Biostatistics I

Benny Chung-Ying Zee, PhD
Division of Biostatistics
School of Public Health and Primary Care
Chinese University of Hong Kong

Outline

- Session I
  - Descriptive statistics
  - Difference between two means
  - Difference between two variances
  - Scatter plots and correlations
  - Regression
  - Chi-square
- Session II
  - Survival analysis
  - Covariate adjustment
  - Subgroup analysis
  - Intention-to-treat

Types of Statistics

- Descriptive Statistics
  - Help to summarize the data relevant to characteristics and responses
- Inferential Statistics
  - Try to determine the reasons for some of the differences in characteristics or responses
  - Determine the extent to which chance, or random variation, might explain apparent differences

Descriptive statistics

- Involve the use of summary values such as means, medians and proportions, and also graphical presentations.
- The raw data to be summarized consist of *variables* (values determined on each subject in the study) that may be *discrete* or *continuous*.
- Check the data for potential problems during data collection and management prior to analysis

Example – type of statistics used

- Reviewed the types of statistics appearing in the Journal of Paediatrics and Child Health, excluding review articles, over the 12 consecutive months from August 1996.
- All original articles contained some descriptive statistics, involving both discrete and continuous variables.
- Inferential statistics appeared in 70% of articles. Most inferential statistics were univariate analyses, with a minority involving multivariate analyses.

Table 1 Statistics appearing in original articles ($n = 74$) in the *Journal of Paediatrics and Child Health* over 12 consecutive months

			n (%)	
Data types	Categorical	Nominal	72 (97.3)	
		Ordinal	43	(58.1)
	Continuous	Interval	39 (52.7)	
		Ratio	60	(81.1)
Inferential statistics	Univariate	Chi-squared Fisher's exact test	33 (44.6)	
		t-test	24	(32.4)
		Mann–Whitney *U*-test	11	(14.9)
		Correlation	14	(18.9)
		Signed rank test	2	(2.7)
	Multivariate	Linear regression/ ANOVA*	12 (16.2)	
		Logistic regression	10	(13.5)
		Other	3	(4.1)

*ANOVA = analysis of variance.

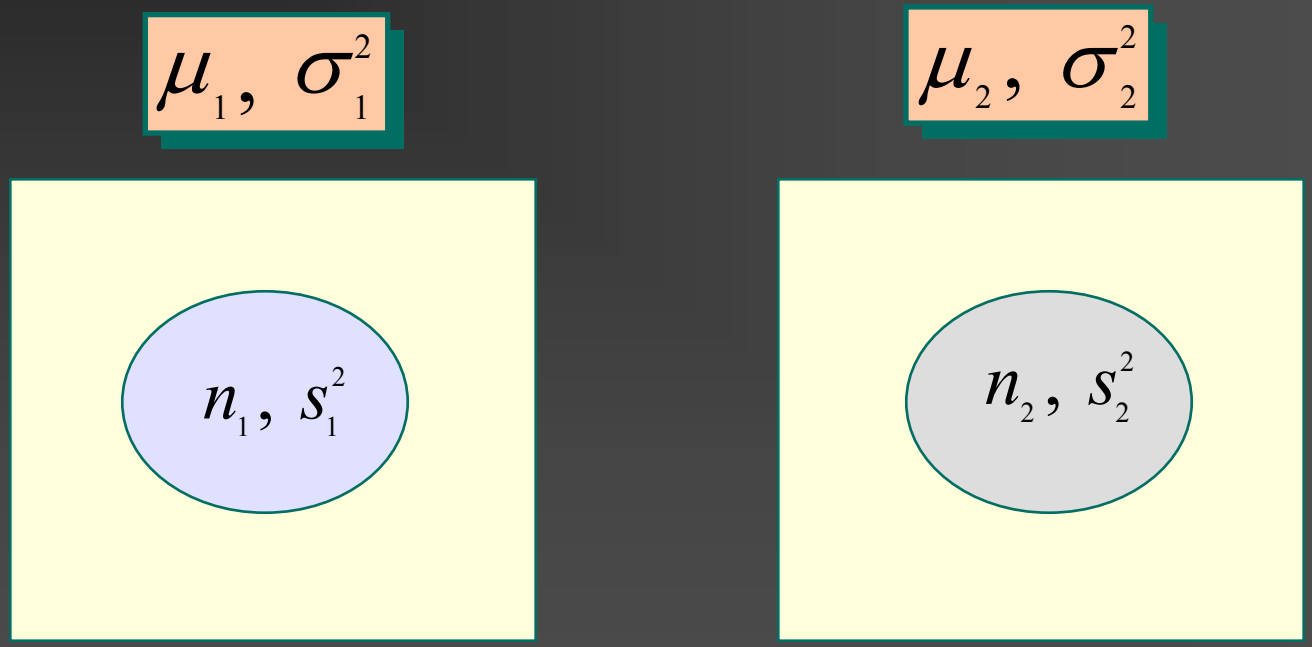
Statistical Inferences

- The essential role of formal statistical analysis is to account for the fact that research studies are performed on finite groups of subjects. The group of patients (or other individuals) in a study is regarded as a *sample* from a larger *population.*
- Statistical inference addresses the question of what can be said about the population based just on the sample, allowing for the crucial fact that another sample or samples would not produce identical results.

Sampling Distribution

- Using a sample mean to estimate a population mean.
- The mean of a variable (x) from a randomly selected sample is the best estimate of the population mean (μ).
- Central Limit Theorem: variation in sample means over repeated samples follows a normal distribution, with a SD determined by the SD of the population (σ) and the sample size (n), given by $SE(x) = \sigma/(\sqrt{n})$. This is called the *standard error of the mean.*
- SEM must itself be estimated by using the sample SD (s) in place of the unknown σ

Testing the Difference between Two Means: Large Samples



Formula for the *z* Test for Comparing Two Means from Independent Populations

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

Formula for Confidence Interval for Difference Between Two Means : Large Samples

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

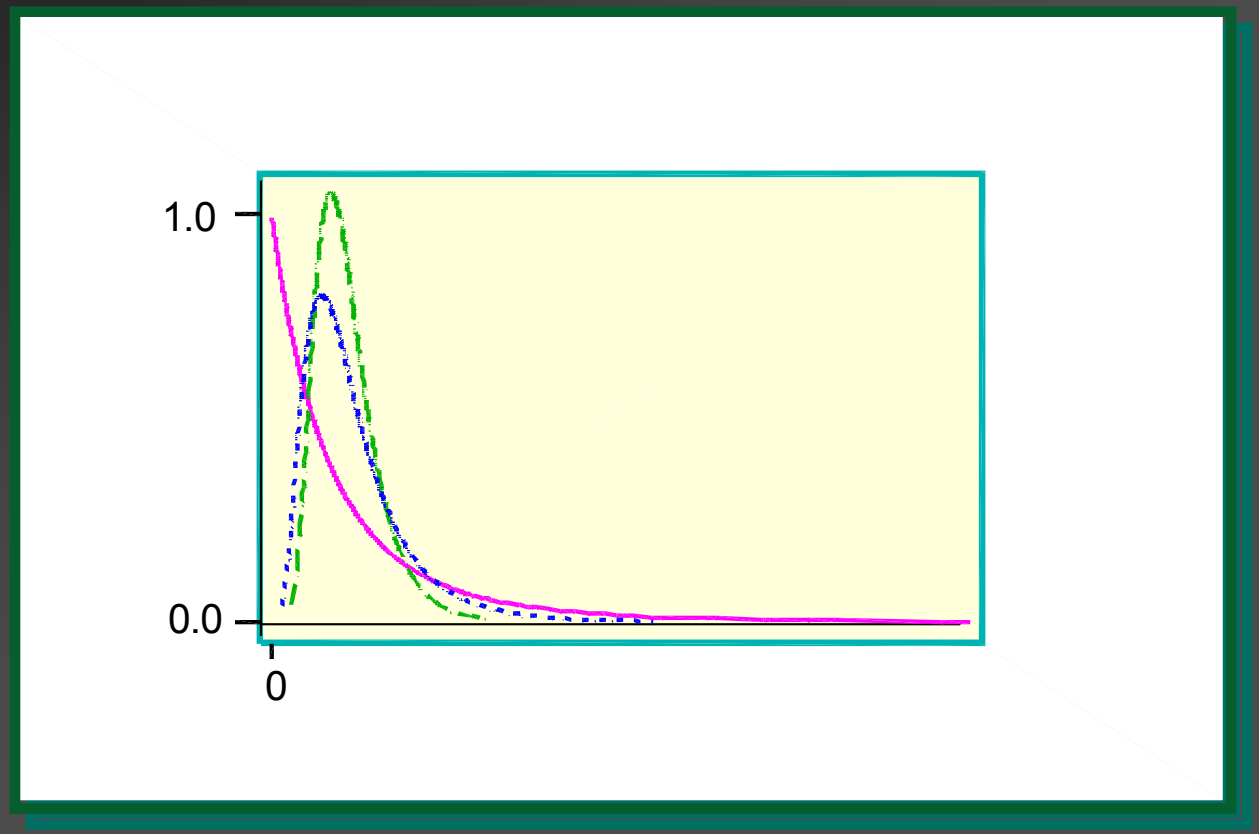
Testing the Difference Between Two Variances

- For the comparison of two variances or standard deviations, an *F* test is used.
- The sampling distribution of the variances is called the *F* distribution.

Characteristics of the *F* Distribution

- The values of *F* cannot be negative.
- The distribution is positively skewed.
- The mean value of *F* is approximately equal to 1.
- The *F* distribution is a family of curves based on the degrees of freedom of the variance of the numerator and denominator.

Curves for the *F* Distribution



Formula for the *F* Test

$$F = \frac{s_1^2}{s_2^2}$$

where s_1^2 *is the larger of the two variances.*

numerator degrees of freedom $= n_1 - 1$

denominator degrees of freedom $= n_2 - 1$

n_1 *is the sample size from which the larger variance was obtained.*

Assumptions for Testing the Difference between Two Variances

- The populations from which the samples were obtained must be normally distributed.
- The samples must be independent of each other.

Testing the Difference between Two Variances - Example

- A researcher wishes to see whether the variances of the heart rates (in beats per minute) of smokers are different from the variances of heart rates of people who do not smoke. Two samples are selected, and the data are given on the next slide. Using $\alpha = 0.05$, is there enough evidence to support the claim?

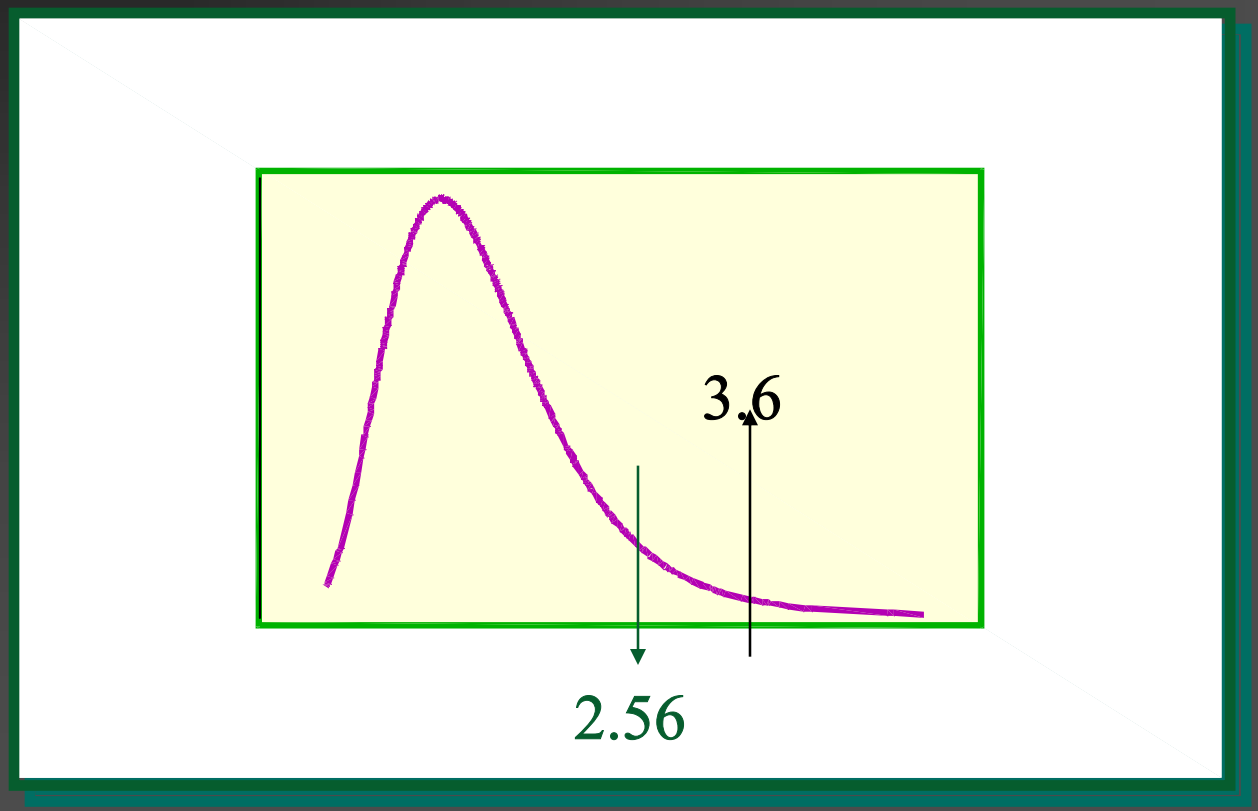
Testing the Difference between Two Variances - Example

- For smokers $n_1 = 26$ and $s_1^2 = 36$; for nonsmokers $n_2 = 18$ and $s_2^2 = 10$.
- Step 1: State the hypotheses and identify the claim.
  $H_0: \sigma_1^2 = \sigma_2^2$ $H_1: \sigma_1^2 \neq \sigma_2^2$ (claim)

Testing the Difference between Two Variances - Example

- Step 2: Find the critical value. Since α = 0.05 and the test is a two-tailed test, use the 0.025 table. Here d.f. N. = 26 – 1 = 25, and d.f.D. = 18 – 1 = 17. The critical value is F = 2.56.
- Step 3: Compute the test value. $F = s_1^2 / s_2^2 = 36/10 = 3.6$.

Testing the Difference between Two Variances - Example

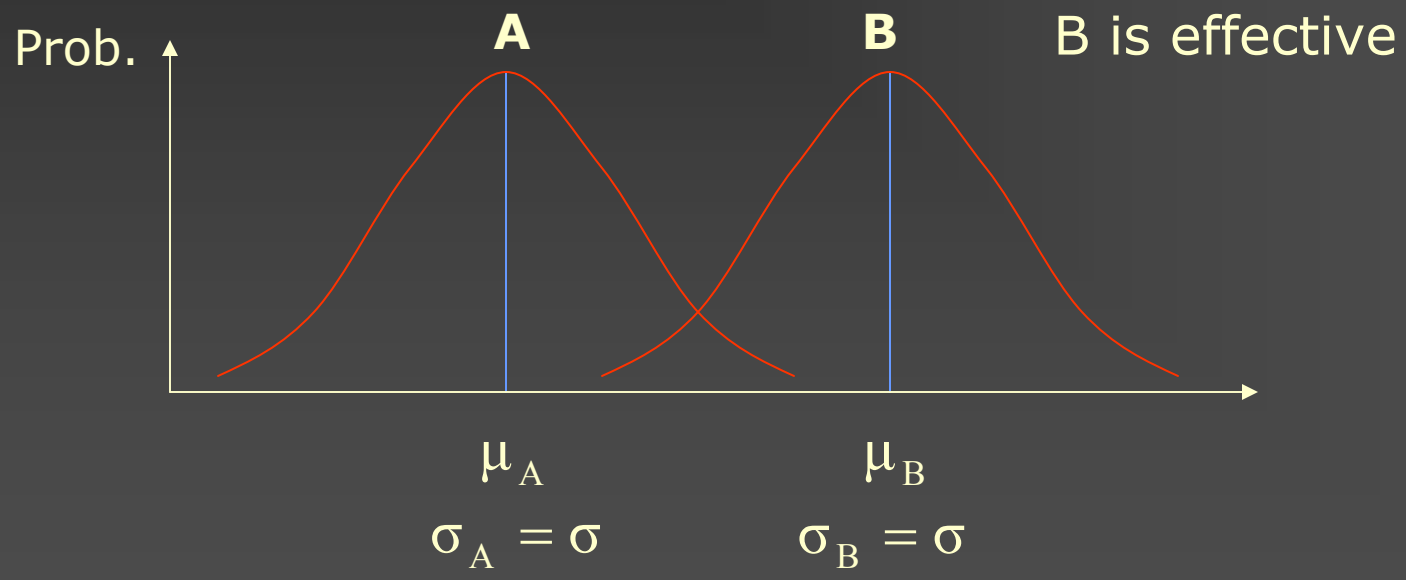


Testing the Difference between Two Variances - Example

- Step 4: Make the decision. Reject the null hypothesis, since 3.6 > 2.56.
- Step 5: Summarize the results. There is enough evidence to support the claim that the variances are different.

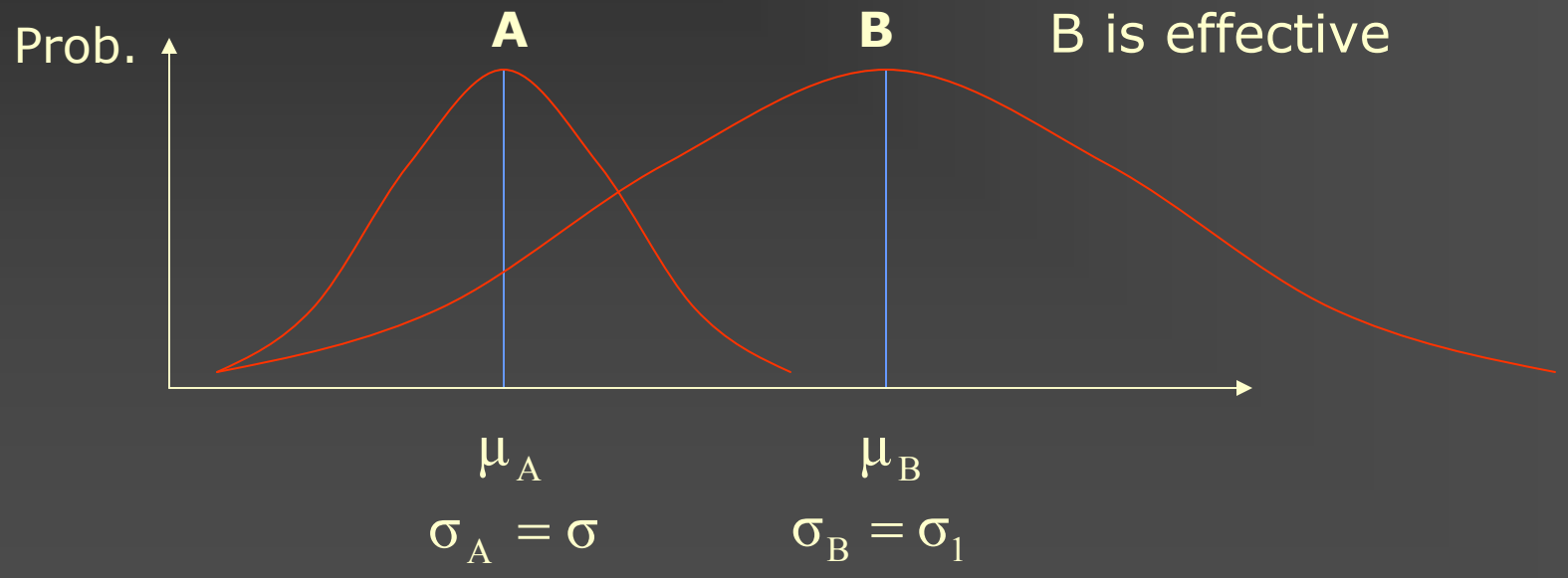
Example

- Suppose Group A received standard treatment and Group B received new treatment with a significant different.



Example

- Group B is still significant but with a significantly larger standard deviation. Is the new treatment still effective to all patients?



Test the difference between two means

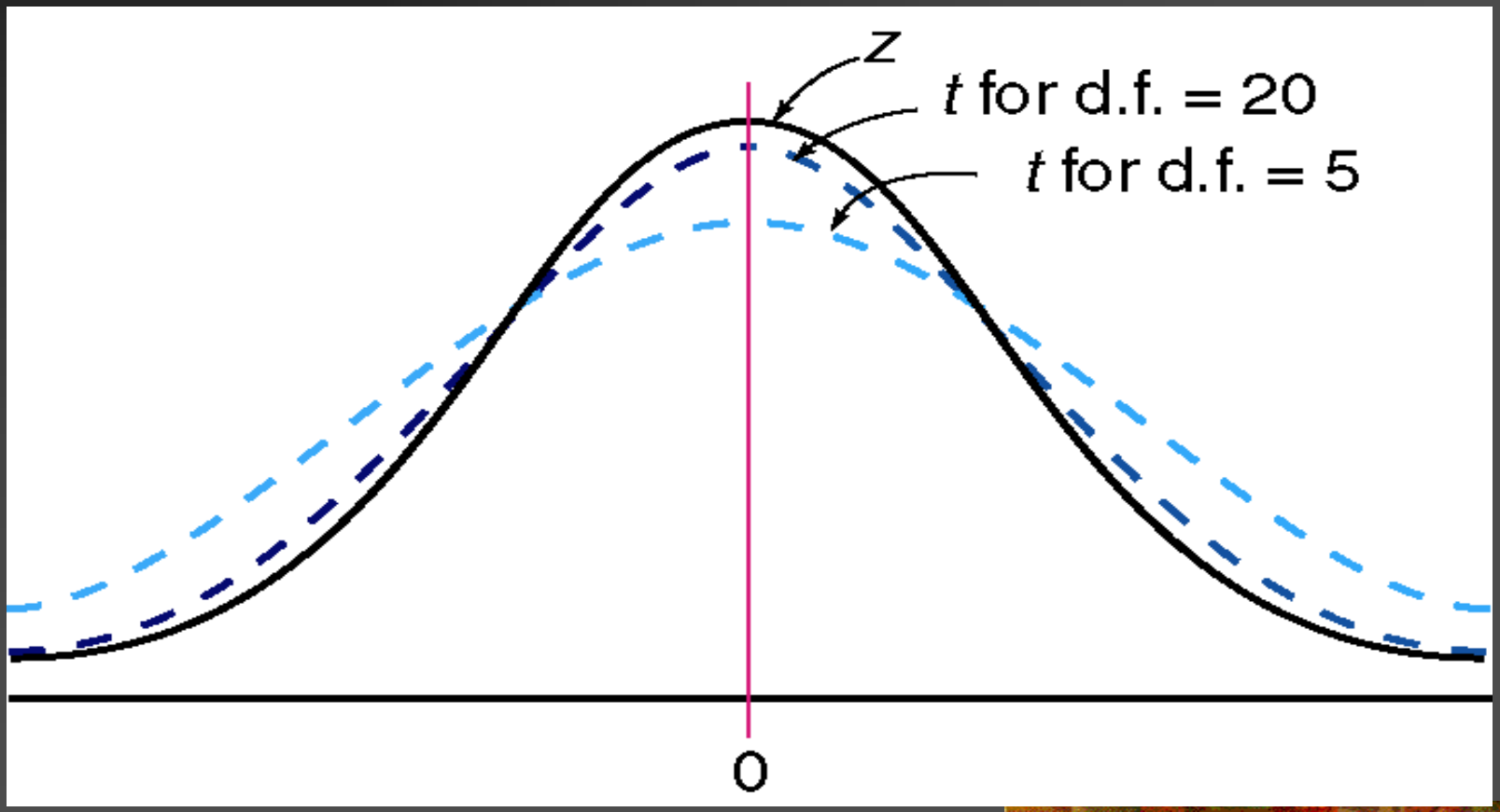
Testing the Difference between Two Means

- When the sample sizes are small (< 30) and the population variances are unknown, a *t* test is used to test the difference between means.
- The two samples are assumed to be independent and the sampling populations are normally or approximately normally distributed.

Characteristics of t-distribution

- The variance is greater than 1
- The *t* distribution is actually a family of curves based on the concept of *degrees of freedom*
- As the sample size increases, the *t* distribution approaches the standard normal distribution

Standard Normal Curve and the *t* Distribution



Testing the Difference between Two Means

- There are two options for the use of the *t* test.
- When the variances of the populations are equal and when they are not equal.
- The *F* test can be used to establish whether the variances are equal or not.

Testing the Difference between Two Means

Unequal Variances

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}};$$

$$d.f. = smaller\ of\ n_1 - 1\ or\ n_2 - 1$$

Testing the Difference between Two Means

Equal Variances

$$s_{pooled} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{s_{pooled}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}; \qquad d.f. = n_1 + n_2 - 2$$

Difference between Two Means: - Example

- The average protein level for a high protein diet is 199, and the average protein level for a normal diet is 191. Assume the data were obtained from two independent samples with standard deviations of 12 and 38 respectively, and the sample sizes are 10 subjects from and 8 subjects respectively. Can it be concluded at α = 0.05 that the average protein levels in the two groups is different?

Difference between Two Means: - Example

- Assume the populations are normally distributed.
- First we need to use the *F* test to determine whether or not the variances are equal.
- The critical value for the *F* test for $\alpha = 0.05$ is 4.20.
- The test value = $38^2/12^2 = 10.03$.
- Since 10.03 > 4.20, the decision is to reject the null hypothesis and conclude the variances are not equal.

Difference between Two Means: - Example

- **Step 1:** State the hypotheses
  $H_0: \mu_1 = \mu_2$ $H_1: \mu \neq \mu_2$
- **Step 2:** Find the critical values. Since $\alpha = 0.05$ and the test is a two-tailed test, the critical values are $t_{0.025,7} = -2.365$ and $+2.365$ with d.f. = 8 – 1 = 7.
- **Step 3:** Compute the test value. Substituting in the formula for the test value when the variances are not equal gives $t = 0.57$.

$$t = \frac{(199 - 191) - 0}{\sqrt{\frac{12^2}{10} + \frac{38^2}{8}}} = 0.57$$

Difference between Two Means: - Example

- Step 4: Make the decision. Do not reject the null hypothesis, since 0.57 < 2.365.
- Step 5: Summarize the results. There is not enough evidence to support the claim that the average protein level is different.
- Note: If the variances were equal - use the other test value formula.

Confidence Intervals for the Difference of Two Means

Unequal Variances

$$(\bar{X}_1 - \bar{X}_2) - t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$< \mu_1 - \mu_2 <$$

$$(\bar{X}_1 - \bar{X}_2) + t_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\text{d. f.} = \textit{smaller} \quad \textit{of} \; n_1 - 1 \quad \textit{or} \quad n_2 - 1$$

Confidence Intervals for the Difference of Two Means

Equal Variances

$$(\bar{X}_1 - \bar{X}_2) - t_{\alpha/2} \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$< \mu_1 - \mu_2 <$$

$$(\bar{X}_1 - \bar{X}_2) + t_{\alpha/2} \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}} \cdot \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

$$\text{d.f.} = n_1 + n_2 - 2.$$

Confidence Intervals for the Difference of Two Means

- The 95% confidence interval for the difference of two means

$$(199-191)-2.365\sqrt{\frac{12^2}{10}+\frac{38^2}{8}}<\mu_1-\mu_2<(199-191)-2.365\sqrt{\frac{12^2}{10}+\frac{38^2}{8}}$$

$$-25.02<\mu_1-\mu_2<41.02$$

SPSS sample output – two sample t-test

Group Statistics

	Smoking habit	N	Mean	Std. Deviation	Std. Error Mean
Pre exercise pulse rate	Yes	16	76.75	11.997	2.999
	No	24	75.42	7.723	1.576

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Pre exercise pulse rate	Equal variances assumed	8.665	.006	.429	38	.671	1.333	3.111	-4.965	7.631
	Equal variances not assumed			.394	23.274	.698	1.333	3.388	-5.671	8.338

Scatter Plots

- A scatter plot is a graph of the ordered pairs *(x, y)* of numbers consisting of the independent variable, *x*, and the dependent variable, *y*.

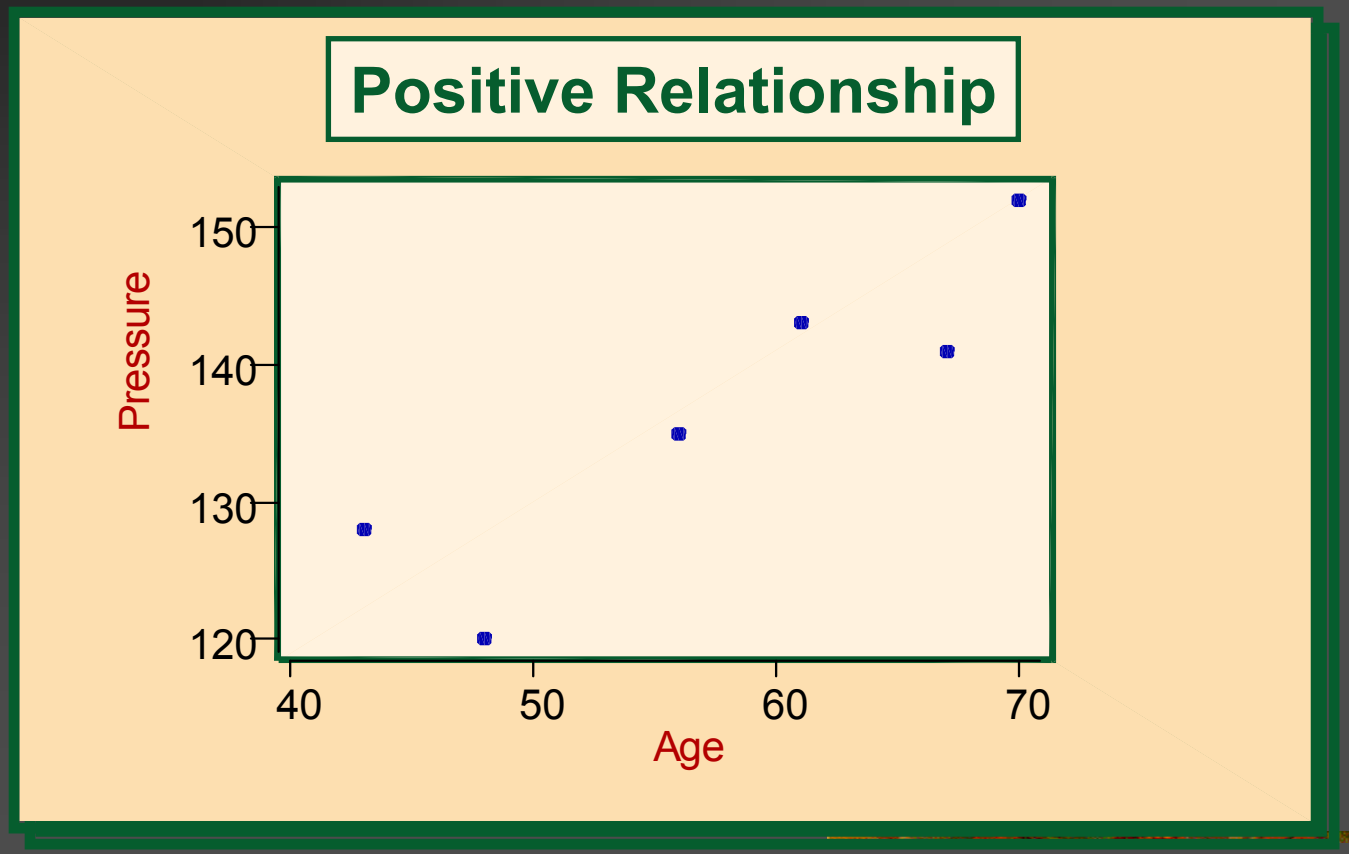
Scatter Plots - Example

- Construct a scatter plot for the data obtained in a study of age and systolic blood pressure of six randomly selected subjects.
- The data is given on the next slide.

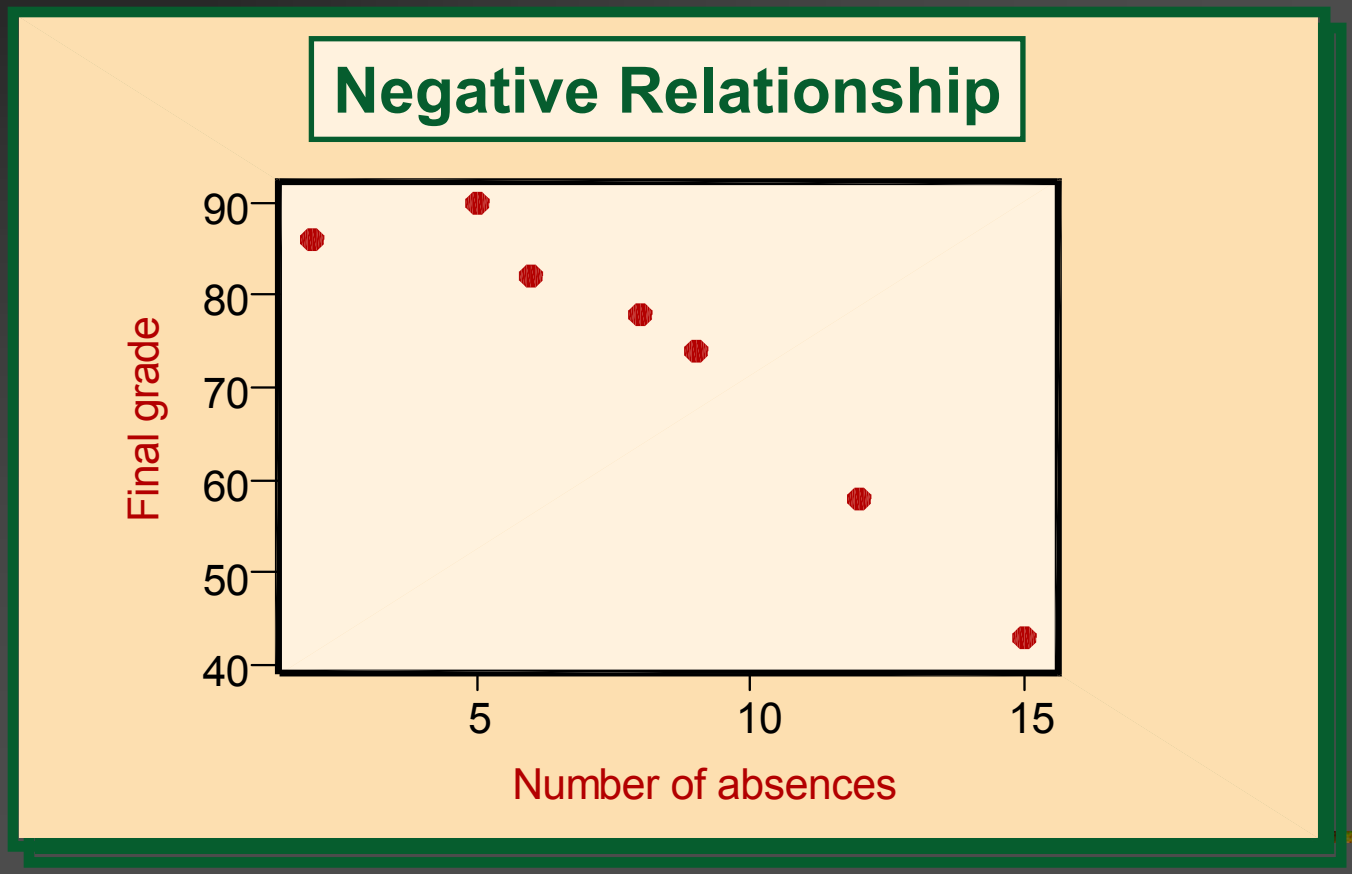
Scatter Plots - Example

Subject	Age, x	Pressure, y
A	43	128
B	48	120
C	56	135
D	61	143
E	67	141
F	70	152

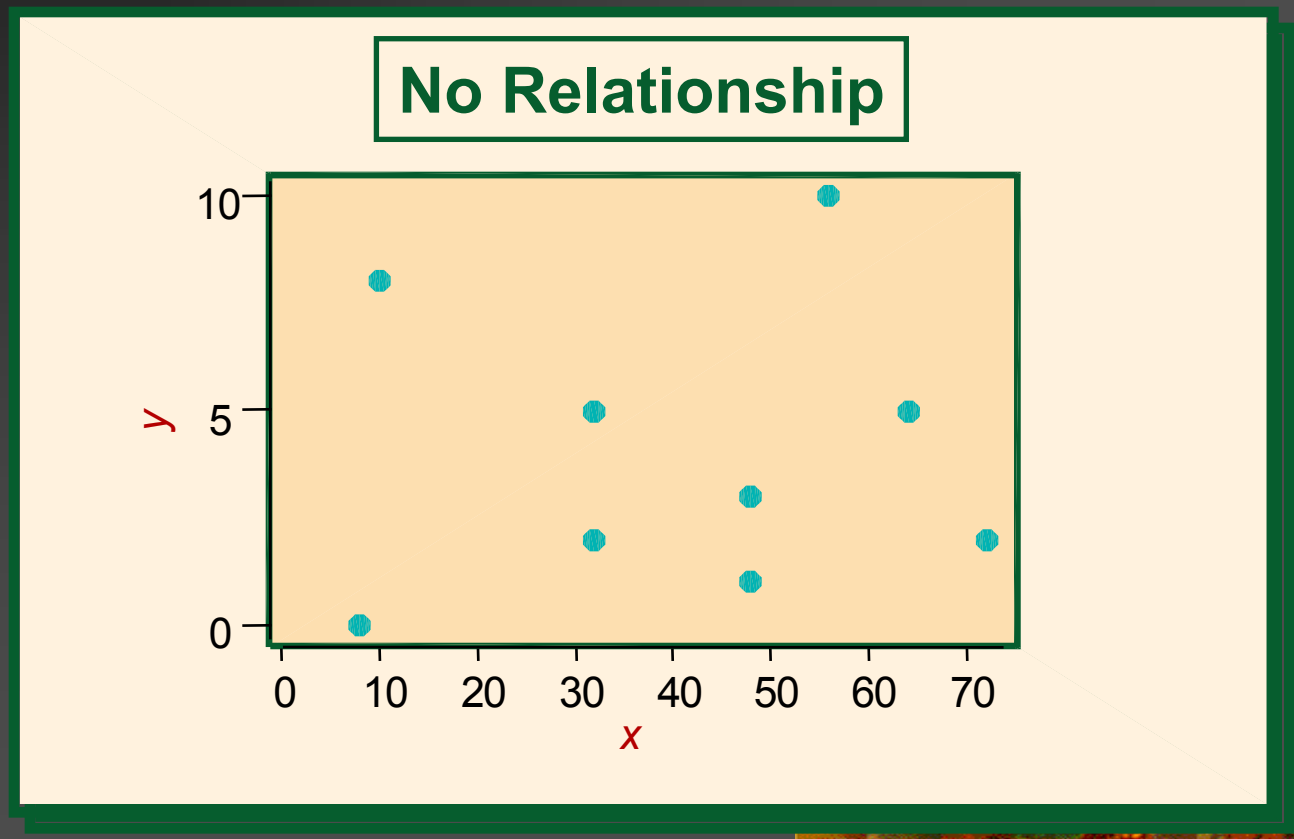
Scatter Plots - Example



Scatter Plots - Other Examples



Scatter Plots - Other Examples



Correlation Coefficient

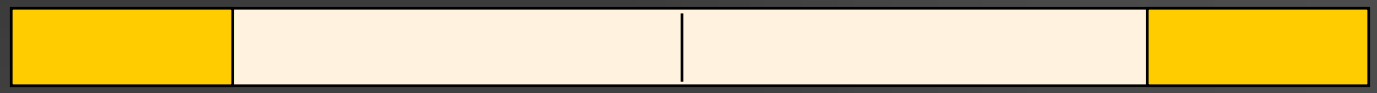
- The correlation coefficient computed from the sample data measures the strength and direction of a relationship between two variables.
- Sample correlation coefficient, r.
- Population correlation coefficient, ρ.

Range of Values for the Correlation Coefficient

Strong negative relationship

No linear relationship

Strong positive relationship



−1 0 +1

Formula for the Correlation Coefficient *r*

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{\left[n(\Sigma x^2) - (\Sigma x)^2\right]\left[n(\Sigma y^2) - (\Sigma y)^2\right]}}$$

Where n is the number of data pairs

Correlation Coefficient - Example

- Compute the correlation coefficient for the age and blood pressure data.

$$\sum x = 345,\ \sum y = 819,\ \sum xy = 47{,}634$$

$$\sum x^2 = 20{,}399,\ \sum y^2 = 112{,}443.$$

Substituting in the formula for r gives

$$r = 0.897.$$

The Significance of the Correlation Coefficient

- The population corelation coefficient, ρ, is the correlation between all possible pairs of data values *(x, y)* taken from a population.

The Significance of the Correlation Coefficient

- $H_0: \rho = 0$ $H_1: \rho \neq 0$
- This tests for a significant correlation between the variables in the population.

Formula for the *t* tests for the Correlation Coefficient

$$t = \sqrt{\frac{n-2}{1-r^2}}$$

$$with \text{ d. f.} = n-2$$

Example

- Test the significance of the correlation coefficient for the age and blood pressure data. Use $\alpha = 0.05$ and $r = 0.897$.
- Step 1: State the hypotheses.
- $H_0: \rho = 0$ $H_1: \rho \neq 0$

Example

- Step 2: Find the critical values. Since $\alpha = 0.05$ and there are 6 – 2 = 4 degrees of freedom, the critical values are $t = +2.776$ and $t = -2.776$.
- Step 3: Compute the test value. $t = 4.059$ (verify).

Example

- Step 4: Make the decision. Reject the null hypothesis, since the test value falls in the critical region (4.059 > 2.776).
- Step 5: Summarize the results. There is a significant relationship between the variables of age and blood pressure.

Regression

- The scatter plot for the age and blood pressure data displays a linear pattern.
- We can model this relationship with a straight line.
- This regression line is called the line of best fit or the regression line.
- The equation of the line is $y = a + bx$.

Formulas for the Regression Line $y = a + bx$.

$$a = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}$$

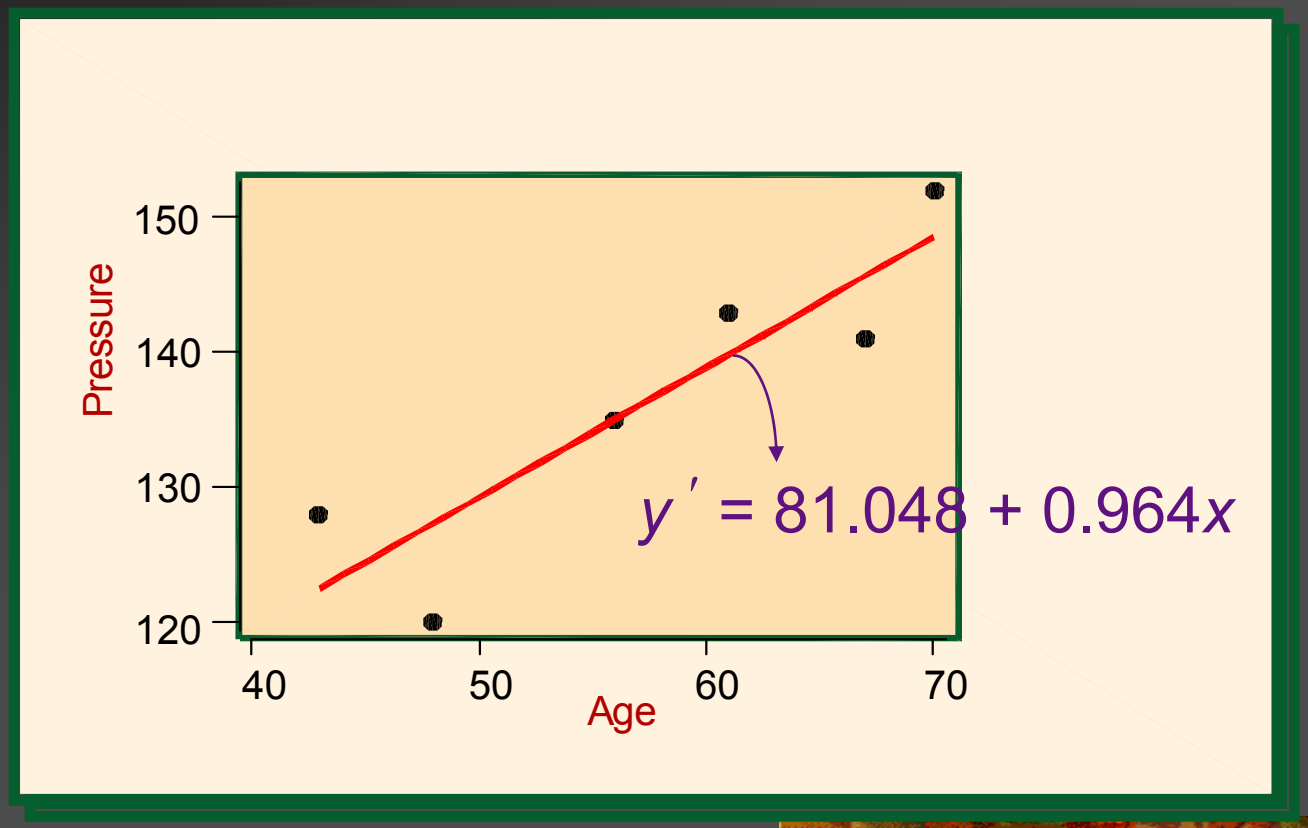
$$b = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}$$

Where a is the y intercept and b is the slope of the line.

Example

- Find the equation of the regression line for the age and the blood pressure data.
- Substituting into the formulas give $a = 81.048$ and $b = 0.964$ (verify).
- Hence, $y = 81.048 + 0.964x$.
- Note, *a* represents the intercept and *b* the slope of the line.

Example



Using the Regression Line to Predict

- The regression line can be used to predict a value for the dependent variable (y) for a given value of the independent variable (x).
- Caution: Use x values within the experimental region when predicting y values.

Example

- Use the equation of the regression line to predict the blood pressure for a person who is 50 years old.
- Since $y = 81.048 + 0.964x$, then $y = 81.048 + 0.964(50) = 129.248 \approx 129$.
- Note that the value of 50 is within the range of x values.

Coefficient of Determination and Standard Error of Estimate

- The coefficient of determination, denoted by r^2, is a measure of the variation of the dependent variable that is explained by the regression line and the independent variable.

Coefficient of Determination and Standard Error of Estimate

- r^2 is the square of the correlation coefficient.
- The coefficient of nondetermination is $(1 - r^2)$.
- Example: If $r = 0.90$, then $r^2 = 0.81$.

Tests Using Contingency Tables

- When data can be tabulated in table form in terms of frequencies, several types of hypotheses can be tested using the chi-square test.
- Two such tests are the independence of variables test and the homogeneity of proportions test.

Tests Using Contingency Tables

- The test of independence of variables is used to determine whether two variables are independent when a single sample is selected.
- The test of homogeneity of proportions is used to determine whether the proportions for a variable are equal when several samples are selected from different populations.

Test for Independence - Example

- Suppose a new postoperative procedure is administered to a number of patients in a large hospital.
- Question: Do the doctors feel differently about this procedure from the nurses, or do they feel basically the same way?
- Data is on the next slide.

Test for Independence - Example

Group	Prefer new procedure	Prefer old procedure	No preference
Nurses	100	80	20
Doctors	50	120	30

Test for Independence - Example

- The null and the alternative hypotheses are as follows:
- H_0: The opinion about the procedure is independent of the profession.
- H_1: The opinion about the procedure is dependent on the profession.

Test for Independence - Example

- If the null hypothesis is not rejected, the test means that both professions feel basically the same way about the procedure, and the differences are due to chance.
- If the null hypothesis is rejected, the test means that one group feels differently about the procedure from the other.

Test for Independence - Example

- Note: The rejection of the null hypothesis does not mean that one group favors the procedure and the other does not.
- The test value is the χ^2 value (same as the goodness-of-fit test value).
- The expected values are computed from: (row sum)×(column sum)/(grand total).

Test for Goodness of Fit - Formula

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$\text{d. f.} = number\ of\ categories - 1$$

$$O = observed\ frequency$$

$$E = expected\ frequency$$

Test for Independence - Example

```
TEST   for   INDEPENDENCE


Expected counts are printed below observed counts

               C1         C2         C3     Total
     1        100         80         20       200
            75.00     100.00     25.00

     2         50        120         30       200
            75.00     100.00     25.00

Total         150        200         50       400

Chi-Sq =   8.333 +   4.000 +   1.000 +
           8.333 +   4.000 +   1.000 =   26.667
DF = 2, P-Value = 0.000
```

Test for Independence - Example

- From the output, the *P*-value < 0.001
  Hence, the null hypothesis will be rejected.
- If the critical value approach is used, the degrees of freedom for the chi-square critical value will be (number of columns – 1)×(number of rows – 1).
- d.f. = (3 –1)(2 – 1) = 2.

Test for Homogeneity of Proportions

- Here, samples are selected from several different populations and one is interested in determining whether the proportions of elements that have a common characteristic are the same for each population.

Test for Homogeneity of Proportions

- The sample sizes are specified in advance, making either the row totals or column totals in the contingency table known before the samples are selected.
- The hypotheses will be:
  H_0: $p_1 = p_2 = \ldots = p_k$
  H_1: At least one proportion is different from the others.

Test for Homogeneity of Proportions

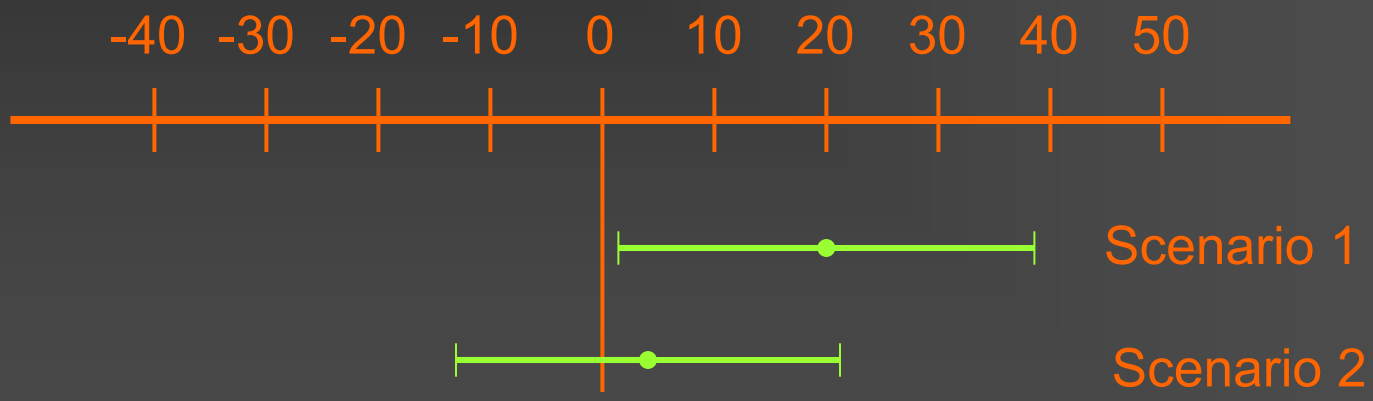
- The computations for this test are the same as that for the test of independence.

Confidence Intervals

- How do we interpret a confidence interval?
- What is the interpretation if the confidence interval for the difference between two groups overlap zero?

95% Confidence Interval of Complete Response Rate

(Difference between arm A and B)



Outline

- Session I
  - Difference between two means
  - Difference between two variances
  - Scatter plots and correlations
  - Regression
  - Chi-square
- Session II
  - Survival analysis
  - Covariate adjustment
  - Subgroup analysis
  - Intention-to-treat