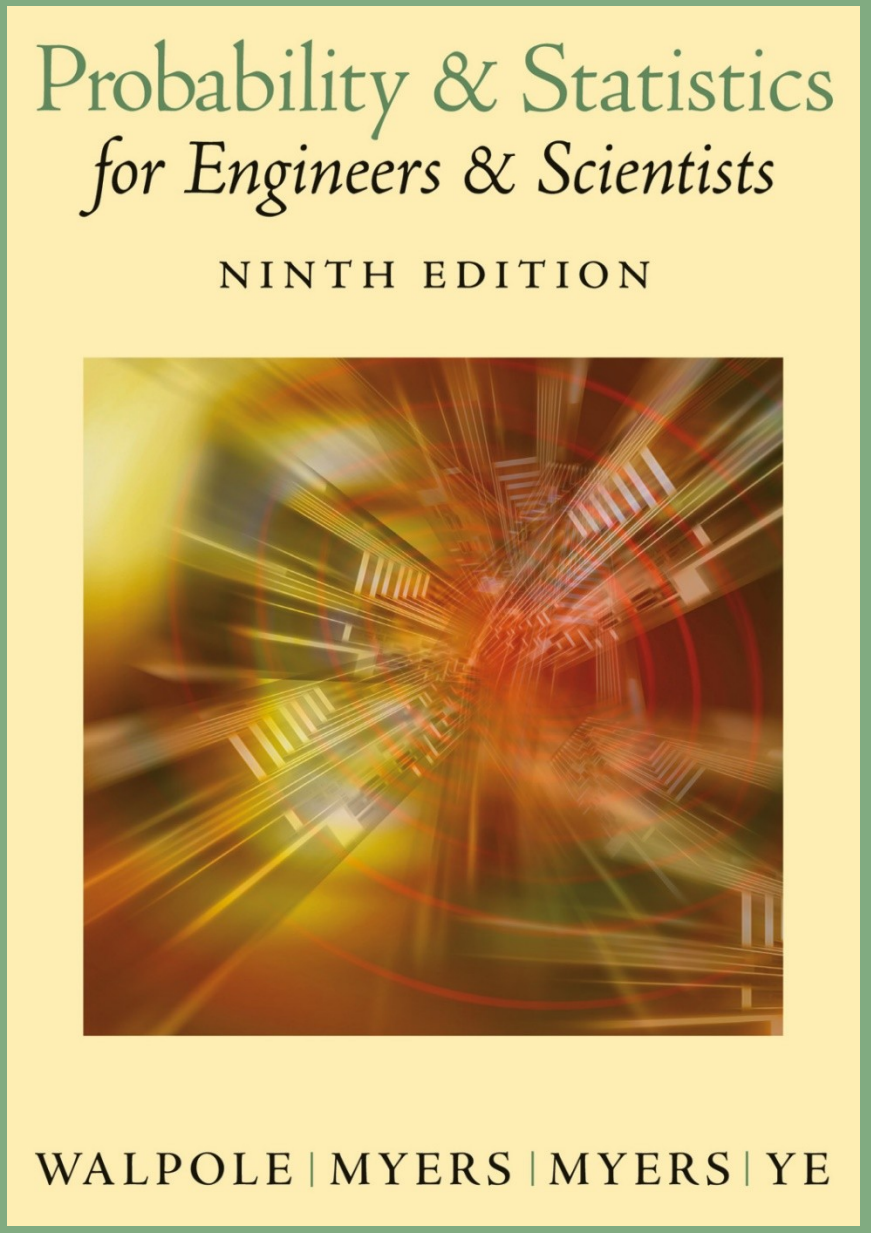
$$= P\left[\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \geq \frac{1 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \right]$$

$$= P\left[Z \geq \frac{1 - 0.5}{0.189} \right]$$

$= P(Z \geq 2.65) = 1 - P(Z < 2.65) = 1 - 0.9960 = 0.0040$

Section 8.5

Sampling Distribution of S^2



PEARSON Addison Wesley

Sample variance
Distribution

Theorem:

Lad $X_1, X_2, \ldots X_n$ be independent normal distributed random variables with mean μ and variance σ^2.

Then

$$\frac{(n-1)\,S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \overline{X})^2 \sim \chi^2(n-1)$$

When calculating S^2, it's usually more convenient to use

$$S^2 = \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n}X_i^2 - \left[\sum_{i=1}^{n}X_i\right]^2\right]$$

Proof

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\mu)^2 &= \sum_{i=1}^{n}[(X_i-\bar{X})+(\bar{X}-\mu)]^2\\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2+\sum_{i=1}^{n}(\bar{X}-\mu)^2+2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\bar{X})\\ &= \sum_{i=1}^{n}(X_i-\bar{X})^2+n(\bar{X}-\mu)^2.\end{aligned}$$

Dividing each term of the equality by σ^2

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2=\frac{(n-1)S^2}{\sigma^2}+\frac{(\bar{X}-\mu)^2}{\sigma^2/n}.$$

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

Theorem 8.4

If S^2 is the variance of a random sample of size n taken from a normal population having the variance σ^2, then the statistic

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

has a chi-squared distribution with $v = n - 1$ degrees of freedom.

Sample variance Example

Problem: Car batteries
A producer of car batteries claims that the life time of their batteries are normal distributed with

mean: $\mu = 3$ **year**
standard deviation: $\sigma = 1$ **year**

Sample of 5 batteries: 1.9 2.4 3.0 3.5 4.2

(a)Calculate sample variance.
(b)Should the manufacturer still be convinced that the batteries have a standard deviation of 1 year?

Solution

Step 1: Find out the sample variance

$$s^2 = \frac{(5)(48.26) - (15)^2}{(5)(4)} = 0.815.$$

Step 2: Calculate the statistic X^2

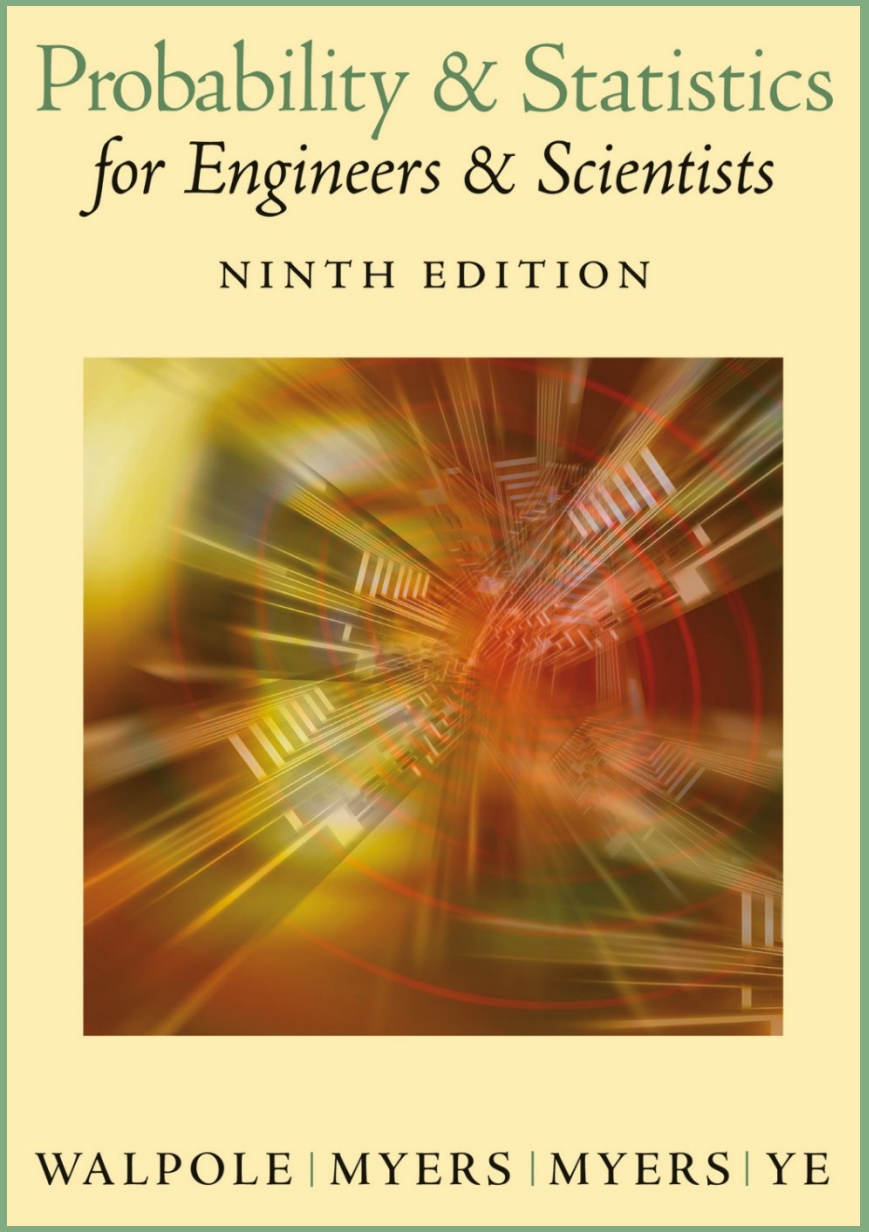
$$\chi^2 = \frac{(4)(0.815)}{1} = 3.26$$

Step 3: Compare the calculated statistic X^2 with critical values

Since 95%of the X^2 values with 4 degrees of freedom fall between 0.484 and 11.143, the computed value with $\sigma^2 = 1$ is reasonable, and therefore the manufacturer has no reason to suspect that the standard deviation is other than 1 year.

Section 8.6

t-Distribution



PEARSON
Addison Wesley

8.7 t-Distribution:

- Recall that, if $X_1, X_2, \ldots, X_n$ is a random sample of size *n* from a normal distribution with mean μ and variance σ^2, i.e. $N(\mu,\sigma)$, then

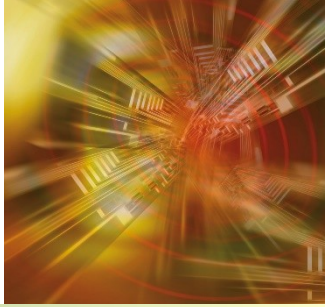
$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

- We can apply this result only when σ^2 is known!

If σ^2 is unknown, we replace the population variance σ^2 with the sample variance $S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$ · to have the following statistic

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

Sample mean
Distribution (unknown varinace)



Typically the variance σ^2 is unknown.
If we replace the unknown variance by s^2 we obtain:

Theorem:
Lad $X_1, X_2, \ldots X_n$ be independent normal distributed random variables with mean μ and variance σ^2 (unknown).

Then

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

Result:

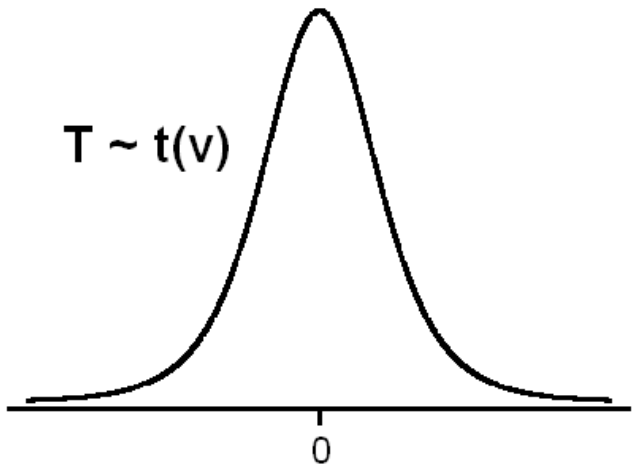
If $X_1, X_2, \ldots, X_n$ is a random sample of size *n* from a normal distribution with mean μ and variance σ^2, i.e. $N(\mu,\sigma)$, then the statistic

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

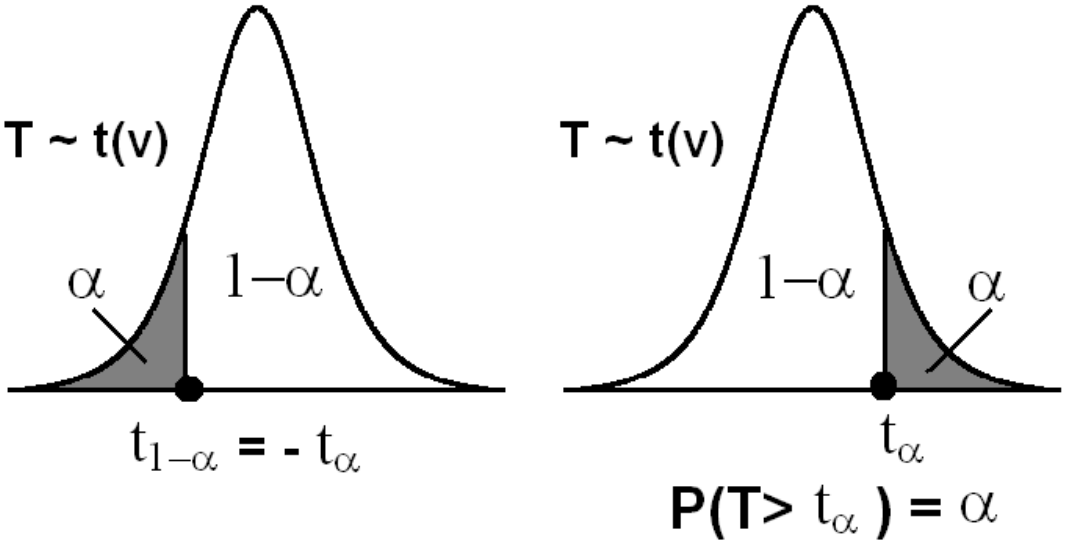
has a t-distribution with ν=*n*-1degrees of freedom (df), and we write T~ t(ν).

Note:

- **t-distribution is a continuous distribution.**
- **The shape of t-distribution is similar to the shape of the standard normal distribution.**



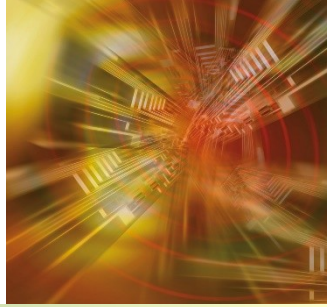
Notation:



- t_α = The t-value above which we find an area equal to α, that is $P(T > t_\alpha) = \alpha$
- Since the curve of the pdf of T~ t(ν) is symmetric about 0, we have

$$t_{1-\alpha} = -t_\alpha$$

- Values of t_α are tabulated in Table A-4 (p.683).



Example:

Find the t-value with ν=14 (df) that leaves an area of:

(a) 0.95 to the left.

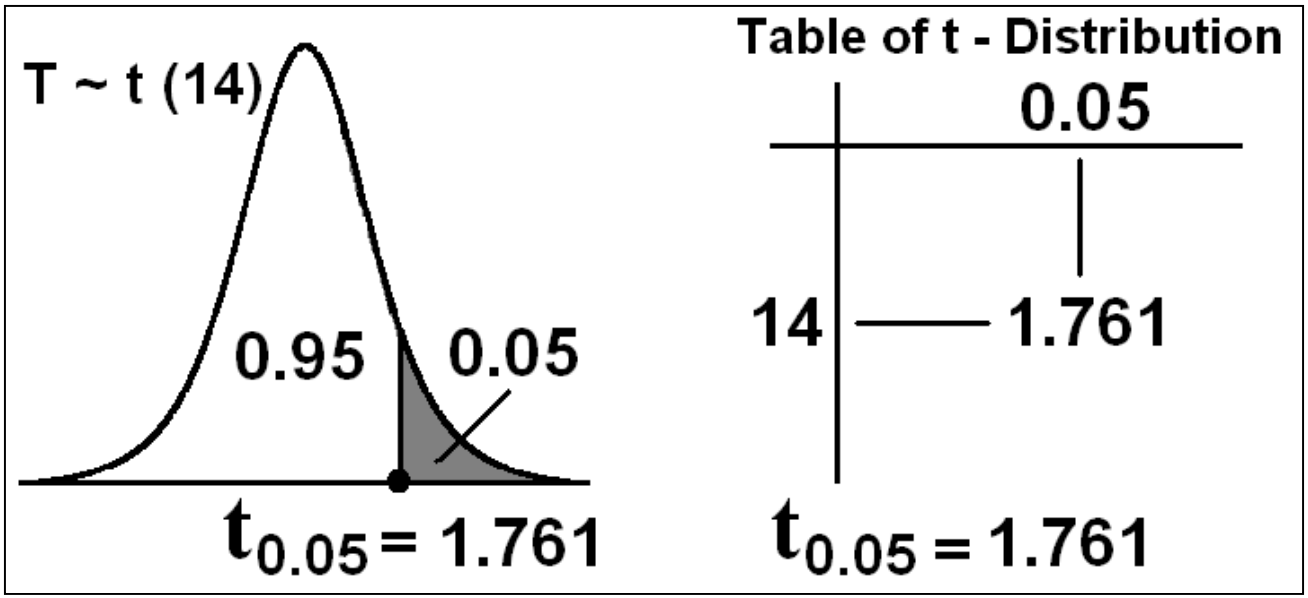
(b) 0.95 to the right.

Solution:

$\nu = 14$ (df); $T \sim t(14)$

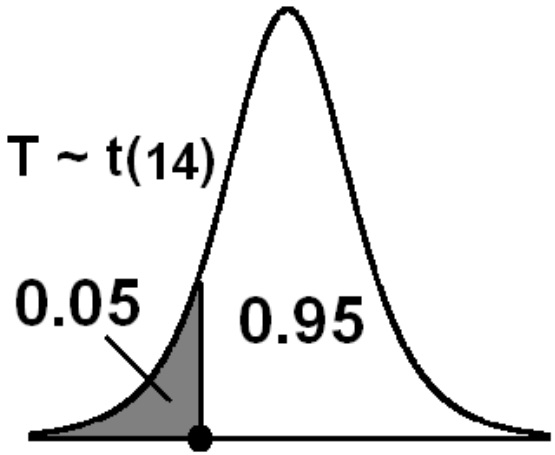
(a) The t-value that leaves an area of 0.95 to the left is

$t_{0.05} = 1.761$



(b) The t-value that leaves an area of 0.95 to the right is

$$t_{0.95} = -\,t_{1-0.95} = -\,t_{0.05} = -\,1.761$$



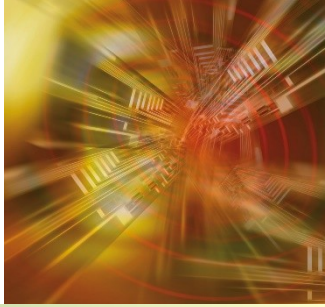
$$t_{0.95} = -\,t_{0.05}$$

$$= -1.761$$

Table of t - Distribution

	0.05
14	1.761

$$t_{0.05} = 1.761$$



Example:

For $\nu = 10$ degrees of freedom (df), find $t_{0.10}$ and $t_{0.85}$.

Solution:

$t_{0.10} = 1.372$

$t_{0.85} = - t_{1-0.85} = -_{t\ 0.15} = -_{1.093}$ $(t_{0.15} = 1.093)$

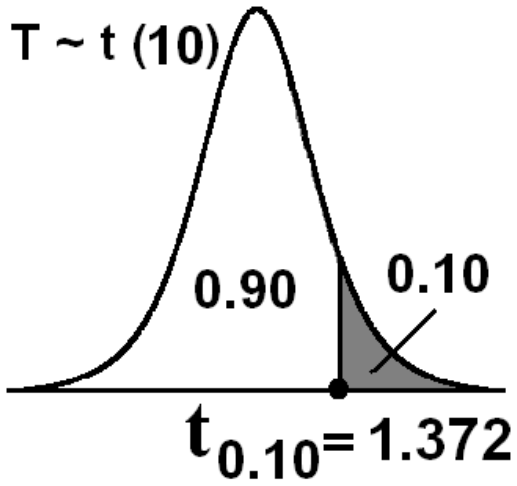
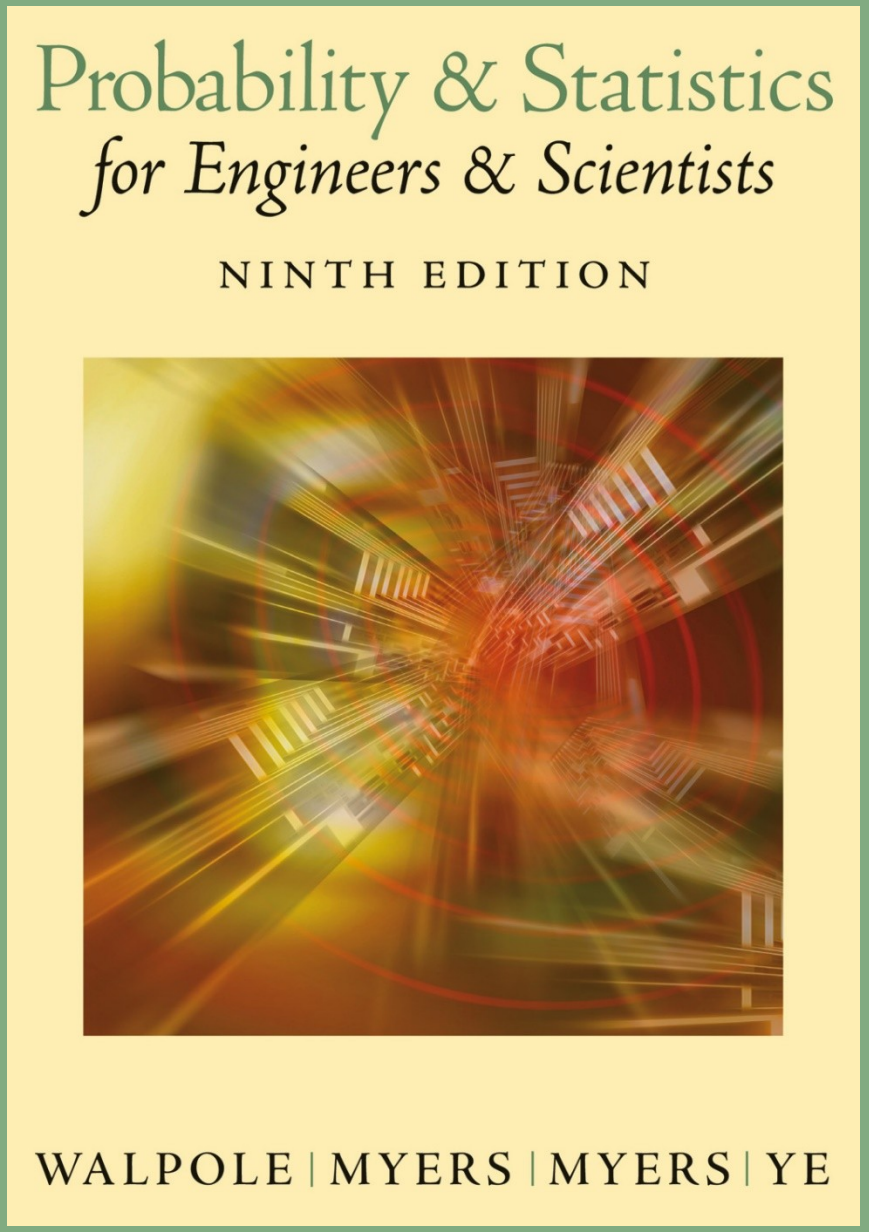


Table of t - Distribution

	0.15	0.10
10	1.093	1.372

Section 8.7

F-Distribution



PEARSON
Addison Wesley

Two sample variances
Comparison

Theorem:

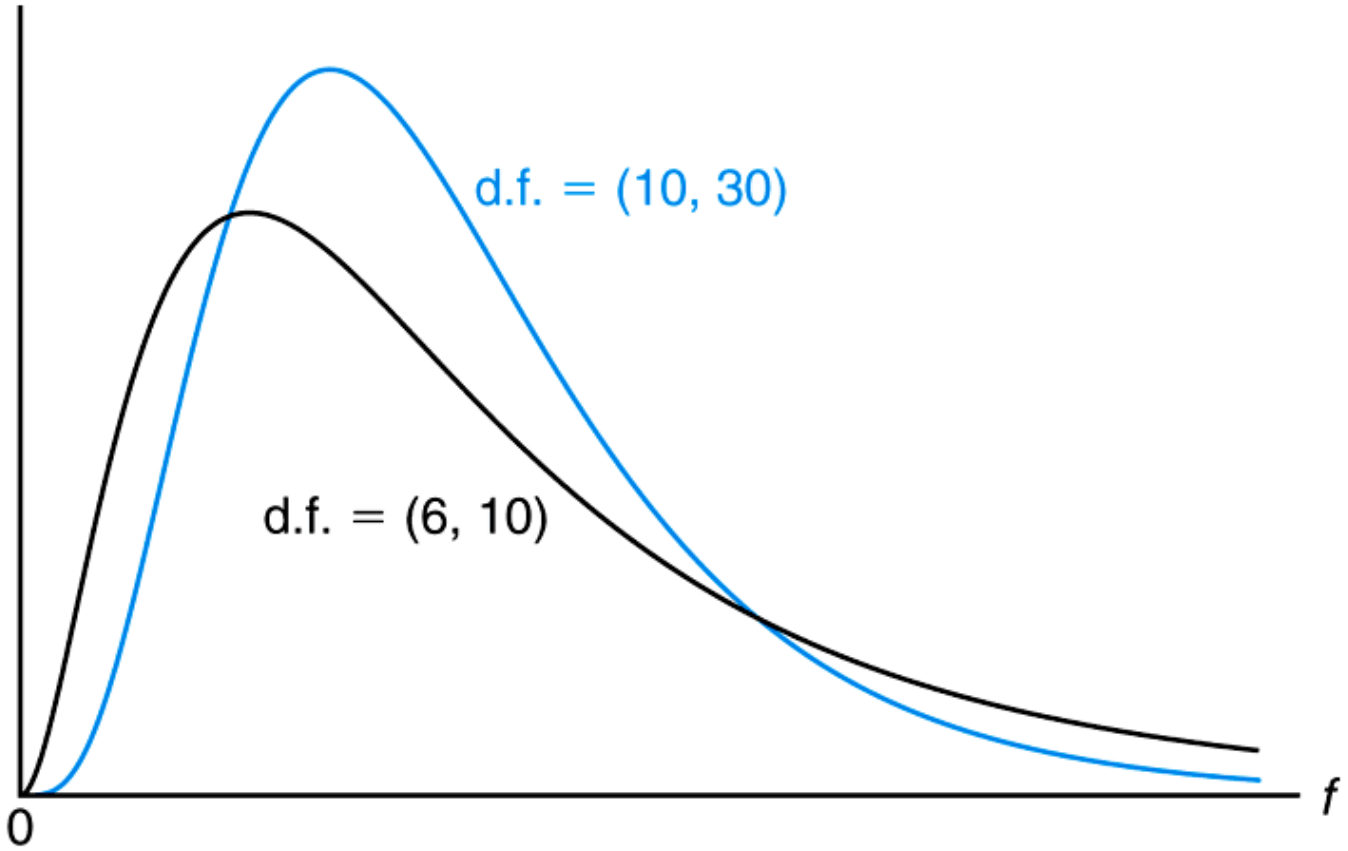
If two independent samples are taken from two normal populations with variances σ_1^2 and σ_2^2, respectively, then

$$\frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} \sim F(n_1 - 1, n_2 - 1)$$

Notice!! $f_{1-\alpha}(n_1, n_2) = \dfrac{1}{f_{\alpha}(n_2, n_1)}$

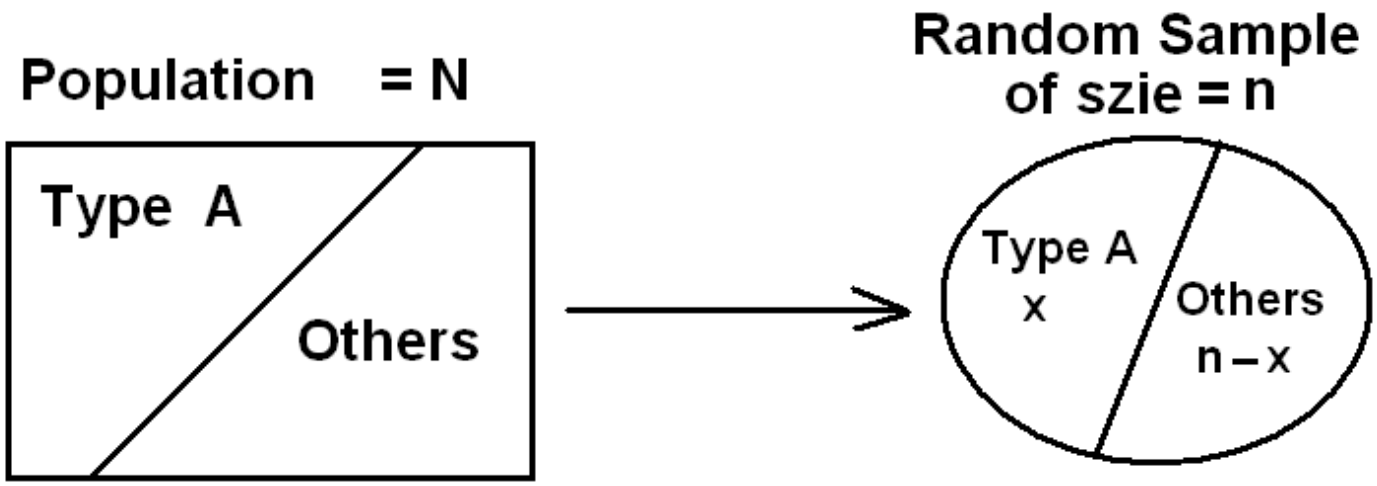
Eg. $f_{0.95}(6, 10) = \dfrac{1}{f_{0.05}(10, 6)}$

Figure 8.11 Typical *F*-distributions



Sampling Distribution of the Sample Proportion:

Suppose that the size of a population is *N*. Each element of the population can be classified as type *A* or non-type *A*. Let *p* be the proportion of elements of type *A* in the population. A random sample of size *n* is drawn from this population. Let $\hat{p}$ be the proportion of elements of type *A* in the sample.

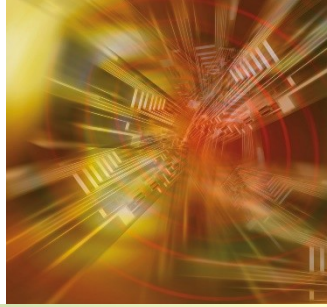


Let X = no. of elements of type *A* in the sample
***p* =Population Proportion**

$$= \frac{\text{no. of elements of type A in the population}}{N}$$

$\hat{p}$ = Sample Proportion

$$= \frac{\text{no. of elements of type A in the sample}}{n} = \frac{X}{n}$$



Result:

(1) $X \sim \text{Binomial}(n, p)$

(2) $E(\hat{p}) = E\left(\frac{X}{n}\right) = p$

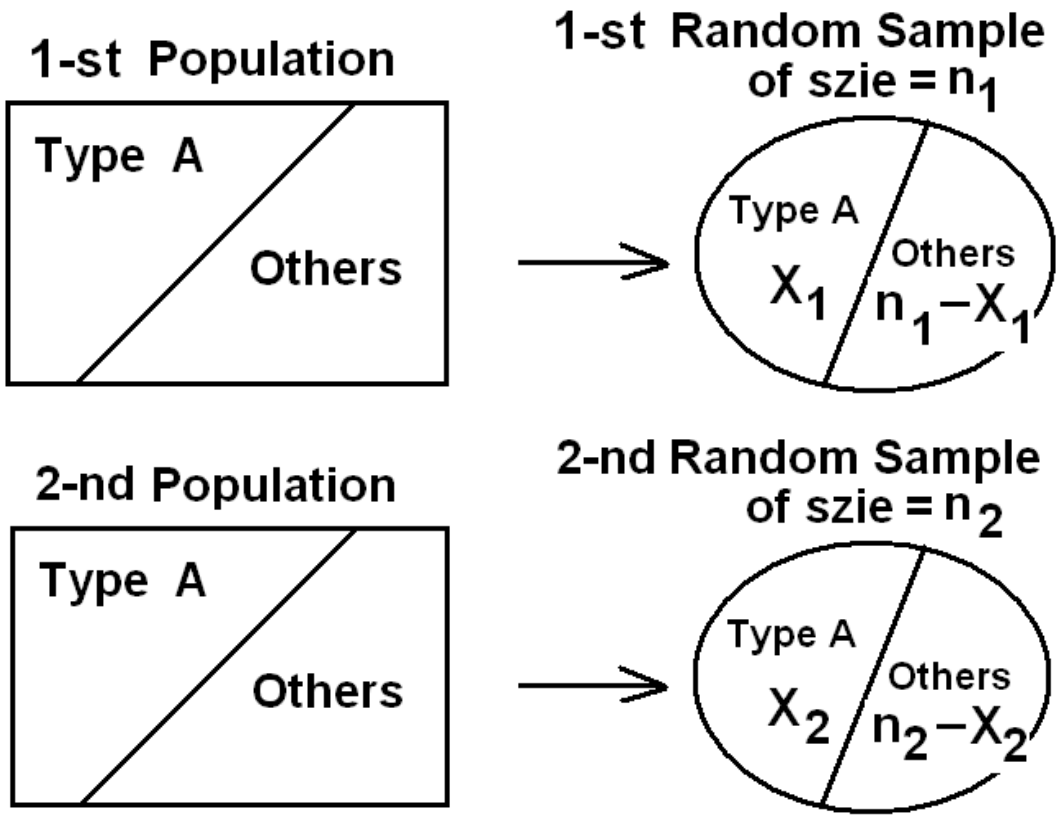
(3) $\text{Var}(\hat{p}) = \text{Var}\left(\frac{X}{n}\right) = \frac{pq}{n}$; $q = 1 - p$

(4) For large n, we have

$$\hat{p} \sim N\left(p, \sqrt{\frac{pq}{n}}\right) \quad \text{(Approximately)}$$

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1) \quad \text{(Approximately)}$$

Sampling Distribution of the Difference between Two Proportions:



Suppose that we have two populations:

- p_1 = proportion of the 1-st population.
- P_2 = proportion of the 2-nd population.
- We are interested in comparing p_1 and p_2, or equivalently, making inferences about p_1- p_2.
- We <u>independently</u> select a random sample of size n_1 from the 1-st population and another random sample of size n_2 from the 2-nd population:
- Let X_1 = no. of elements of type *A* in the 1-st sample.
- Let X_2 = no. of elements of type *A* in the 2-nd sample.
- $\hat{p}_1 = \frac{X_1}{n_1}$ = proportion of the 1-st sample
- $\hat{p}_2 = \frac{X_2}{n_2}$ = proportion of the 2-nd sample

- The sampling distribution of $\hat{p}_1 - \hat{p}_2$ is used to make inferences about p_1- p_2.

Result:

(1) $E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$

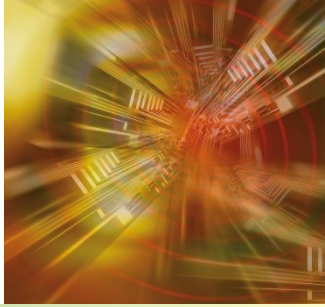
(2) $Var(\hat{p}_1 - \hat{p}_2) = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$; $q_1 = 1 - p_1, q_2 = 1 - p_2$

(3) **For large n_1 and n_2, we have**

$$\hat{p}_1 - \hat{p}_2 \sim N\left(p_1 - p_2, \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right)$$ **(Approximately)**

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \sim N(0,1)$$ **(Approximately)**

Example



In a certain area of a large city it is hypothesized that 40 percent of the houses are in a dilapidated condition. A random sample of 75 houses from this section and 90 houses from another section yielded a difference, $\hat{p}_1 - \hat{p}_2$, of .09. If there is no difference between the two areas in the proportion of dilapidated houses, what is the probability of observing a difference this large or larger?

Solution

1) Write the given information

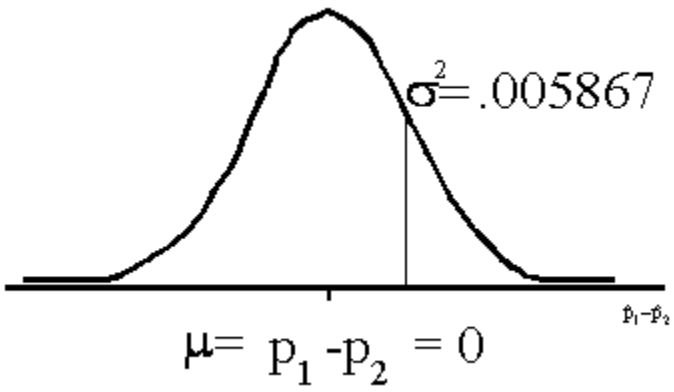
n1 = 75, p1 = .40

n2 = 90, p2 = .40

difference between two sample proportions = .09

Find P(p1 - p2 greater than or equal to .09)

(2) Sketch a normal curve



Solution

(3) Find the z score

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

$$z = \frac{(.09) - (0)}{\sqrt{\frac{.4(.6)}{75} + \frac{.4(.6)}{90}}}$$

$$z = 1.17$$

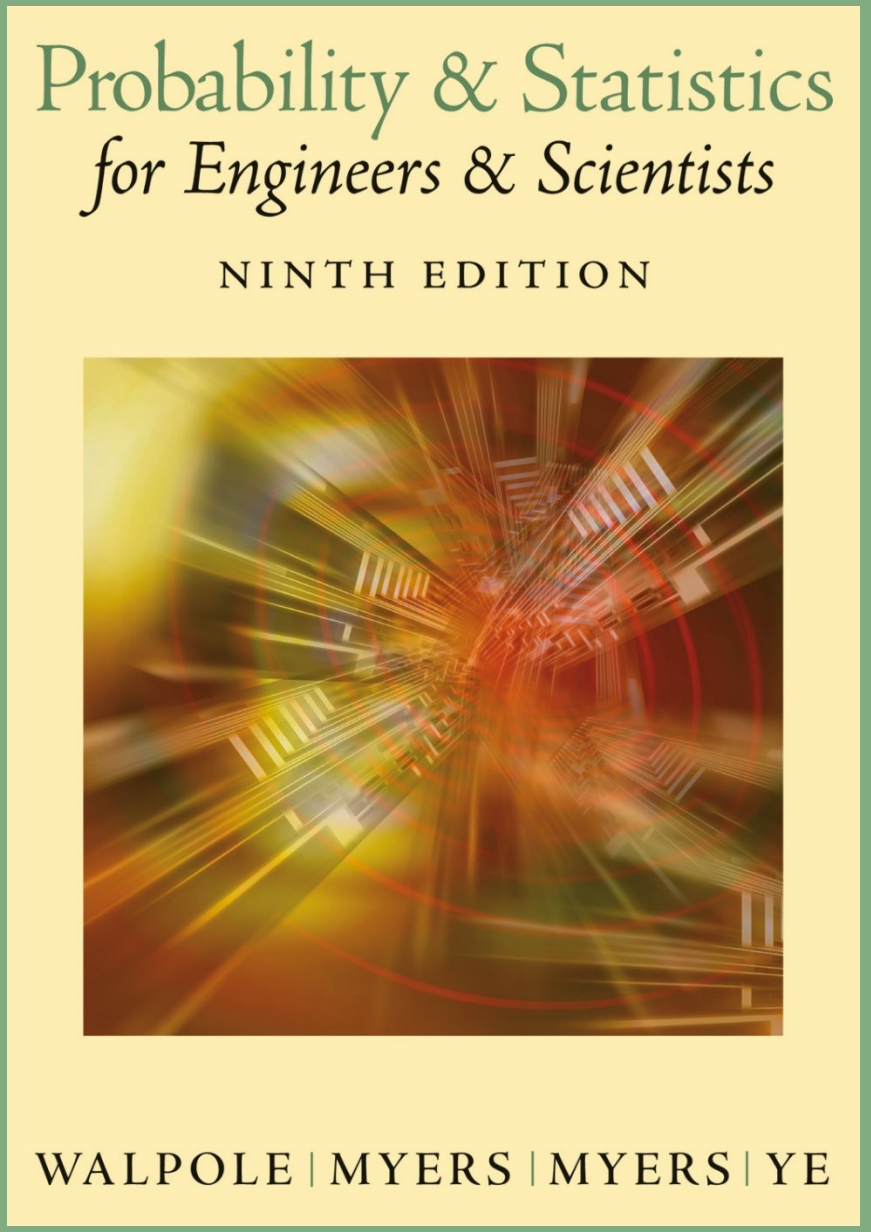
4) Find the appropriate value(s) in the table

A value of z = 1.17 gives an area of .8790 which is subtracted from 1 to give the probability

P (z > 1.17) = .121

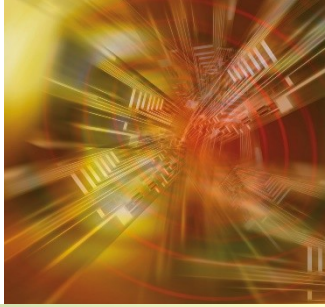
Section 8.8

Descriptive data



PEARSON
Addison Wesley

The Five-Number Summary (FNS)



- Minimum
- Quartile One (Q1)
- Median (Q2)
- Quartile Three (Q3)
- Maximum

The Five-Number Summary (FNS)

- Median

1	0.6
2	1.2
3	1.5
4	1.6
5	1.9
6	2.1
7	2.3
8	2.3
9	2.5
10	2.8
11	2.9
12	3.3
13	3.4
14	3.6
15	3.7
16	3.8
17	3.9
18	4.1
19	4.2
20	4.5
21	4.7
22	4.9
23	5.3
24	5.6
25	**6.1**

1) Sort observations from smallest to largest.
***n* = number of observations**
2) The location of the median is (*n* + 1)/2 in the sorted list

---

← If *n* is odd, the median is the value of the center observation

***n* = 25**
(*n*+1)/2 = 13
Median = 3.4

If *n* is even, the median is the mean of the two center observations →

(*n+1*)/2 = 12.5
Median = (3.3+3.4)/2 = 3.35

1	0.6
2	1.2
3	1.5
4	1.6
5	1.9
6	2.1
7	2.3
8	2.3
9	2.5
10	2.8
11	2.9
12	3.3
13	3.4
14	3.6
15	3.7
16	3.8
17	3.9
18	4.1
19	4.2
20	4.5
21	4.7
22	4.9
23	5.3
24	5.6

The Five-Number Summary (FNS)

- The median, Q1 and Q3 are **robust or resistant** to outliers.

- **NOTE:** Different technology platforms may use slightly different definitions for the quartiles.

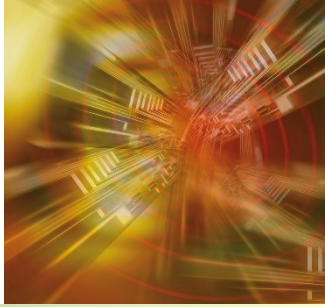
1	0.6
2	1.2
3	1.5
4	1.6
5	1.9
6	2.1
7	2.3
8	2.3
9	2.5
10	2.8
11	2.9
12	3.3
13	3.4
14	3.6
15	3.7
16	3.8
17	3.9
18	4.1
19	4.2
20	4.5
21	4.7
22	4.9
23	5.3
24	5.6
25	6.1

Q_1= first quartile = 2.2

***M* = median = 3.4**

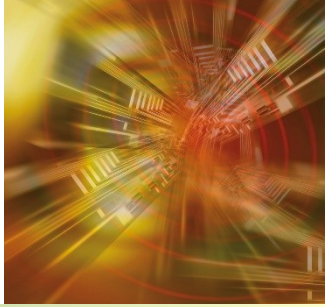
Q_3= third quartile = 4.35

Box Plots(Box-and-Whisker Plots)



- Purpose: To give a quick display of some important features of the data.
- Note: The box plot represents a distillation of the data.
- The box plot loses some of the information in the data.
- However, under several very reasonable assumptions, the information lost is of little or no value.

Let's Talk about Outliers

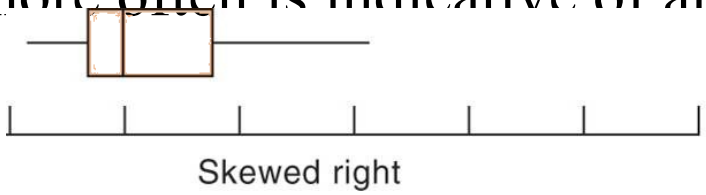


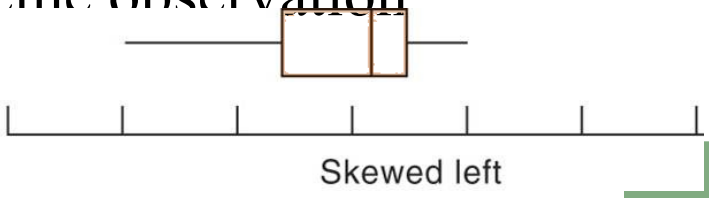
Any time you see outliers, you need to ask why the outlier is there! It could be any of the following:

- A typo or transcription error (like accidentally shifting a decimal point)
- A result of the measuring apparatus malfunctioning or changing with respect to time (like taking the operating temperature of an engine that isn't warmed up yet)
- No explanation is available, in which case we have to assume the value is valid but rare.

Boxplot Interpretation

- A boxplot can indicate the shape and spread of the distribution just as a stemplot can.
- A symmetric distribution has quartiles that are approximately equidistant from the median.
- One long quartile probably means that the distribution is skewed in some manner.
- A longer whisker can sometimes indicate the same thing but more often is indicative of an extreme observation





Boxplot

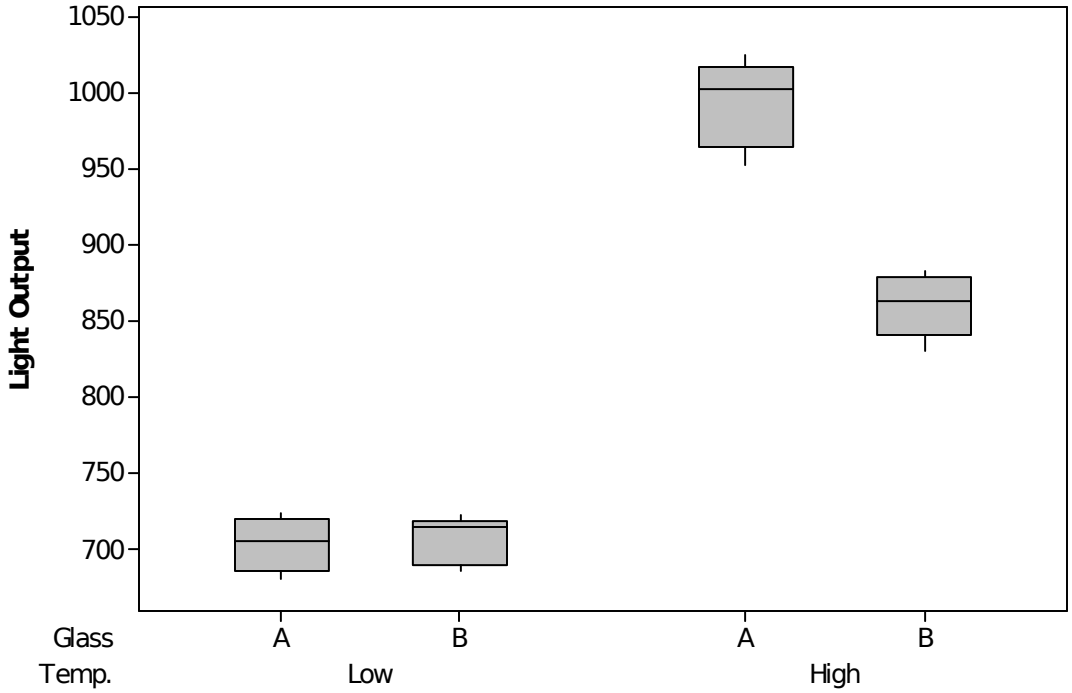
- Box Plots can also be used to analyze designed experiments. When there are categorical factors, the design can be analyzed using parallel box plots.

- Example: Consider an experiment to study the influence of operating temperature and glass type on light output.

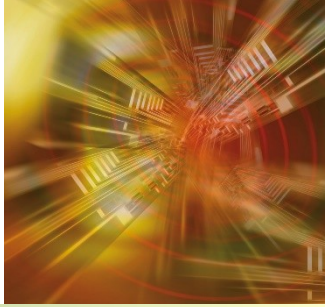
	Temp.	
Glass Type	Low	High
A	550	1380
	565	1365
	540	1384
	575	1374
	584	1379
B	545	880
	582	891
	576	864
	553	875
	574	883

Boxplot

- The resulting box plot is given below.



Histograms

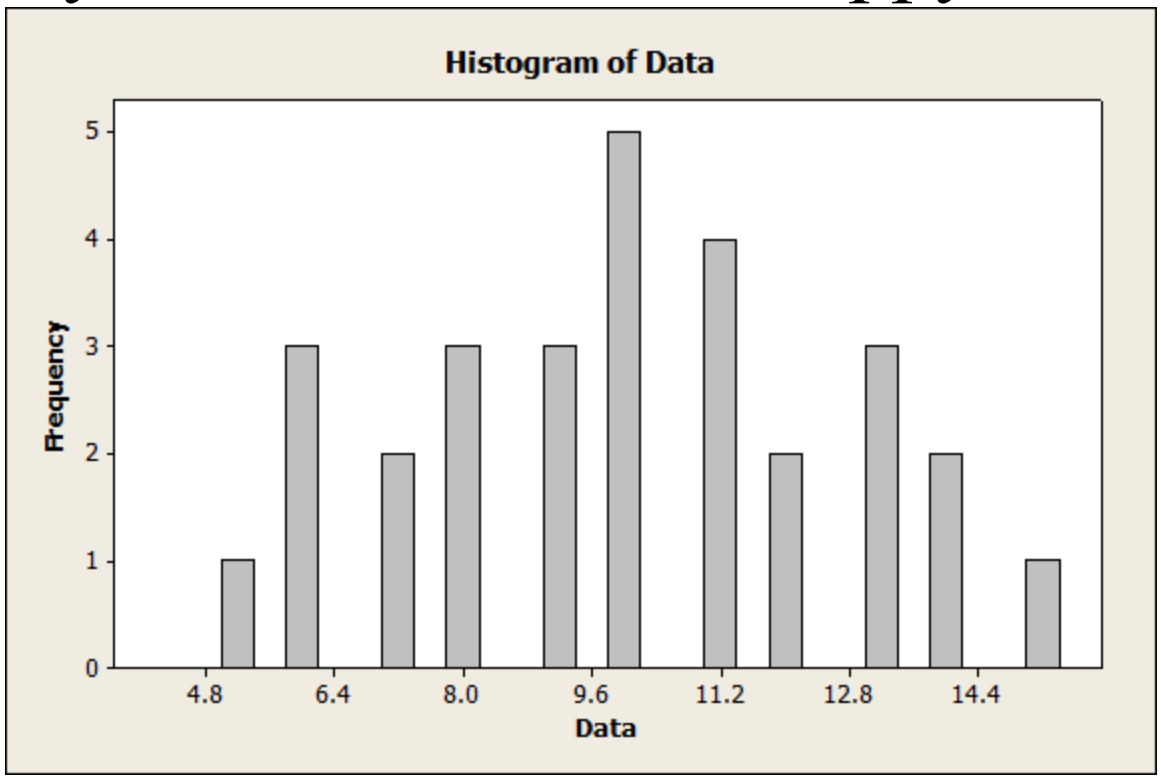


- A histogram:
  - Is a graph
  - Groups the data into *bins* of the same length
  - Displays the frequency of the data in each bin (small data sets)
  - Displays the relative proportion of observations that fall in each bin (for large data sets)
  - Easily shows the shape of a distribution

- Downsides: we lose the individual data values; bin width can drastically affect the way the graph looks

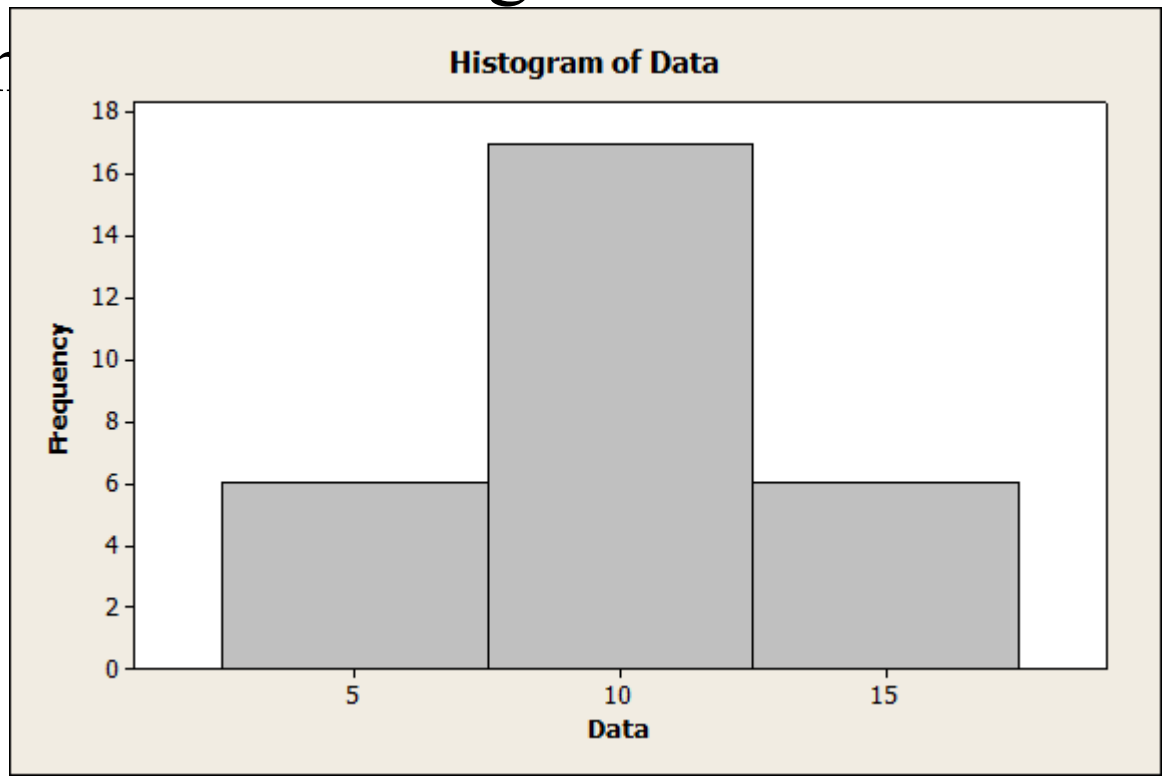
Histograms

- Too many bins – data looks choppy



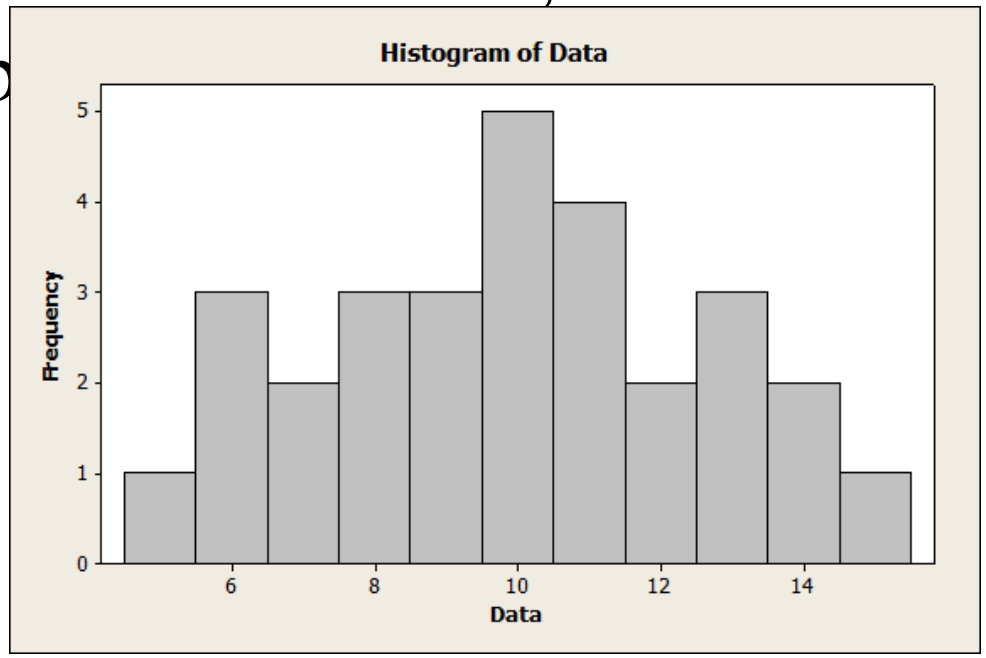
Histograms

- Too few bins – data gives us minimal inform



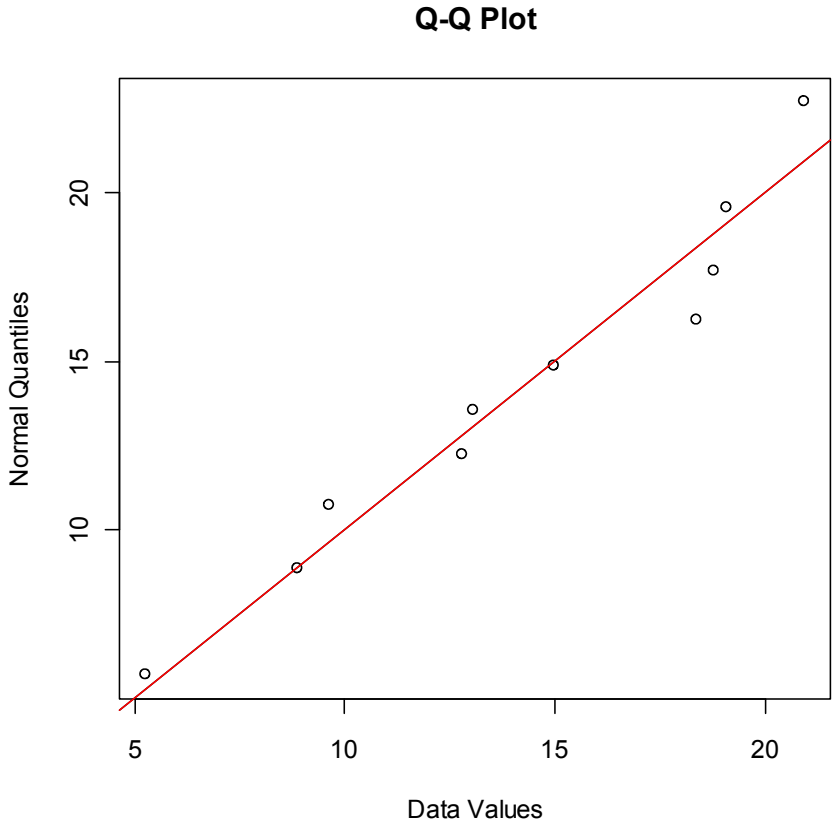
Histograms

- "Just Right" – nothing is ever perfect but you should experiment with the bins; the default option may not b



Q-Q Plots:
Normal Probability Plot

- Using software, make a *scatter plot* of the data values against the normal quantiles.

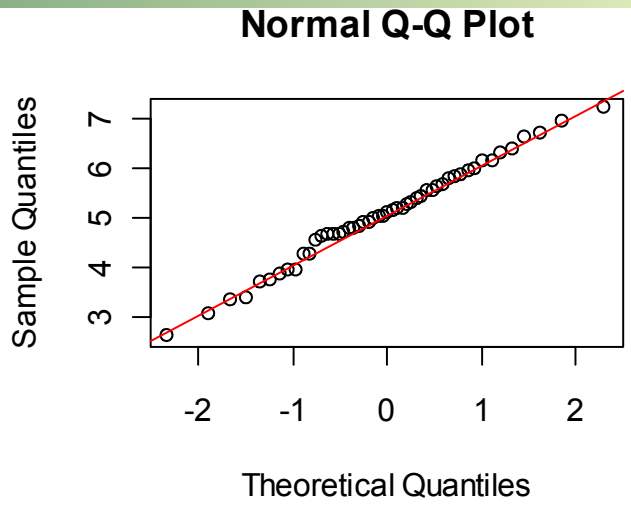


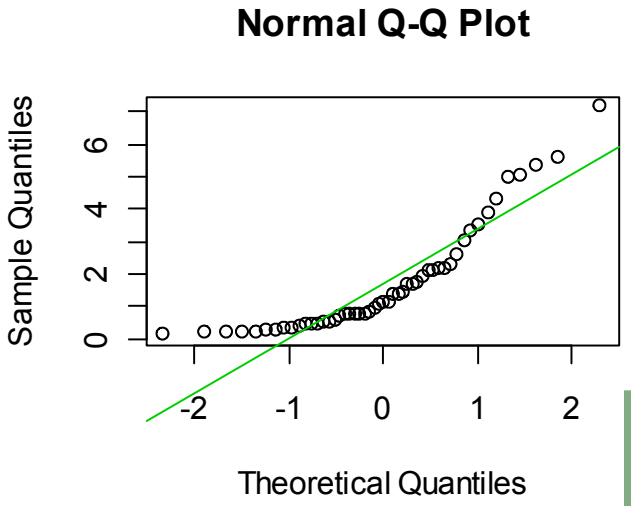
Definition 8.7

The **normal quantile-quantile plot** is a plot of $y_{(i)}$ (ordered observations) against $q_{0,1}(f_i)$, where $f_i = \frac{i-\frac{3}{8}}{n+\frac{1}{4}}$.

Q-Q Plots

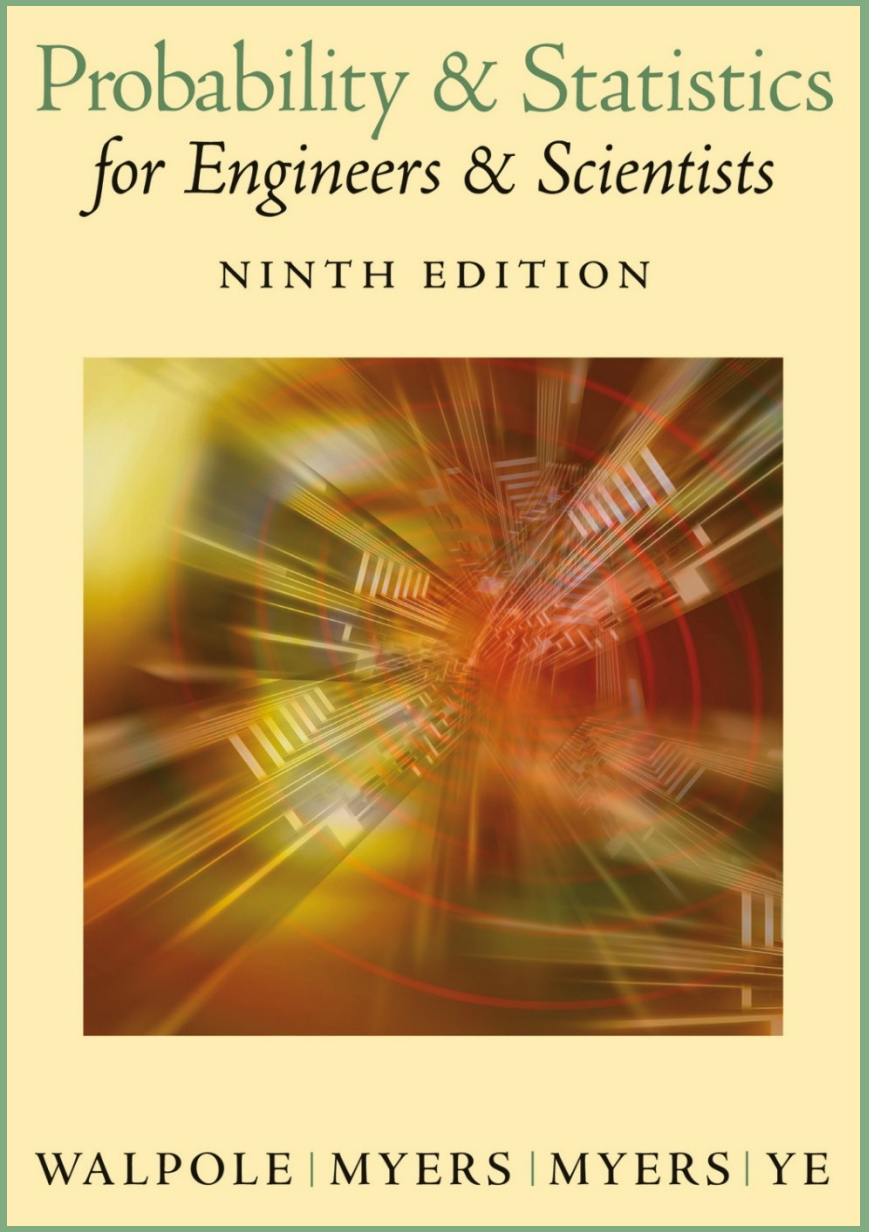
- The top graph is from data that are approximately normal.
- The bottom graph is from data that are exponential (clearly not normal!)





Section 8.9

Review



PEARSON
Addison Wesley

Definition 8.4

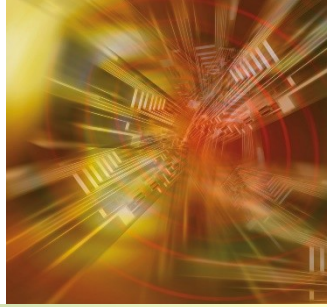
Any function of the random variables constituting a random sample is called a **statistic**.

Theorem 8.1

If S^2 is the variance of a random sample of size n, we may write

$$S^2 = \frac{1}{n(n-1)}\left[n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2\right].$$

Definition 8.5



The probability distribution of a statistic is called a **sampling distribution.**

Theorem 8.2

Central Limit Theorem: If $\bar{X}$ is the mean of a random sample of size n taken from a population with mean μ and finite variance σ^2, then the limiting form of the distribution of

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

as $n \to \infty$, is the standard normal distribution $n(z; 0, 1)$.

Theorem 8.3

If independent samples of size n_1 and n_2 are drawn at random from two populations, discrete or continuous, with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively, then the sampling distribution of the differences of means, $\bar{X}_1 - \bar{X}_2$, is approximately normally distributed with mean and variance given by

$$\mu_{\bar{X}_1-\bar{X}_2} = \mu_1 - \mu_2 \text{ and } \sigma^2_{\bar{X}_1-\bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}.$$

Hence,

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(\sigma_1^2/n_1) + (\sigma_2^2/n_2)}}$$

is approximately a standard normal variable.

Corollary 8.1

Let $X_1, X_2, \ldots, X_n$ be independent random variables that are all normal with mean μ and standard deviation σ. Let

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and} \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Then the random variable $T = \frac{\bar{X}-\mu}{S/\sqrt{n}}$ has a t-distribution with $v = n - 1$ degrees of freedom.

Theorem 8.6

Let U and V be two independent random variables having chi-squared distributions with v_1 and v_2 degrees of freedom, respectively. Then the distribution of the random variable $F = \frac{U/v_1}{V/v_2}$ is given by the density function

$$h(f) = \begin{cases} \frac{\Gamma[(v_1+v_2)/2](v_1/v_2)^{v_1/2}}{\Gamma(v_1/2)\Gamma(v_2/2)} \frac{f^{(v_1/2)-1}}{(1+v_1 f/v_2)^{(v_1+v_2)/2}}, & f > 0, \\ 0, & f \leq 0. \end{cases}$$

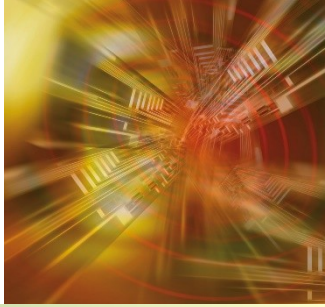
This is known as the ***F*-distribution** with v_1 and v_2 degrees of freedom (d.f.).

Theorem 8.7

Writing $f_\alpha(v_1, v_2)$ for f_α with v_1 and v_2 degrees of freedom, we obtain

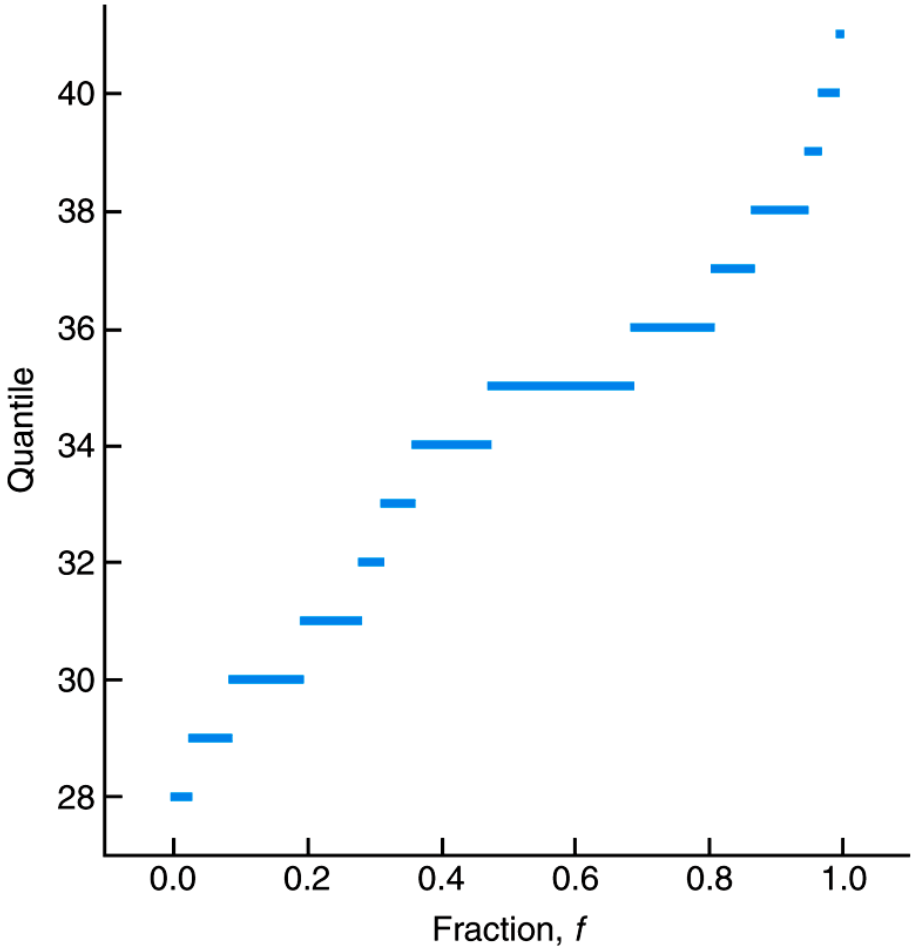
$$f_{1-\alpha}(v_1, v_2) = \frac{1}{f_\alpha(v_2, v_1)}.$$

Definition 8.6



A **quantile** of a sample, $q(f)$, is a value for which a specified fraction f of the data values is less than or equal to $q(f)$.

Figure 8.15 Quantile plot for paint data



Definition 8.7

The **normal quantile-quantile plot** is a plot of $y_{(i)}$ (ordered observations) against $q_{0,1}(f_i)$, where $f_i = \frac{i-\frac{3}{8}}{n+\frac{1}{4}}$.

Figure 8.16 Normal quantile-quantile plot for paint data