Lecture 24
Numerical Summaries of Quantitative Variables

STAT 225 Introduction to Probability Models
April 14, 2014

Whitney Huang
Purdue University

Agenda

1 Population vs. Sample
2 Numerical Summaries
3 Graphical Summary: Boxplots

Notes
Population vs. Sample

- The term "population" is used in Statistics to represent all possible outcomes that are of interest in a particular study.
- The term "sample" refers to a portion of the population that is representative of the population.
- We use parameters to describe the population.
- We use statistics to describe the sample with respect to the population.

Statistics provides a way to make inferences of the population by using sample data.

Numerical Summaries of data

- **Mean**: the average/expected value of a set of numbers.
  - Population mean: \( \mu_x = \frac{\sum_{i=1}^{N} x_i}{N} \)
  - Sample mean: \( \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \)

- **Variance**: measures how far a set of numbers is spread out.
  - Population variance: \( \sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N} \)
  - Sample variance: \( s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} \)

- **Mode**: the value that appears most often in a set of numbers.

- **Range**: the largest value – the smallest value in a set of numbers.
Example 59

Suppose we have the data set 1, 2, 3, 4, and 5. Find the mean of the data. Also compute variance in 2 ways (one assuming that this is a sample, the other assuming that this represents the entirety of the population)

Solution.

Numerical Summaries: Percentiles

- **Percentile**: The $p_{th}$ percentile is a value of the data set such that at least $p\%$ of the data set is less than or equal to this value
- Calculation of Percentiles using the indexing method:
  1. Sort the set of numbers in an increasing order
  2. For $p_{th}$ percentile, compute $i = \frac{np}{100}$ where $n$ is the sample size
  3. If $i$ is an integer then $p_{th}$ percentile is the average of $i_{th}$ value and $(i + 1)_{th}$ value, otherwise take the $(i + 1)_{th}$ value
- **Quartiles**:
  - $Q1$: first quartile
  - $M$ or $Q2$: median or second quartile
  - $Q3$: third quartile
  - **Interquartile range or IQR**: $Q3 - Q1$
**Regular Boxplots**

A boxplot is a visual representation of the 5 number summary: Min, Q1, Median, Q3, Max

**Modified Boxplots**

- The modified boxplot will highlight if there are outliers.
- **Outliers**: an outlier is a number that is far from other numbers.
- **LL (Lower Limit)**: \( LL = Q1 - 1.5 \times IQR \)
- **UL (Upper Limit)**: \( UL = Q3 + 1.5 \times IQR \)
- A number is considered as an outlier if it is \( \leq LL \) or \( \geq UL \)
Example 60

Hank Aaron hit an astounding 755 home runs in his career. His career spanned from 1954 through 1976. In those 23 seasons he hit 13, 27, 26, 44, 30, 39, 40, 34, 45, 44, 24, 32, 44, 39, 29, 44, 38, 47, 34, 40, 20, 12, 10 home runs.

- What is the mode of the data set?
- What is the range of the data set?
- Create both a regular and a modified boxplot for the number of home runs that Hank Aaron hit in a season
- Find the 61st percentile

Example 60 cont’d

Solution.