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

Credit: wikipedia

Hongye Liu, Teaching Assistant Prof, CS361, UIUC, 8.27.2021
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?
Objectives

- Histograms
- **Grasp** Summary Statistics
Visualizing Data with Histogram

- **Histogram**

  A set of bars that are organized by bins that contains numerical data (discrete or continuous)

  Data: “iris”
Visualizing Data with Histogram (II)

Conditional histogram

Histogram generated by subsets of the data

Data: “iris”
Which group has the higher total scores?
Which group has the higher total scores?
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

\( N \) items
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:
  \[
  \text{mean}\left(\{K x_i + c\}\right) = K \cdot \text{mean}\left(\{x_i\}\right) + c
  \]

- 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
  \[ \text{argmin}_\mu \sum_{i=1}^{N} (x_i - \mu)^2 = \text{mean}(\{x_i\}) \]
Proof: $\sum_{i=1}^{n} (x_i - \text{mean}(\{x_i\})) = 0$

$LHS = \sum_{i=1}^{n} x_i - \sum_{i=1}^{n} \text{mean}(\{x_i\})$

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

$= \sum_{i=1}^{n} x_i - \frac{\sum_{i=1}^{n} x_i}{N}$

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

\( \text{Argmin} \): Argument \( \mu \) that minimizes the function that follows:

\[ \text{LHS} = \mu \rightarrow \text{the special} \ \mu \ \text{that minimizes} \ f(\mu) = \frac{1}{m} \sum_{i=1}^{N} (x_i - \mu)^2 \]

To find \( \hat{\mu} \), set \( \frac{\text{d} f(\mu)}{\text{d} \mu} = 0 \) and solve it.

One way is to use the Chain rule:

\[ f(\mu) = \sum_{i=1}^{N} h(\mu) = \sum_{i=1}^{N} g^2(\mu) \]

\[ g = x_i - \mu \]

\[ \frac{\text{d} f}{\text{d} \mu} = \sum_{i=1}^{N} \frac{\text{d} h}{\text{d} \mu} = \sum_{i=1}^{N} \frac{\text{d} h}{\text{d} g} \cdot \frac{\text{d} g}{\text{d} \mu} \]
Proof: Argmin \( \mu \) \( \left( \sum_{i=1}^{N} (x_i - \mu)^2 \right) = \text{mean}(\{x_i\}) \)

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

\[
h = g^2
\]

\[
g = x_i - \mu
\]

\[
\Rightarrow \sum g = 0
\]

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

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

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

\[
\frac{d^2f(\mu)}{d\mu^2}
\]
Q1: What is the answer for

$$mean(\{mean(\{x_i\})\})$$?

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

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

\[
= \sqrt{mean(\{(x_i - mean(\{x_i\}))^2\})}
\]
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

\[
\begin{align*}
&\frac{0.5N}{k^2} \quad \frac{N - N}{k^2} \quad \frac{0.5N}{k^2} \\
&-k\sigma \quad 0 \quad k\sigma
\end{align*}
\]

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

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

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

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

  $$var(\{k \cdot x_i\}) = k^2 \cdot var(\{x_i\})$$
  $$var(\{x_i + c\}) = var(\{x_i\})$$
Q3: Standard deviation

What is the value of

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

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:

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

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

mean(\{\hat{x}_i\}) is:

A. 1  B. 0  C. unsure

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

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

\[ \{ \text{AX} \} \]

\[ \geq 90\% \]

\[ v = k\sigma \]

\[ k\sigma = k \]

\[ \therefore \sigma \left( \hat{\epsilon}, \hat{\gamma} \right) = 1 \]
Standard Coordinates/normalized data to $\mu=0$, $\sigma=1$, $\sigma^2=1$

- Data in standard coordinates always has mean = 0; standard deviation = 1;
  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

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

Translating data translates the median

$$median(\{x_i + c\}) = median(\{x_i\}) + c$$
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.

Given the set $\{1, 2, 3, 4, 5, 6, 7, 12\}$, the 75th percentile is $6$, which is $75\%$ or $0.75$.
Interquartile range

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

- Scaling data scales the interquartile range

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

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

  \[ \text{iqr}(\{x_i + c\}) = \text{iqr}(\{x_i\}) \]
Assignments

- HW1 due Mon. Sept. 6.
- Quiz 1 (open 4:30pm 9/1 next Wed until Sat.9/4)
- Reading upto Chapter 2.1
- Next time: more summary statistics and correlation coefficient
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!