Statistical Description of Data

- *Cf. NRiC*, Chapter 14.
- 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 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*, Chapter 6.

Moments of a Distribution

- *Cf. NRiC* §14.1.
- 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:

$$\mathrm{Var} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$$

  or its square root, the standard deviation:

$$\sigma = \sqrt{\mathrm{Var}}$$

  - Why *N* -1? If the mean is known *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!

More on Moments

- A clever way to minimize round-off error when computing the variance is to use the *corrected two-pass algorithm*. First compute <*x*>, then do:

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

  - The second sum would be zero if <x> were exact, but otherwise it does a good job of correcting RE in Var.
- Higher moments, like skewness (3rd moment) and kurtosis (4th moment) are also sometimes used.

Distribution Functions

- A distribution function (DF) $p(x)$ gives the probability of finding value between x & $x + dx$.
  - 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 & smaller values of x are equally probable:

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

- For discrete values, sort in ascending order, then:

$$x_{\text{med}} = \begin{cases} x_{(N+1)/2}, & N \text{ odd} \\ \frac{1}{2}\left(x_{N/2} + x_{(N/2)+1}\right), & N \text{ even} \end{cases}$$

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 no. of *standard errors* ~ $\sigma/N^{1/2}$ between two means.
     b) Compute statistic using nasty formula.
     c) Small numerical value indicates significant difference.
  2. *F*-test for significantly different variances.
     a) Compute $F = \mathrm{Var}_1/\mathrm{Var}_2$ and plug into nasty formula.
     b) Small value indicates significant difference.

Comparing Distributions, Cont'd

- Given two sets of data, can generalize to a single question: Are the sets drawn from the same DF?
- Recall can only disprove, 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 χ^2 test. For continuous data, use the KS test. *NRiC* §14.3.

Chi-Square (χ^2) Test

- Suppose have N_i events in ith bin but expect n_i:

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

  - Large value of χ^2 indicates unlikely match.
  - Compute probability $Q(\chi^2|\nu)$ from *incomplete gamma function*, where ν is *# degrees of freedom*.

- For two binned data sets with events R_i and S_i:

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

Kolmogorov-Smirnov (KS) Test

- Appropriate for unbinned distributions.
- 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 $S_N(x_{\min}) = 0$, $S_N(x_{\max}) = 1$.
    - Behavior between $x_{\min}$ & $x_{\max}$ distinguishes distributions.

KS Test, Cont'd

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

$$D = \max_{x_{\min} \le x \le x_{\max}} \left| S_N(x) - P(x) \right|$$

- To compare two unknown data sets:

$$D = \max_{x_{\min} \le x \le x_{\max}} \left| S_{N_1}(x) - S_{N_2}(x) \right|$$

- Plug D and N (or $N_e = N_1 N_2 / (N_1 + N_2)$) into nasty formula to get numerical value of significance.