THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Online Review Course of
Undergraduate Probability and Statistics

Review Lecture 2
Descriptive Statistics, part 1

Chris A. Mack
Adjunct Associate Professor

Course Website: www.lithoguru.com/scientist/statistics/review.html
Data sets accompanying this lecture: StatReview_Lecture2&3.xlsx

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Descriptive Statistics

- Descriptive statistics
  - Describe or summarize a large set of data with a graph and/or just a few numbers
  - Applies to univariate data
  - The statistics that can be used depend on the measurement scale (nominal, ordinal, interval, or ratio)

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Nominal/Ordinal Scales

- Also called categorical data
- Three-step process
  - Decide on the categories (all categories are arbitrary, some categories are useful)
  - Count number in each category
  - Calculate statistics, graph counts/frequency (e.g., bar chart)

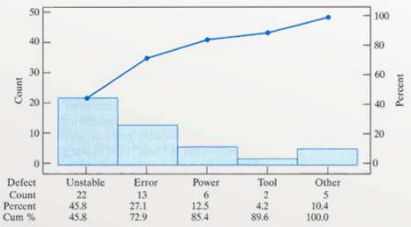
THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Nominal/Ordinal Statistics

- Nominal Scale statistics
  - Number of cases per category (frequency)
  - Mode (category with largest frequency)
  - Contingency table and correlation (for multivariate data)
- Ordinal scale, add these statistics:
  - Median
  - Percentiles

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Plotting Nominal Scale Data



Defect	Unstable	Error	Power	Tool	Other
Count	22	13	6	2	5
Percent	45.8	27.1	12.5	4.2	10.4
Cum %	45.8	72.9	85.4	89.6	100.0

Richard A. Johnson, "Miller and Freund's Probability and Statistics for Engineers", 8th edition, Prentice Hall, Figure 2.1, p. 13 (2011).

Note the nice combination of table and bar chart (Pareto diagram) in one.

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Interval/Ratio Statistics

- We call interval/ratio scale data "quantitative" data
- Interval Scale statistics
  - Mean (measure of central tendency)
  - Standard deviation (measure of spread)
  - Correlations (for multivariate data)
- Ratio scale, add these statistics:
  - Relative standard deviation (coefficient of variation): (standard deviation)/mean

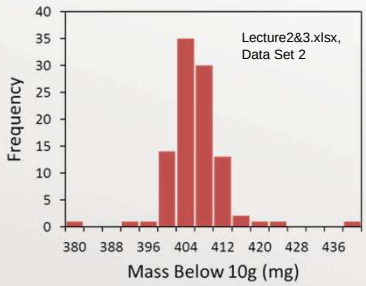
THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Plotting Univariate Quantitative Data

- The most common means of plotting univariate quantitative data is with the *histogram*
  - Separate the full range of data into equal-sized, non-overlapping bins
  - Count the number of data points in each bin
  - To avoid overlap, give intervals one open (< or >) and one closed (≤ or ≥) boundary – e.g., (5,10].
- Two arbitrary plotting choices:
  - How many bins to use
  - Where the first bin starts

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Histograms



- Shape may change dramatically depending on bin settings
- Bins with few counts have high statistical uncertainty
- Interpretation can be difficult without huge amounts of data
- It is often useful to plot the cumulative distribution as well
- Plotting % frequency is also common

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Practice

- Using the two data sets in Lecture2&3.xlsx, practice using Excel to make a histogram plot
  - Install the Analysis Toolpak Add-in (if you haven't already done so)
  - Data Tab, Data Analysis button, select Histogram
  - Create your own bins using the bin range feature
    - Rule of thumb: number of bins = SQRT(sample size)
  - Make a well-formated bar-chart plot
  - Compare to histograms already in the spreadsheet

THE UNIVERSITY OF TEXAS AT AUSTIN — WHAT STARTS HERE CHANGES THE WORLD

Review #2: What have we learned?

- What two measurement scales are generally called "categorical data"?
- What statistics apply to categorical data?
- How is univariate categorical data generally plotted?
- Why are histograms for quantitative data problematic?

