“The statement that “The average US family has 2.6 children” invites mockery” – Prof. Forsyth reminds us about critical thinking

Credit: wikipedia
Last lecture

- Welcome/Orientation
- Big picture of the contents
- Lecture 1 - Data Visualization & Summary (I)
- Orientation quiz due today
Warm up question:

What kind of data is a letter grade?

What do you ask for usually about the stats of an exam with numerical scores?

1) A: Categorical   2) write

[ ] B: ordinal

C: Continuous
Objectives

- **Grasp** Summary Statistics
- Learn more Data Visualization for Relationships
Summarizing 1D continuous data

For a data set \{x\} or annotated as \{x_i\}, we summarize with:

- **Location Parameters**
  - Mean (\(\mu\)), Median, Mode

- **Scale parameters**
  - Standard deviation (\(\sigma\)), Variance (\(\sigma^2\)), Interquartile range

\(\hat{N}\) items
Summarizing 1D continuous data

Mean

\[ mean(\{x_i\}) = \frac{1}{N} \sum_{i=1}^{N} x_i \]

It’s the centroid of the data geometrically, by identifying the data set at that point, you find the center of balance.
\{ x_i \} \quad i \in \{1, 8\}

\{ x_i \} = 1, 2, 3, 4, 5, 6, 7, 12

\text{Mean}(\{ x_i \}) = 5
Properties of the mean

-scaling data scales the mean

\[ \text{mean}\left(\{k \cdot x_i\}\right) = k \cdot \text{mean}\left(\{x_i\}\right) \]

-translating the data translates the mean

\[ \text{mean}\left(\{x_i + c\}\right) = \text{mean}\left(\{x_i\}\right) + c \]
Less obvious properties of the mean

The signed distances from the mean sum to 0

\[ \sum_{i=1}^{N} (x_i - \text{mean}(\{x_i\})) = 0 \]

The mean minimizes the sum of the squared distance from any real value

\[ \arg\min_{\mu} \sum_{i=1}^{N} (x_i - \mu)^2 = \text{mean}(\{x_i\}) \]
Proof: \[ \sum_{i=1}^{N} (x_i - \text{mean}(\{x_i\})) = 0 \]

\[
\text{LHS} = \sum_{i=1}^{N} x_i - \sum_{i=1}^{N} \text{mean}(\{x_i\})
\]

\[
= \sum_{i=1}^{N} x_i - N \cdot \text{mean}(\{x_i\})
\]

\[
= N \sum_{i=1}^{N} x_i - N \cdot \frac{\sum_{i=1}^{N} x_i}{N}
\]

\[
= \sum_{i=1}^{N} x_i - \frac{N}{N} \sum_{i=1}^{N} x_i = 0
\]
Proof: \( \text{Argmin}_{\mu} \left( \sum_{i=1}^{N} (x_i - \mu)^2 \right) = \text{mean} \left( \{x_i\} \right) \)

\text{Argmin}_{\mu}: \text{Argument } \mu \text{ that minimizes } \\
\text{the function that follows } \\
\text{LHS} = \hat{\mu} \rightarrow \text{the special } \mu \text{ that } \\
\text{minimizes } f(\mu) = \sum_{i=1}^{N} (x_i - \mu)^2 \\
\text{To find } \hat{\mu}, \text{ set } \frac{df(\mu)}{d\mu} = 0 \text{ and solve it.} \\
\text{One way is to use the chain rule:} \\
f(\mu) = \sum_{i=1}^{N} h(\mu) = \frac{N}{2} g^{2}(\mu) \quad g = x_i - \mu \\
\frac{df}{dn} = \frac{d \sum h}{d \mu} = \sum_{i=1}^{N} \frac{dh}{d \mu} = \frac{N}{2} \left( \frac{dh}{dg} \cdot \frac{dg}{d\mu} \right)
Proof: \( \text{Argmin}_\mu \left( \sum_{i=1}^{N} (x_i - \mu)^2 \right) = \text{mean}(\{x_i\}) \)

\[
\frac{df(\mu)}{d\mu} = \sum \frac{dh}{dy} \frac{dg}{d\mu} = \sum 2g \cdot (-1) = 0
\]

\[
h = g^2
\]

\[
g = x_i - \mu
\]

\[
\Rightarrow \sum g = 0
\]

\[
\Rightarrow \sum_{i=1}^{N} (x_i - \mu) = 0
\]

\[
\Rightarrow \sum_{i=1}^{N} x_i - N \cdot \mu = 0
\]

\[
\Rightarrow \hat{\mu} = \frac{\sum x_i}{N}
\]

\[
= \text{mean}(\{x_i\})
\]
Q1:

What is the answer for

\[ \text{mean}(\text{mean}(\{x_i\})) \] ?

A. \( \text{mean}(\{x_i\}) \)  B. unsure  C. 0
The standard deviation

\[
\text{std} \left( \{x_i\} \right) = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \text{mean} \left( \{x_i\} \right))^2}
\]

\[= \sqrt{\text{mean} \left( \{(x_i - \text{mean} \left( \{x_i\} \right))^2\} \right)}
\]
How much the data spreads out wrt mean

\[
\text{Std} = \sqrt{\frac{1}{4} \sum_{i=1}^{4} d_i^2}
\]
Q2. Can a standard deviation of a dataset be -1?

A. YES
B. NO
Properties of the standard deviation

Scaling data scales the standard deviation

\[ \text{std}(\{k \cdot x_i\}) = |k| \cdot \text{std}(\{x_i\}) \]

Translating the data does **NOT** change the standard deviation

\[ \text{std}(\{x_i + c\}) = \text{std}(\{x_i\}) \]
Standard deviation: Chebyshev’s inequality (1st look)

- At most \( \frac{N}{k^2} \) items are \( k \) standard deviations (\( \sigma \)) away from the mean.

- Rough justification: Assume mean = 0.

\[
std = \sqrt{\frac{1}{N} \left[ \left( N - \frac{N}{k} \right) 0^2 + \frac{N}{k^2} (k\sigma)^2 \right]} = \sigma
\]
Variance ($\sigma^2$)

- **Variance** = (standard deviation)$^2$

$$\text{var}(\{x_i\}) = \frac{1}{N} \sum_{i=1}^{N} (x_i - \text{mean}(\{x_i\}))^2$$

- **Scaling and translating similar to standard deviation**

$$\text{var}(\{k \cdot x_i\}) = k^2 \cdot \text{var}(\{x_i\})$$

$$\text{var}(\{x_i + c\}) = \text{var}(\{x_i\})$$
Q3: Standard deviation

What is the value of $std(\{mean(\{x_i\})\})$?

A. 0  B. 1  C. unsure
The mean tells where the data set is and the standard deviation tells how spread out it is. If we are interested only in comparing the shape, we could define:

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

for every i

We say \( \{\hat{x}_i\} \) is in standard coordinates.
Q4: Mean of standard coordinates

\[ \hat{x}_i = \frac{x_i - \text{mean}(\{x_i\})}{\text{std}(\{x_i\})} \]
Q5: Standard deviation ($\sigma$) of standard coordinates

$\text{Std}(\{\hat{x}_i\})$ is:

A. 1  B. 0  C. unsure

$$\hat{x}_i = \frac{x_i - \text{mean}(\{x_i\})}{\text{std}(\{x_i\})}$$
Q6: Variance of standard coordinates

Variance of $\{\hat{x}_i\}$ is:

A. 1  B. 0  C. unsure

\[ \hat{x}_i = \frac{x_i - \text{mean}(\{x_i\})}{\text{std}(\{x_i\})} \]
Q7: Estimate the range of data in standard coordinates

Estimate as close as possible, 90% data is within:

A. [-10, 10]
B. [-100, 100]
C. [-1, 1]
D. [-4, 4]
E. others

\[ \hat{x}_i = \frac{x_i - \text{mean}\{x_i\}}{\text{std}\{x_i\}} \]
\[ \frac{N}{k^2} = \frac{1}{k^2} \leq 10\% \]

\[ \geq 90\% \]

\[ v = k\sigma \]

\[ k\sigma = k \]

\[ \therefore \sigma (\hat{\omega}) \]

\[ = 1 \]
Data in standard coordinates always has

\[ \text{mean} = 0; \text{ standard deviation} = 1; \]
\[ \text{variance} = 1. \]

Such data is unit-less, plots based on this sometimes are more comparable

We see such normalization very often in statistics
We first sort the data set \( \{x_i\} \)

Then *if* the number of items \( N \) is *odd*

\[
\text{median} = \text{middle item's value}
\]

*if* the number of items \( N \) is *even*

\[
\text{median} = \text{mean of middle 2 items' values}
\]
Properties of Median

 Scaling data scales the median

$$\text{median}(\{k \cdot x_i\}) = k \cdot \text{median}(\{x_i\})$$

 Translating data translates the median

$$\text{median}(\{x_i + c\}) = \text{median}(\{x_i\}) + c$$
Percentile

- $k^{th}$ percentile is the value relative to which $k\%$ of the data items have smaller or equal numbers.

- Median is roughly the 50$^{th}$ percentile.

\{1, 2, 3, 4, 5, 6, 7, 12\}

75$^{th}$ percentile = 6 \neq 0.75 \times \frac{12}{75}\%

**Interquartile range**

- $iqr = (75\text{th percentile}) - (25\text{th percentile}) > 0$

- Scaling data scales the interquartile range

  $iqr\left(\{k \cdot x_i\}\right) = |k| \cdot iqr\left(\{x_i\}\right)$

- Translating data does **NOT** change the interquartile range

  $iqr\left(\{x_i + c\}\right) = iqr\left(\{x_i\}\right)$
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
Boxplots details, outliers

How to define outliers? (the default)

- Whisker
- Box
- Median
- Outlier

> 1.5 iqr
< 1.5 iqr

Interquartile Range (iqr)

(default)
Q. TRUE or FALSE

mean is more sensitive to outliers than median

A. True
B. False
interquartile range is more sensitive to outliers than std.

A. True
B. False
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

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

Left Skew

- mean, median, mode
- left tail

Right Skew

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

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

Median = 47

Mean = 46

Mean < Median => Left
Assignments

- **HW1** due Thurs. Feb. 4.
- **Quiz 1** (open 4:30pm today until Mon. next week)
- Reading upto Chapter 2.1
- Next time: the quantitative part of correlation coefficient
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!