

Introduction to Statistics for Traffic Crash Reconstruction

Jeremy Daily
Jackson Hole Scientific Investigations, Inc.

©2003

www.jhscientific.com

Why Use and Learn Statistics?

1. We already do when ranging solutions for traffic crash reconstruction.
2. All human performance data is reported statistically.
3. Real life is statistical so statistics are used everywhere.
4. People lie with statistics (knowingly or unknowingly) and understanding is the key to the truth.
5. Some one opposing you may use statistical arguments to make you doubt your reconstruction work.

What is statistics?

Statistics The mathematics of the collection, organization, and interpretation of numerical data, especially the analysis of population characteristics by inference from sampling. (THE AMERICAN HERITAGE COLLEGE DICTIONARY)

The important phrase is at the end: *population characteristics by inference from sampling.*

Population is the complete set of data for all the subject in a characteristic group.

Sample is a portion of the population.

Example of Population v Sample

- ✘ Polls report the opinions of a sample of Americans, not all Americans.
- ✘ Perception-Reaction times are based on a sample of capable drivers– not all drivers
- ✘ Pedestrian walking times have been recorded for a handful of people, not all people who walk the streets.
- ✘ Different notation is used for a population or a sample
 - ✘ Sample: \bar{x} and s
 - ✘ Population: μ and σ

Misused Statistics– Case #1

Many sample come from a limited population and are then assumed to reflect on the total population. For example, if SAT test scores are gathered from West Virginia are shown to decrease over the years, then the scores for all American students have been decreasing.

- False Assumption: West Virginians are not representative of all Americans
- Remedy: Make sure the statistic describes the population you are interested in. No one would believe a drag factor of .8 on wet icy roads based on 30 skid tests done on dry pavement.

- Once again, a sample of tests only apply to the specific quantity being tested. Most qualifying conditions are not reported.

Uncertainty

- There are three interrelated fields used to quantify uncertainty:
 - Descriptive Statistics** is the gathering, reducing, summarizing, and reporting of data.
 - Probability** is the mathematics of chance developed first for gambling.
 - Statistical Inference** is the science of making informed decisions based on uncertain information.
- Either a specific point is unknown or
- The population characteristics are unknown.

Key Understanding

1. The quantity we are trying to determine is fixed but is unknown.
2. Knowing the exact value is impossible, but quantifying the uncertainty is done using statistical methods.
3. For example, the value of the drag factor for a sliding vehicle in a specific crash is fixed– it does not vary. However, the testing for that particular drag factor do vary, which enables us to treat the drag factor as a varying quantity.

Descriptive Statistics (Data Analysis)

- Measures of Central Tendency
- Measures of Spread or Variation
- Understanding Percentile
- Statistical Graphing
- Quiz

Descriptive Statistics Case Study

Two traffic crash reconstruction classes participated in a field study during which each member was asked to walk 100 feet. The time for each person was recorded from which speed could be calculated. The data are as follows (in mph):

Pedestrian Walking Speeds (mph)									
3.34	3.13	3.06	3.24	2.92	2.87	2.93	3.31	3.14	3.36
3.18	3.01	3.14	3.08	3.67	3.56	3.24	3.21	3.57	3.76
3.63	3.13	3.62	3.28	3.00	3.79	2.51	3.82	3.25	3.12
2.95	3.25	3.41	2.63	3.13	2.97	3.17	2.97	3.17	2.95
3.16	3.26	3.29	3.10	2.56	4.00	3.29			

Sorting

- ✓ Data needs to be sorted in order to make sense of it.
- ✓ Sort from the bottom up

index refers to the position of the sorted data starting with the lowest value having an *index* = 1

- ✓ Sort from the top down

rank refers to the position of the sorted data starting with the highest value having a *rank* = 1

Sorting Example

Index	1	2	3	4	5	6	7	8	9
x	25.1	26.3	27.0	28.8	31.5	35.2	38.7	40.1	45.2
Rank	9	8	7	6	5	4	3	2	1

- Computers work well for sorting as long as care is taken when interpreting the results.
- Remember “Garbage in– Garbage out”

Summation Notation

Summation Notation is a shorthand of writing long addition problems. It is used extensively in statistics and other types of math. It uses the Greek capital letter for sigma Σ .

$$\sum_{i=1}^n x_i = x_1 + x_2 + x_3 + \cdots + x_n$$

where x is the variable, i is the index and n is the total number of data points.

Measures of Central Tendency

First thing people ask– “What is the average?”

Average is any number that typifies the data.

Arithmetic Mean is the most common average having the formula:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

Median is the middle value when the data are sorted.

Mode is the value that occurs the most frequently.

Geometric Mean is an average based on the products of each data point with the formula.

$$\text{geometric mean} = \sqrt[n]{x_1 \times x_2 \times \cdots \times x_n}$$

Example: $X = \{1, 2, 3, 3, 4, 5, 6\}$

- Arithmetic mean: $\bar{x} = \frac{1+2+3+3+4+5+6}{7} = \frac{24}{7} \approx 3.429$
- Median: the 4th index and the 4th rank is 3, thus making it the middle.
- Mode: 3 (it occurs twice)
- Geometric mean: $\text{geometric mean} = \sqrt[7]{1 \times 2 \times 3 \times 3 \times 4 \times 5 \times 6} = \sqrt[7]{2160} \approx 2.995$

Measures of Spread

Range is the difference between the maximum and the minimum values.

Interquartile Range is the difference between the third quartile (75th percentile) and the first quartile (25th percentile).

Variance is the average of the sum of the deviation squared. Here is the formula:

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

The Standard Deviation

Standard Deviation is the square root of the variance. It has the same units as the mean with the following formula:

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

Coefficient of Variation is a way to compare the variation of different things. It is defined as the ratio of the standard deviation to the mean:

$$CV = \frac{s}{\bar{x}}$$

Understanding Percentile

- Percentile gives relative standing to the data broken down in 1/100ths.
- The median is the 50th percentile.
- Can also have quartiles where the data is broken into quarters.
 - Q_1 – the first quartile (25th percentile)
 - Q_3 – the third quartile (75th percentile)
- Breaking up data into tenths gives deciles.

Determining Percentile

- Percentile is determined by “looking up” the value.
- Depends on how the data was sorted

- If indexed:

$$index = \frac{P}{100}(n - 1) + 1$$

- If ranked:

$$rank = n - \frac{P}{100}(n - 1)$$

Linear Interpolation

Take the calculated quantity *index* and split it at the decimal.

1. Call the whole number i and call the fraction or remainder R .
2. Now x_i is the lower value before the index and x_{i+1} is the higher value after the index.
3. The percentile point (PP) follows:

$$PP = x_i + R(x_{i+1} - x_i)$$

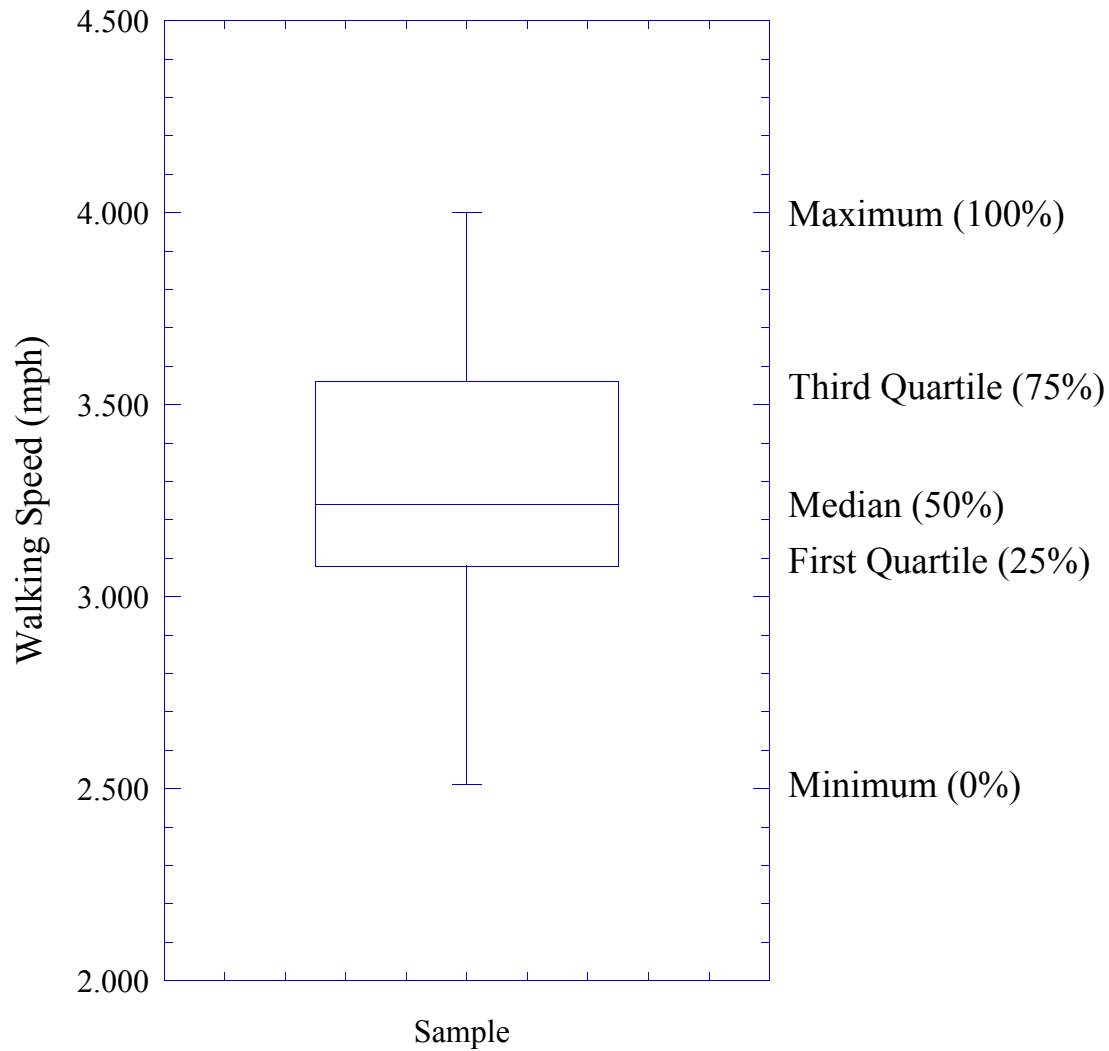
The Box and Whisker Plot

A box and whisker plot gives a five number summary and visualization of the data.

1. Maximum
2. Minimum
3. Median
4. First Quartile
5. Third Quartile

All based on relative standing (Percentile) while giving a good picture of the data distribution.

**Box and Whisker Plot of
Walking Speeds of Two Crash Reconstruction Classes**



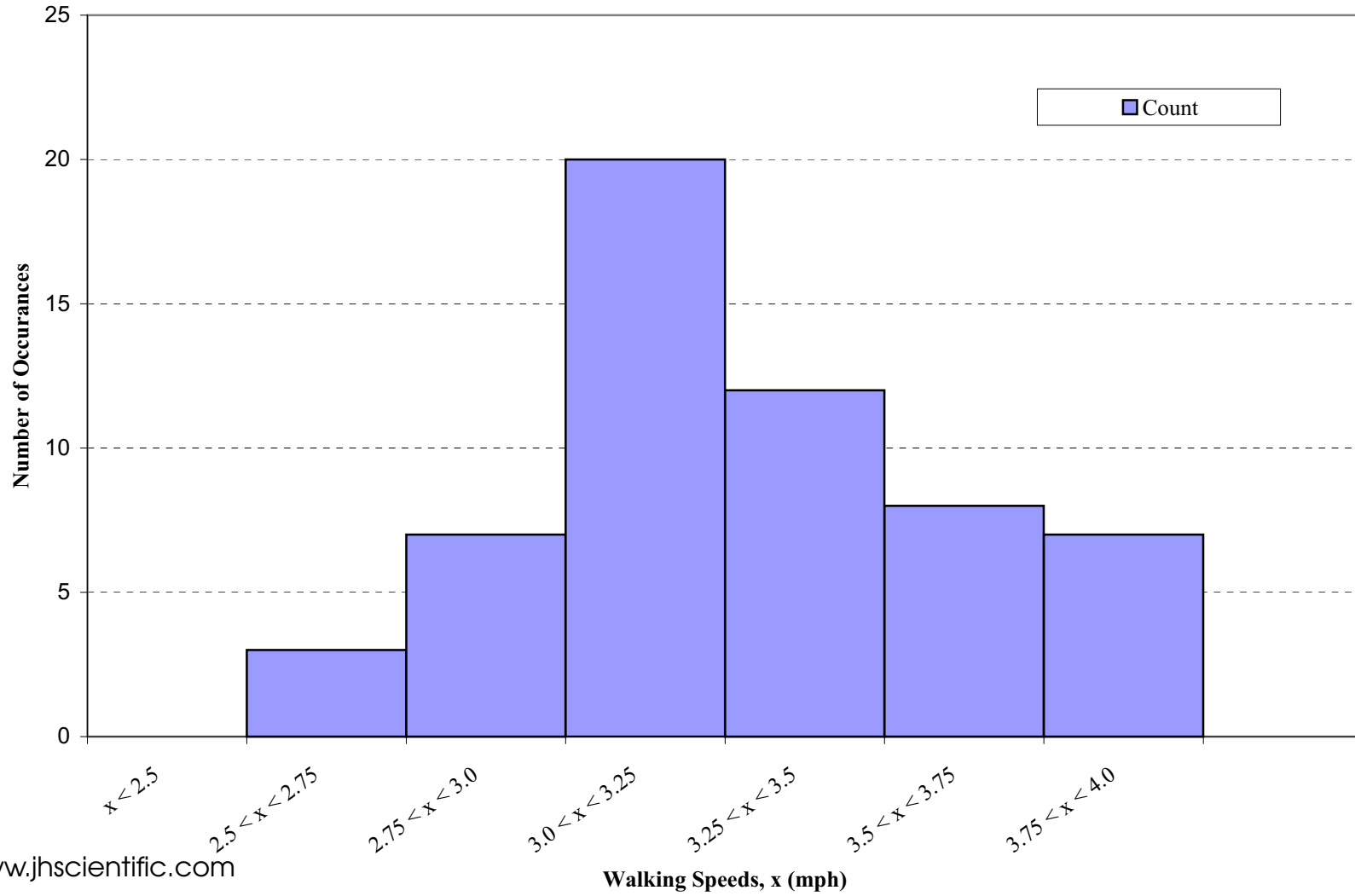
Histogram

A Histogram is a frequency chart showing how the data is distributed.

To construct a histogram:

1. Divide the x-axis into into bins, usually of a fixed width.
2. Sort the data and place it in the appropriate bin.
3. Count how many points are in each bin to determine the frequency.
4. Make a bar chart of the frequency vs. bin range.

Histogram of Walking Speeds of Two Crash Reconstruction Classes

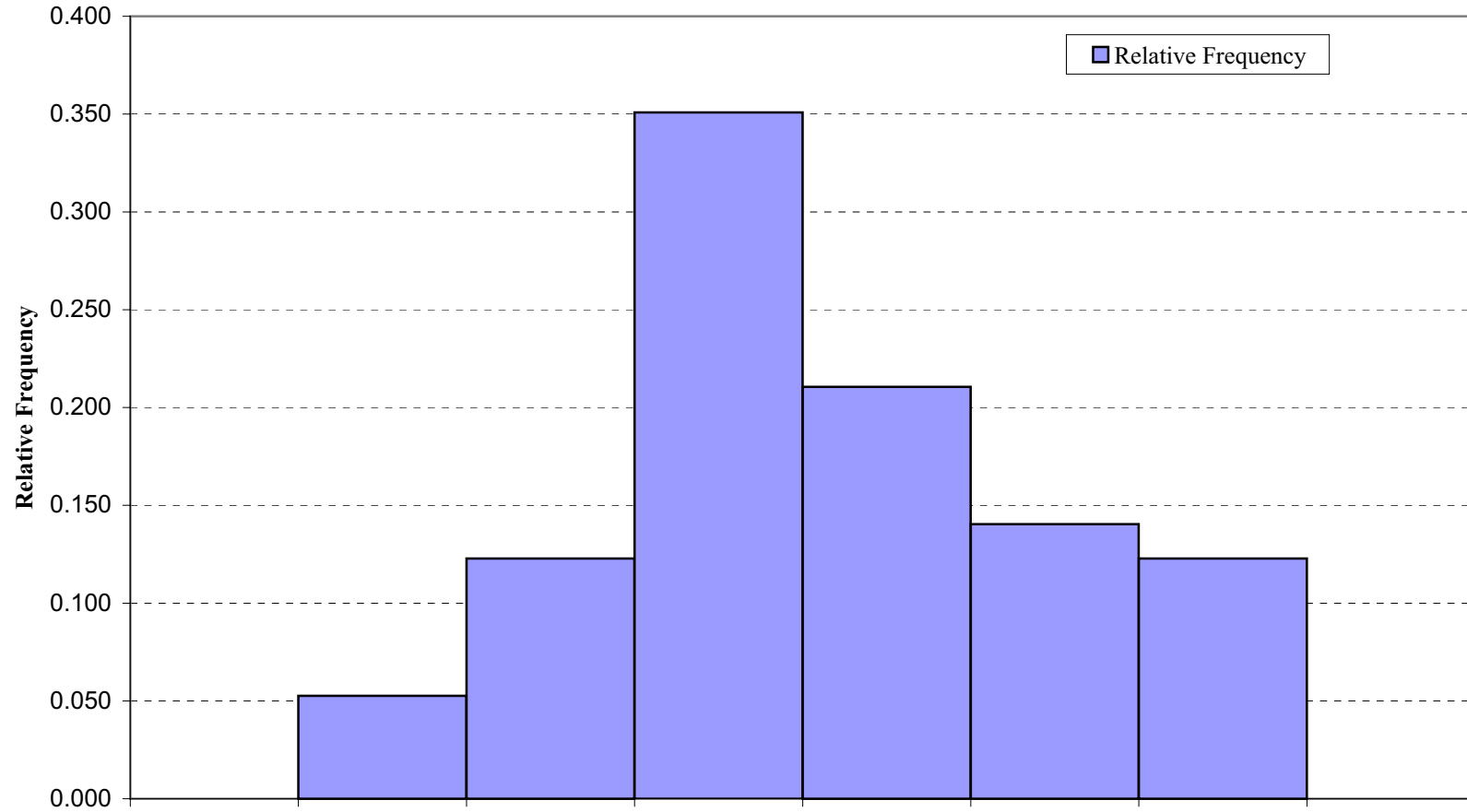


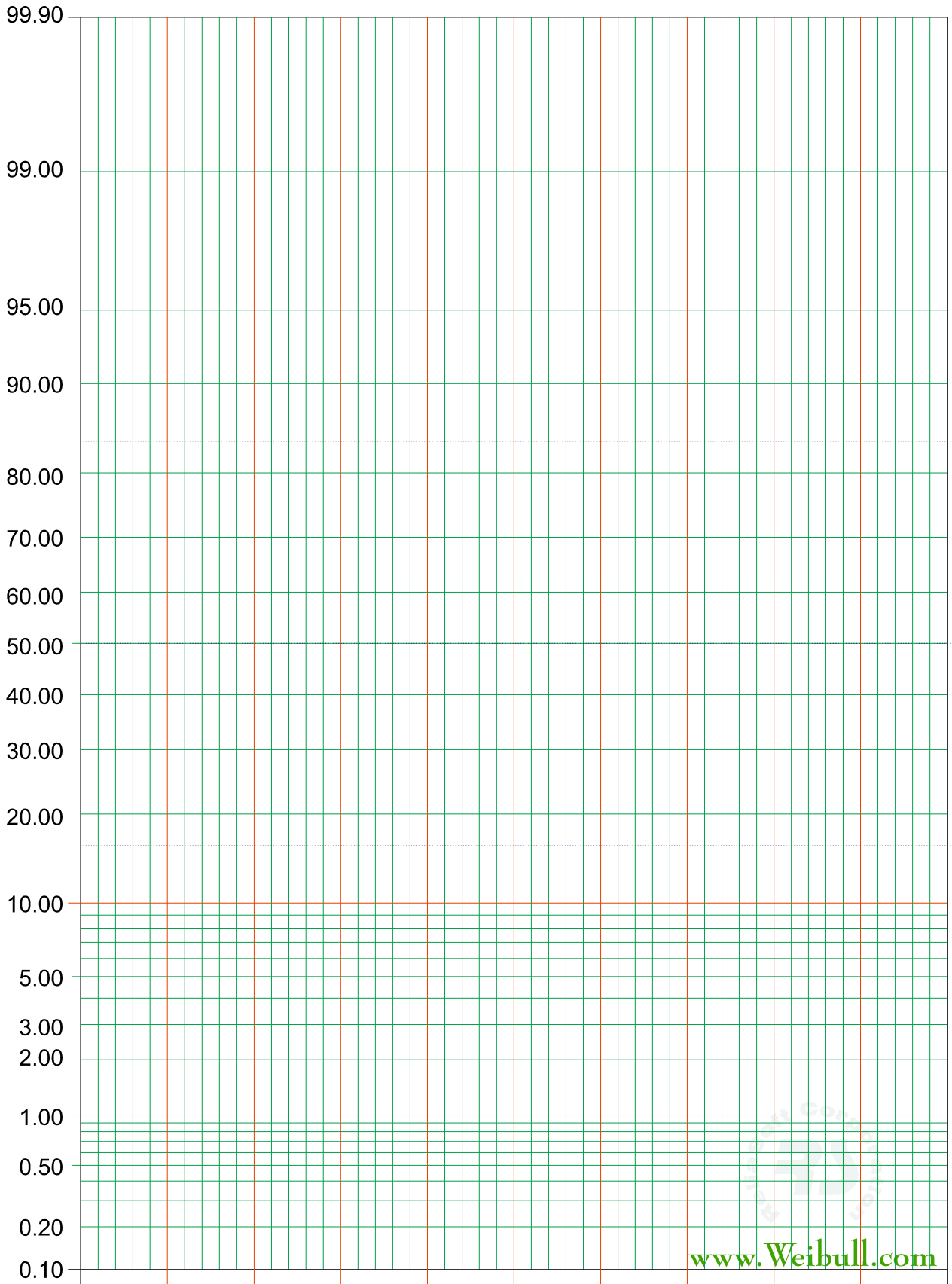
Relative Frequency Diagram

The relative frequency diagram looks the same as a histogram except the y-axis is different. Instead of using the total frequency, the relative frequency is used, which means:

$$rel. freq. = \frac{frequency}{n}$$

**Relative Frequency Graph (Percent) of
Walking Speeds of Two Crash Reconstruction Classes**





Descriptive Statistics Worksheet

Please ask questions!!!