MATH 3339 - 03 15951
Statistics for the Sciences
Numerical description; Jointly Distributed Variables

Wendy Wang
wwang60@central.uh.edu

Lecture 10 - 3339
Outline

1. Measures of Variability
2. Jointly Distributed Data
3. Scatterplots for Jointly Distributed Variables
4. Covariance and Correlation
Popper Set Up

- Fill in all of the proper bubbles.
- Make sure your ID number is correct.
- Make sure the filled in circles are very dark.
- This is popper number 05.
Sample Standard Deviation of Section A test scores

- Sample standard deviation is $s = 4.77$.
- This implies that from the sample of the 10 students from section A the tests scores has a spread, on average, of 4.77 points from the mean of 71.50 points.
Multiply by 2, Popper question 1

For the following dataset the mean is \( \bar{x} = 4.5 \), the variance is \( s^2 = 3.5 \) and the standard deviation is \( s = 1.870829 \).

\[
3, \quad 6, \quad 2, \quad 7, \quad 4, \quad 5
\]

Now, multiply each value by 2. What is the new mean, new variance and the new standard deviation?

a) 4.5, 14, 3.5

b) 4.5, 3.5, 1.870829

c) 9, 14, 3.7416

d) 9, 7, 3.7416
Calculating Standard Deviation

- For larger data sets use a calculator or computer software.
- Each calculator is different if you cannot determine how to compute standard deviation from your calculator ask your instructor.
- For this course we will be using R as the software.
- The function for the sample standard deviation in R is \( \text{sd}(\text{data name}$variable name\)).
2. This is a standard deviation contest, which list of numbers have the largest standard deviation? No calculations are required.
   a) 10, 10, 10, 10
   b) 20, 20, 20, 20
   c) 10, 10, 20, 20
   d) 10, 15, 15, 20
Measures of Variability: Coefficient of Variation

- This is to compare the variation between two groups.

- The **coefficient of variation** (cv) is the ratio of the standard deviation to the mean.

\[
\text{cv} = \frac{sd}{\text{mean}}
\]

- A smaller ratio will indicate less variation in the data.
## CV of test scores

<table>
<thead>
<tr>
<th></th>
<th>Section A</th>
<th>Section B</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sample Size</strong></td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td><strong>Sample Mean</strong></td>
<td>71.5</td>
<td>71.5</td>
</tr>
<tr>
<td><strong>Sample Standard Deviation</strong></td>
<td>4.770</td>
<td>18.22</td>
</tr>
<tr>
<td><strong>CV</strong></td>
<td>( \frac{4.77}{71.5} = 0.066 )</td>
<td>( \frac{18.22}{71.5} = 0.2548 )</td>
</tr>
</tbody>
</table>
The following statistics were collected on two different groups of stock prices:

<table>
<thead>
<tr>
<th></th>
<th>Portfolio A</th>
<th>Portfolio B</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample size</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td>Sample mean</td>
<td>$52.65</td>
<td>$49.80</td>
</tr>
<tr>
<td>Sample standard deviation</td>
<td>$6.50</td>
<td>$2.95</td>
</tr>
</tbody>
</table>

What can be said about the variability of each portfolio?
Motivating examples

- In the housing market, for a larger size house, the price of the house increases.

- In a stadium, is the number of hot dogs sold related to the number of sodas sold?

- A survey wants to know if there is a relationship between age and health care cost.

- An insurance wants to know if there is a relationship between color of a car and the number of accidents.
Jointly Distributed Data

- **Jointly Distributed Data** is data for two different variables and we want to know the relationship between these two variables.

- Several applications looks at the relationship between two variables.

- The data can be quantitative or categorical.

- If both variables are categorical we can look at bar graphs and cross tabulations to determine if there is a relationship.

- If one variable is quantitative and another variable is categorical we can use the side-by-side box plot to look at the relationship.

- If both variables are quantitative we look at the scatter plot to determine if there is a relationship.
A **response variable** measures an outcome of a study. Sometimes called a dependent variable. Usually the $y$-variable.

An **explanatory variable** explains or influences changes in a response variable. Sometimes called an independent variable or predictor or factor. Usually the $x$-variable.
In the housing market for a larger size house, the price of the house increases.

- The **response variable** is the price of the house.
- The **explanatory variable** is the size of the house.
For each of the scenarios determine the response variable.

3. A survey wants to know if there is a relationship between age and health care cost.
   a) age  b) health care cost  c) Not enough information

4. In a stadium, is the number of hot dogs sold related to the number of sodas sold?
   a) Number of hot dogs sold  b) Number of sodas sold  c) Not enough information
Plotting the data: Scatterplots

- The best way to initially observe the relationship between two **quantitative** variables measured on the same individuals.

- The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis.

- Each individual in the data appears as the point in the plot fixed by the values of both variables for that individual.

- Scatterplots can show if there is some kind of association between the two quantitative variables.

- To create a scatterplot in R `plot(explanatory,response)`
We want to look at the relationship between the estimated cost of the car and age of the vehicle.

<table>
<thead>
<tr>
<th>Vehicle</th>
<th>Estimated Cost</th>
<th>Age</th>
</tr>
</thead>
<tbody>
<tr>
<td>Honda Insight</td>
<td>$5,555</td>
<td>8</td>
</tr>
<tr>
<td>Toyota Prius</td>
<td>$17,888</td>
<td>3</td>
</tr>
<tr>
<td>Toyota Prius</td>
<td>$9,963</td>
<td>6</td>
</tr>
<tr>
<td>Toyota Echo</td>
<td>$6,793</td>
<td>5</td>
</tr>
<tr>
<td>Honda Civic Hybrid</td>
<td>$10,774</td>
<td>5</td>
</tr>
<tr>
<td>Honda Civic Hybrid</td>
<td>$16,310</td>
<td>2</td>
</tr>
<tr>
<td>Chevrolet Prizm</td>
<td>$2,475</td>
<td>8</td>
</tr>
<tr>
<td>Mazda Protege</td>
<td>$2,808</td>
<td>10</td>
</tr>
<tr>
<td>Toyota Corolla</td>
<td>$7,073</td>
<td>9</td>
</tr>
<tr>
<td>Acura Integra</td>
<td>$8,978</td>
<td>8</td>
</tr>
<tr>
<td>Scion xB</td>
<td>$11,213</td>
<td>2</td>
</tr>
<tr>
<td>Scion xA</td>
<td>$9,463</td>
<td>3</td>
</tr>
<tr>
<td>Mazda3</td>
<td>$15,055</td>
<td>2</td>
</tr>
<tr>
<td>Mini Cooper</td>
<td>$20,705</td>
<td>2</td>
</tr>
</tbody>
</table>
Creating a Scatterplot in R

\texttt{plot(x,y)}

\texttt{> plot(carsreg$Age,carsreg$Cost)}
Interpreting scatterplots

- **Look for Direction**: A pattern that runs from the upper left to the lower right is said to have a negative direction. A trend that is running the other way has a positive direction.

- **Look for Form**: If there is a straight line relationship, it will appear as a cloud or swarm of points stretched out in a generally consistent, straight form. This is linear form.

- **Look for Strength**: How much scatter the plot has. Do the points appear to follow a single stream? This is a very strong association. Or does the swarm of points seem to form a vague cloud through which we can barely discern any trend or pattern? This is a weak association. Look for the unexpected.
Example 2

Suppose we want to know if there is an association between the number of spaces a property is from GO and the cost of the property in a monopoly game.
Example 3

Suppose we want to know if there is an association between the height of the father and their child.
Covariance

If we have two quantitative variables from a sample, then their covariance is calculated by:

\[
\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})
\]

In R: `cov(x,y)`

> cov(carsreg$Age,carsreg$Cost)
[1] -13254.98

However, the strength of the relationship cannot be measured by the covariance, thus we divide each deviation in the sum by the standard deviation of that variable. The result is call the correlation

\[
\text{cor}(x, y) = \frac{\text{cov}(x, y)}{sd(x) \times sd(y)}
\]

In R: `cor(x,y)`
Facts About Correlation

- The correlation coefficient is a value that measures the **direction** and **strength** of a **straight-line** relationship between two quantitative variables.

- Correlation makes no distinction between explanatory and response variables.

- Correlation requires that both variables be quantitative.

- Because $r$ uses the standardize values of the observations, $r$ does not change when we change the units of measurement of $x$, $y$ or both.

- Positive $r$ indicates positive association between the variables, and negative $r$ indicates negative association.
Facts About Correlation

- The $r$ is always a number between $-1$ and 1. Values close to 0 indicate a very weak linear relationship. If $r$ is close to $-1$ the association is a very strong negative linear relationship. If $r$ is close to 1 the association is a very strong positive association.

- Correlation measures the strength of only a linear relationship.

- $r$ is strongly affected by a few outlying observations.
Scatterplot of Height vs Income

Income

Height

$\text{Scatterplot of Height vs Income}$

$r = -0.078$
$r = -0.707$

Scatterplot of Reliability vs Price

- Price: 32000, 34000, 36000, 38000, 40000, 42000
- Reliability: 1, 2, 3, 4, 5

Wendy Wang  wwang60@central.uh.edu

MATH 3339  Lecture 10 - 3339  28 / 29
$r = 0.925$