"Correlation is not Causation" but **Correlation** is so beautiful!

Credit: wikipedia

Hongye Liu, Teaching Assistant Prof, CS361, UIUC, 9.1.2021
Last time

- Mean
- Standard deviation
- Variance
- Standardizing data
- Median,
- Interquartile
- Mode
Objectives

- Boxplots, comparison btw summary statistics
- Scatter plots, Correlation Coefficient
- Visualizing relationships
  Heatmap, 3D bar, Time series plots,
Box plots

- Boxplots
  - Simpler than histogram
  - Good for outliers
  - Easier to use for comparison

Vehicle death by region

Data from https://www2.stetson.edu/~jrasp/data.htm
How to define outliers? (the default)

Boxplots details, outliers

- Whisker
- Box
- Median

Interquartile Range (iqr)

Outlier

< 1.5 iqr

> 1.5 iqr

25 percentile
\[ \sum_{i=1}^{N} \frac{x_i}{N} \]

\[ \sum_{i=1}^{N+1} x_i + x_i \]

\[ 251 \]
Sensitivity of summary statistics to outliers

- mean and standard deviation are very sensitive to outliers
- median and interquartile range are not sensitive to outliers
Modes

- Modes are peaks in a histogram.
- If there are more than 1 mode, we should be curious as to why.
Multiple modes

We have seen the “iris” data which looks to have several peaks

Data: “iris” in R
Example Bi-modes distribution

Modes may indicate multiple populations

Data: Erythrocyte cells in healthy humans

Piagnerelli, JCP 2007
Tails and Skews

Symmetric Histogram (smoothed)

- mode, median, mean, all on top of one another

Left Skew

- left tail
- mean
- median
- mode
- right tail

Right Skew

- left tail
- mode
- median
- mean
- right tail

Credit: Prof. Forsyth
Q. How is this skewed?

A. Left
B. Right

Median = 47

mean = 46
Relationship between data features

Example: Does the weight of people relate to their height?

<table>
<thead>
<tr>
<th>IDNO</th>
<th>BODYFAT</th>
<th>DENSITY</th>
<th>AGE</th>
<th>WEIGHT</th>
<th>HEIGHT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12.6</td>
<td>1.0708</td>
<td>23</td>
<td>154.25</td>
<td>67.75</td>
</tr>
<tr>
<td>2</td>
<td>6.9</td>
<td>1.0853</td>
<td>22</td>
<td>173.25</td>
<td>72.25</td>
</tr>
<tr>
<td>3</td>
<td>24.6</td>
<td>1.0414</td>
<td>22</td>
<td>154.00</td>
<td>66.25</td>
</tr>
<tr>
<td>4</td>
<td>10.9</td>
<td>1.0751</td>
<td>26</td>
<td>184.75</td>
<td>72.25</td>
</tr>
<tr>
<td>5</td>
<td>27.8</td>
<td>1.0340</td>
<td>24</td>
<td>184.25</td>
<td>71.25</td>
</tr>
<tr>
<td>6</td>
<td>20.6</td>
<td>1.0502</td>
<td>24</td>
<td>210.25</td>
<td>74.75</td>
</tr>
<tr>
<td>7</td>
<td>19.0</td>
<td>1.0549</td>
<td>26</td>
<td>181.00</td>
<td>69.75</td>
</tr>
<tr>
<td>8</td>
<td>12.8</td>
<td>1.0704</td>
<td>25</td>
<td>176.00</td>
<td>72.50</td>
</tr>
<tr>
<td>9</td>
<td>5.1</td>
<td>1.0900</td>
<td>25</td>
<td>191.00</td>
<td>74.00</td>
</tr>
<tr>
<td>10</td>
<td>12.0</td>
<td>1.0722</td>
<td>23</td>
<td>198.25</td>
<td>73.50</td>
</tr>
</tbody>
</table>

x : HIGHT, y: WEIGHT
Scatter plot

Body Fat data set
Scatter plot

Scatter plot with density
Scatter plot

_removed of outliers & standardized_
Correlation seen from scatter plots

Zero Correlation

Positive correlation

Negative correlation

Credit: Prof. Forsyth
What kind of Correlation?

- Line of code in a database and number of bugs +
- Frequency of hand washing and number of germs on your hands -
- GPA and hours spent playing video games -
- earnings and happiness 0

Credit: Prof. David Varodayan
Correlation doesn’t mean causation

» Shoe size is correlated to reading skills, but it doesn’t mean making feet grow will make one person read faster.
Correlation Coefficient

Given a data set \( \{(x_i, y_i)\} \) consisting of items \((x_1, y_1) \ldots (x_N, y_N)\),

- Standardize the coordinates of each feature:
  \[
  \hat{x}_i = \frac{x_i - \text{mean}(\{x_i\})}{\text{std}(\{x_i\})} \\
  \hat{y}_i = \frac{y_i - \text{mean}(\{y_i\})}{\text{std}(\{y_i\})}
  \]

- Define the correlation coefficient as:
  \[
  \text{corr}(\{(x_i, y_i)\}) = \frac{1}{N} \sum_{i=1}^{N} \hat{x}_i \hat{y}_i
  \]
Correlation Coefficient

\[ \hat{x}_i = \frac{x_i - \text{mean}(\{x_i\})}{\text{std}(\{x_i\})} \quad \hat{y}_i = \frac{y_i - \text{mean}(\{y_i\})}{\text{std}(\{y_i\})} \]

\[ \text{corr}(\{(x_i, y_i)\}) = \frac{1}{N} \sum_{i=1}^{N} \hat{x}_i \hat{y}_i \]

\[ = \text{mean}(\{\hat{x}_i \hat{y}_i\}) \]
Q: Correlation Coefficient

Which of the following describe(s) correlation coefficient correctly?

A. It’s unitless

B. It’s defined in standard coordinates

C. Both A & B

\[ \text{corr}(\{(x_i, y_i)\}) = \frac{1}{N} \sum_{i=1}^{N} \hat{x}_i \hat{y}_i \]
A visualization of correlation coefficient


In a data set \{ (x_i, y_i) \} consisting of items

\[(x_1, y_1) \ldots (x_N, y_N),\]

\[corr(\{(x_i, y_i)\}) > 0 \text{ shows positive correlation}\]

\[corr(\{(x_i, y_i)\}) < 0 \text{ shows negative correlation}\]

\[corr(\{(x_i, y_i)\}) = 0 \text{ shows no correlation}\]
The Properties of Correlation Coefficient

The correlation coefficient is symmetric

\[ \text{corr}\left(\{(x_i, y_i)\}\right) = \text{corr}\left(\{(y_i, x_i)\}\right) \]

Translating the data does **NOT** change the correlation coefficient
Scaling the data may change the sign of the correlation coefficient

\[
corr\left(\left\{ (a x_i + b, c y_i + d) \right\}\right) = \text{sign}(a \times c) \times corr\left(\left\{ (x_i, y_i) \right\}\right)
\]
The Properties of Correlation Coefficient

The correlation coefficient is bounded within \([-1, 1]\)

\[
corr(\{(x_i, y_i)\}) = 1 \quad \text{if and only if} \quad \hat{x}_i = \hat{y}_i
\]

\[
corr(\{(x_i, y_i)\}) = -1 \quad \text{if and only if} \quad \hat{x}_i = -\hat{y}_i
\]
Which of the following has correlation coefficient equal to 1?

A. Left and right
B. Left
C. Middle
\[ y = ax \]

\[ \hat{y} = \frac{\sum x \cdot y}{\sum x} = \frac{\sum x - \mu_x \sum y}{\sum x} = ax - \mu_y \cdot \mu_x \]
The correlation coefficient can be written as

$$corr(\{(x_i, y_i)\}) = \frac{1}{N} \sum_{i=1}^{N} \hat{x}_i \hat{y}_i$$

It's the inner product of two vectors

$$\left\langle \frac{\hat{x}_1}{\sqrt{N}}, \ldots, \frac{\hat{x}_N}{\sqrt{N}} \right\rangle \quad \text{and} \quad \left\langle \frac{\hat{y}_1}{\sqrt{N}}, \ldots, \frac{\hat{y}_N}{\sqrt{N}} \right\rangle$$
Inner product

**Inner product’s geometric meaning:**

\[ |\nu_1| |\nu_2| \cos(\theta) \]

**Lengths of both vectors**

\[ \nu_1 = \left\langle \frac{x_1}{\sqrt{N}}, \ldots, \frac{x_N}{\sqrt{N}} \right\rangle \quad \nu_2 = \left\langle \frac{y_1}{\sqrt{N}}, \ldots, \frac{y_N}{\sqrt{N}} \right\rangle \]

are 1
Bound of correlation coefficient

\[ |\text{corr}(\{(x_i, y_i)\})| = |\cos(\theta)| \leq 1 \]

\[ \mathbf{v}_1 = \left\langle \frac{\hat{x}_1}{\sqrt{N}}, \ldots, \frac{\hat{x}_N}{\sqrt{N}} \right\rangle \quad \mathbf{v}_2 = \left\langle \frac{\hat{y}_1}{\sqrt{N}}, \ldots, \frac{\hat{y}_N}{\sqrt{N}} \right\rangle \]
The Properties of Correlation Coefficient

- Symmetric
- Translating invariant
- Scaling only may change sign
- bounded within [-1, 1]
Using correlation to predict

Caution! Correlation is **NOT** Causation

Math doctorates awarded correlates with
Uranium stored at US nuclear power plants

Credit: Tyler Vigen
How do we go about the prediction?

- Removed of outliers & standardized
Using correlation to predict

Given a correlated data set \( \{ (x_i, y_i) \} \)
we can predict a value \( y_0 \) that goes with \( x_0 \) a value.

In standard coordinates \( \{ (\hat{x}_i, \hat{y}_i) \} \)
we can predict a value \( \hat{y}_0 \) that goes with \( \hat{x}_0 \) a value.
Which coordinates will you use for the predictor using correlation?

A. Standard coordinates
B. Original coordinates
C. Either
We will assume that our predictor is linear

\[ \hat{y}^p = a \hat{x} + b \]

We denote the prediction at each \( \hat{x}_i \) in the data set as \( \hat{y}_i^p \)

\[ \hat{y}_i^p = a \hat{x}_i + b \]

The error in the prediction is denoted \( u_i \)

\[ u_i = \hat{y}_i - \hat{y}_i^p = \hat{y}_i - a \hat{x}_i - b \]
Require the mean of error to be zero

We would try to make the mean of error equal to zero so that it is also centered around 0 as the standardized data:

\[
\text{mean}\left(\{u_i\}\right) = 0 \quad \checkmark
\]

\[
\text{mean}\left(\{u_i\}\right) = \text{mean}(\{\hat{y} - \hat{y}^p\})
\]

Recall:

\[
\text{mean}\left(\{kx + b\}\right) = k \text{mean}(\{x\}) + b
\]

\[
\begin{align*}
= \text{mean}(\{\hat{y}\} - a\text{mean}(\{\hat{x}\}) - b) \\
= \text{mean}(\{\hat{y}\} - a\hat{x} - b) \\
= -b = 0 \\
\Rightarrow \quad b = 0
\end{align*}
\]
Require the variance of error is minimal

\[
\text{var}(\{ u_i^2 \}) = \text{mean}( \{ u_i - \text{mean}( \{ u_i \} )^2 \} = \text{mean}( \{ u_i^2 \} )
\]

\[
= \text{mean}( \{ (\hat{y} - \hat{y}_0)^2 \} ) \quad u_i = \hat{y} - \hat{y}_0
\]

\[
= \text{mean}( \{ (\hat{y} - \hat{a}\hat{x})^2 \} ) = \hat{y} - \hat{a}\hat{x}
\]

\[
= \text{mean}( \{ \hat{y}^2 - 2\hat{a}\hat{x}\hat{y} + \hat{a}^2\hat{x}^2 \} )
\]

\[
= \text{mean}(\{\hat{y}^2\} - 2\hat{a}\text{mean}(\{\hat{x}\hat{y}\}) + \hat{a}^2\text{mean}(\{\hat{x}^2\} )
\]

\[
= \text{mean}(\{\hat{y}^2\} )
\]

\[
= \text{mean}(\{(\hat{y} - \hat{y})^2\})
\]

\[
= \text{mean}(\{(\hat{y} - \text{mean}(\{\hat{y}\})\}^2\}
\]

\[
= \text{var}(\{\hat{y}\}) = 1
\]
Require the variance of error is minimal

\[
\text{var}\{u^3\} = \text{mean}\{\hat{y}^2\} - 2a\text{mean}\{\hat{x}\hat{y}\} \\
+ a^2\text{mean}\{\hat{x}^2\} \\
= 1 - 2a\text{mean}\{\hat{x}\hat{y}\} + a^2 \\
= 1 - 2\times\text{corr}\{\hat{x}, \hat{y}\} + a^2 \\
r = \text{corr}\{\hat{x}, \hat{y}\} \\
= 1 - 2ar + a^2 \\
\frac{\partial \text{var}\{u^3\}}{\partial a} = 0 \Rightarrow 2a - 2r = 0 \Rightarrow a = r
\]
Require the variance of error is minimal

\[ j = a \hat{x} + b = r \hat{x} \]

\[ a = r \]
\[ b = 0 \]
Here is the linear predictor!

\[ \hat{y}^p = r \hat{x} \]

Correlation coefficient
Prediction Formula

In standard coordinates

\[ \hat{y}_0^p = r \hat{x}_0 \quad \text{where} \quad r = \text{corr}(\{(x_i, y_i)\}) \]

In original coordinates

\[ \frac{y_0^p - \text{mean}(\{y_i\})}{\text{std}(\{y_i\})} = r \frac{x_0 - \text{mean}(\{x_i\})}{\text{std}(\{x_i\})} \]
Root-mean-square (RMS) prediction error

Given $\text{var}(\{u_i\}) = 1 - 2ar + a^2$

& $a = r$

$\text{var}(\{u_i\}) = 1 - r^2$

$\text{RMS error} = \sqrt{\text{mean}(\{u_i^2\})}$

$= \sqrt{\text{var}(\{u_i\})}$

$= \sqrt{1 - r^2}$

when $|r| = 1$
$\text{var}(\{u_i\}) = 0$

when $r = 0$
$\text{RMS} = 1$
See the error through simulation

https://rpsychologist.com/d3/correlation/
Example: Body Fat data

\[ r = 0.513 \]
Example: remove 2 more outliers

$r = 0.556$
Heatmap

- Display matrix of data via gradient of color(s)

**Figure 2-4.** Monthly normal mean temperatures for four locations in the US. Data source: NOAA.

Summarization of 4 locations’ annual mean temperature by month
3D bar chart

- Transparent 3D bar chart is good for small # of samples across categories.
Relationship between data feature and time

Example: How does Amazon’s stock change over 1 years?

take out the pair of features

x: Day

y: AMZN

<table>
<thead>
<tr>
<th>Day</th>
<th>AMZN</th>
<th>DUK</th>
<th>KO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>38.700001</td>
<td>34.971017</td>
<td>17.874906</td>
</tr>
<tr>
<td>2</td>
<td>38.900002</td>
<td>35.044103</td>
<td>17.882263</td>
</tr>
<tr>
<td>3</td>
<td>38.369999</td>
<td>34.240172</td>
<td>17.757161</td>
</tr>
<tr>
<td>6</td>
<td>37.5</td>
<td>34.294985</td>
<td>17.871225</td>
</tr>
<tr>
<td>7</td>
<td>37.779999</td>
<td>34.130544</td>
<td>17.885944</td>
</tr>
<tr>
<td>8</td>
<td>37.150002</td>
<td>33.984374</td>
<td>17.9117</td>
</tr>
<tr>
<td>9</td>
<td>37.400002</td>
<td>34.075731</td>
<td>17.933777</td>
</tr>
<tr>
<td>10</td>
<td>38.200001</td>
<td>33.91129</td>
<td>17.863866</td>
</tr>
<tr>
<td>14</td>
<td>38.66</td>
<td>34.020917</td>
<td>17.845469</td>
</tr>
<tr>
<td>15</td>
<td>37.880001</td>
<td>33.966104</td>
<td>17.882263</td>
</tr>
<tr>
<td>16</td>
<td>36.98</td>
<td>34.130544</td>
<td>17.790276</td>
</tr>
<tr>
<td>17</td>
<td>37.02</td>
<td>34.240172</td>
<td>17.757161</td>
</tr>
<tr>
<td>20</td>
<td>36.950001</td>
<td>34.057458</td>
<td>17.672533</td>
</tr>
<tr>
<td>21</td>
<td>36.43</td>
<td>34.112272</td>
<td>17.705649</td>
</tr>
<tr>
<td>22</td>
<td>37.259998</td>
<td>34.258442</td>
<td>17.709329</td>
</tr>
<tr>
<td>23</td>
<td>37.080002</td>
<td>34.569051</td>
<td>17.639418</td>
</tr>
<tr>
<td>24</td>
<td>36.849998</td>
<td>34.861392</td>
<td>17.598945</td>
</tr>
</tbody>
</table>
Time Series Plot: Stock of Amazon
Assignments

• Quiz1 open at 4:30pm today on PL
• Finish reading Chapter 2 of the textbook
• Work on the Week 2 module on Canvas
• Next time: Probability a first look
Additional References

Charles M. Grinstead and J. Laurie Snell
"Introduction to Probability”

Morris H. Degroot and Mark J. Schervish
"Probability and Statistics”
See you next time

See You!