

# An Introduction to Statistics

- What does a statistician do?
- What *is/are* statistics???
- Data?
- Terminology

# What does a statistician do?

- A statistician makes sense of information collected about the world

# What does a statistician do?

- A statistician makes sense of information collected about the world
  - Analyze effectiveness of drug/treatment
  - Determining behavior patterns
  - Analyze spread of West Nile Virus
  - Analyze performance of individuals, businesses, stocks, ...

- **Economics - about money**
- **Psychology - about why we think what we think (we think)**
- **Biology - about life**
- **Anthropology - about who**
- **History - about what, where, when**
- **Philosophy - about why**
- **Engineering - about how**
- **Accounting - about how much**
- **Statistics -**

- Economics - about money
- Psychology - about why we think what we think (we think)
- Biology - about life
- Anthropology - about who
- History - about what, where, when
- Philosophy - about why
- Engineering - about how
- Accounting - about how much
- Statistics - *about variation*

**What is Variation?**

# What is Variation?

- The act or process of varying
- An instance of the act or process of varying
- Amount or degree of change
- A deviation in structure or character from others of the same species

*Random House Webster's College Dictionary ©1995*

# What is Statistics?

- Statistics *is* a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical/categorical data

# What is Statistics?

- Statistics *is* a way of reasoning together with collection tools and methods designed to help us to understand data and information about the world

# What *are* statistics?

- Statistics *are* calculations made from data.

# What are data?

- Systematically recorded information *together with context*
- Note: *Data* is plural and *datum* is singular.

# Context

- Tells
  - *Who* was measured
  - *What* was measured
  - *Where* data were collected
  - *When* data were collected
  - *Why* study was performed
  - *How* data were collected

**Data is useless without ...**

# Data is useless without ...

- **Context**

- ◊ Who
- ◊ What
- ◊ Where
- ◊ When
- ◊ Why
- ◊ How

# Data can be ...

- Numbers
- Dates
- Words
- Codes or other labels

# Data Table

- An arrangement of data

# Data Table

- Rows represent
  - Individuals, i.e. *who was studied/examined*
    - Respondents of a survey
    - Subjects of an experiment
    - Participants of a study
    - Experimental units (animals, plants, websites, inanimate objects)
- Rows are called
  - Records (for a database)
  - Most generally, cases

# Data Table

- Columns contain
  - Characteristics/information collected/recorded about each case
- The variables for the study are listed in the columns of a data table.
  - The variables are the characteristics or information collected about each individual in the study, i.e. *what was measured*

# Variables

- **Categorical**
  - **Words**
    - ◉ **Codes**
    - ◉ **Names**
- **Quantitative**
  - **Numerical values**
  - **Measure**
- **Ordinal**
  - **Ordered rating**

You are hired to analyze the per capita income, in dollars, as well as the percentages of the labor force employed in agriculture, industry, and service occupations for the twenty (20) OECD countries for 1960.

COUNTRY	Per Capita Income	Agriculture	Industry	Service
<i>U. KINGDOM</i>	1105	4	56	40
<i>BELGIUM</i>	1005	6	52	42
<i>NETHERLANDS</i>	810	11	49	40
<i>SWITZERLAND</i>	1361	11	56	33
<i>CANADA</i>	1536	17	43	45
<i>SWEEDEN</i>	1644	17	53	33
<i>LUXEMBOURG</i>	1242	17	51	34
<i>W. GERMANY</i>	1035	19	60	25
<i>DENMARK</i>	1049	19	45	37
<i>FRANCE</i>	1013	20	44	36
<i>NORWAY</i>	977	20	49	32
<i>AUSTRIA</i>	681	20	47	30
<i>ICELAND</i>	839	23	47	29
<i>ITALY</i>	504	23	46	28
<i>JAPAN</i>	344	24	35	32
<i>IRELAND</i>	529	36	30	34
<i>SPAIN</i>	290	42	37	21
<i>PORTUGAL</i>	238	44	33	23
<i>GREECE</i>	324	71	24	20
<i>TURKEY</i>	177	79	12	9

For your summer internship with U.S. News and World Report, you are asked to analyze data on the Class of 1999 for the top twenty (20) universities in the United States. From your analysis, you obtain information on the percentage of incoming freshmen who graduate within four years (the freshman retention rate), the percentage of students admitted as freshmen who graduated (the graduation rate), determining both the predicted graduation rate and the actual graduation rate, the percentage of freshman applicants who were accepted (the acceptance rate), and the percentage of alumni who made donations to the university after graduation (the alumni giving rate) for this class.

School	Freshman Retention	Predicted Graduation Rate	Actual Graduation Rate	Acceptance Rate	Alumni Giving Rate
University of California – Berkeley	95	81	82	27	18
Johns Hopkins University	95	89	87	33	28
Northwestern University	96	85	92	32	29
Columbia University	97	89	91	14	32
University of Chicago	94	89	81	48	35
Cornell University	96	87	91	33	36
Stanford University	98	91	90	15	37
Washington University in St. Louis	96	83	86	34	37
Rice University	95	91	88	27	39
Emory University	92	83	86	42	39
University of Pennsylvania	96	87	90	26	40
California Institute of Technology	92	96	82	18	41
Brown University	97	87	93	17	43
Massachusetts Inst. of Technology	97	94	91	19	43
Duke University	97	89	92	28	45
Harvard University	96	94	97	11	47
University of Notre Dame	98	80	95	35	48
Yale University	98	93	94	16	49
Dartmouth College	96	90	94	21	52
Princeton University	99	92	96	11	66

# Variables

- **Categorical or Qualitative variable**
  - is a classification or categorization
- **Quantitative variable**
  - is a measurement, amount, or count
- **Ordinal variable**
  - is an ordered rating or ranking

# Variables

- **Categorical or Qualitative variable**
  - is a classification or categorization
    - Something that is a name (sub-classification - Nominal Variable), word, or code
- **Quantitative variable**
  - is a measurement, amount, or count
    - Can be indicated by words "amount" or "number of"
- **Ordinal variable**
  - is an ordered rating or ranking
    - Can be indicated by words "rating", "ranking", or "evaluation"

# Variables

- **Categorical or Qualitative variable**
  - **is a classification or categorization**
    - ◉ **Something that is a name (sub-classification - Nominal Variable), word, or code**
  - **Examples:**
    - ◉ **Name of country (Nominal Variable)**
    - ◉ **Type of bill in your wallet**
    - ◉ **Type of shoe**
    - ◉ **Class year**
    - ◉ **Area code**
    - ◉ **Zip code**

# Variables

- **Quantitative variable**
  - is a measurement, amount, or count
    - Can be indicated by words "amount" or "number of"
  - **Examples:**
    - Amount of money in your wallet
    - Height
    - Amount of gasoline in your gas tank
    - Amount of time spent studying
    - Number of bills in your wallet
    - Distance traveled

# Variables

- **Ordinal variable**
  - is an ordered rating or ranking
    - Can be indicated by words “rating”, “ranking”, or “evaluation”
  - **Examples:**
    - J.D. Powers customer service rating for cellular phone service
    - National Highway Traffic Safety Administration rating for car safety
    - Course grade
    - Customer service rating for Dell Computers

# Variables

- **Caution**
  - **A thing is not a variable**

# Variables

- **Caution**
  - A thing is not a variable
  - Time is a thing. The *amount of time* is a quantitative variable.
  - The bills in your wallet are things. The *number of bills in your wallet* is a quantitative variable.
  - Shoes are things. The *number of shoes* is a quantitative variable. The *type of shoes* is a qualitative variable.

# Quantitative Variables have units

- Units are quantities or amounts adopted as standard measurements

# Quantitative Variables have units

- Units are quantities or amounts adopted as standard measurements
- Examples
  - Dollars
  - Hours
  - Years
  - Feet

# Quantitative Variables

- Quantitative variables can be
  - Discrete
  - Continuous

# Quantitative Variables

- **Continuous variable** - a quantitative variable that has an infinite number of possible values that are not countable
  - Note: Suppose a continuous variable can take on all values between 0 and 1, inclusive. Then, the values 0.9, 0.99, 0.999, ..., 0.9999999999999999, etc., are all possible values for the variable

# Quantitative Variables

- Discrete variable - a quantitative variable that has either a finite number of possible values or a countable number of possible values
  - Note: We say that a set is countable if there is a bijection from the set to the Natural Numbers.

# Quantitative Variables

- Discrete variable - a quantitative variable that has either a finite number of possible values or a countable number of possible values
  - Note: We say that a set is countable if there is a *bijection* from the set to the Natural Numbers.

WHAT?

# Quantitative Variables

- Discrete variable - a quantitative variable that has either a finite number of possible values or a countable number of possible values
  - Note: We say that the values for a variable are countable if you can count all the possibilities, for example, 1, 2, 3, 4, 5, .... We can say that there are 0 values as well if there are no possible values for the variable.

# Quantitative Variables

- A quantitative variable is discrete *if the variable involves a count*
  - Examples:
    - ◉ The number of students
    - ◉ The number of bills in your wallet
    - ◉ The amount of money in your wallet
    - ◉ The number of minutes
    - ◉ The number of shoes

# Quantitative Variables

- A quantitative variable is continuous *if the variable involves a measurement*
  - Examples:
    - Height
    - Distance traveled
    - The amount of cereal in the box
    - Amount of gasoline in your gas tank
    - The weight of a box

# Ordinal Variable

- **Note:** Some ordinal variables can be treated as categorical or quantitative variables
- **Why?**

# Ordinal Variable

- **Note:** Some ordinal variables can be treated as categorical or quantitative variables
- **Why?** Ordinal values can act as classifications/categories *and* as numerical measures/values

# Cautions

- Important to consider what values *mean* and how they are being *used*
- Do not treat all numerical data as quantitative
  - The average of categorical data is meaningless
  - Example - the average of hair color data is meaningless even if provided on a five-point scale

# Cautions

- Important to consider what values *mean* and how they are being *used*
- Do not treat all numerical data as quantitative
  - The average of categorical data is meaningless
  - Example - the average zip code makes no sense

# Cautions

- Important to consider what values *mean* and how they are being *used*
- Do not treat all numerical data as quantitative
  - The average of categorical data is meaningless
  - Example - the average area code makes no sense

- For the following two examples, is each column heading a variable?
  - If the column heading is a variable, classify the column heading.
  - If the column heading is not a variable, state the variable that is represented by the column heading and classify the variable.

You are hired to analyze the per capita income, in dollars, as well as the percentages of the labor force employed in agriculture, industry, and service occupations for the twenty (20) OECD countries for 1960.

COUNTRY	Per Capita Income	Agriculture	Industry	Service
<i>U. KINGDOM</i>	1105	4	56	40
<i>BELGUIM</i>	1005	6	52	42
<i>NETHERLANDS</i>	810	11	49	40
<i>SWITZERLAND</i>	1361	11	56	33
<i>CANADA</i>	1536	17	43	45
<i>SWEEDEN</i>	1644	17	53	33
<i>LUXEMBOURG</i>	1242	17	51	34
<i>W. GERMANY</i>	1035	19	60	25
<i>DENMARK</i>	1049	19	45	37
<i>FRANCE</i>	1013	20	44	36
<i>NORWAY</i>	977	20	49	32
<i>AUSTRIA</i>	681	20	47	30
<i>ICELAND</i>	839	23	47	29
<i>ITALY</i>	504	23	46	28
<i>JAPAN</i>	344	24	35	32
<i>IRELAND</i>	529	36	30	34
<i>SPAIN</i>	290	42	37	21
<i>PORTUGAL</i>	238	44	33	23
<i>GREECE</i>	324	71	24	20
<i>TURKEY</i>	177	79	12	9

For your summer internship with U.S. News and World Report, you are asked to analyze data on the Class of 1999 for the top twenty (20) universities in the United States. From your analysis, you obtain information on the percentage of incoming freshmen who graduate within four years (the freshman retention rate), the percentage of students admitted as freshmen who graduated (the graduation rate), determining both the predicted graduation rate and the actual graduation rate, the percentage of freshman applicants who were accepted (the acceptance rate), and the percentage of alumni who made donations to the university after graduation (the alumni giving rate) for this class.

School	Freshman Retention	Predicted Graduation Rate	Actual Graduation Rate	Acceptance Rate	Alumni Giving Rate
University of California – Berkeley	95	81	82	27	18
Johns Hopkins University	95	89	87	33	28
Northwestern University	96	85	92	32	29
Columbia University	97	89	91	14	32
University of Chicago	94	89	81	48	35
Cornell University	96	87	91	33	36
Stanford University	98	91	90	15	37
Washington University in St. Louis	96	83	86	34	37
Rice University	95	91	88	27	39
Emory University	92	83	86	42	39
University of Pennsylvania	96	87	90	26	40
California Institute of Technology	92	96	82	18	41
Brown University	97	87	93	17	43
Massachusetts Inst. of Technology	97	94	91	19	43
Duke University	97	89	92	28	45
Harvard University	96	94	97	11	47
University of Notre Dame	98	80	95	35	48
Yale University	98	93	94	16	49
Dartmouth College	96	90	94	21	52
Princeton University	99	92	96	11	66

# Cautions

- Use meaningful identifiers for
  - Variables in data table
    - May not remember what heading represents later
    - Keep a code book for data for software that does not allow many characters for column names
    - Make sure that you have the context for data
- Be skeptical
  - Information, context, and testing questions can be slanted

# To do Statistics right,

- **Think** about the scenario that you are examining
- **Show** through calculating statistics and making necessary displays
- **Tell** what you have learned through your analysis so that *anyone can understand*

# Caution

- In statistics, there can be *more than one right answer* since a statistician interprets the data and the statistics.
- Important for “right answer”
  - Analysis
  - Explanation
  - Tools
    - ◉ Graphs/tables
    - ◉ Calculations

# Tabular representations of data

- **Frequency table**
  - Lists counts for categorical data
  - Records totals and category names
- **Relative frequency table**
  - Gives percentages rather than counts for each category name
- **Frequency and relative frequency tables describe the distribution of a categorical variable**

# Tabular representations of data

- Frequency table
  - Lists counts for categorical data
  - Records totals and category names
- Relative frequency table
  - Gives percentages rather than counts for each category name
- Frequency and relative frequency tables describe the distribution of a categorical variable

# Graphical representations of data

- Display of data reveals things that one will most likely not see in data table
  - Patterns
  - Groupings
  - Relationships
- Helpful for telling others about data