STAT 3355
Introduction to Data Analysis

Lecture 08: Summaries for Univariate Data II

Created by: Qiwei Li
Assistant Professor of Statistics
Presented by: Octavious Smiley
Assistant Professor of Instruction

Department of Mathematical Sciences
The University of Texas at Dallas

Last Class

- Summarize a univariate discrete data $\boldsymbol{x}$

	Discrete
Numerical	`table(`x`)`
Graphical	`barplot(table(`x`))`
	`pie(table(`x`))`
	`dotchart2(table(`x`))`

Learning Goals

- Summarize a univariate continuous data $\boldsymbol{x}$

	Discrete	Continuous
Numerical	`table(`x`)`	
Graphical	`barplot(table(`x`))`	
	`pie(table(`x`))`	
	`dotchart2(table(`x`))`	

Learning Goals

- Summarize a univariate data in three ways
  - Center
  - Spread
  - Shape
- Numerical summaries for continuous data
  - Center: The sample mean and the sample median
  - Spread: The sample variance (standard deviation) and the IQR

Continuous Data

- Unlikely for a pair of samples to share the same value
- Data type
  - Integer (if the number of unique values is large)
  - Numeric data
- Examples
  - The height of person in cm
  - The weight of a person in lb
  - The age of a person in year
  - The weekly self-learning time for STAT3355 in minute

The Sample Mean

- Denote a univariate continuous dataset by

$$\boldsymbol{x} = [x_1, \ldots, x_i, \ldots, x_n], \text{ where } x_i \in \mathbb{R}$$

- The sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\left(x_1 + \ldots + x_i + \ldots + x_n\right)$$

- Interpretation
  - The balance point
  - "Centering": Average out the data so that $\bar{x} = 0$

$$\hat{\boldsymbol{x}} = [x_1 - \bar{x}, \ldots, x_i - \bar{x}, \ldots, x_n - \bar{x}]$$

The Sample Mean

- Implementation in R
  - x is a numeric vector
    - `mean(`x`, na.rm = TRUE)`
  - x is a numeric variable in a data frame X
    - `mean(`X`$x_name, na.rm = TRUE)`
  - Calculate the mean for each column in a numeric matrix or a data frame X
    - `colMeans(`X`, na.rm = TRUE)`
    - `apply(`X`, MARGIN = 2, mean, na.rm = TRUE)`

The Sample Mean

- Examples

```
library(UsingR)

# Load data babies
data("babies")

# Birth weight variable
mean(babies$wt)

# Mother age variable
x <- babies$age
mean(x)
index_99 <- which(x == 99)
x[index_99] <- NA
mean(x, na.rm = TRUE)
```

Other Types of Mean

- Geometric mean

$$\bar{x}_{\mathsf{GM}} = \left(\prod_{i=1}^{n} x_i\right)^{\frac{1}{n}} = (x_1 \ldots x_i \ldots x_n)^{\frac{1}{n}}$$

- Rates of growth

Year	GDP (Trillion)	Annual growth	Ratio
2017	$59,915$	NA	NA
2018	$62,805$	4.8%	1.048
2019	$65,095$	3.7%	1.037
2020	$63,028$	-3.2%	0.968
2021	$69,288$	9.9%	1.099

Other Types of Mean

- Harmonic mean

$$\bar{x}_{\mathsf{HM}} = n\left(\sum_{i=1}^{n} \frac{1}{x_i}\right)^{-1} = \frac{n}{\left(\frac{1}{x_1} + \ldots + \frac{1}{x_i} + \ldots + \frac{1}{x_n}\right)}$$

  - Ratios, e.g. speed (distance per unit of time)

Date	Flight no.	Speed (mph)	Departure	Arrival
Sep 22	WN5	532	DAL	HOU
Sep 22	WN4	500	HOU	DAL
Sep 21	WN5	492	DAL	HOU
Sep 21	WN4	550	HOU	DAL
Sep 20	WN5	513	DAL	HOU

- Relationship: $\bar{x} \geq \bar{x}_{\mathsf{GM}} \geq \bar{x}_{\mathsf{HM}}$, where equality holds if and only if all x_i's are equal

The Sample Median

- Denote a univariate continuous dataset by

$$\boldsymbol{x} = [x_1, \ldots, x_i, \ldots, x_n], \text{ where } x_i \in \mathbb{R}$$

- Sort the n values in an ascending order

$$\boldsymbol{x}_{\text{sorted}} = \left[x_{[1]}, \ldots, x_{[i]}, \ldots, x_{[n]}\right], \text{ where } x_{[i+1]} \geq x_{[i]}$$

- The sample median

$$M = \begin{cases} x_{[k+1]} & \text{if } n = 2k+1 \\ \left(x_{[k]} + x_{[k+1]}\right)/2 & \text{if } n = 2k \end{cases}$$

- Interpretation
  - The center by count
  - A point that splits the data in half
  - Resistant to the extremely small or large values in $\boldsymbol{x}$

The Sample Median

- Implementation in R
  - x is a numeric vector
    - `median(`x`, na.rm = TRUE)`
    - `quantile(`x`, probs = 0.5, na.rm = TRUE)`
    - `summary(`x`)["Median"]`
  - x is a numeric variable in a data frame $\boldsymbol{X}$
    - `median(`X`$x_name, na.rm = TRUE)`
    - `quantile(`X`$x_name, probs = 0.5, na.rm = TRUE)`
    - `summary(`X`$x_name)["Median"]`
  - Calculate the median for each column in a numeric matrix or a data frame $\boldsymbol{X}$
    - `apply(`X`, MARGIN = 2, median, na.rm = TRUE)`

The Sample Median

- Examples

```
library(UsingR)

# Load data babies
data("babies")

# Birth weight variable
mean(babies$wt)
median(babies$wt)
summary(babies$wt)

# Load data CEO compensation
data("exec.pay")
mean(exec.pay)
median(exec.pay)
```

The Sample Median

- Mean and median can give different senses of center
- Examples: Fuel efficiency by year (https://fueleconomy.gov/)
  - Highway MPG

Year	Median	Mean	Ratio
1989	22	22.47	1.02
1992	22	22.44	1.02
1995	22	22.67	1.03
1998	23	23.55	1.02
2001	23	23.33	1.01
2004	23	23.06	1.00
2007	23	23.08	1.00
2010	25	24.97	1.00

The Sample Median

- Mean and median can give different senses of center
- Examples: Household net worth in U.S. by year
  - Income

Year	Median ($)	Mean ($)	Ratio
1989	$79,100$	$313,600$	4.0
1992	$75,100$	$282,900$	3.8
1995	$81,900$	$300,400$	3.7
1998	$95,600$	$377,300$	3.9
2001	$106,100$	$487,000$	4.6
2004	$107,200$	$517,100$	4.8
2007	$126,400$	$584,600$	4.6
2010	$77,300$	$498,800$	6.5

  - Real-estate prices
  - Waiting times for auto repairs/maintenance

The p-th Sample Quantile

- Denote a univariate continuous dataset by

$$\boldsymbol{x} = [x_1, \ldots, x_i, \ldots, x_n], \text{ where } x_i \in \mathbb{R}$$

- Sort the n values in an ascending order

$$\boldsymbol{x}_{\text{sorted}} = \left[x_{[1]}, \ldots, x_{[i]}, \ldots, x_{[n]}\right], \text{ where } x_{[i+1]} \geq x_{[i]}$$

- The p-th quantile, where $p \in [0, 1]$

$$Q(p) = \begin{cases} x_{[k]} & \text{if } k = p(n-1) + 1 \in \mathbb{N} \\ (1-p)x_{[k]} + p x_{[k+1]} & \text{if } (n-1)p < k \leq (n-1)p + 1 \end{cases}$$

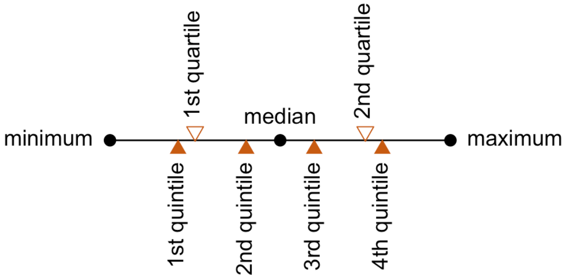
- Interpretation
  - $100p\%$ of the data is less than the value of $Q(p)$
  - $100(1-p)\%$ of the data is more than the value of $Q(p)$

The p-th Sample Quantile

- Implementation in `R`
  - $\boldsymbol{x}$ is a numeric vector
    - `quantile(`x`, probs = c(...) na.rm = TRUE)`
    - `summary(`x`)`
  - $\boldsymbol{x}$ is a numeric variable in a data frame $\boldsymbol{X}$
    - `quantile(`X`$x_name, probs = c(...), na.rm = TRUE)`
    - `summary(`X`$x_name)`
  - Calculate the p-th sample quantile for each column in a numeric matrix or a data frame $\boldsymbol{X}$
    - `apply(`X`, MARGIN = 2, quantile, probs = c(...), na.rm = TRUE)`

The p-th Sample Quantile

- Special cases
  - $Q(0.5)$: Median
  - $Q(0)$ and $Q(1)$: Minimum and maximum
  - $Q(0.25)$ and $Q(0.75)$: 1st (lower) and 2nd (upper) quartiles
  - $Q(0.2)$, $Q(0.4)$, $Q(0.6)$, and $Q(0.8)$: 1st, 2nd, 3rd, and 4th quintiles



The p-th Sample Quantile

- Examples

```
# Get Q(0), Q(1)
range(exec.pay)

# Get Q(0), Q(0.25), Q(0.5), Q(0.75), Q(1)
summary(exec.pay)

# Get Q(0.2), Q(0.4), Q(0.6), Q(0.8)
quantile(exec.pay, probs = seq(0.2, 0.8, by
    = 0.2))

# Get any p-th quantile
p <- 0.15
quantile(exec.pay, probs = p)
```

The Trimmed Mean

- Denote a univariate continuous dataset by

$$\boldsymbol{x} = [x_1, \dots, x_i, \dots, x_n], \text{ where } x_i \in \mathbb{R}$$

- The trimmed mean

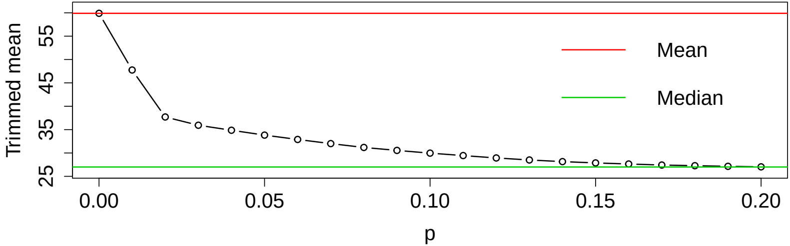
$$\bar{x}_{\mathsf{TM}}(p) = \frac{\sum_{i=1}^{n} x_i I(Q(p) \leq x_i \leq Q(1-p))}{\sum_{i=1}^{n} I(Q(p) \leq x_i \leq Q(1-p))}$$

- Interpretation
  - The “bulk” point after ignoring extreme points at both ends
- Implementation in `R`
  - `mean(`x`, trim = `p`, na.rm = TRUE)`, where $p \in [0, 0.5]$

The Trimmed Mean

- Examples

```
mean(exec.pay)
median(exec.pay)
mean(exec.pay, trim = 0.05)
mean(exec.pay, trim = 0.2)
```



Your Turn

- The data set `rivers` in the package `UsingR` contains the lengths (in miles) of 141 major rivers in North America
  - What proportion are less than the median length?
  - What proportion are less than the mean length?
  - Compare the mean, median, and 25%-trimmed mean. Is there a big difference among the three numbers?

Your Turn

- Solutions

```
# Load data
data("rivers")
x <- rivers
n <- length(x)

# What proportion are less than the mean
    length
x_bar <- mean(x)
print(sum(x < x_bar)/n)
```

Your Turn

- Solutions

```
# Compare the mean, median, and 25%-trimmed
   mean
print(mean(x))
print(median(x))
print(mean(x, trim = 0.25))
```

The Sample Variance

- Denote a univariate continuous dataset by

$$\boldsymbol{x} = [x_1, \ldots, x_i, \ldots, x_n], \text{ where } x_i \in \mathbb{R}$$

- The sample variance

$$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

- The sample standard deviation s
- Interpretation
  - Large values indicate more spread-out data

The Sample Variance

- Interpretation
  - "Centering": Average out the data so that $\bar{x} = 0$

$$\hat{\boldsymbol{x}} = [x_1 - \bar{x}, \ldots, x_i - \bar{x}, \ldots, x_n - \bar{x}]$$

  - "Scaling": Average out and normalized the data so that $\bar{x} = 0$ and $s = 1$

$$\boldsymbol{z} = \left[\frac{x_1 - \bar{x}}{s}, \ldots, \frac{x_i - \bar{x}}{s}, \ldots, \frac{x_n - \bar{x}}{s}\right]$$

  - Empirical rule: If the data is bell-shaped, then 68%, 95%, and 99.7% of the data have a z-score in $[-1, 1]$, $[-2, 2]$, and $[-3, 3]$

The Sample Variance

- Implementation in R
  - x is a numeric vector
    - `var(`x`, na.rm = TRUE)`
  - x is a numeric variable in a data frame $\boldsymbol{X}$
    - `var(`X`$x_name, na.rm = TRUE)`
  - Calculate the sample variance for each column in a numeric matrix or a data frame $\boldsymbol{X}$
    - `apply(`X`, MARGIN = 2, var, na.rm = TRUE)`

The Sample Variance

- Examples

```
# Sample variance
var(babies$wt)

# Sample standard deviation
sqrt(var(babies$wt))

# Data scaling (calculating z-scores)
z <- c(scale(babies$wt, center = TRUE, scale
    = TRUE))
z <- (babies$wt - mean(babies$wt))/sqrt(var(
   babies$wt))
sum(abs(z) <= 1) / length(z)
sum(abs(z) <= 2) / length(z)
sum(abs(z) <= 3) / length(z)
```

The InterQuartile Range (IQR)

- Denote a univariate continuous dataset by

$$\boldsymbol{x} = [x_1, \ldots, x_i, \ldots, x_n], \text{ where } x_i \in \mathbb{R}$$

- The lower and upper quartiles of $\boldsymbol{x}$ are $Q(0.25)$ and $Q(0.75)$
- The interquartile range is

$$\text{IQR} = Q(0.75) - Q(0.25)$$

- Interpretation
  - The range of the middle 50% of $\boldsymbol{x}$
  - Resistant to the extremely small or large values in $\boldsymbol{x}$
  - Range $= Q(1) - Q(0)$

The InterQuartile Range (IQR)

- Implementation in R
  - x is a numeric vector
    - `IQR(x, na.rm = TRUE)`
    - `diff(quantile(x, probs = c(0.25, 0.75), na.rm = TRUE))`
  - x is a numeric variable in a data frame $\boldsymbol{X}$
    - `IQR(X$x_name, na.rm = TRUE)`
    - `diff(quantile(X$x_name, probs = c(0.25, 0.75), na.rm = TRUE))`
  - Calculate the IQR for each column in a numeric matrix or a data frame $\boldsymbol{X}$
    - `apply(X, MARGIN = 2, IQR, na.rm = TRUE)`

The InterQuartile Range (IQR)

- Examples

```
# IQR
IQR(babies$wt)

# Range
range(babies$wt)
```

The Median Absolute Deviation (MAD)

- Denote a univariate continuous dataset by

$$\boldsymbol{x} = [x_1, \ldots, x_i, \ldots, x_n], \text{ where } x_i \in \mathbb{R}$$

- The median is $Q(0.5)$ or M
- Subtract each entry in $\boldsymbol{x}$ by M, take the absolute value

$$\boldsymbol{y} = [|x_1 - M|, \ldots, |x_i - M|, \ldots, |x_n - M|]$$

and sort the n values in an ascending order

- The median absolute deviation is

$$\text{MAD} = \begin{cases} 1.4826 \cdot y_{[k+1]} & \text{if } n = 2k+1 \\ 1.4826 \cdot \left(y_{[k]} + y_{[k+1]}\right)/2 & \text{if } n = 2k \end{cases}$$

- Interpretation
  - Resistant to the extreme (especially larger) values

The Median Absolute Deviation (MAD)

- Implementation in R
  - x is a numeric vector
    - `mad(`x`, na.rm = TRUE)`
  - x is a numeric variable in a data frame X
    - `mad(`X`$x_name, na.rm = TRUE)`
  - Calculate the MAD for each column in a numeric matrix or a data frame X
    - `apply(`X`, MARGIN = 2, mad, na.rm = TRUE)`

The Median Absolute Deviation (MAD)

- Examples

```
x <- babies$wt

# Standard deviation
sqrt(var(x))

# Self-defined without the adjustment
median(abs(x - median(x)))

# MAD
mad(x)
```

Your Turn

- The data set `rivers` in the package `UsingR` contains the lengths (in miles) of 141 major rivers in North America
  - Compare the standard deviation, IQR, and MAD. Is there a big difference among the three numbers?
  - Scale the data so that the data has zero-mean and unit variance
  - Verify the empirical rule

Your Turn

- Solutions

```
# Load data
data("rivers")
x <- rivers
n <- length(x)

# Compare the standard deviation, IQR, and
   MAD
print(IQR(x))
print(mad(x))

# Obtain the z-scores
x_bar <- mean(x)
s <- sqrt(var(x))
z <- (x - x_bar) / s
```

Your Turn

- Solutions

```
# Verify the empirical rule
sum(abs(z) <= 1) / length(z)
sum(abs(z) <= 2) / length(z)
sum(abs(z) <= 3) / length(z)

# Verify the empirical rule in log scale
z <- (log(x) - mean(log(x))) / sd(log(x))
sum(abs(z) <= 1) / length(z)
sum(abs(z) <= 2) / length(z)
sum(abs(z) <= 3) / length(z)
```

After-class Reading

- *Using R for Introductory Statistics (1st Ed.)* by John Verzani
- Chapter 2 Univariate data
  - Section 2.2 Numeric data
    - Subsection 2.2.3 The center: mean, median, and mode
    - Subsection 2.2.4 Variation: the variance, standard deviation, and IQR

After-class Reading

- *Using R for Introductory Statistics (2nd Ed.)* by John Verzani
- Chapter 2 Univariate data
  - Section 2.3 Numeric summaries