Descriptive Statistics©

I. **Frequency Distribution**: a tallying of the number of times (**frequency**) each score value (or interval of score values) is represented in a group of scores.

A. **Ungrouped**: frequency of each score value is given

Statophobia	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	9	1.51	9	1.51
1	17	2.86	26	4.37
2	15	2.52	41	6.89
3	35	5.88	76	12.77
4	38	6.39	114	19.16
5	93	15.63	207	34.79
6	67	11.26	274	46.05
7	110	18.49	384	64.54
8	120	20.17	504	84.71
9	47	7.90	551	92.61
10	43	7.23	594	99.83
11	1	0.17	595	100.00

B. **Grouped**: total range of scores divided into several (usually equal in width) intervals, with frequency given for each interval.

1. Summarizes data, but involves loss of info, possible distortion
2. Usually 5-20 intervals

Nucophobia

		Frequency	Percent	Cumulative Percent
Valid	0-9	13	2.1	2.1
	10-19	17	2.8	4.9
	20-29	30	4.9	9.8
	30-39	45	7.3	17.1
	40-49	36	5.9	23.0
	50-59	144	23.5	46.5
	60-69	129	21.0	67.5
	70-79	81	13.2	80.8
	80-89	57	9.3	90.0
	90-100	61	10.0	100.0
	Total	613	100.0	

C. **Percent**: the percentage of scores at a given value or interval of values.

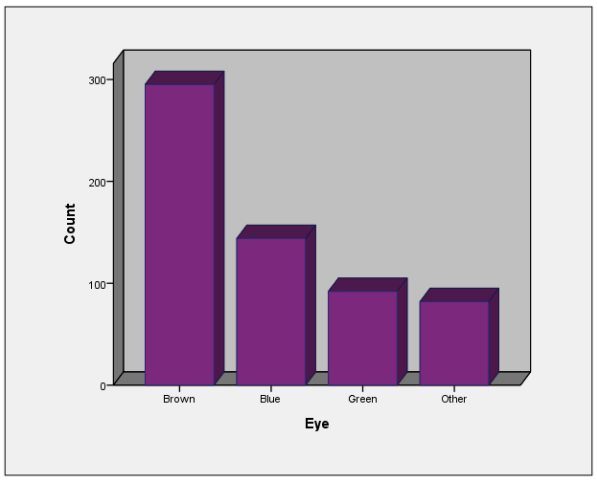
D. **Cumulative Frequency**: the number of scores at or below a given value or interval of values

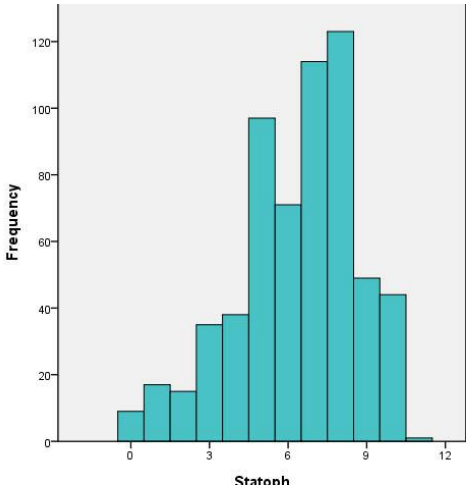
E. **Cumulative Percent**: the percentage of scores at or below a given value or interval of values. This is also known as the **Percentile Rank**.

II. Graphing

A. **Bar Graph** (below, left)

1. Bars should be separated, indicating the variable is discrete
2. Plot frequencies, percents, cumulative frequencies, or cumulative percents on the ordinate, values of the variable on the abscissa



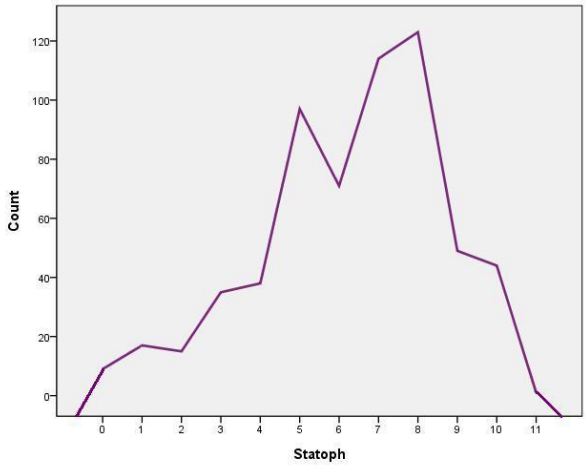


B. **Histogram** (above, right)

1. Continuously distributed variable; The bars are connected, indicating that the variable is continuous.
2. Plot frequencies, percents, cumulative frequencies, or cumulative percents on the ordinate, values of the variable on the abscissa

C. **Frequency Polygon (left)**→

1. As if you took a histogram, placed a dot at the middle of each bar, erased the bars, and joined the dots with straight line segments.
2. Connect ends to abscissa to form a polygon.



See SPSS Output with Frequency Distributions

D. **Three-Quarter Rule**

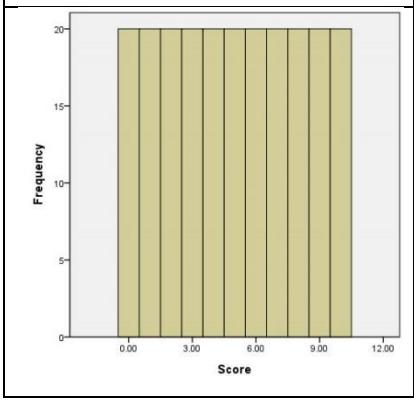
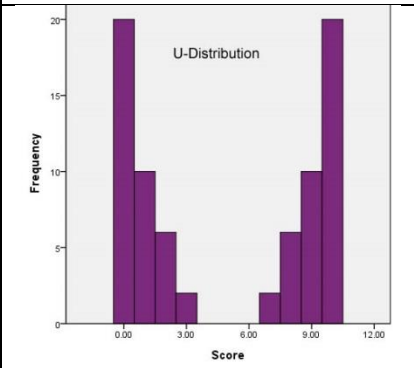
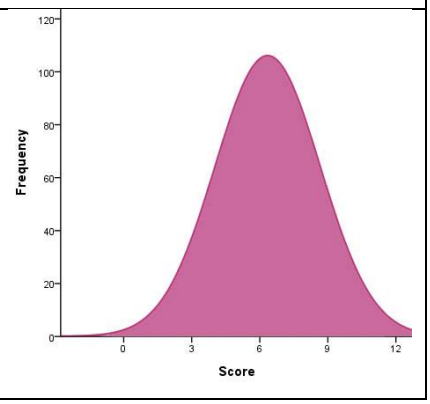
1. Height of highest point on the ordinate should be equal to three quarters of the length of the abscissa.
2. Violating this rule may distort the data.

E. **Gee-Whiz** - a way to distort the data

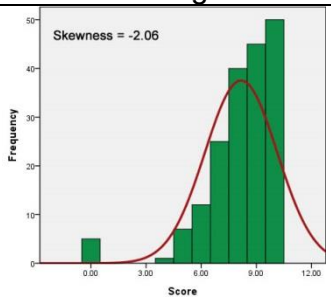
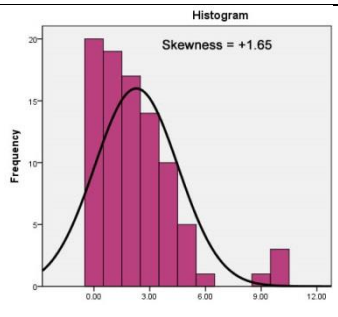
1. Ordinate starts with a frequency greater than zero
2. Exaggerates differences
3. Examples of the effects of violating the ¾ rule and of the Gee-Whiz technique can be viewed here.

F. **Shapes of Frequency Distributions**

1. Symmetrical – the left side is the mirror image of the right side
   a. rectangular or uniform - for example, ranked data – within some range, every score has the same frequency
   b. U and inverted U
   c. bell shaped

Uniform Distribution	U Distribution	Bell-Shaped Distribution
		

2. Skewed Distributions – they are lopsided
   a. Negative skewness: most of the scores are high, but there are a few very low scores.
   b. Positive skewness: most of the scores are low, but there are a few very high scores.
   c. In the plots below, I have superimposed a normal bell-shaped distribution with mean and standard deviation the same as that of the plotted distribution.

Negative Skewness	Positive Skewness
	

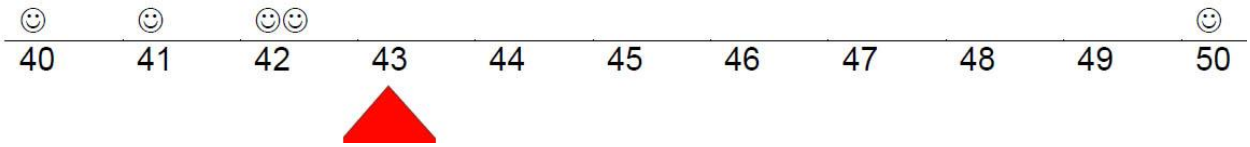
III. Measures of Central Tendency (Location)

A. **Mean** - three definitions (M = sample mean, μ = population mean)

1. $\mu = \Sigma Y \div N \Rightarrow$ just add up the scores and divide by number of scores
2. $\Sigma (Y - \mu) = 0$ – the mean is the point that makes the sum of deviations about it exactly zero – that is, it is a balance point. It is the mean that balances (makes equal) the sum of the negative deviations above it and the sum of the positive deviations about it.
3. $\Sigma (Y - \mu)^2$ is minimal – the mean is the point that makes the sum of squared deviations about it as small as possible. This definition of the mean will be very important later.

Illustration of (2). Five happy youngsters (smiley faces), four girls and one boy, are playing on a see-saw (aka teeter-totter). The girls sit on one side, the boy on the other. The play is best when the fulcrum is at the balance point, the point that makes the absolute differences between the weights and mean weight the same on one side as on the other, in which case the see-saw will balance. Where should we put the fulcrum to achieve such balance?

The answer is the mean of the children's weights. In this case the fulcrum should be placed at position 43.



The deviations from the mean (43) on the left side are 3 + 2 + 1 + 1 = 7 and on the right side 7.

B. **Median** - the preferred measure with markedly skewed distributions

1. The middle point in the distribution, the score or point which has half of the scores above it, half below it
2. Arrange the scores in order, from lowest to highest. The median location (*ml*) is (n + 1)/2, where n is the number of scores. Count in ml scores from the top score or the bottom score to find the median. If *ml* falls between two scores, the median is the mean of those two scores.
   10, 6, 4, 3, 1: *ml* = 6/2 = 3, median = 4.
   10, 8, 6 ,4, 3, 1: *ml* = 7/2 = 3.5, median = 5

C. **Mode** - the score with the highest frequency. A **bimodal distribution** is one which has two modes. A multimodal distribution has three or more modes.

D. **Skewness**

1. the mean is very sensitive to extreme scores and will be drawn in the direction of the skew – see IQ_Lake-Wobegon for an extreme example of how the mean is drawn in the direction of the skew.
2. the median is not sensitive to extreme scores
3. if the mean is greater than the median, positive skewness is likely

4. if the mean is less than the median, negative skewness is likely
5. one simple measure of skewness is (mean - median) / standard deviation
6. statistical packages typically compute ***g₁***, an estimate of Fisher's skewness, based up the sum of cubed deviations of scores from their mean, $\sum(Y-\mu)^3$. The value 0 represents the absence of skewness. Values between -1 and +1 represent trivial to small skewness.

IV. Measures of Variability (Dispersion)

A. Each of the four distributions in the table below has a mean of 3, but these distributions obviously differ from one another. Our description of them can be sharpened by measuring their variability.

X	Y	Z	V
3	1	0	-294
3	2	0	-24
3	3	15	3
3	4	0	30
3	5	0	300

B. **Range** = highest score minus lowest score

X: 3 - 3 = 0

Y: 5 - 1 = 4

Z: 15 - 0 = 15

V: 300 - (-294) = 594

C. **Interquartile Range** = $Q_3 - Q_1$, where Q_3 is the third quartile (the value of Y marking off the upper 25% of scores) and Q_1 is the first quartile (the value of Y marking off the lower 25%). The interquartile range is the range of the middle 50% of the scores.

D. **Semi-Interquartile Range** = $(Q_3 - Q_1)/2$. This is how far you have to go from the middle in both directions to mark off the middle 50% of the scores. This is also known as the probable error. Half of the scores in a distribution will be within one probable error of the middle of the distribution and half will not. Astronomers have used this statistic to estimate by how much one is likely to be off when estimating the value of some astronomical parameter.

E. **Mean Absolute Deviation** = $\Sigma |Y - \mu| \div N$ (this statistic is rarely used)

X: 0; Y: 6/5; Z: 24/5; V: 648/5

F. **Population Variance**: $\sigma^2 = \frac{\Sigma(Y-\mu)^2}{N} = \frac{SS_y}{N}$.

$$SS_y = \Sigma(Y-\mu)^2 = \Sigma Y^2 - \frac{(\Sigma Y)^2}{N}$$

"SS" stands for "sum of squares," more explicitly, the sum of squared deviations of scores from their mean.

G. **Population Standard Deviation**: $\sigma = \sqrt{\sigma^2}$

X: 0; Y: $\sqrt{2}$; Z: 6; V: $\sqrt{35575}$

H. **Estimating Population Variance from Sample Data**

1. computed from a sample, *SS / N* tends to underestimate the population variance
2. s^2 is an unbiased estimate of population variance

$$s^2 = \frac{\sum(Y-M)^2}{N-1} = \frac{SS_y}{N-1}$$

3. *s* is a relatively unbiased estimate of population standard deviation

$s = \sqrt{s^2}$

4. Since in a bell-shaped (normal) distribution nearly all of the scores fall within plus or minus 3 standard deviations from the mean, when you have a moderately sized sample of scores from such a distribution the standard deviation should be approximately one-sixth of the range.

I. **Example Calculations**

	Y	(Y-*M*)	(Y-*M*)^2	*z*
	5	+2	4	1.265
	4	+1	1	0.633
	3	0	0	0.000
	2	-1	1	-0.633
	1	-2	4	-1.265
Sum	15	0	10	0
Mean	3	0	4	0

Notice that the sum of the deviations of scores from their mean is zero (as always). If you find the mean of the squared deviations, 10/5 = 2, you have the variance, assuming that these five scores represent the entire population. The population standard deviation is $\sqrt{2} = 1.414$. Usually we shall consider the data we have to be a sample. In this case the sample variance is 10/4 = 2.5 and the sample standard deviation is $\sqrt{2.5} = 1.581$.

Alternatively, one can compute the sum of squares with this computational formula:

$$SS_y = \sum(Y - M)^2 = \sum Y^2 - \frac{(\sum Y)^2}{N}$$

For example, Y has values 1, 2, 3, 4, and 5. The sum of those scores is 15. The squared scores are 1, 4, 9, 16, and 25. The uncorrected sum of those squared scores is 55. To get the corrected sum of squares we subtract from the uncorrected sum of squares the correction for the mean. $\sum Y^2 - \frac{(\sum Y)^2}{N} = 55 - \frac{15^2}{5}$ = 55 – 45 = 10. If one wants the sample variance, we divide the corrected sum of squares by the degrees of freedom, *N* – 1, yielding 10/4 = 2.5.

Notice that for this distribution the mean is 3 and the median is also 3. The distribution is perfectly symmetric. Watch what happens when I replace the score of 5 with a score of 40.

Distribution of Y: 40, 4, 3, 2, 1

Median = 3, Mean = 10. The mean is drawn in the direction of the (positive) skew. The mean is somewhat deceptive here – notice that 80% of the scores are below average (below the mean) – that sounds fishy, eh?

V. Standard Scores

A. Take the scores from a given distribution and change them such that the new distribution has a standard mean and a standard deviation

B. This transformation does not change the shape of the distribution

C. *Z* - Scores: a mean of 0, standard deviation of 1

$Z = \frac{Y - \mu}{\sigma} \rightarrow$ how many standard deviations the score is above or below the mean In the table above, I have computed z for each score by subtracting the sample mean and dividing by the sample standard deviation.

D. Standard Score = Standard Mean + (*z* score)(Standard Standard Deviation). Examples

- Suzie Cueless has a z score of -2 on a test of intelligence. We want to change this score to a standard IQ score, where the mean is 100 and the standard deviation is 15. Suzie's IQ = 100 - (2)(15) = 70.
- Gina Genius has a z score of +2.5 on a test of verbal aptitude. We want to change this score to the type of standard score that is used on the SAT tests, where the mean is 500 and the standard deviation is 100. Suzie's SAT-Verbal score is 500 + (2.5)100 = 750.

Return to Wuensch's Page of Statistics Lessons

Exercises Involving Descriptive Statistics