Topic 2: Basic concepts – part I

Montgomery: chapter 2

Prof. Lingling An
University of Arizona

Outline

- Two major chapters –
  - Descriptive: graphical & numerical
  - Inferential: hypothesis testing & confidence interval
- Review basic distributions

Statistics: two major chapters

Descriptive Statistics

- Gives *numerical* and *graphic* procedures to summarize a collection of data in a clear and understandable way

Inferential Statistics

- Provides procedures to draw inferences about a population from a sample

Some Basic Statistical Concepts

- Describing sample data
  - Random samples
  - Sample mean, variance, standard deviation
  - Populations versus samples
  - Population mean, variance, standard deviation
  - Estimating parameters
- Simple **comparative** experiments
  - The hypothesis testing framework
  - The two-sample *t*-test
  - Checking assumptions, validity

Descriptive Statistics - Graphical approach

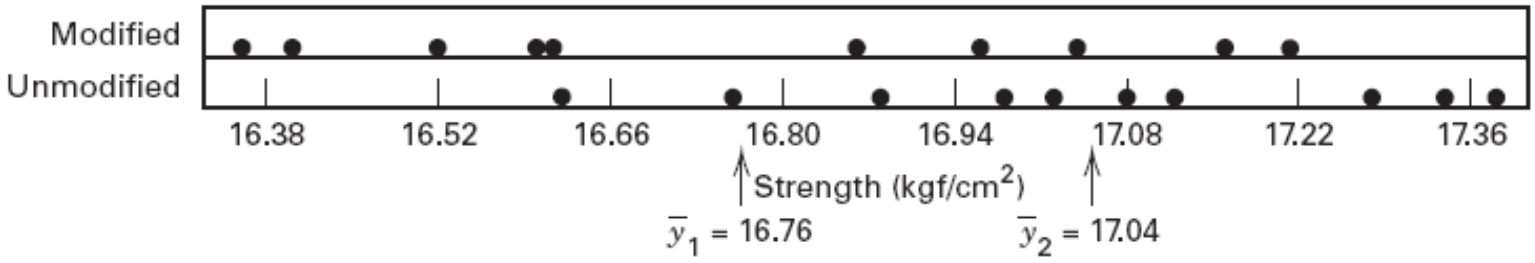
Portland Cement Formulation (page 24)

■ TABLE 2.1

Tension Bond Strength Data for the Portland Cement Formulation Experiment

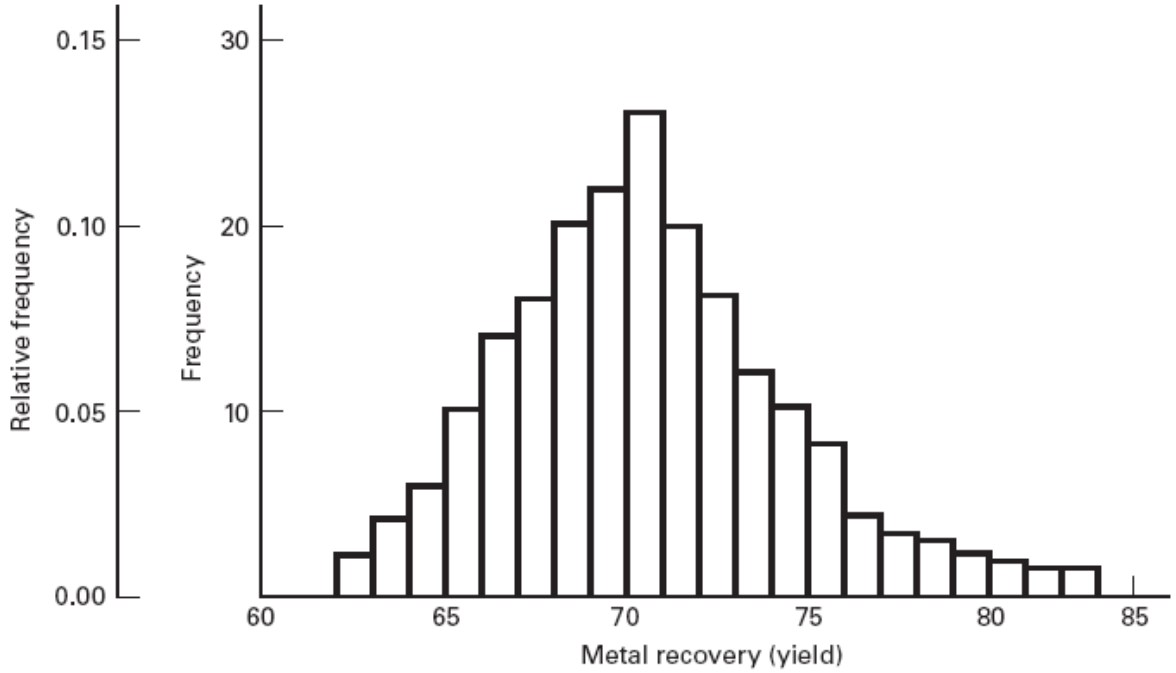
j	Modified Mortar y_{1j}	Unmodified Mortar y_{2j}
1	16.85	16.62
2	16.40	16.75
3	17.21	17.37
4	16.35	17.12
5	16.52	16.98
6	17.04	16.87
7	16.96	17.34
8	17.15	17.02
9	16.59	17.08
10	16.57	17.27

Graphical View of the Data - dot diagram



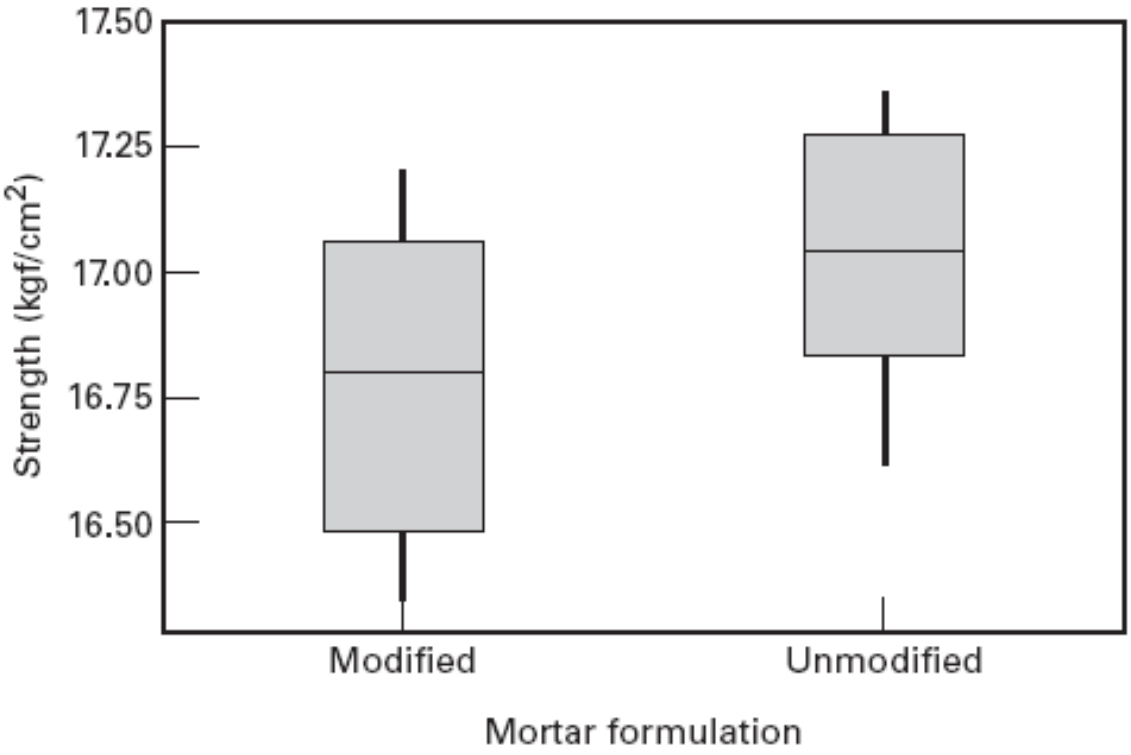
■ FIGURE 2.1 Dot diagram for the tension bond strength data in Table 2.1

If you have a large sample, a histogram may be useful



■ FIGURE 2.2 Histogram for 200 observations on metal recovery (yield) from a smelting process

Box Plots



■ FIGURE 2.3 Box plots for the Portland cement tension bond strength experiment

Numerical approach – Descriptive Measures

- **Central tendency measures**
  compute to give a "center" around which the measurements in the data are distributed.
- **Variation or Variability measures**
  describe "data spread" or how far away the measurements are from the center.
- **Skewness**
  describe how much the data is skewed to one side
- **Kurtosis**
  measure the peak of the data

Random Variable and Probability Distribution

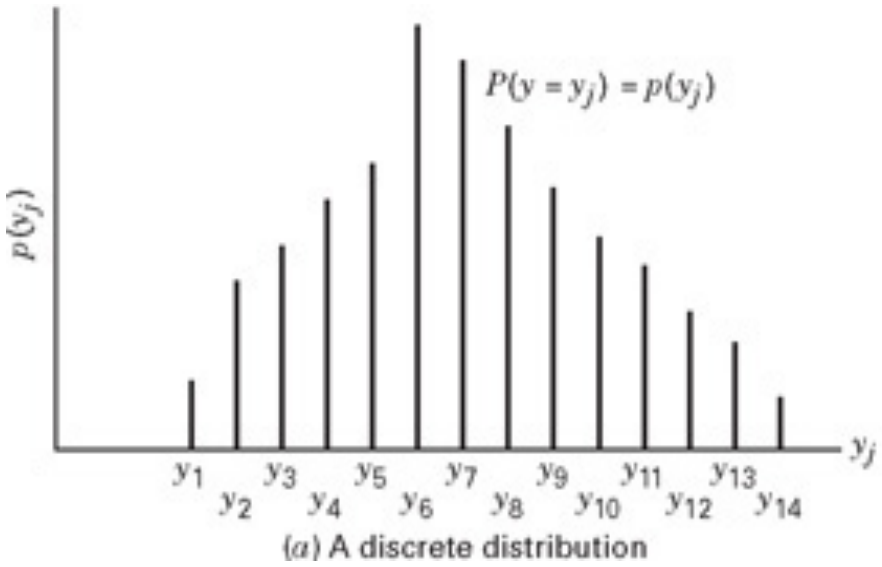
- Discrete random variable Y:
  - Finite possible values $\{y_1, y_2, y_3, \ldots, y_k\}$
  - Probability mass function $\{p(y_1), p(y_2), \ldots p(y_k)\}$ satisfying

$$p(y_i) \geq 0 \text{ and } \sum_{i=1}^{k} p(y_i) = 1.$$

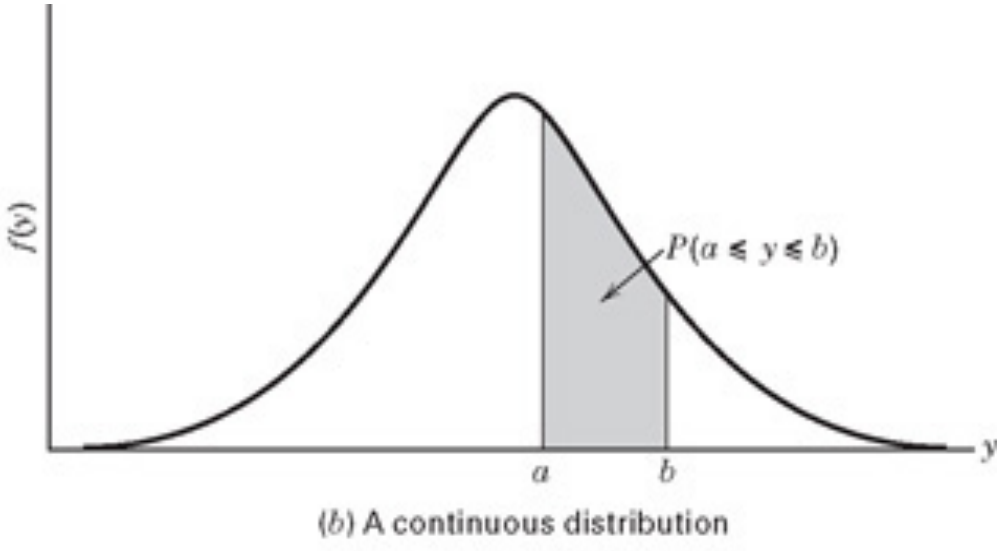
- Continuous random variable Y:
  - Possible values form an interval
  - Probability density function $f(y)$ satisfying

$$f(y) \geq 0 \text{ and } \int f(y)dy = 1.$$

Probability distributions



(a) A discrete distribution



(b) A continuous distribution

Mean, variance, formulas

- Mean and variance

Mean $\mu = \mathrm{E}(Y)$: center, location, etc.

Variance $\sigma^2 = \mathrm{Var}(Y)$: spread, dispersion, etc.

Discrete Y:

$$\mu = \sum_{i=1}^{k} y_i p(y_i); \quad \sigma^2 = \sum_{i=1}^{k} (y_i - \mu)^2 p(y_i)$$

Continuous Y:

$$\mu = \int y f(y) dy; \quad \sigma^2 = \int (y - \mu)^2 f(y) dy$$

- Formulas for calculating mean and variance for two variables

If Y_1 and Y_2 are **independent**, then

- $\mathrm{E}(Y_1 Y_2) = \mathrm{E}(Y_1)\mathrm{E}(Y_2)$
- $\mathrm{Var}(aY_1 \pm bY_2) = a^2 \mathrm{Var}(Y_1) + b^2 \mathrm{Var}(Y_2)$

More formula on p28

Inferential Statistics

- Draw conclusion or make generalizations about a population via sample data
- Contain two main components:
  - Confidence interval
  - Hypothesis testing

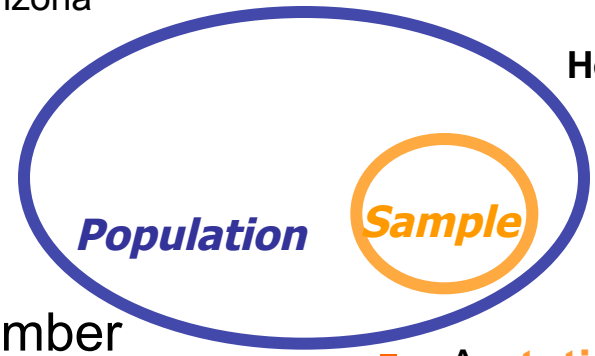
Population vs. sample

- **Population:** The entire group of individuals in which we are interested but can't assess directly

  Example: all adult males in Arizona

- **Sample:** A part of the population we actually examine and for which we do have data

  Example: Take a sample of 100 adult males in Arizona



How well the sample represents the population depends on the *sampling design*.

- A **parameter** is a number describing a characteristic of the population.
  - Example: average height of the adult males in Arizona
  - Usually parameter is unknown and need to be estimated by using ----

- A **statistic** is a number describing a characteristic of a sample.
  - Example: average height of these 100 adult males
  - Different sample results in different statistic

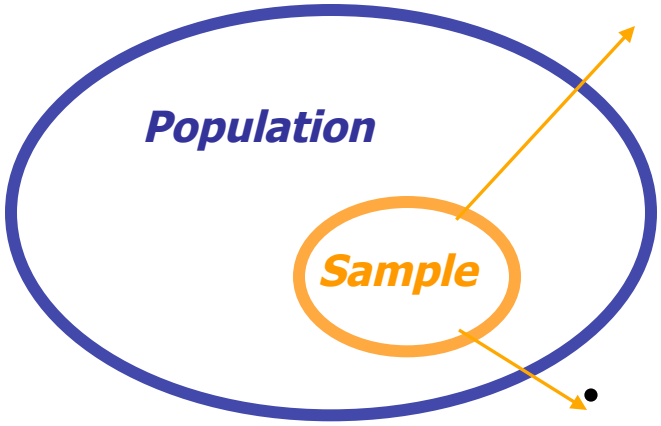
Parameter – statistic- estimate

More explanations:

- Parameter, which is unknown, is our interest,
- A statistic, distinct from an unknown parameter, can be computed from a sample.
- Very often, a statistic used to estimate a parameter is also called an **estimate**.
  - For instance, the *sample mean* is a statistic and is an estimate for the *population mean*, which is a parameter.

Mean and standard deviation from a population and from a sample

- **Population:**
  - Population mean μ
  - Population standard deviation σ



- **Sample:**
  - Sample mean

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

  - Sample standard deviation: variation around the mean

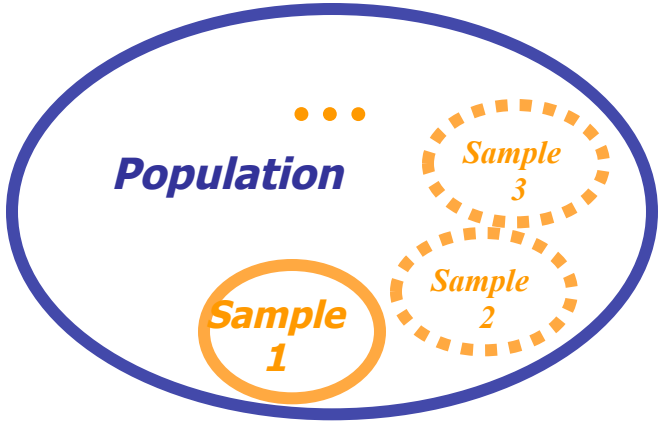
$$s = \sqrt{\frac{1}{n-1}\sum_{1}^{n}(x_i - \bar{x})^2}$$

Why sampling distribution?

We have discussed (point) estimates:

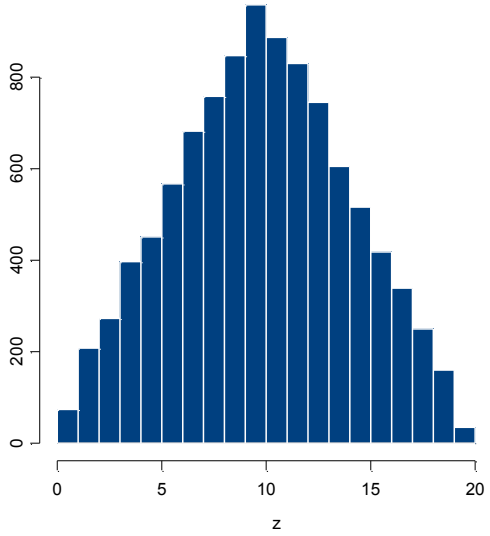
- as an estimate of population mean, μ
- as an estimate of population standard deviation, σ

- These (point) estimates ($\bar{x}$ and *s*) are almost never exactly equal to the true values they are estimating.
- In order for the point estimate to be useful, it is necessary to describe just how far off from the true value it is likely to be.



Sampling Distributions

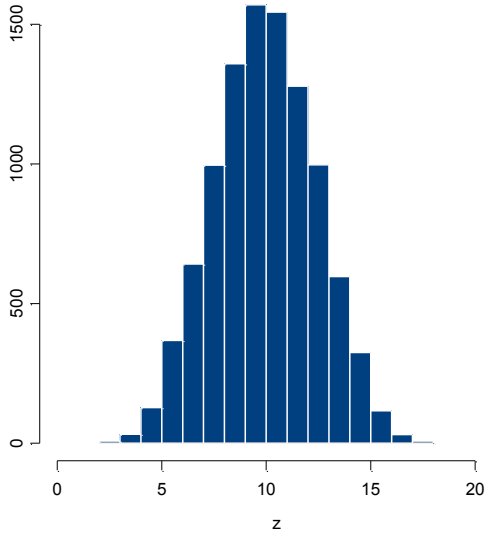
- By drawing random samples of size *n* from a population with a specific mean and variance, we can learn
  - (a) how much error we can expect on average and
  - (b) how much variation there will be on average in the errors observed
- **Sampling distribution**: the distribution of a sample statistic (e.g., a mean) when sampled under known sampling conditions from a known population.



n = 2

mean of sample means = 10

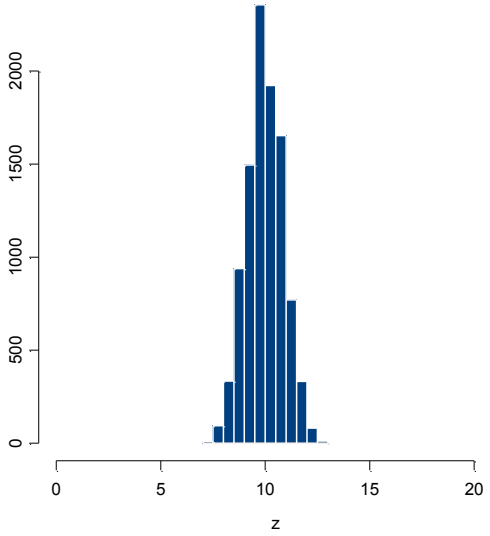
SD of sample means = 4.16



n = 5

mean of sample means = 10

SD of sample means = 2.41

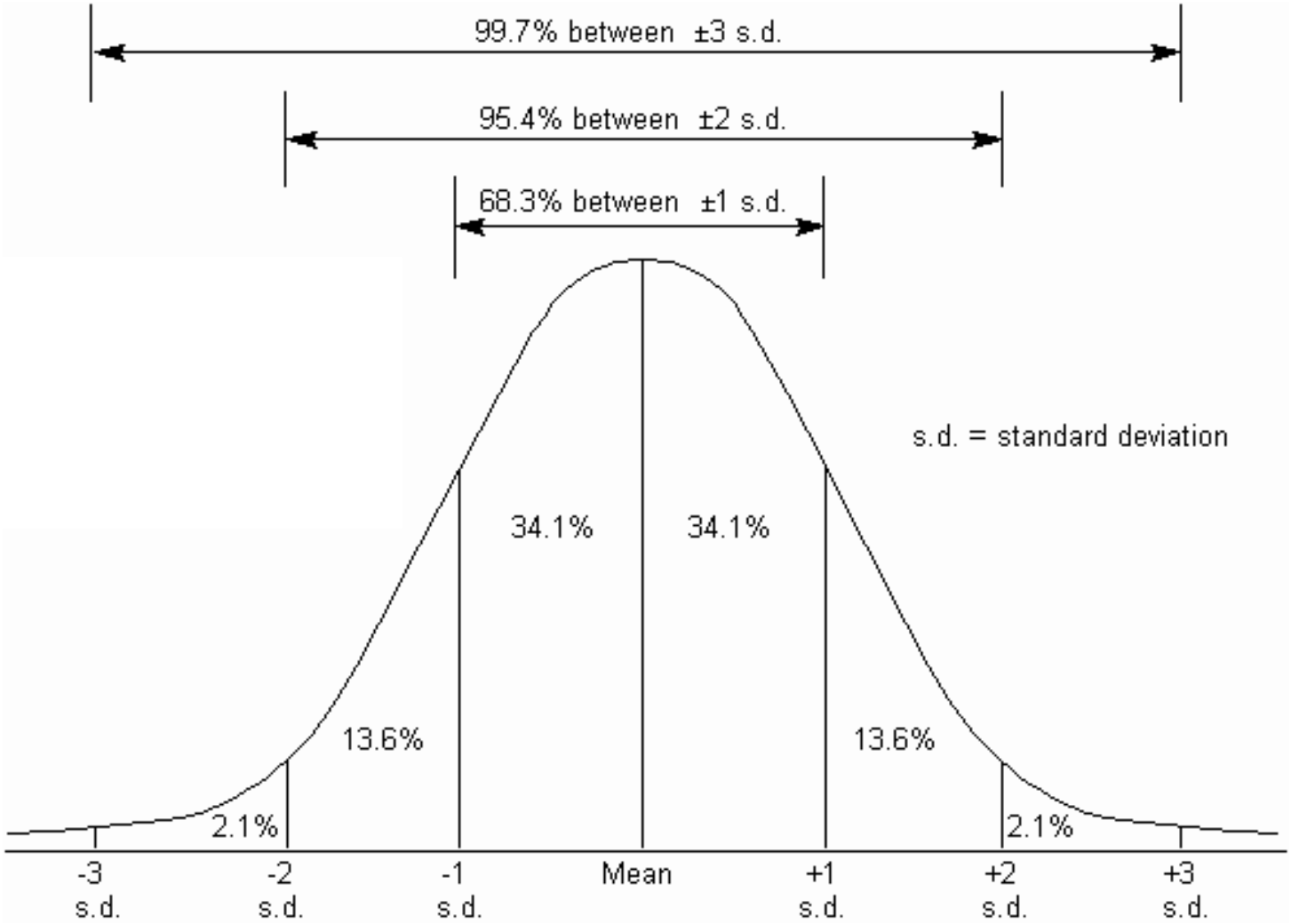


n = 15

mean of sample means = 10

SD of sample means = 0.87

Normal distribution



Sampling Distribution - example

Random sample: $Y_1, Y_2, \ldots, Y_n \sim N(\mu, \sigma^2)$

Sample mean $\overline{Y} = (Y_1 + Y_2 + \cdots + Y_n)/n$

$$E(\overline{Y}) = E(\frac{1}{n}\sum Y_i) = \frac{1}{n}\sum E(Y_i) = \frac{1}{n} n\mu = \mu$$

$$\text{Var}(\overline{Y}) = \text{Var}(\frac{1}{n}\sum Y_i) = \frac{1}{n^2}\sum \text{Var}(Y_i) = \frac{1}{n^2} n\sigma^2 = \sigma^2/n$$

$\overline{Y}$ follows $N(\mu, \frac{\sigma^2}{n})$

Central Limit Theorem

$Y_1, Y_2, \ldots, Y_n$ are n independent and identically distributed random variables with $E(Y_i) = \mu$ and $\text{Var}(Y_i) = \sigma^2$. Then

$$Z_n = \frac{\overline{Y} - \mu}{\sigma / \sqrt{n}}$$

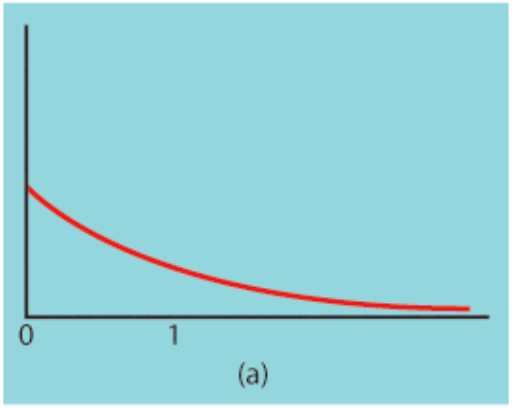
approximately follows the standard normal distribution $N(0, 1)$.

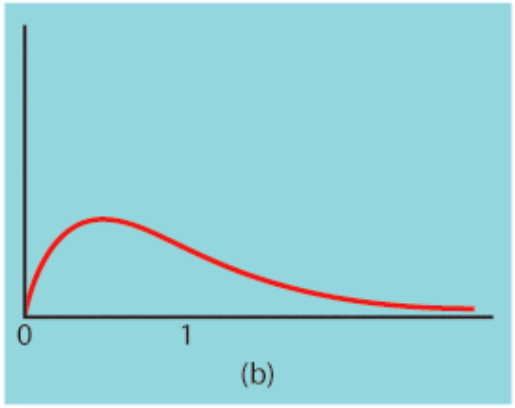
- Remark:

1. Do not need to assume the original population distribution is normal
2. When the population distribution is normal, then Z_n exactly follows $N(0, 1)$.

The central limit theorem (illustration)

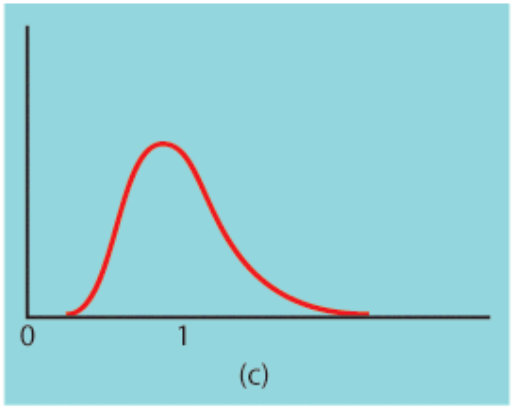
Population with strongly skewed distribution

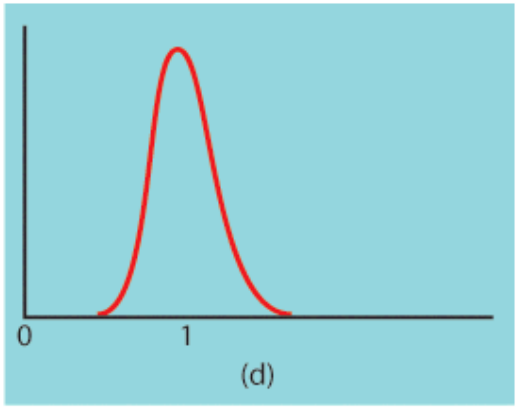




Sampling distribution of $\bar{x}$ for n = 2 observations

Sampling distribution of $\bar{x}$ for n = 10 observations





Sampling distribution of $\bar{x}$ for n = 25 observations

Sampling distribution: sample variance

$$S^2 = \frac{(Y_1 - \overline{Y})^2 + (Y_2 - \overline{Y})^2 + \cdots + (Y_n - \overline{Y})^2}{n-1}$$

$$E(S^2) = \sigma^2$$

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{\sigma^2} \sim \chi^2_{n-1}$$

Chi-squared distribution

If $Z_1, Z_2, \ldots, Z_k$ are i.i.d as $N(0,1)$, then

$$W = Z_1^2 + Z_2^2 + \cdots + Z_k^2$$

follows a Chi-squared distribution with degree of freedom k, denoted by χ_k^2

Density functions of χ_k^2

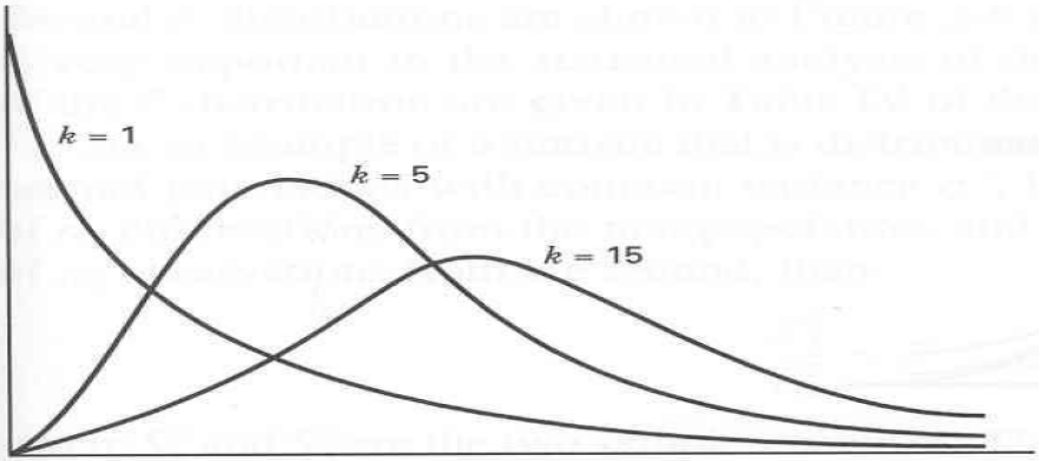


Figure 2-6 Several chi-square distributions.

t-distribution

If $Z \sim N(0,1)$, $W \sim \chi^2_k$ and Z and W independent, then

$$T_k = \frac{Z}{\sqrt{W/k}}$$

follows a t-distribution with d.f. k, i.e., $t(k)$.

For example, in t-test:

$$T = \frac{\overline{Y} - \mu_0}{S/\sqrt{n}} = \frac{\sqrt{n}(\overline{Y} - \mu_0)/\sigma}{\sqrt{S^2/\sigma^2}} = \frac{Z}{\sqrt{W/(n-1)}} \sim t(n-1)$$

Remark:

As n goes to infinity, $t(n-1)$ converges to $N(0,1)$.

Density function of $t_{(k)}$ distributions

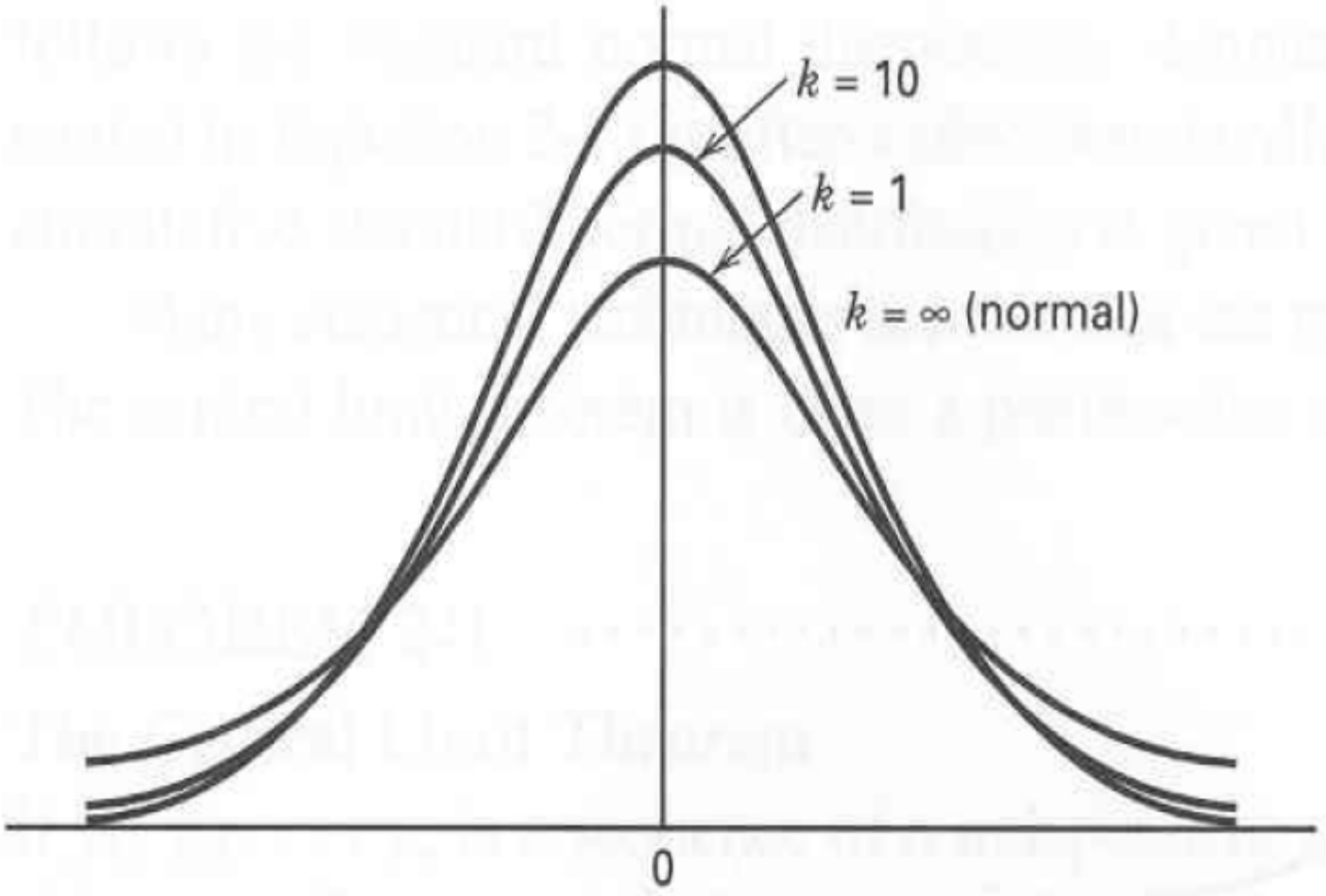


Figure 2-7 Several *t* distributions.

F-distribution

- **F-distributions:** F_{k_1,k_2}

  Suppose random variables $W_1 \sim \chi^2_{k_1}$, $W_2 \sim \chi^2_{k_2}$, and W_1 and W_2 are independent, then

$$F = \frac{W_1/k_1}{W_2/k_2}$$

  follows F_{k_1,k_2} with numerator d.f. k_1 and denominator d.f. k_2.

- **Example:** $H_0 : \sigma_1^2 = \sigma_2^2$, the test statistic is

$$F = \frac{S_1^2}{S_2^2} = \frac{S_1^2/\sigma^2}{S_2^2/\sigma^2} = \frac{W_1/(n_1-1)}{W_2/(n_2-1)} \sim F_{n_1-1,n_2-1}$$

  Refer to Section 2.6 for details.

Density function of F-distribution

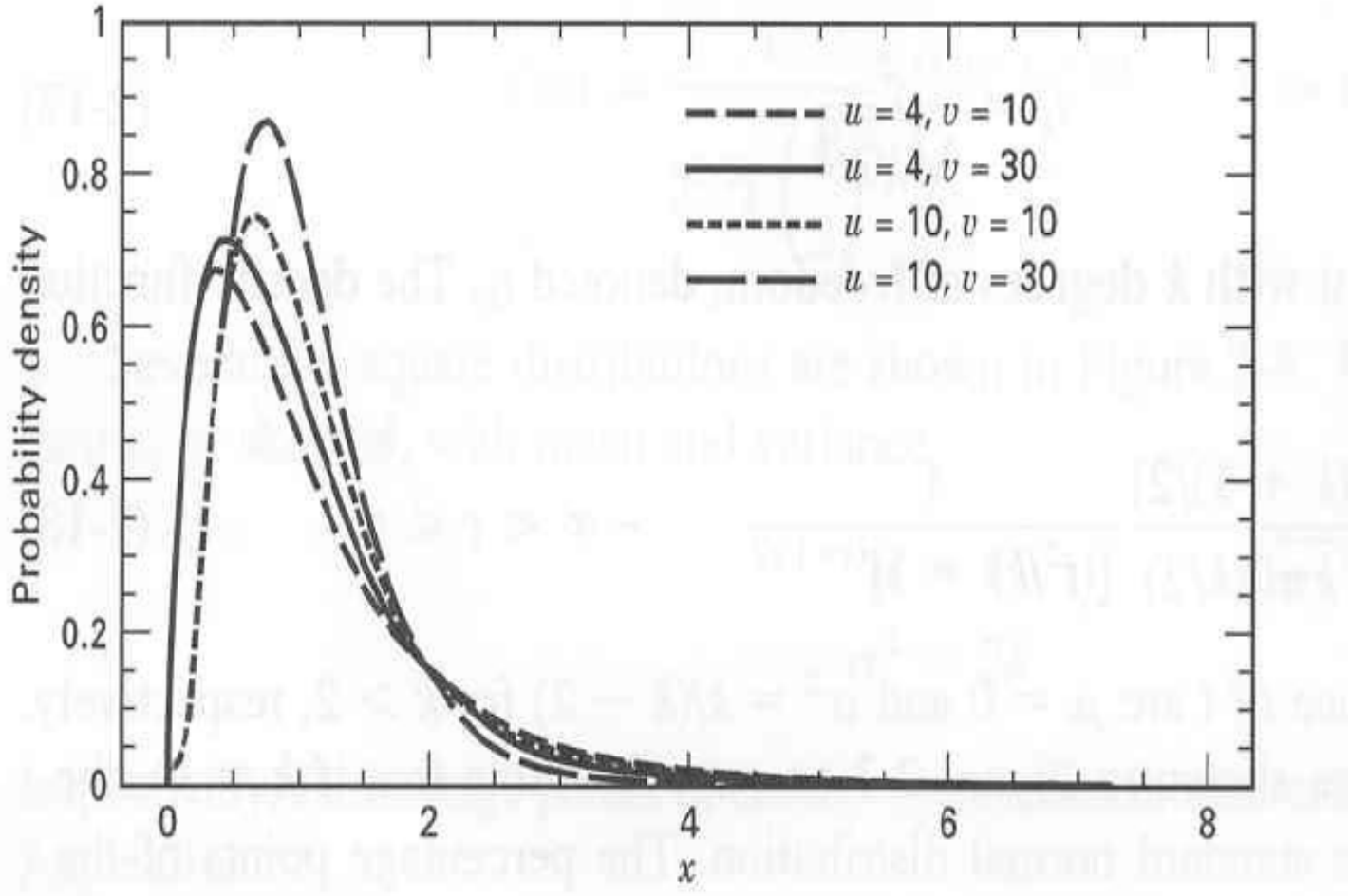
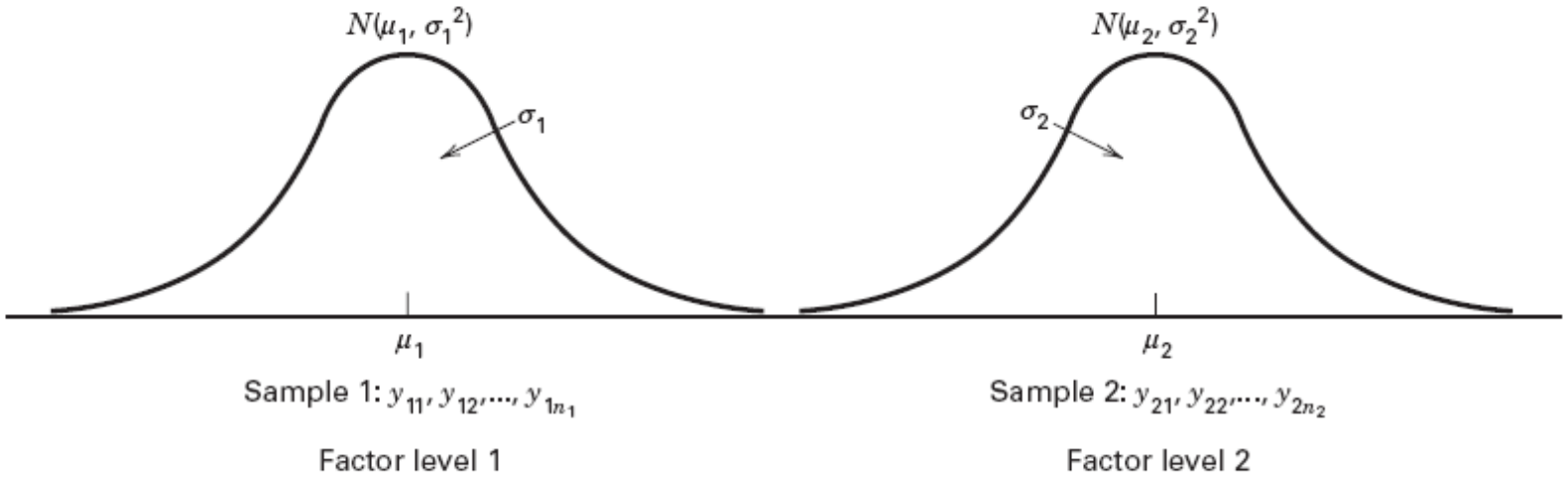


Figure 2-8 Several *F* distributions.

Statistical inference - Hypothesis Testing

- **Statistical hypothesis testing** is a useful framework for many experimental situations
- Origins of the methodology date from the early 1900s
- We will use a procedure known as the **two-sample *t*-test**

The Hypothesis Testing Framework



■ FIGURE 2.9 The sampling situation for the two-sample *t*-test

- Sampling from two **normal** distributions
- Statistical hypotheses: $H_0 : \mu_1 = \mu_2$

$$H_1 : \mu_1 \neq \mu_2$$

Hypothesis testing: Estimation of Parameters

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \text{ estimates the population mean } \mu$$

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n} (y_i - \bar{y})^2 \text{ estimates the variance } \sigma^2$$

Summary Statistics (p36)

Formulation 1

"New recipe":

modified mortar

$\bar{y}_1 = 16.76$

$S_1^2 = 0.100$

$S_1 = 0.316$

$n_1 = 10$

Formulation 2

"Original recipe":

Unmodified mortar

$\bar{y}_2 = 17.04$

$S_2^2 = 0.061$

$S_2 = 0.248$

$n_2 = 10$

How the Two-Sample *t*-Test Works:

Use the sample means to draw inferences about the population means

$$\bar{y}_1 - \bar{y}_2 = 16.76 - 17.04 = -0.28$$

$$\frac{\text{Difference in sample means}}{\text{Standard deviation of the difference in sample means}}$$

$$\sigma^2_{\bar{y}} = \frac{\sigma^2}{n}$$

This suggests a statistic:

$$Z_0 = \frac{\bar{y}_1 - \bar{y}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

How the Two-Sample *t*-Test Works -2

Use S_1^2 and S_2^2 to estimate σ_1^2 and σ_2^2

The previous ratio becomes $\dfrac{\bar{y}_1 - \bar{y}_2}{\sqrt{\dfrac{S_1^2}{n_1} + \dfrac{S_2^2}{n_2}}}$

However, we have the case where $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Pool the individual sample variances:

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

How the Two-Sample *t*-Test Works -3

The test statistic is

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- Values of t_0 that are near zero are consistent with the null hypothesis
- Values of t_0 that are very different from zero are consistent with the alternative hypothesis
- t_0 is a "distance" measure-how far apart the averages are expressed in standard deviation units
- Notice the interpretation of t_0 as a **signal-to-noise** ratio

The Two-Sample (Pooled) *t*-Test: example

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} = \frac{9(0.100) + 9(0.061)}{10 + 10 - 2} = 0.081$$

$$S_p = 0.284$$

$$t_0 = \frac{\bar{y}_1 - \bar{y}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{16.76 - 17.04}{0.284\sqrt{\frac{1}{10} + \frac{1}{10}}} = -2.20$$

The two sample means are a little over two standard deviations apart

Is this a "large" difference?

William Sealy Gosset (1876 -1937)

- Gosset's interest in barley cultivation led him to speculate that design of experiments should aim, not only at improving the average yield, but also at breeding varieties whose yield was insensitive (robust) to variation in soil and climate.

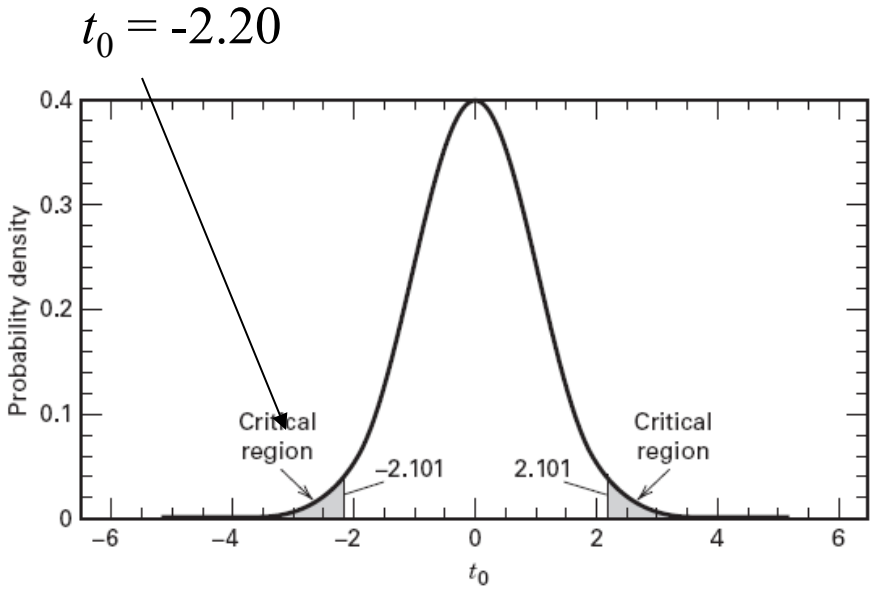
- Developed the *t*-test (1908) using his pen name “student”

- Gosset was a friend of both Karl Pearson and R.A. Fisher, an achievement, for each had a monumental ego and a loathing for the other.

'Student' in 1908

The Two-Sample (Pooled) *t*-Test -1

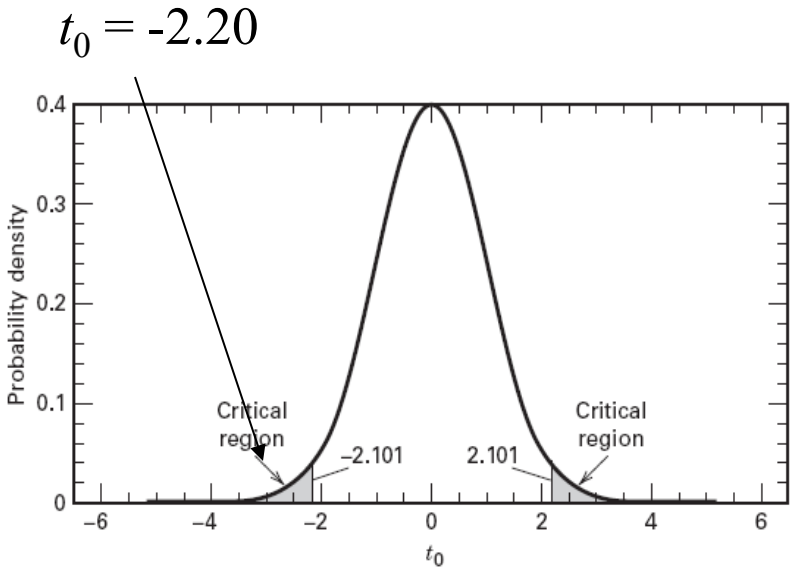
- We need an **objective** basis for deciding how large the test statistic t_0 really is.
- In 1908, W. S. Gosset derived the **reference distribution** for t_0 … called the *t* distribution
- Tables of the *t* distribution – see textbook appendix



■ **FIGURE 2.10** The *t* distribution with 18 degrees of freedom with the critical region $\pm t_{0.025,18} = \pm 2.101$

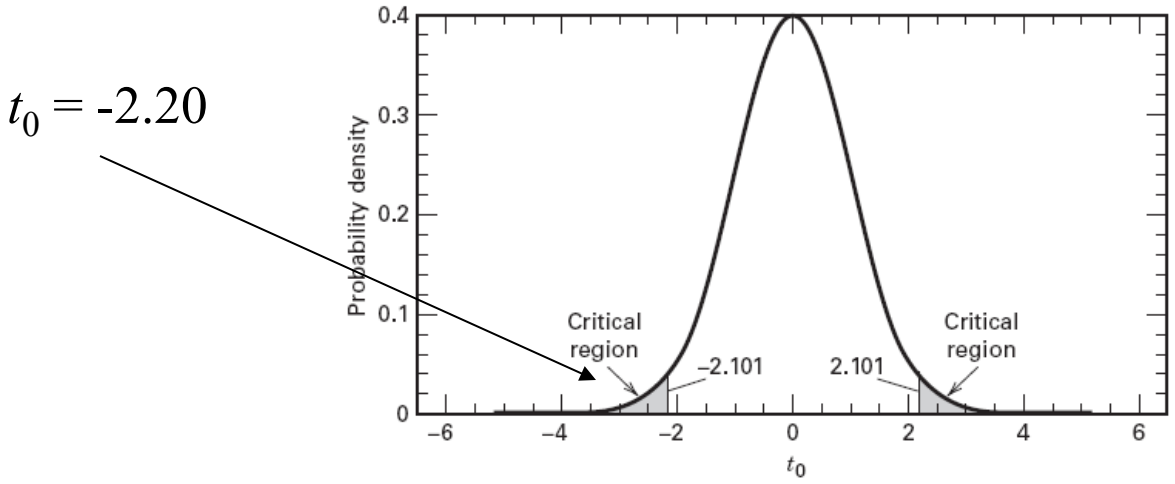
The Two-Sample (Pooled) *t*-Test - 2

- A value of t_0 between –2.101 and 2.101 is consistent with equality of means.
- It is possible for the means to be equal and t_0 to exceed either 2.101 or –2.101, but it would be a "**rare event**" … leads to the conclusion that the means are different
- Could also use the ***P*-value** approach



■ FIGURE 2.10 The t distribution with 18 degrees of freedom with the critical region $\pm t_{0.025,18} = \pm 2.101$

The Two-Sample (Pooled) *t*-Test - 3



■ **FIGURE 2.10** The t distribution with 18 degrees of freedom with the critical region $\pm t_{0.025,18} = \pm 2.101$

- The ***P*-value** is the area (probability) in the tails of the *t*-distribution beyond -2.20 + the probability beyond +2.20 (it's a two-sided test)
- The *P*-value is a measure of how unusual the value of the test statistic is given that the null hypothesis is true
- The *P*-value is the risk of **wrongly rejecting** the null hypothesis of equal means (it measures rareness of the event)
- The *P*-value in our problem is $P = 0.042$

Hypothesis testing: decision rules

- Given significance level , there are two approaches:

  - Compare observed test statistic with critical value
  - Compute the P-value of observed test statistic
    - Reject H_0, if the P-value $\leq \alpha$.

Last slide

- Read Sections 2.1 – 2.4.1