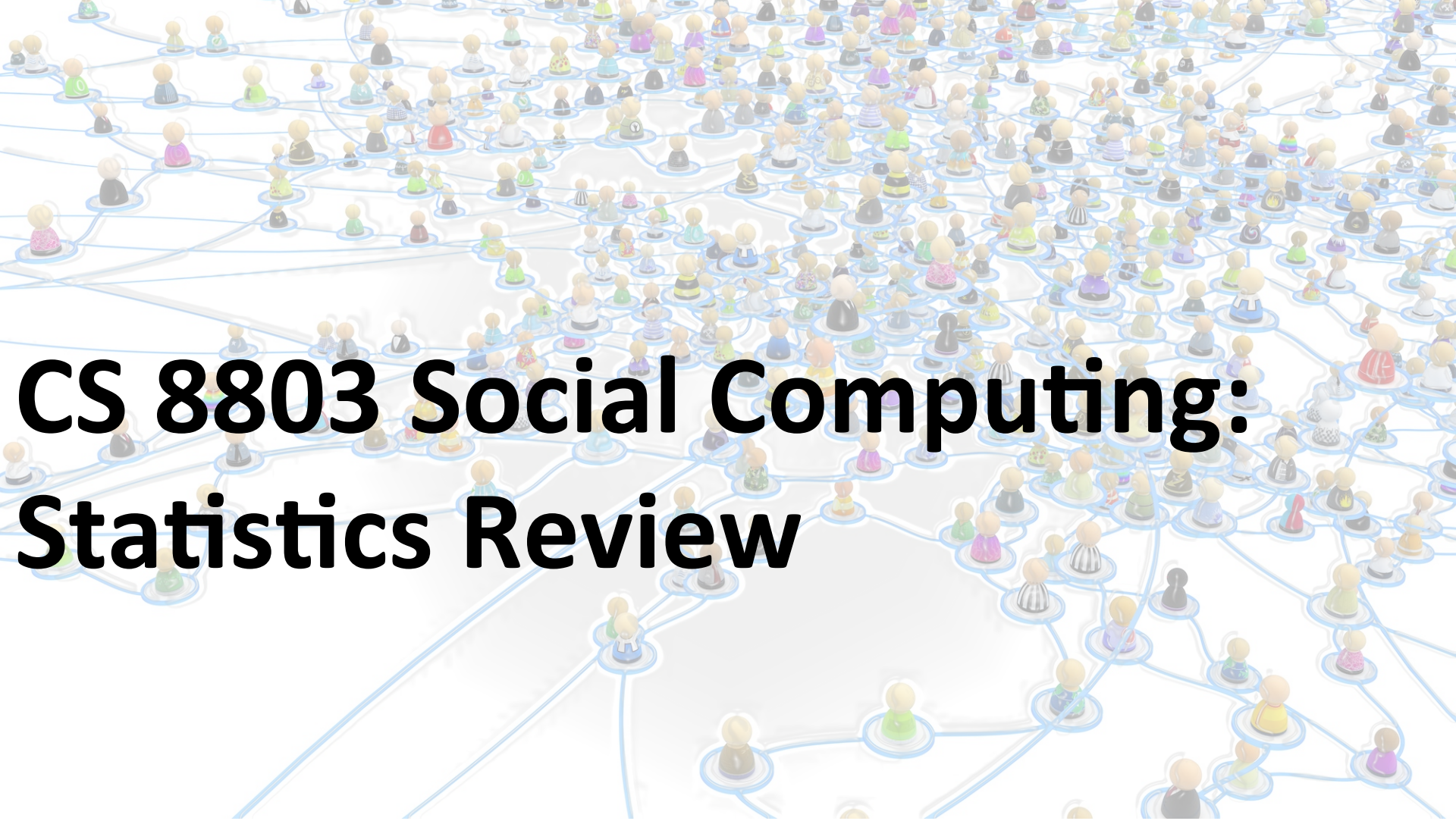


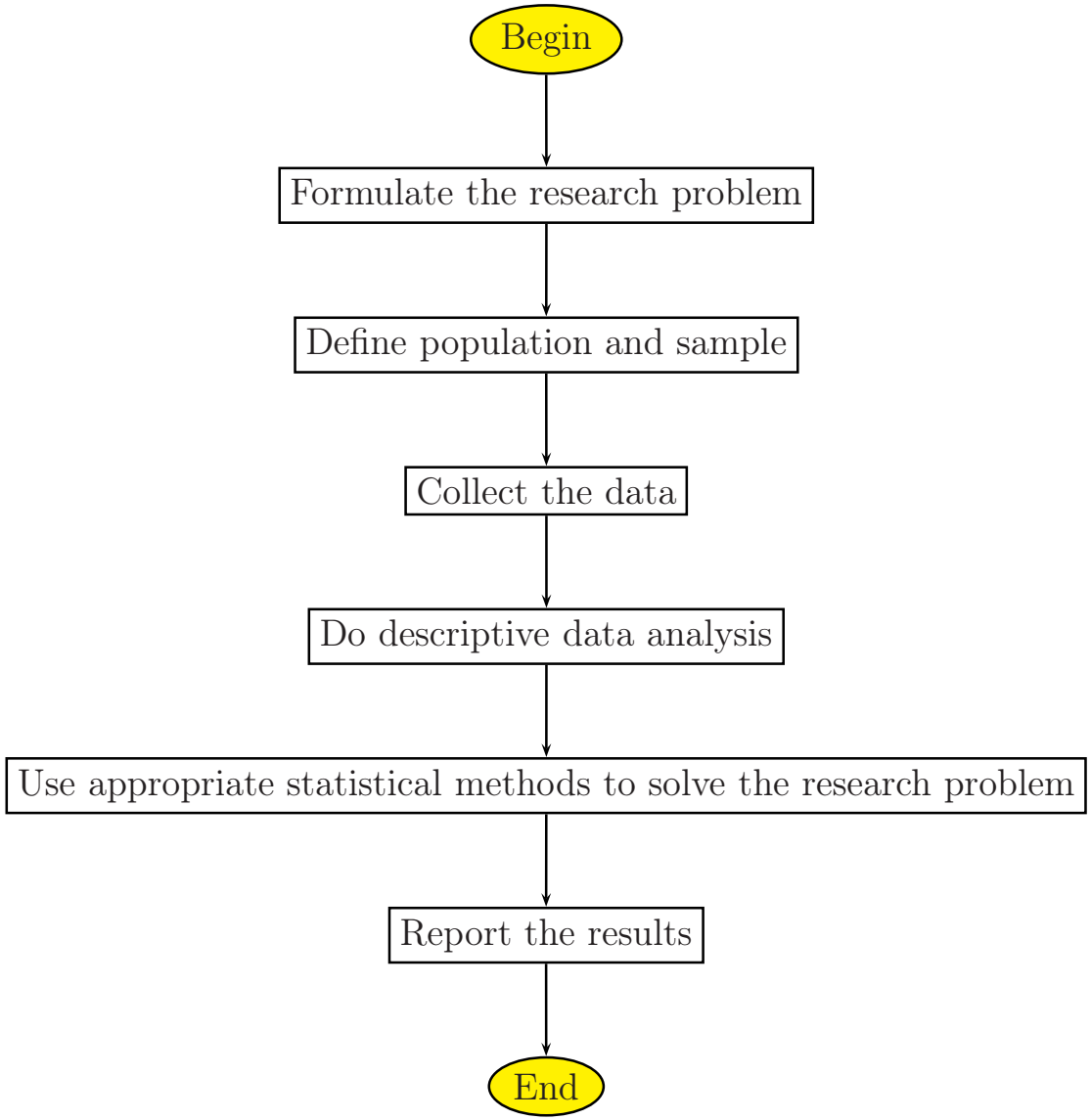
CS 8803 Social Computing: Statistics Review

Munmun De Choudhury

munmund@gatech.edu

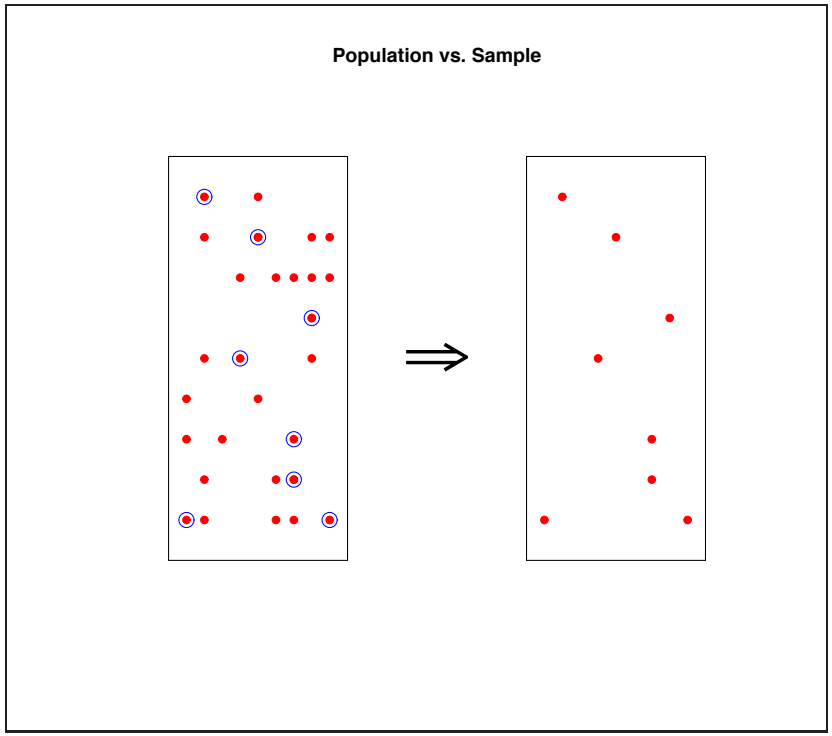
Week 4 | September 10, 2014

Why Statistics?



Definitions

- *Population.* A (statistical) population is the set of measurements (or record of some qualitative trait) corresponding to the entire col- lection of units for which inferences are to be made. (Johnson & Bhattacharyya, 1992)
- *Sample.* A sample from statistical population is the set of measurements that are actually collected in the course of an investigation. (Johnson & Bhattacharyya, 1992)



Definitions

- *Descriptive Statistics*. Descriptive statistics consist of methods for organizing and summarizing information (Weiss, 1999)
- *Inferential Statistics.* Inferential statistics consist of methods for drawing and measuring the reliability of conclusions about population based on information obtained from a sample of the population. (Weiss, 1999)

- Descriptive and inferential statistics are interrelated
- The preliminary descriptive analysis of a sample often reveals features that lead to the choice of the appropriate inferential method to be later used.

- *Example: describing/inferring Twitter use between two cohorts.*

Definitions

- *Parameters and Statistics.* A parameter is an unknown numerical summary of the population. A statistic is a known numerical sum- mary of the sample which can be used to make inference about parameters. (Agresti & Finlay, 1997)

- Example:
  - *Parameter*: The proportion *p* of 18-30 year-olds who use Twitter at least once a day.
  - *Statistic*: The proportion p' of 18-30 year-olds using Twitter at least once a day, calculated from a sample of 18-30 year-olds.

Definitions

- Quantitative variables – numerical values
- Qualitative variables – categorical values

- Discrete variables – a variable is is discrete if it can assume only a finite numbers of values or as many values as there are integers e.g. tweet count per day
- Continuous variables – a variable that can assume any real value e.g. time difference between two consecutive tweet postings

Measures of central tendency

- The *mean* of a variable is the sum of observed values in a data divided by the number of observations.

- PS: Often affected by extreme values in the data (i.e., "outliers").

```
>>> import numpy
>>> numpy.mean(numpy.array([1,2,3,1,3,3,54,6]))
9.125
```

Measures of central tendency

- *Median* of a quantitative variable is that value of the variable in a data set that divides the set of observed values in half, so that the observed values in one half are less than or equal to the median value and the observed values in the other half are greater or equal to the median value.

- To compute median, arrange the observed values of variable in a data in increasing order.
  - If the number of observation is odd, then the sample median is the observed value exactly in the middle of the ordered list.
  - If the number of observation is even, then the sample median is the number halfway between the two middle observed values in the ordered list.

```
>>> import numpy
>>> numpy.median(numpy.array([1,2,3,1,3,3,54,6]))
3
```

Measures of central tendency

Mode of a qualitative or a discrete quantitative variable is that value of the variable which occurs with the greatest frequency in a data set.

PS: If the greatest frequency is 1 (i.e. no value occurs more than once), then the variable has no mode.

```
>>> from statistics import mode
>>> mode([1, 1, 2, 3, 3, 3, 3, 4])
 3
```

Measures of variation

- The sample *range* of the variable is the difference between its maximum and minimum values in a data set:
- Range = Max – Min.

- PS: In using the range, a great deal of information is ignored, that is, only the largest and smallest values of the variable are considered; the other observed values are disregarded.

Measures of variation

- To understand variability in data, one of the most commonly used measures are percentiles or quartiles. The quartiles of the variable divide the observed values into quarters (4 parts).

- Let *n* denote the number of observations in a data set. Arrange the observed values of variable in a data in increasing order.
  - The first quartile Q1 is at position (n+1) / 4
  - The second quartile Q2 (the median) is at position (n+1) / 2
  - The third quartile Q3 is at position 3(n+1) / 4

  in the ordered list.

- The *interquartile range* of a variable is the difference between the first and third quartiles of the variable, that is,

  IQR = Q3 – Q1.

  Roughly speaking, the IQR gives the range of the middle 50% of the observed values.

Measures of variation

- The sample interquartile range represents the length of the interval covered by the center half of the observed values of the variable.
- This measure of variation is not disturbed if a small fraction the observed values are very large or very small.

- *Five-number summary*. The five-number summary of the variable consists of minimum, maximum, and quartiles written in increasing order:

  Min, Q1, Q2, Q3, Max.

Measures of variation

- Standard deviation is measure of a distribution's deviation from its mean. It is the square root of variance.

- For a variable *x*, the sample standard deviation is

$$\sigma = \sqrt{\frac{1}{N}\left[(x_1-\mu)^2+(x_2-\mu)^2+\cdots+(x_N-\mu)^2\right]}, \quad \text{where} \quad \mu = \frac{1}{N}(x_1+\cdots+x_N),$$

- Since the standard deviation is defined using the sample mean of the variable *x*, it is preferred measure of variation when the mean is used as the measure of center (i.e. in case of symmetric distribution).
- Note that the standard deviation is always positive number

Measures of variation

- The more variation there is in the observed values, the larger is the standard deviation for the variable in question.
- However it can be strongly affected by a few extreme observations.

```
>>> import numpy
>>> numpy.std([0.1,2.8,3.7,2.6,5,3.4])
1.4851
```

Estimation

- Statistical inference uses sample data to form two types of estimators of parameters.
- A *point estimate* consists of a single number, calculated from the data, that is the best single guess for the unknown parameter.
- A *interval estimate* consists of a range of numbers around the point estimate, within which the parameter is believed to fall.

- A good point estimator of a parameter is one with sampling distribution that is centered around parameter, and has small standard error as possible.
- However, it is often more desirable to produce an interval of values that is likely to contain the true value of the unknown parameter.

Estimation

- A *confidence interval estimate* of a parameter consists of an interval of numbers obtained from a point estimate of the parameter together with a percentage that specifies how confident we are that the parameter lies in the interval.
- The confidence percentage is called the *confidence level*.

Comparing two distributions

Hypothesis testing

- *Hypothesis*. A hypothesis is a statement about some characteristic of a variable or a collection of variables. (Agresti & Finlay, 1997)

- When a hypothesis relates to characteristics of a population, such as population parameters, one can use statistical methods with sample data to test its validity.

- Example: Are men Twitter users and women Twitter users distinct in their use of the site?

Hypothesis testing

- A *significance test* is a way of statistically testing a hypothesis by comparing the data to values predicted by the hypothesis.
- Data that fall far from the predicted values provide evidence against the hypothesis.
- All significance tests have five elements: assumptions, hypotheses, test statistic, *p*-value, and conclusion.

Hypothesis testing

- A significance test considers two hypotheses about the value of a population parameter: the *null hypothesis* and the *alternative hypothesis*.

- *Null and alternative hypotheses*. The null hypothesis H_o is the hypothesis that is directly tested. This is usually a statement that the parameter has value corresponding to, in some sense, no effect. The alternative hypothesis H_a is a hypothesis that contradicts the null hypothesis. This hypothesis states that the parameter falls in some alternative set of values to what null hypothesis specifies. (Agresti & Finlay, 1997)

Hypothesis testing

- The *test statistics* is a statistic calculated from the sample data to test the null hypothesis. This statistic typically involves a point estimate of the parameter to which the hypotheses refer.

- *t*-statistic common; measures difference of means when comparing two distributions
- We calculate the *p*-value under assumption that H_0 is true. That is, we give the benefit of the doubt to the null hypothesis, analyzing how likely the observed data would be if that hypothesis were true.
- The *p*-value is the probability that the true *t*-statistic is at least as large in absolute value as the observed value *t*.

Hypothesis testing

- *p*-value. The *p*-value is the probability, when H_0 is true, of a test statistic value at least as contradictory to H_0 as the value actually observed. The smaller the *p*-value, the more strongly the data contradict H_0. (Agresti & Finlay, 1997)

- *p*-value is often defined with respect to a chosen confidence level, α

Hypothesis testing

- The *t*-test has important assumptions that must be satisfied in order for the associated *p*-value to be valid.
  - The samples are independent.
  - Each sample is from a normally distributed population.

Hypothesis testing

- Calculating *t*-test for the means of TWO INDEPENDENT samples of scores:

```
>>> from scipy import stats


>>> rvs1 = stats.norm.rvs(loc=5,scale=10,size=500)
>>> rvs2 = stats.norm.rvs(loc=5,scale=10,size=500)

>>> stats.ttest_ind(rvs1,rvs2)
(0.26833823296239279, 0.78849443369564776)
```

```
>>> rvs4 = stats.norm.rvs(loc=5, scale=20, size=100)
```

```
>>> stats.ttest_ind(rvs1, rvs4, equal_var = False)
(-0.69712570584654099, 0.48716927725402048)
```

Hypothesis testing

- Sometimes one sample can be "greater" than the other – in an ordinal sense
- The Mann–Whitney U test (also called the Mann–Whitney–Wilcoxon (MWW), Wilcoxon rank-sum test, or Wilcoxon–Mann–Whitney test) is a nonparametric test of the null hypothesis that two populations are the same against an alternative hypothesis, especially that a particular population tends to have larger values than the other.

```
>>> from scipy import stats
>>> rvs2 = stats.norm.rvs(loc=5,scale=10,size=500)
>>> rvs5 = stats.norm.rvs(loc=50,scale=10,size=200)
>>> stats.ranksums(rvs2,rvs5)
(-17.268624879732251, 8.1050738020911939e-67)
```

Hypothesis testing

- *Kolmogorov-Smirnov statistic* – a two-sided test for the null hypothesis that 2 independent samples are drawn from the same **continuous** distribution.

```
>>> from scipy import stats
>>> rvs1 = stats.norm.rvs(size=200, loc=0., scale=1)
>>> rvs2 = stats.norm.rvs(size=200, loc=0.5, scale=1.5)
>>> stats.ks_2samp(rvs1, rvs2)
(0.20833333333333337, 4.667497551580698e-005)
```

Comparing more than two distributions

Hypothesis testing

- ANOVA provides a statistical test of whether or not the means of several groups are equal, and therefore generalizes the t-test to more than two groups.
- As doing multiple two-sample t-tests would result in an increased chance of committing a statistical type I error, ANOVAs are useful in comparing (testing) three or more means (groups or variables) for statistical significance.
- The F-test is used for comparing the factors of the total deviation. For example, in one-way, or single-factor ANOVA, statistical significance is tested for by comparing the F test statistic

$$F = \frac{\text{variance between treatments}}{\text{variance within treatments}}$$

Hypothesis testing

- The ANOVA test has important assumptions that must be satisfied in order for the associated *p*-value to be valid.
  - The samples are independent.
  - Each sample is from a normally distributed population.
  - The population standard deviations of the groups are all equal. This property is known as homoscedasticity.

Hypothesis testing

- 1-way ANOVA for three independent samples:

```
>>> from scipy import stats
>>> rvs1 = stats.norm.rvs(loc=5,scale=10,size=500)
>>> rvs2 = stats.norm.rvs(loc=5,scale=10,size=200)
>>> rvs3 = stats.norm.rvs(loc=5,scale=10,size=240)
>>> [f_value, p_value] = stats.f_oneway(rvs1, rvs2, rvs3)
(1.6144352794299781, 0.19955607421980850)
```

- Note that rejecting the null hypothesis does not indicate which of the groups differs.

Hypothesis testing

- If any of the three assumptions of ANOVA are not true, it is recommended to use the Kruskal Wallis H test, which is a non-parametric version of ANOVA.
- The Kruskal-Wallis H-test tests the null hypothesis that the population median of all of the groups are equal.
- The test works on 2 or more independent samples, which may have different sizes.

```
>>> from scipy import stats
>>> rvs1 = stats.norm.rvs(loc=5,scale=10,size=500)
>>> rvs2 = stats.norm.rvs(loc=5,scale=10,size=200)
>>> rvs4 = stats.norm.rvs(loc=5,scale=20,size=240)
>>> stats.kruskal(rvs1,rvs2,rvs4)
(2.9637587853571858, 0.22721026954861492)
```

Association between distributions

Pearson correlation coefficient

- *Correlation coefficient*. The sample correlation coefficient is a measure of the strength of the linear relation between two variables x and y.
- 1 is total positive correlation, 0 is no correlation, and –1 is total negative correlation

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

```
>>> from scipy import stats
>>> rvs1 = stats.norm.rvs(loc=5,scale=10,size=500)
>>> rvs6 = stats.norm.rvs(loc=5,scale=10,size=500)
>>> stats.pearsonr(rvs1,rvs6)
(-0.028991426987572115, 0.5177734067731351)
```

Spearman's correlation

- Pearson's correlation assumes normality of the distributions being compared
- The Spearman correlation is a nonparametric measure of the monotonicity of the relationship between two datasets.

```
>>> from scipy import stats
>>> rvs1 = stats.norm.rvs(loc=5,scale=10,size=500)
>>> rvs6 = stats.norm.rvs(loc=5,scale=10,size=500)
>>> stats.spearmanr(rvs1,rvs6)
(-0.045163188652754614, 0.3135186410886480 2)
```

Kendall's tau

- A measure of the correspondence between two *rankings*. Values close to 1 indicate strong agreement, values close to -1 indicate strong disagreement.

```
>>> from scipy import stats
>>> rvs1 = stats.norm.rvs(loc=5,scale=10,size=500)
>>> rvs6 = stats.norm.rvs(loc=5,scale=10,size=500)
>>> stats.kendalltau(rvs1,rvs6)
(-0.03191983967935870, 0.2860220143222619 3)
```

Exercise

A. Compare number of Facebook likes of two friends Alice and Bob

B. Compare the #followers of celebrities, news organizations, and regular users

- Which measure of central tendency will you use?
- What is a good measure of variation?
- What statistical test will you use to compare the two groups?

Next class

- Monday 9/15 (topic: Data Mining Review)
- **No assigned readings**