Lecture 30
Statistical description of data

- Moments of a distribution.
- Do two distributions have the same mean and variance?
- Are two distributions the same?

Moments of a distribution:

When sets of data have enough tendency to "clump" around a certain value, it is useful to characterize them by the "moments" measuring how well they clump around the value. The most elementary indicator of clumping is the mean, which gives an idea of around which value the data are clumped,

$$\overline{x} = \frac{1}{N} \sum_{j=1}^{N} x_j$$

Although quite common, the mean might not be the best estimator of the value in question. For data sets that are "broad" or have very long tails, the mean might be very slow to converge to the correct value. Other estimators, like the median and the mode, might be better.

The median of a probability distribution is the value of x for which it is equally probable to find values larger or smaller,

$$\int_{-\infty}^{x_{\rm med}} p(x)\,dx = \frac{1}{2} = \int_{x_{\rm med}}^{\infty} p(x)\,dx$$

To compute it for a given data set, one finds the value of x that has equal number of points larger and smaller than it. Of course, this is not possible if the number of points is even. In that case one takes the mean of the two central points.

If values are sorted in ascending (or descending) order then the median is given by,

$$x_{\rm med} = \begin{cases} x_{(N+1)/2}, \\ \frac{1}{2}(x_{N/2} + x_{(N/2)+1}), \end{cases}$$

If most of the area of a distribution is below a central peak, the median gives an idea of the position of the peak. It is more robust than the mean in the sense that it fails as an indicator if the area below the tails of the distribution is large. The mean fails if the moment of the tails is large. One can build many examples of distributions with large moments for the tails but small areas below them.

The **mode** of a distribution is the point where it attains its maximum. It is only useful for very peaked distributions.

To measure the "width" of a set of data, we use the variance or the standard deviation,

$$\mathrm{Var}(x_1 \ldots x_N) = \frac{1}{N-1}\sum_{j=1}^{N}(x_j - \overline{x})^2 \qquad \sigma(x_1 \ldots x_N) = \sqrt{\mathrm{Var}(x_1 \ldots x_N)}$$

In many situations these quantities (which depend on the second moment of the data) may fail to converge. A more robust measure is the absolute deviation,

$$\mathrm{ADev}(x_1 \ldots x_N) = \frac{1}{N}\sum_{j=1}^{N}|x_j - \overline{x}|$$

This measure is less popular among mathematicians because the non analyticity of the absolute value makes proving theorems hard. But it is a more robust measure than the variance (it converges in more cases than the second moment). In general measures with higher powers of the data have to be used with even more caution.

The next order measure of the moments yields the **skewness** of the distribution.

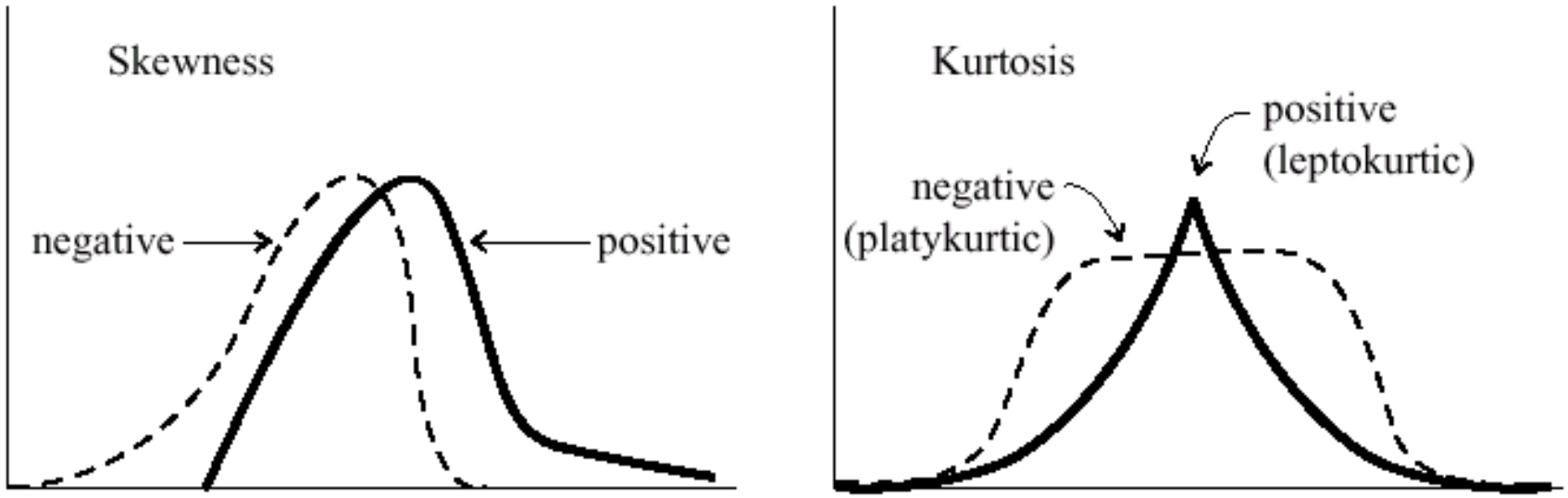
$$\mathrm{Skew}(x_1 \ldots x_N) = \frac{1}{N} \sum_{j=1}^{N} \left[\frac{x_j - \overline{x}}{\sigma} \right]^3$$

In real sets of data this quantity will almost always be non-zero (it is a dimensionless number).

How much skewness is really relevant? One should compare the value of the skewness with *its* standard deviation. Unfortunately, this depends on the distribution (in a detailed way). For a Gaussian th deviation is $(15/N)^{1/2}$. As a rule of thumb, values that are larger than that are of some significance.

The **kurtosis** measures how peaked a distribution is, compared with a Gaussian,

$$\mathrm{Kurt}(x_1 \ldots x_N) = \left\{ \frac{1}{N} \sum_{j=1}^{N} \left[\frac{x_j - \overline{x}}{\sigma} \right]^4 \right\} - 3$$



The standard deviation of the kurtosis for the Gaussian distribution is $(96/N)^{1/2}$. In real life application, measurement of such a high moment is exceedingly difficult and its variance will in general be very large, rendering it almost useless as an indicator.

The formulas for variance, etc, are not hard to implement numerically. One has to be careful how one writes them. Some textbooks rearrange things,

$$\mathrm{Var}(x_1 \ldots x_N) = \frac{1}{N-1}\left[\left(\sum_{j=1}^{N} x_j^2\right) - N\overline{x}^2\right] \approx \overline{x^2} - \overline{x}^2$$

This implementation is bad since one adds up numbers that are close and nevertheless their difference is important (and therefore gets swamped by roundoff). It is much better to rearrange as,

$$\mathrm{Var}(x_1 \ldots x_N) = \frac{1}{N-1}\left\{\sum_{j=1}^{N}(x_j - \overline{x})^2 - \frac{1}{N}\left[\sum_{j=1}^{N}(x_j - \overline{x})\right]^2\right\}$$

Where one adds up small numbers encoding the differences one is trying to assess.

Do two distributions have the same mean and variance?

On many occassions one wishes to know if two distributions of data have the same mean. A way to ask this question meaningfully is to ask how many standard deviations are the means apart.

This number assesses how strong is the difference in the means. It can be important, if the difference is genuine. One has to distinguish between *strength* and *significance*. A small difference in means (measured with respect to the standard deviation) can be very significant if one has lots of points. A large difference can be insignificant if one has few points.

A quantity that measures the significance of a difference of means is to ask “how many standard errors are they apart”. The standard error is a measurement of how accurately the sample mean represents the “true” mean. It is given by the standard deviation divided by the square root of the number of points in the sample.

Applying this concept, one way of testing for the significance of the difference of means is given by Student's t-test. One starts by evaluating the standard error of the difference of the means,

$$s_D = \sqrt{\frac{\sum_{i\in A}(x_i - \overline{x_A})^2 + \sum_{i\in B}(x_i - \overline{x_B})^2}{N_A + N_B - 2}\left(\frac{1}{N_A} + \frac{1}{N_B}\right)}$$

And constructing a variable t that compares the actual difference with the standard error,

$$t = \frac{\overline{x_A} - \overline{x_B}}{s_D}$$

One then evaluates the significance using the Student-t distribution,

$$A(t|\nu) = \frac{1}{\nu^{1/2}B(\frac{1}{2}, \frac{\nu}{2})}\int_{-t}^{t}\left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}} dx$$

ν is the number of bins.

The significance is a number between zero and one which expresses what is the chance that such large a difference in means could have happened by chance. Small numbers are therefore very significant.

There are generalizations of Student's t test for distributions with different variances. However, if the variances are quite different, it might also indicate that the distributions are different in shape and therefore knowing if their means is the same is not very useful.

Are two distributions the same?

The questions we've been asking can be generalized to the all encompassing question: given two sets of data can we prove that they come from different probability distributions?

Notice that we phrased the question in the negative. The positive question is impossible to answer: one cannot, with finite amounts of data, determine if two distributions are the same.

There are two ways to address this problem, depending on if the data is binned or is continuous.

If the data are binned, one applies the chi-square test. If N_i is the number of points in a bin and n_i is the theoretically expected number, one evaluates,

$$\chi^2 = \sum_i \frac{(N_i - n_i)^2}{n_i}$$

A large value of chi implies that it is unlikely that the distributions are the same. If $n_i=N_i=0$ one should omit the point from the sum. If N_i is non-zero but n_i is, chi is infinite since then it is impossible that the distributions are the same.

There is a probability distribution (Chi squared) that says how likely it is that a sum of normal distributions will yield squared sums that have a given value. Its explicit functional form is related to the incomplete gamma functions.

For continuous data a common test to determine if two distributions are the same is the Kolmogorov-Smirnov test. One starts by evaluating the difference between the two distributions in the sup norm,

$$D = \max_{-\infty<x<\infty} |S_{N_1}(x) - S_{N_2}(x)|$$

The significance of this number is given by,

$$\text{Probability}\ (D > \text{observed}\ ) = Q_{KS}\left(\left[\sqrt{N_e} + 0.12 + 0.11/\sqrt{N_e}\right]\ D\right)$$

Where,

$$N_e = \frac{N_1 N_2}{N_1 + N_2} \qquad Q_{KS}(\lambda) = 2\sum_{j=1}^{\infty}(-1)^{j-1}\ e^{-2j^2\lambda^2}$$

The Kolmogorov test gives preferential weigth to points near the mean, since the variance of D is maximum near the mean. It is therefore good to detect shifts, but not so good at detecting spreads. This can be remedied by using other measures instead of D. (Anderson-Darling, Kuiper, etc).

Measures of correlation between distributions:

A task of usual interest is: given a set of events characterized by two Numbers and given knowledge of the value of one of these two variables, can one predict something about the value of the other?

Example: which percentage of bagpipers teach computational physics?

Let N_{ij} be the number of cases in which the first variable takes the value i and the second the value j. Let N be the total number of events. Let N_i be the number of events in which the first variable takes the value i irrespective of the value of j.

$$N_{i.} = \sum_j N_{ij} \qquad N_{.j} = \sum_i N_{ij}$$

$$N = \sum_i N_{i.} = \sum_j N_{.j}$$

Suppose the two varibles are uncorrelated. In that case, the probability of getting a value i for a given j should be the same as the probability of getting the value irrespective of j. If we call n_{ij} the expected value of N_{ij} one has,

$$\frac{n_{ij}}{N_{.j}} = \frac{N_{i.}}{N} \qquad \text{which implies} \qquad n_{ij} = \frac{N_{i.}N_{.j}}{N}$$

We can therefore run a chi squared test,

$$\chi^2 = \sum_{i,j} \frac{(N_{ij} - n_{ij})^2}{n_{ij}}$$

Which will tell us if there is or not a correlation, but will not quantify it for us. To fix this it is good to remap chi to a variable that takes values in the interval [0,1] and is independent of the number of data considered. Two such possibilities are given by Cramer's V and the contingency coefficient C.

$$V = \sqrt{\frac{\chi^2}{N \min(I-1, J-1)}}$$

$$C = \sqrt{\frac{\chi^2}{\chi^2 + N}}$$

The contingency coefficient should not be used to compare data between sets with different rows and columns. Cramer's V is unity only if there is only one non-zero column element in each row of the data, meaning perfect correlation.

Neither of these two indicators have absolute meaning. Is V=0.23 good or bad?

When one has sets of data that are ordinal (not “how many events that satisfy condition x” but a variable y that can take values with certain probability) a common measure of correlation between two datasets is the linear correlation.

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2} \sqrt{\sum_i (y_i - \overline{y})^2}}$$

The regression coefficient r is 1 if the data are just proportional to each other (with a positive proportionality constant) and is -1 if they are proportional with a negative proportionality constant.

If the distributions have enough convergent moments and N is large, then the standard deviation of r is $1/N^{1/2}$ and the probability that two uncorrelated data sets yield a value of r is given by,

$$\operatorname{erfc}\left(\frac{|r|\sqrt{N}}{\sqrt{2}}\right)$$

Summary

- Student t test tells us if two distributions have the same mean and variance.
- Kolmogorov-Smirnov tells us if two distributions are the same.
- Chi square gives us a measure of the correlation of two distributions.