Descriptive Statistics

URBPL 5/6010: Urban Research

University of Utah

Pam Perlich

Statistics

- Etymology: German *Statistik* study of political facts and figures, from New Latin *statisticus* of politics, from Latin *status* state
  1 : a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of masses of numerical data
  2 : a collection of quantitative data

Merriam-Webster Online Dictionary, 11/04/2005

Description vs. Inferential

- Descriptive Statistics
  - Procedures to describe and summarized a data set
  - Concise information to characterize data
- Inferential Statistics
  - Use sample data to draw inferences about the population
  - Comparing groups
  - Discovering relationships

Descriptive Statistics

- Generally description
- Measures of
  - Central tendency
  - Distribution
  - Size of data set
  - Skewness
- Excel
  - Tools => Data Analysis => Descriptive Statistics

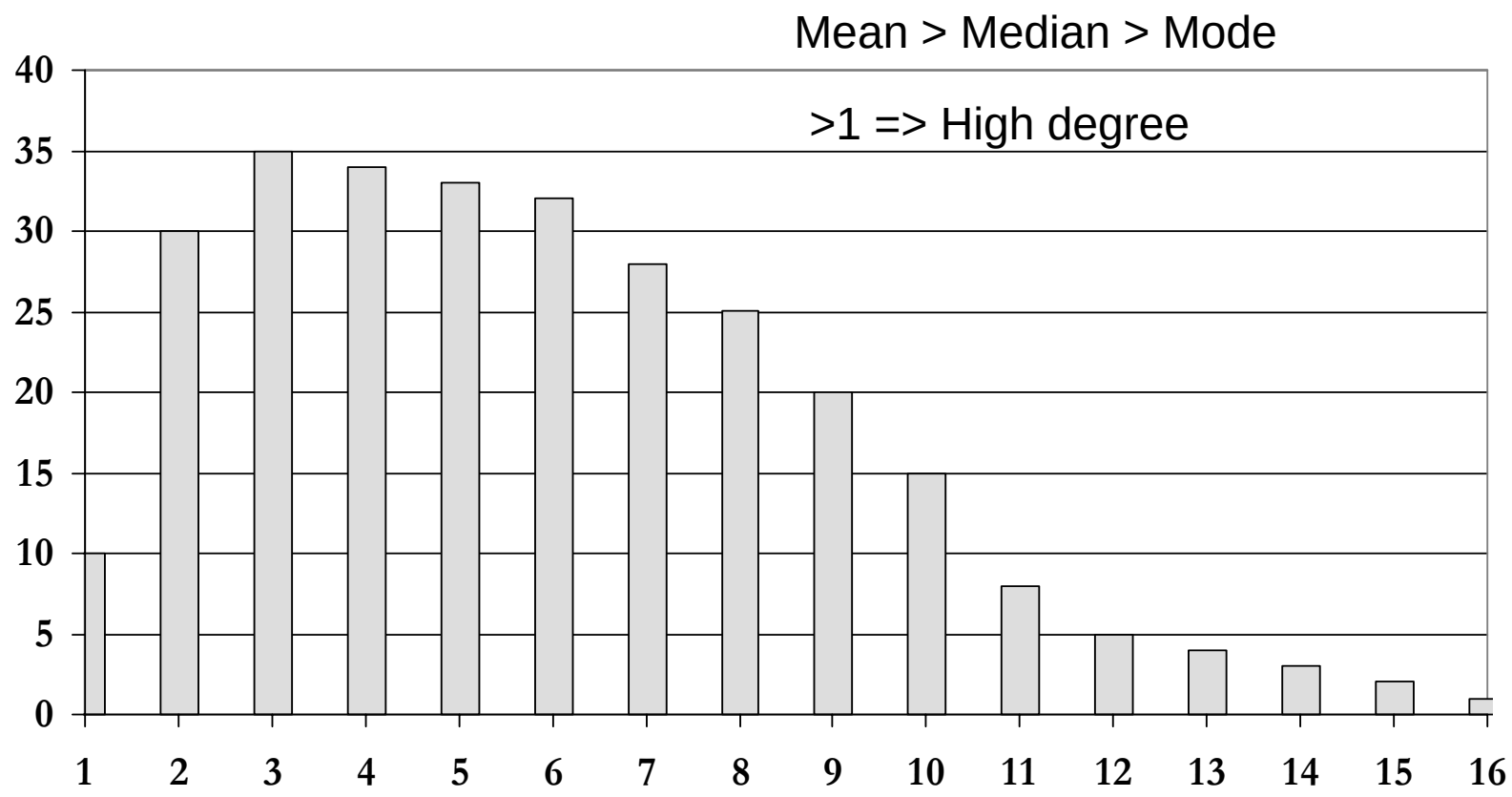
Central Tendency

- Mean – simple average -
  - n is the number of observations
  - AVERAGE function (Excel)
  - Outliers affect mean

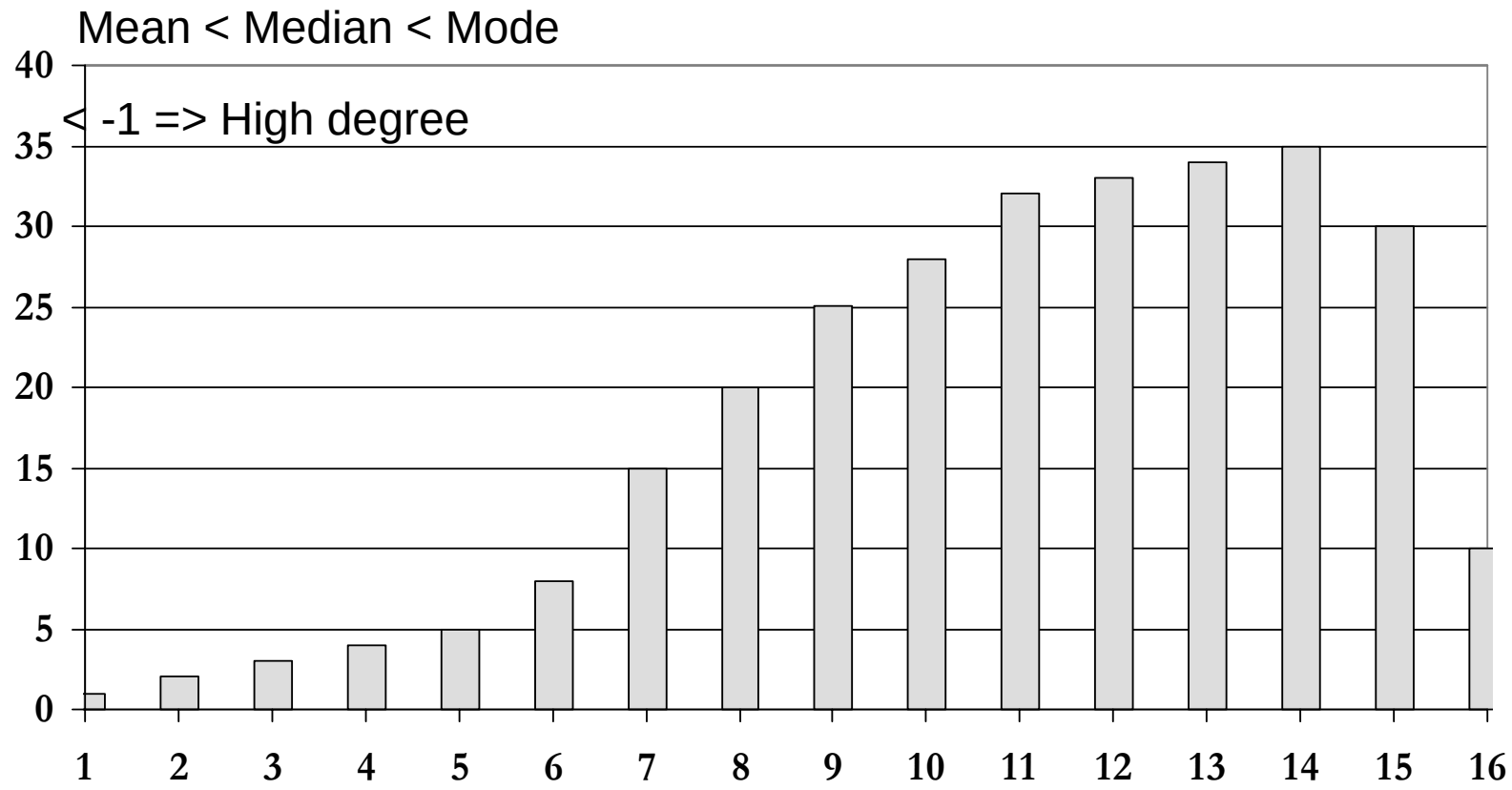
$$\bar{x} = \frac{\sum_1^n x_i}{n}$$

- Median – middle observation
  - Array data from low to high
  - MEDIAN function (Excel)
- Mode – most frequently occurring
  - MODE function (Excel)
  - Seldom used
- Symmetric data => Mean = Median = Mode

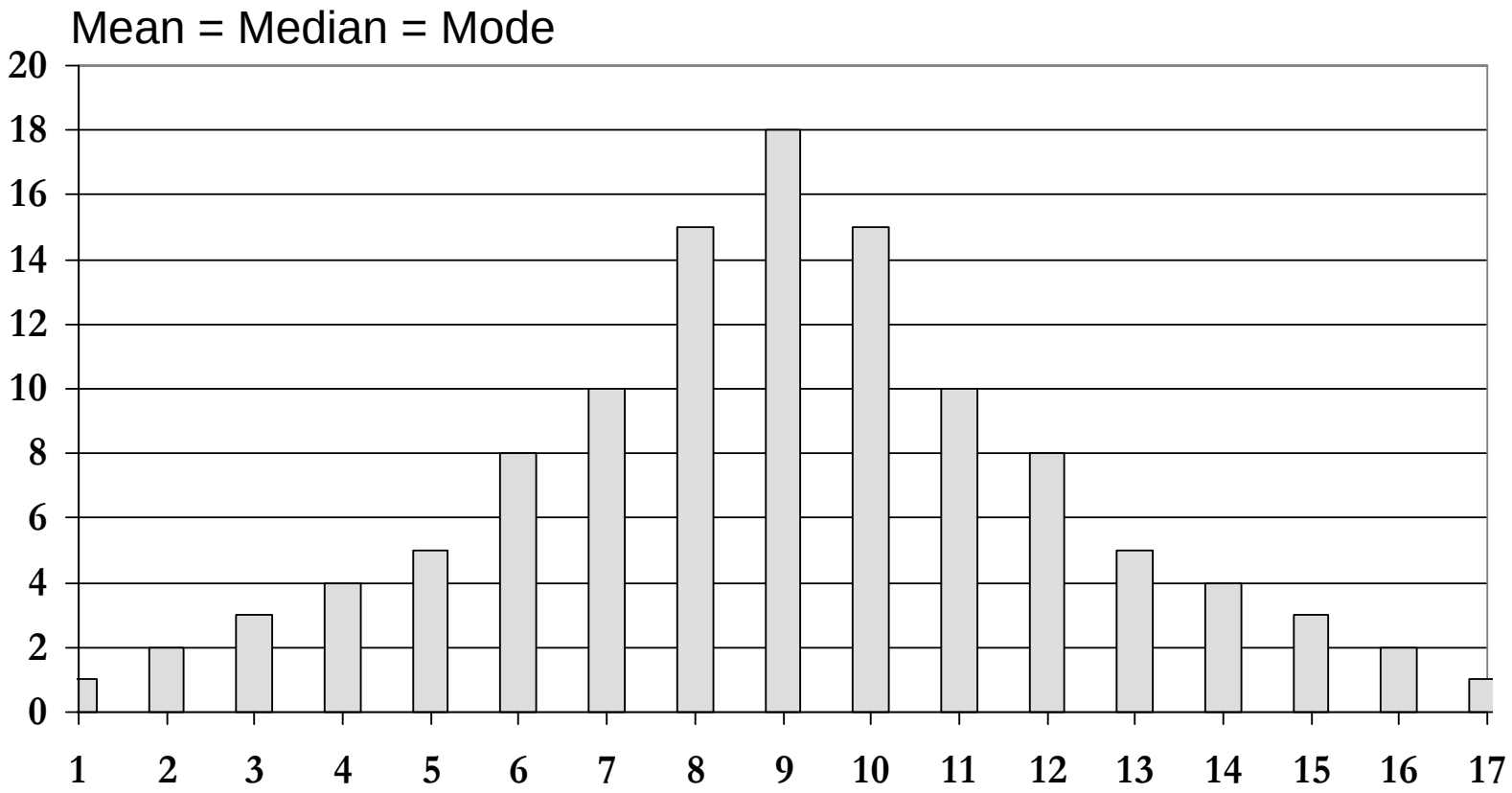
Positive Skewness: Skewed to the Right



Negative Skewness: Skewed to the Left



Symmetric


Mean = Median = Mode
20
18
16
14
12
10
8
6
4
2
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17

Kurtosis

- Measure of how “peaked” the distribution is
- Compares to a normal distribution
- Very Peaked => Kurtosis is positive
- Very Flat => Kurtosis is negative

Sampling

- Sample → population
  - Representative
  - Random
  - Unbiased
- Statistics about samples

Sample Standard Deviation & Variance

- Variance:

$$s^2 = \left(\frac{\sum_1^n (x_i - \bar{x})^2}{n-1} \right)$$

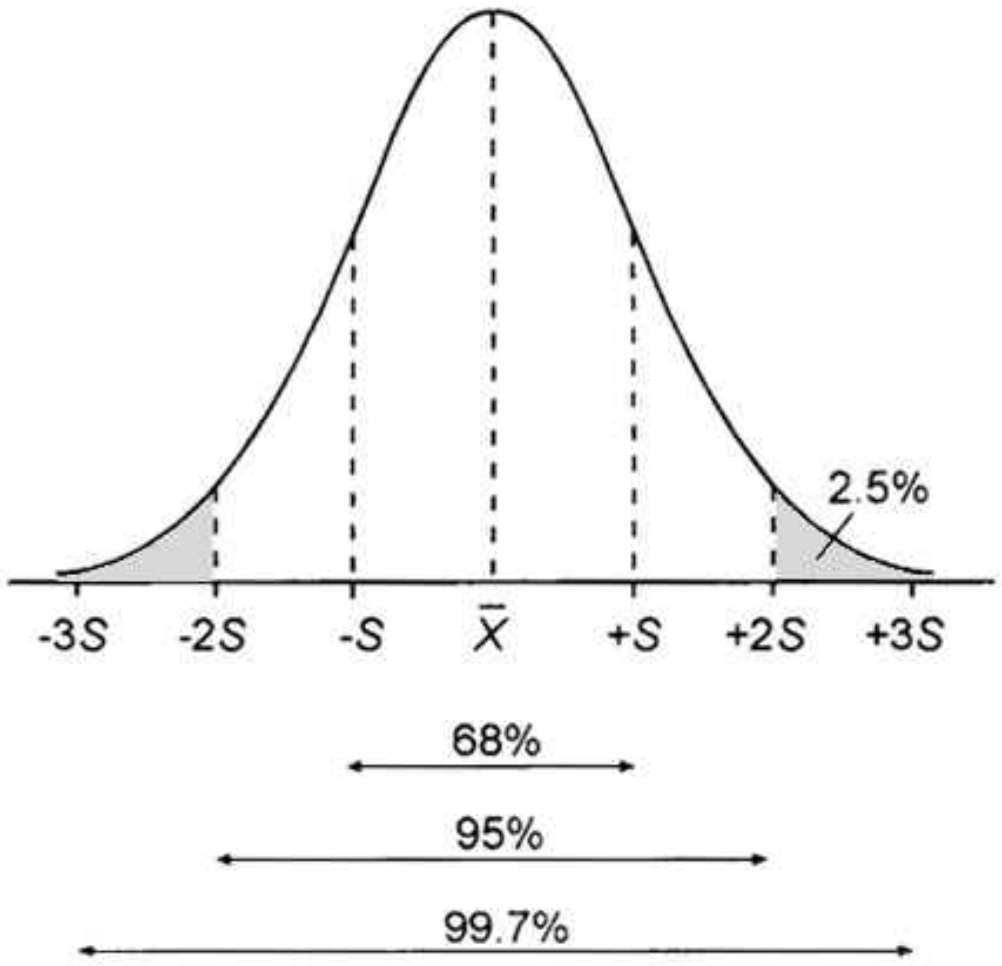
- Standard Deviation:

$$s = \sqrt{\left(\frac{\sum_1^n (x_i - \bar{x})^2}{n-1} \right)}$$

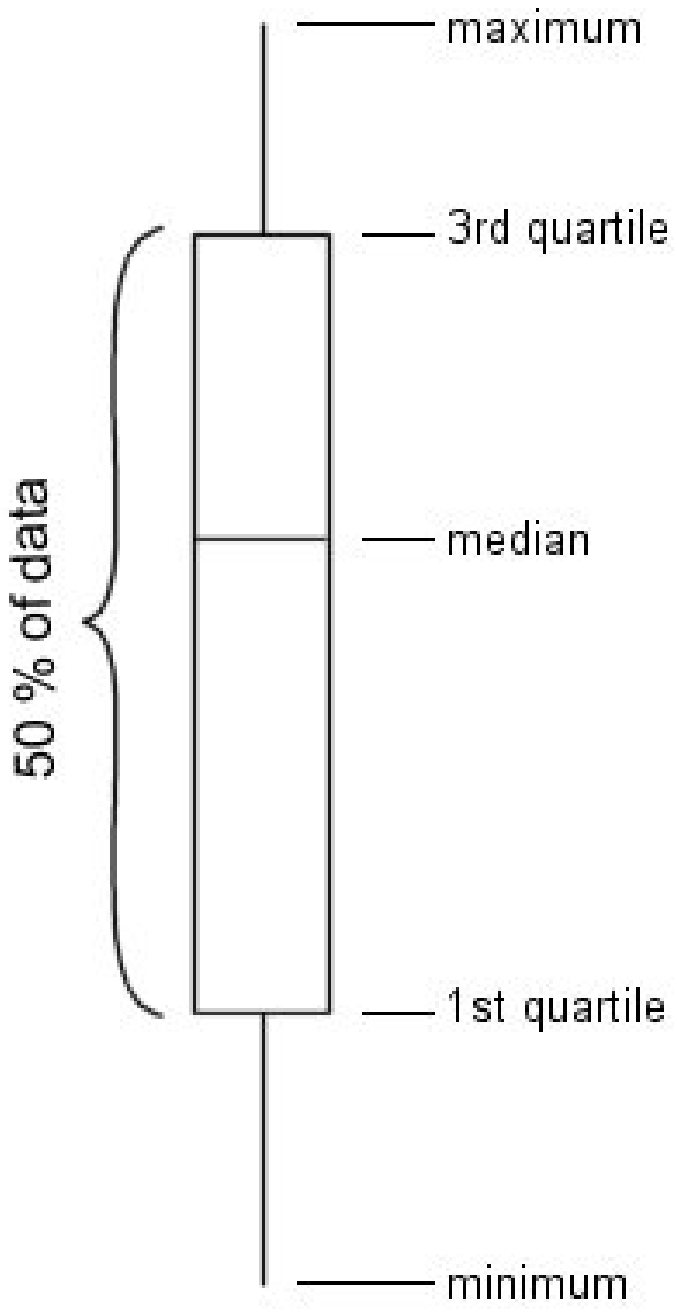
Distribution of Sample

- About 68% of observations are within 1 standard deviation of the mean
- About 95% of observations are within 2 standard deviation of the mean
- About 99.7% of observations are within 3 standard deviation of the mean

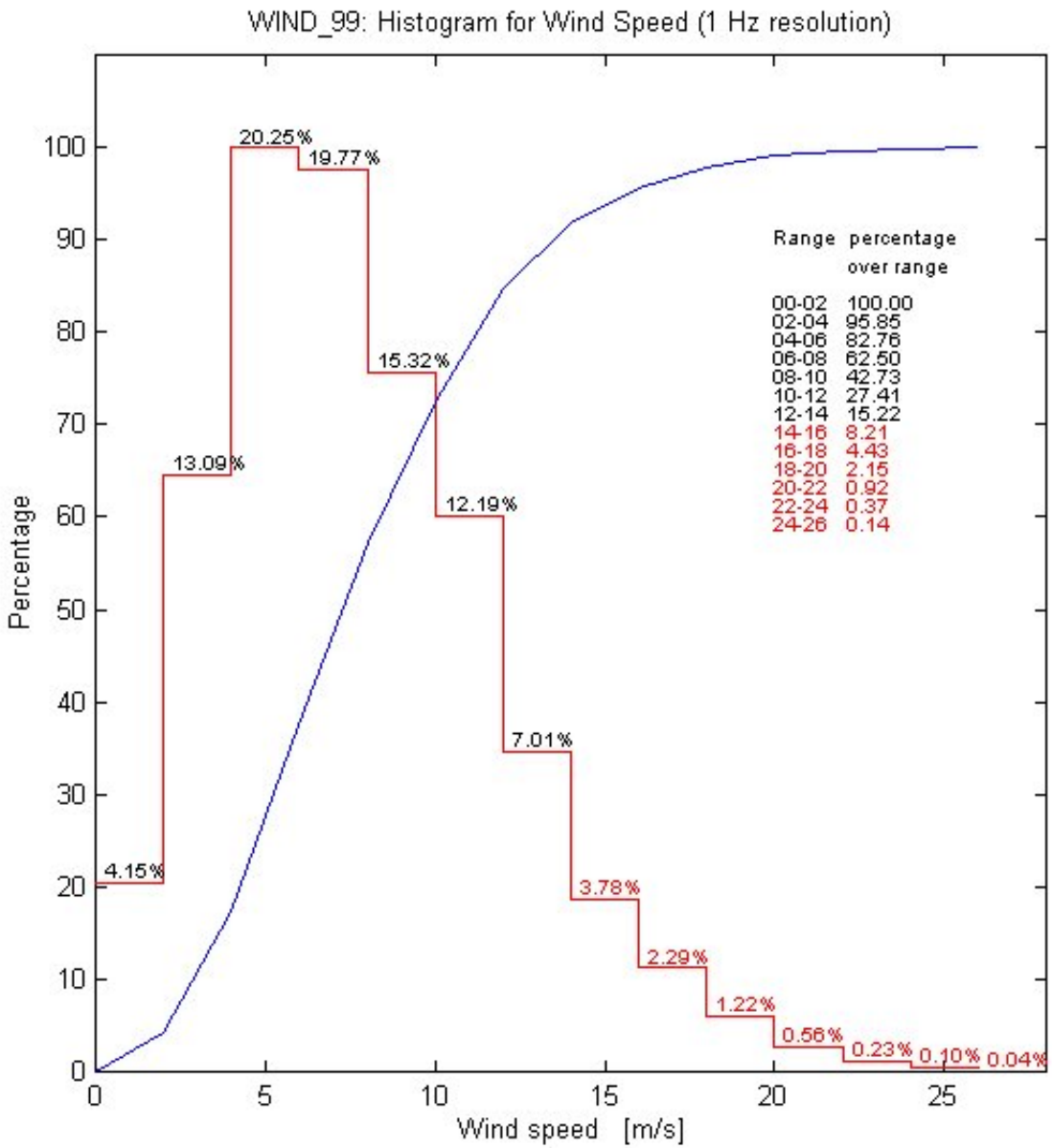
Normal Distribution



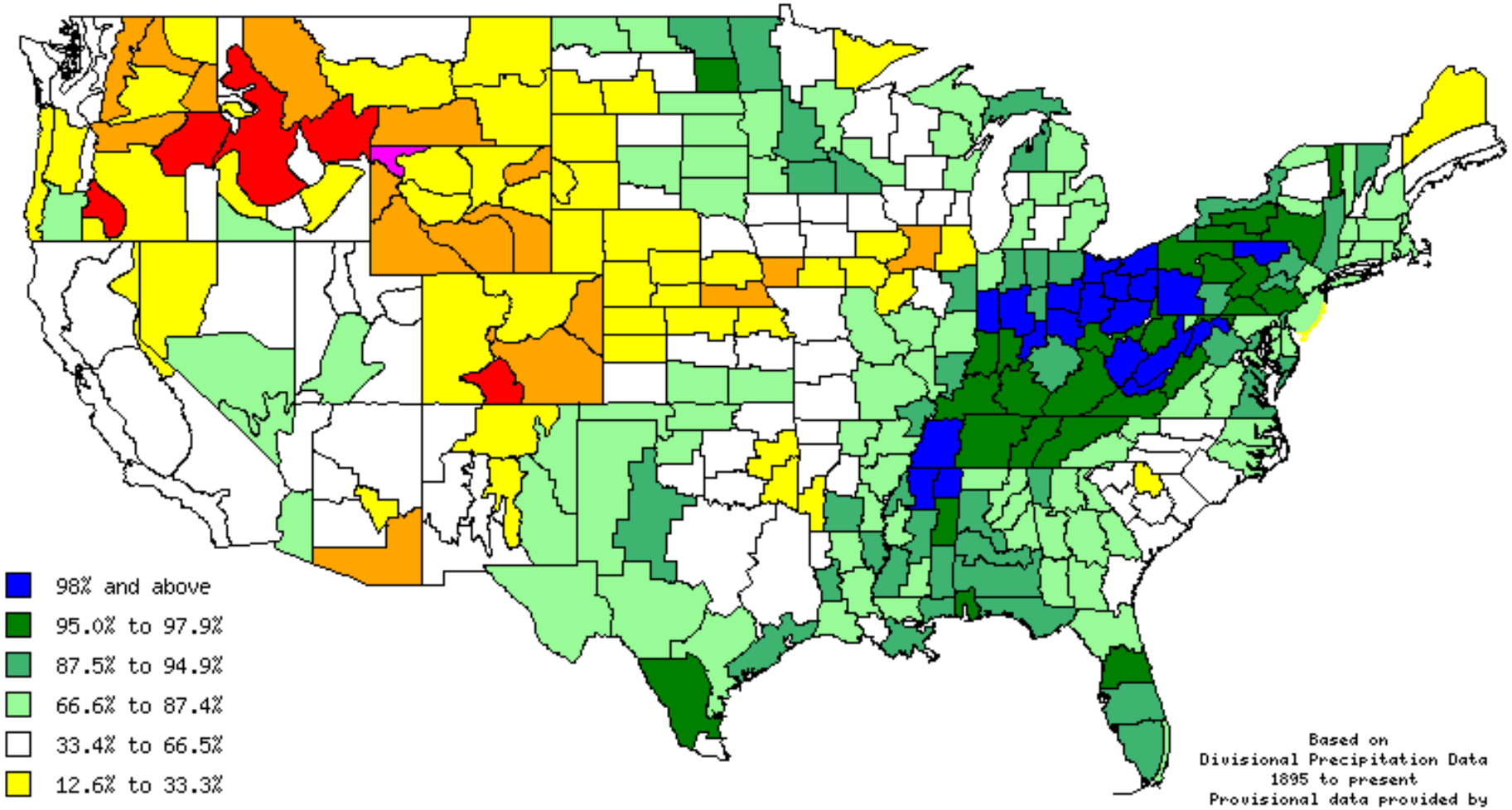
Box Plot


maximum
3rd quartile
median
1st quartile
minimum
50 % of data

Histogram


WIND_99: Histogram for Wind Speed (1 Hz resolution)
Percentage
Wind speed [m/s]
Range percentage over range
00-02 100.00
02-04 95.85
04-06 82.76
06-08 62.50
08-10 42.73
10-12 27.41
12-14 15.22
14-16 8.21
16-18 4.43
18-20 2.15
20-22 0.92
22-24 0.37
24-26 0.14
4.15%
13.09%
20.25%
19.77%
15.32%
12.19%
7.01%
3.78%
2.29%
1.22%
0.56%
0.23%
0.10%
0.04%

48-month Precipitation Percentile (non-exceedance) through the end of September 2005



Based on
Divisional Precipitation Data
1895 to present
Provisional data provided by
NOAA/NWS/CPC & NOAA/NESDIS/NCDC

Western Regional Climate Center
Desert Research Institute
Reno, Nevada

Excel

- Tools => Data Analysis
  - Descriptive Statistics
  - Histogram
    - Pareto ranks by frequency
    - Bin values –sets ranges for histogram
  - Rank and Percentile