To Do

Read Sections 6.1 – 6.3.

Assignment 4 is due Friday November 25.

See detailed information regarding Tutorial Test 3 (Wednesday November 30) posted on Learn.

Last Class

(1) General Form of a Gaussian Response Model

(2) Linear Regression Models

(3) Checking the Assumptions of the Simple Linear Regression Model

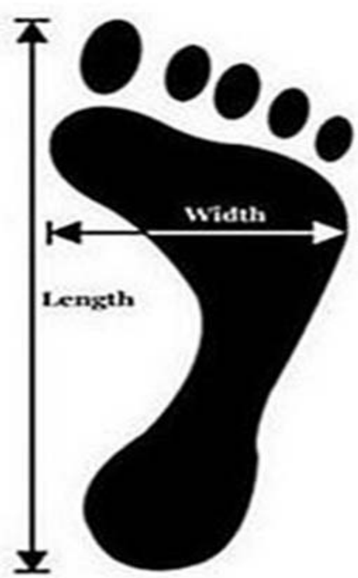
Today's Class

(1) Comparing the Means of Two Gaussian Populations

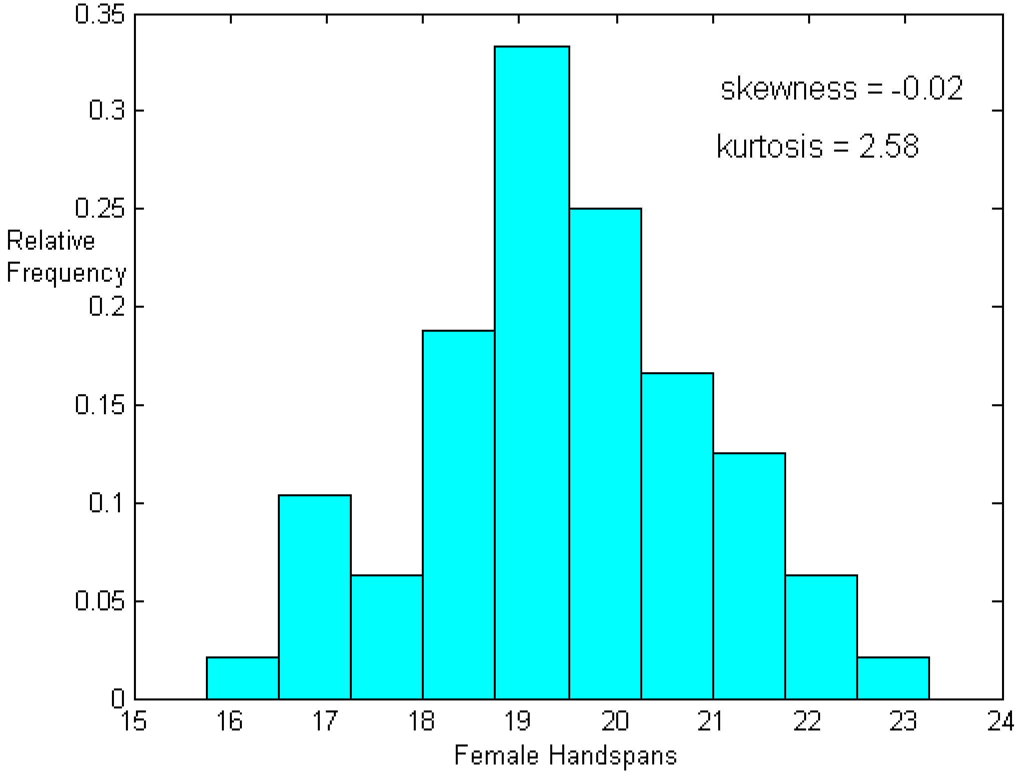
(2) Comparison of Two Means, Equal Variances

Section 6.3: Comparing the Means of Two Populations

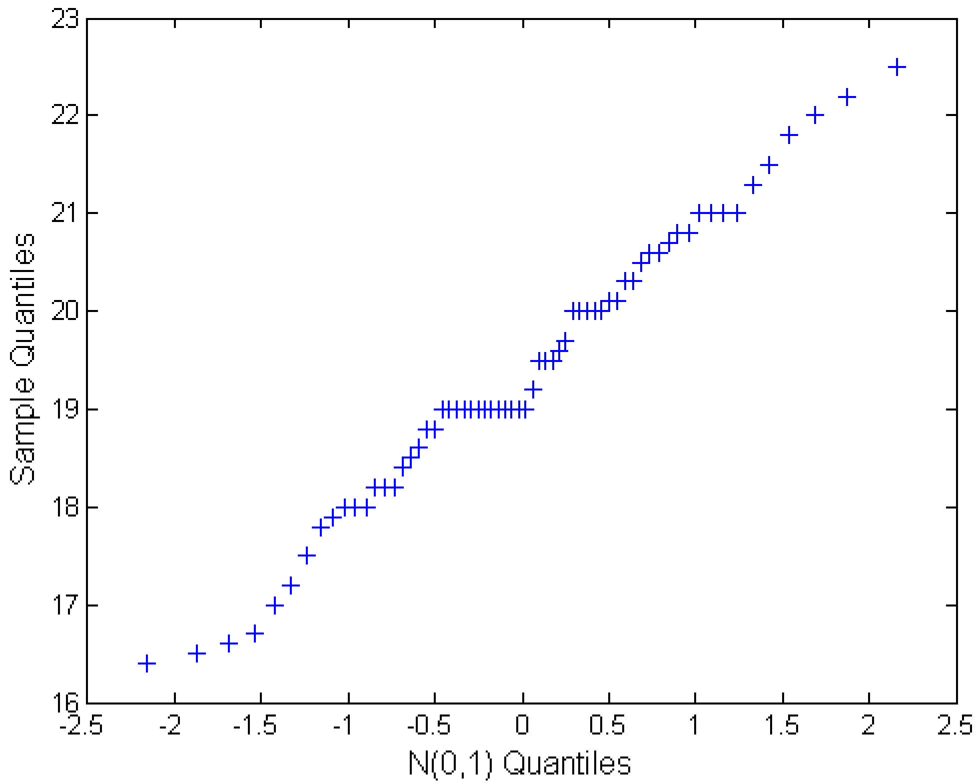
Recall the handspan example.



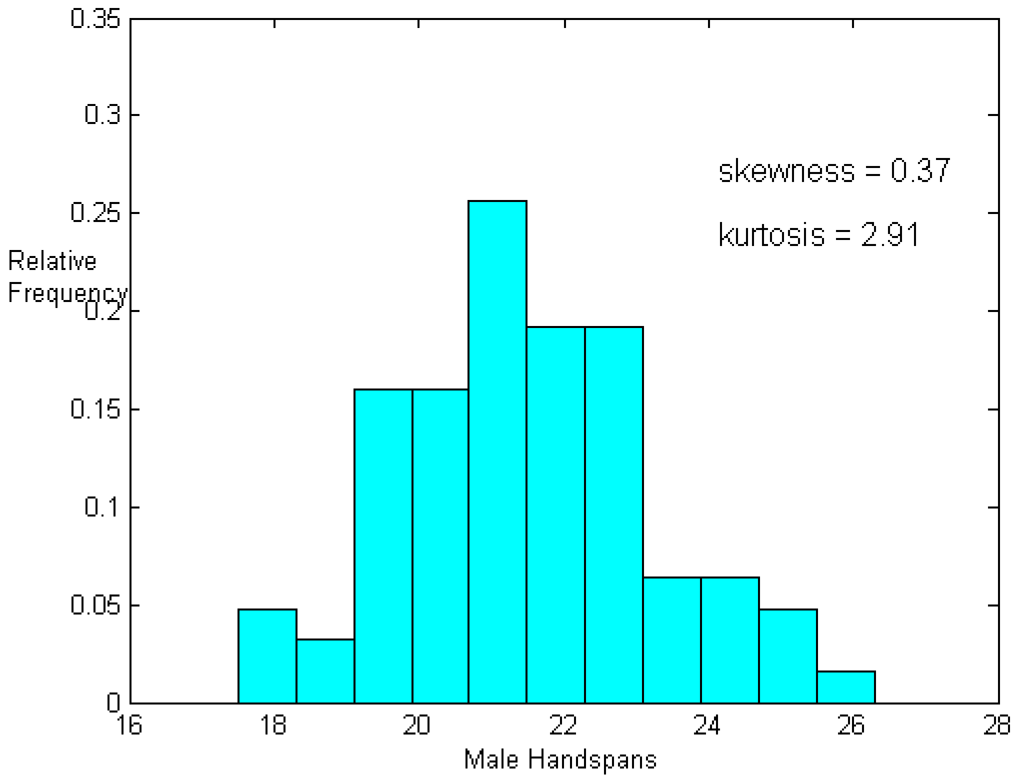
Relative Frequency Histogram of Female Handspans



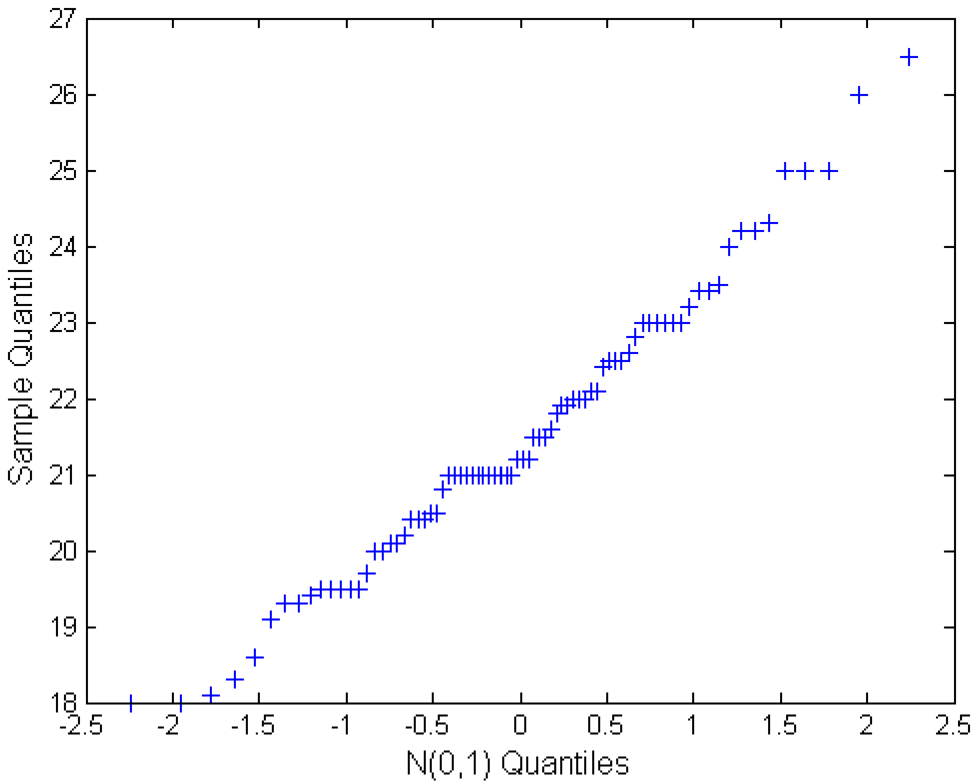
Qqplot of Female Handspans



Relative Frequency Histogram of Male Handspans



Qqplot of Male Handspans



Questions

Suppose we wanted to answer the question:

Are the hand spans of females enrolled in STAT 231 in Fall 2016 different on average from the hand span of males enrolled in STAT 231 in Fall 2016?

Model

To answer this question we need a model.

Let Y_{1i} = the handspan of the *i*th male, *i* =1,2,…,78 and let Y_{2i} = be the handspan of the *i*th female, *i* = 1,2,…,64

Based on the frequency histograms and the qqplots, it seems reasonable to assume Gaussian models for the Y_{1i}'s and the Y_{2i}'s.

Model

Assume

$Y_{1i} \backsim G(\mu_1, \sigma)$ **for** $i = 1,2,\ldots,78$ **independently**

and independently

$Y_{2i} \backsim G(\mu_2, \sigma)$ **for** $i = 1,2,\ldots,64$ **independently**

Note that we have assumed both Gaussian populations have the same standard deviation σ**.**

Unknown Parameters: μ_1, μ_2 and σ

The parameter μ_1 represents the mean handspan in centimeters for males enrolled in STAT 231 in Fall 2016 (the study population).

The parameter μ_2 represents the mean handspan in centimeters for females enrolled in STAT 231 in Fall 2016 (the study population).

(Note that we are assuming there is no bias in the measurements.)

The parameter σ represents the standard deviation of handspans in the study population.

Model – General Case

Assume

$Y_{1i} \backsim G(\mu_1, \sigma)$ for $i = 1,2,\ldots,n_1$ independently

and independently

$Y_{2i} \backsim G(\mu_2, \sigma)$ for $i = 1,2,\ldots,n_2$ independently

We call this a two sample Gaussian problem.

It can be shown to be a special case of the Gaussian Response Model.

Point Estimators of μ_1 and μ_2

The maximum likelihood estimator of μ_1 is

$$\tilde{\mu}_1 = \overline{Y_1} = \frac{1}{n_1}\sum_{i=1}^{n_1} Y_{1i}$$

and the maximum likelihood estimator of μ_2 is

$$\tilde{\mu}_2 = \overline{Y_2} = \frac{1}{n_2}\sum_{i=1}^{n_2} Y_{2i}$$

A point estimator of the difference μ_1 - μ_2 is

$$\tilde{\mu}_1 - \tilde{\mu}_2 = \overline{Y_1} - \overline{Y_2}$$

Point Estimator of σ

To define the point estimator of σ define

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (Y_{1i} - \overline{Y_1})^2$$

which is the point estimator of σ based on only the Y_{1i}'s and

$$S_2^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_1} (Y_{2i} - \overline{Y_2})^2$$

which is the point estimator of σ based on only the Y_{2i}'s.

Point Estimator of *σ*

The point estimator of *σ* is

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

$$= \frac{1}{n_1 + n_2 - 2}\left[\sum_{i=1}^{n_1}(Y_{1i} - \overline{Y}_1)^2 + \sum_{i=1}^{n_2}(Y_{2i} - \overline{Y}_2)^2\right]$$

is called the pooled estimator of variance, since it is obtained by "pooling" the estimators S_1^2 and S_2^2 of σ^2 from the two samples.

Point Estimator of σ

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

$$= \frac{1}{n_1 + n_2 - 2}\left[\sum_{i=1}^{n_1}(Y_{1i} - \overline{Y}_1)^2 + \sum_{i=1}^{n_2}(Y_{2i} - \overline{Y}_2)^2\right]$$

S_p^2 is not the maximum likelihood estimator.

We use S_p^2 as the estimator of σ^2 since $E(S_p^2) = \sigma^2$.

Pivotal Quantity $\mu_1 - \mu_2$ if σ Known

Since

$$\tilde{\mu}_1 = \bar{Y}_1 \sim G\left(\mu_1, \frac{\sigma}{\sqrt{n_1}}\right) \quad \text{and} \quad \tilde{\mu}_2 = \bar{Y}_2 \sim G\left(\mu_2, \frac{\sigma}{\sqrt{n_2}}\right)$$

independently we have

$$\tilde{\mu}_1 - \tilde{\mu}_2 = \bar{Y}_1 - \bar{Y}_2 \sim G\left(\mu_1 - \mu_2, \sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right)$$

or

$$\frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim G(0,1)$$

Pivotal Quantity for σ

$$\frac{(n_1+n_2-2)S_p^2}{\sigma^2} \sim \chi^2(n_1+n_2-2)$$

Pivotal Quantity μ_1 - μ_2 if σ Unknown

Since

$$\frac{\overline{Y_1} - \overline{Y_2} - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim G(0,1)$$

and

$$\frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} \sim \chi^2(n_1 + n_2 - 2)$$

independently we have

$$\frac{\overline{Y_1} - \overline{Y_2} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

Confidence Interval for μ_1 - μ_2

A 100p% confidence interval for μ_1 - μ_2 is given by

$$\bar{y}_1 - \bar{y}_2 \pm a s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where

$$P\left(T \le a\right) = \frac{1+p}{2} \text{ and } T \sim t\left(n_1 + n_2 - 2\right)$$

Test of Hypothesis for No Difference in Means

To H_0: $\mu_1 = \mu_2$ or H_0: $\mu_1 - \mu_2 = 0$ we use the test statistic

$$D = \frac{\left|\overline{Y}_1 - \overline{Y}_2 - 0\right|}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

with observed value

$$d = \frac{\left|\overline{y}_1 - \overline{y}_2 - 0\right|}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Test of Hypothesis for No Difference in Means

The *p*-value is given by

$$p\text{-value} = 2[1 - P(T \leq d)]$$

where $T \backsim t(n_1+n_2 - 2)$ **and**

$$d = \frac{\left|\bar{y}_1 - \bar{y}_2 - 0\right|}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Handspan Data

Males:

$$n_1 = 78,\ \hat{\mu}_1 = \bar{y}_1 = 21.50,\ s_1^2 = 3.4309$$

Females:

$$n_2 = 64,\ \hat{\mu}_2 = \bar{y}_2 = 19.37,\ s_2^2 = 2.055$$

Therefore

$$\hat{\mu}_1 - \hat{\mu}_2 = \bar{y}_1 - \bar{y}_2 = 21.50 - 19.37 = 2.14$$

and

$$s_p = \sqrt{\frac{73(3.4309) + 63(2.055)}{78 + 63 - 2}} = 1.6768$$

Confidence Interval for μ_1 - μ_2

Since P($T \leq$ 1.97705) = 0.975 for $T \backsim t(140)$ a 95% confidence interval for μ_1 - μ_2 is

$$\bar{y}_1 - \bar{y}_2 \pm a s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$= 2.14 \pm (1.97705)(1.6768)\sqrt{\frac{1}{78} + \frac{1}{64}}$$

$$= 2.14 \pm 0.5591$$

$$= [1.58, 2.70]$$

Test of H_0: $\mu_1 - \mu_2 = 0$

Since the value $\mu_1 - \mu_2 = 0$ is not in the interval [1.58,2.70] we know the *p*-value for testing H_0: $\mu_1 - \mu_2 = 0$ is less than 0.05.

Since

$$d = \frac{|2.14 - 0|}{1.67687\sqrt{\frac{1}{78} + \frac{1}{64}}} = 7.56$$

***p*-value = 2[1 - P($T \leq 7.56$)] $\approx$ 0 where $T \sim t(140)$.**

Test of H_0: $\mu_1 - \mu_2 = 0$

Since p-value ≈ 0 there is very strong evidence to contradict the hypothesis H_0: $\mu_1 - \mu_2 = 0$ based on the data.

The difference is statistically significant.

Is the difference of practical significance?

Clicker Question 2

A statistics instructor wants to determine whether
μ_1 = mean grade of the students in STAT 231 in W16 equals μ_2 = mean grade of students in STAT 231 in W15.

Based on data collected in her class she determines that the *p*-value for testing H_0: $\mu_1 - \mu_2 = 0$ is equal to 0.003. Which statement is TRUE?

A: There is a 0.3% chance that the mean grade for W16 is equal to the mean grade for W15.

B: There is a 0.3% chance that the mean grade for W16 is different from the mean grade for W15.

C: It is very unlikely that the instructor would see results like these if the mean grade for W16 was equal to the mean grade for W15.

D: There is a 0.3% chance that another sample will give these same results.