Statistics and the K-S Test

Massimo Ricotti

ricotti@astro.umd.edu

University of Maryland
**Statistical Description of Data**


- Statistics provides tools for understanding data.
  - In the wrong hands these tools can be dangerous!

- Here’s a typical data analysis cycle:
  1. Apply some formula to data to compute a “statistic.”
  2. Find where that value falls in a probability distribution computed on the basis of some “null hypothesis.”
  3. If it falls in an unlikely spot (on distribution tail), conclude null hypothesis is *false* for your data set.
Statistics

- Statistics and probability theory are closely related. Statistics can never prove things, only disprove them by ruling out hypotheses.

- Distinguish between *model-independent* statistics (this class, e.g., mean, median, mode) and *model-dependent* statistics (next class, e.g., least-squares fitting).

- Will make use of special functions (e.g., gamma function) described in NRiC §6.
Moments of a Distribution

- The **mean**, **median**, and **mode** of distributions are called *measures of central tendency*.

- The most common description of data involves its *moments*, sums of integer powers of the values.

- The most familiar moment is the **mean**:

\[
\bar{x} = <x> = \frac{1}{N} \sum_{i=1}^{N} x_i.
\]
Variance

The width of the central value is estimated by its second moment, called the variance,

\[ \text{Var} = \frac{1}{N - 1} \sum_{i=1}^{N} (x_i - \bar{x})^2, \]

or its square root, the standard deviation,

\[ \sigma = \sqrt{\text{Var}}. \]

Why \( N - 1 \)? If the mean is known \textit{a priori}, i.e., if it’s not measured from the data, then use \( N \), else \( N - 1 \). If this matters to you, then \( N \) is probably too small!
A clever way to minimize round-off error when computing the variance is to use the *corrected two-pass algorithm*. First compute \( \bar{x} \), then do:

\[
\text{Var} = \frac{1}{N - 1} \left\{ \frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2 - \frac{1}{N} \left[ \sum_{i=1}^{N} (x_i - \bar{x}) \right]^2 \right\}.
\]

The second sum would be zero if \( \bar{x} \) were exact, but otherwise it does a good job of correcting RE in \( \text{Var} \). **Proof**: EFTS (hint: set \( \bar{x} \to \bar{x} + \epsilon \)).
Other moments

- Higher moments, like skewness (3rd moment) and kurtosis (4th moment) are also sometimes used, but can be unreliable.

Distribution Functions

A distribution function (DF) \( p(x) \) gives the probability of finding a value between \( x \) and \( x + dx \), e.g., the familiar “normal” (Gaussian) distribution

\[
p(x) \, dx = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx.
\]

The expected mean data value is:

\[
\langle x \rangle = \frac{\int_{-\infty}^{\infty} x \, p(x) \, dx}{\int_{-\infty}^{\infty} p(x) \, dx}.
\]

For a discrete DF:

\[
\langle x \rangle = \frac{\sum_i x_i \, p_i}{\sum_i p_i}.
\]

Similar to weighted means, e.g., center of mass.
Median

- The **median** of a DF is the value $x_{med}$ for which larger and smaller values of $x$ are equally probable:

\[
\int_{-\infty}^{x_{med}} p(x) \, dx = \frac{1}{2} = \int_{x_{med}}^{\infty} p(x) \, dx.
\]

- For discrete values, sort in ascending order ($i = 1, 2, \ldots, N$), then:

\[
x_{med} = \begin{cases} 
  x(N+1)/2, & \text{if } N \text{ is odd,} \\
  \frac{1}{2}(x_{N/2} + x_{N/2+1}), & \text{if } N \text{ is even.}
\end{cases}
\]
Mode

The **mode** of a probability DF $p(x)$ is the value of $x$ where the DF takes on a maximum value.

Most useful when there is a single, sharp max, in which case it estimates the central value.

Sometimes a distribution will be *bimodal*, with two relative maxima. In this case the mean and median are not very useful since they give only a “compromise” value between the two peaks.
Comparing Distributions

- Often want to know if two distributions have different means or variances (NRiC §14.2):
  1. Student's $t$-test for significantly different means.
     (a) Find number of standard errors $\sim \sigma / N^{1/2}$ between two means.
     (b) Compute statistic using nasty formula: probability that the two means are different by chance.
     (c) Small numerical value indicates significant difference.
  2. $F$-test for significantly different variances.
     (a) Compute $F = \text{Var}_1 / \text{Var}_2$ and plug into nasty formula (the distribution of $F$ in the case that the variances are the same—the null hypothesis—is related to the incomplete beta function).
     (b) Small value indicates significant difference.
Given two sets of data, can generalize to a single question: Are the sets drawn from the same DF? E.g., are stars distributed uniformly in the sky? Do two brands of lightbulbs have the same distribution of burn-out times?

Recall can only disprove (to a certain confidence level), not prove.

May have continuous or binned data.

May want to compare one data set with known DF, or two unknown data sets with each other.

Popular technique for binned data is the $\chi^2$ test. For continuous data, use the KS test. Cf. *NRiC* §14.3.
Chi-square ($\chi^2$) test

Suppose have $N_i$ events in $i$th bin but expect $n_i$:

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

Large value of $\chi^2$ indicates unlikely match (i.e., $N_i$’s probably not drawn from population represented by $n_i$’s).

Compute probability $Q(\chi^2|\nu)$ from incomplete gamma function, where $\nu$ is the number of degrees of freedom.

- Typically $\nu = N_B$, where $N_B$ is the number of bins, or $N_B - 1$, if the $n_i$’s are normalized such that $\sum_i n_i = \sum_i N_i$.
- Null hypothesis assumes differences $N_i - n_i$ are standard normal random variables of unit variance and zero mean.
For two binned data sets with events $R_i$ and $S_i$:

$$\chi^2 = \sum_i \frac{(R_i - S_i)^2}{R_i + S_i}.$$ 

Have sum in denominator, rather than average, because variance of difference of two normal quantities is sum of individual variances.
Kolmogorov-Smirnov (KS) test

- Appropriate for unbinned distributions of single independent variable.

- From sorted list of data points, construct estimate $S_N(x)$ of the cumulative DF of the probability DF from which it was drawn...

  - $S_N(x)$ gives fraction of data points to the left of $x$.

  - Constant between $x_i$’s, jumps $1/N$ at each $x_i$.

  - Note $S_N(x_{\text{min}}) = 0$, $S_N(x_{\text{max}}) = 1$.

    - Behaviour between $x_{\text{min}}$ and $x_{\text{max}}$ distinguishes distributions.

- Cf. NRiC Fig. 14.3.1.

- Statistic is maximum value of absolute difference between two cumulative DFs.
To compare data set to known cumulative DF:

\[ D = \max_{x_{\min} \leq x \leq x_{\max}} |S_N(x) - P(x)|. \]

To compare two unknown data sets:

\[ D = \max_{x_{\min} \leq x \leq x_{\max}} |S_{N_1}(x) - S_{N_2}(x)|. \]

Plug \( D \) and \( N \) (or \( N_e = N_1 N_2 / (N_1 + N_2) \)) into nasty formula to get numerical value of significance. As usual, a small value indicates a significant difference.