NOTES FROM PHYS 600 CLASS BASED ON FEIGELSON and BABU BOOK

The following summaries were compiled after having gone through the Statistical Astronomy class in Fall of 2013 and Spring of 2015.

1. Preliminaries

1.1. Definitions

Probability distribution function: p(x)

Sample distribution function: f(x)

Mean: Expected value of x = E(x) = $\mu = \int_{-\infty}^{\infty} xp(x)dx$

Variance: E(x^2) − E^2(x) = E(x−μ)^2 = σ^2

Moment Generating Function: E(e^{tx}). Expand this in powers of t to get moments of the distribution, like E(x), E(x^2), etc.

Standard deviation: σ

Empirical Distribution Function: e.d.f. = $\widehat{F}(x) = \int_{-\infty}^{x} f(x)dx$

Cumulative Distribution Function: c.d.f. = F(x) = $\int_{-\infty}^{x} p(x)dx$

Trimmed Mean: Like a mean but you reject a certain number of high and low points. IRAF does the same thing with the imcombine task. With the maximum rejections you end up with the median.

Interquartile Range: Where 25% and 75% of the data fall. Indicated with the box part of a box plot.

Box-whiskers: Box usually defined as above. No consensus on whiskers. Outlier points are plotted individually.

MAD: median absolute deviation = Med$|X_i - Med|$ is an excellent measure of dispersion in contaminated data.

Heteroscedastic/Homoscedastic: Errors or scatter vary/don't vary with the value of the variable.

1.2. Concepts

Robust: Results that are insensitive to outlier points. For example, the median is robust but the mean is not.

Breakdown: Fraction of data that need to be contaminated to wreck a statistic. For the median it is 50% .

Confidence Intervals: Typically used to reject at some level of confidence that some condition exists in the data. For example, if getting a correlation as good as observed occurs less than 1% of the time for data that are really uncorrelated so the apparent correlation is caused by noise, then we reject with 99% confidence the hypothesis that the data are uncorrelated. Another example: you can reject the hypothesis that the mean of the data $\mu > 100$ with 99% certainty.

1.3. Probability Distribution Functions

Binomial: Like a series of coin flips, this distribution is the probability of getting x successes in n attempts when the probability of a success is θ for each attempt. The mean $\mu = \theta$ and the variance $\sigma^2 = \theta(1-\theta)$.

$$p(x) = \frac{n!}{x!(n-x)!}\theta^x(1-\theta)^{n-x}$$

Poisson: Limit of the binomial when n $\rightarrow \infty$, $\theta \rightarrow$0 such that nθ=λ is a constant. Also occurs in counting statistics where the chance of success in a time interval is λ, each count is independent of the previous history, and the probability of more than one count in a time interval is negligible. This is the distribution of shot noise, where the signal-to-noise ratio $\mu/\sigma = \lambda^{1/2} = \mu^{1/2}$, so S/N $\sim$ C$^{1/2}$ where C are the counts. $\mu = \sigma^2 = \lambda$.

$$p(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Normal: The usual bell-curve. The half-width of the bell curve is something close to the standard deviation. A standard Normal has $\mu = 0$ and $\sigma = 1$. The distribution of $\bar{x}$ is normal. That is, if you repeatedly sample a population and measure $\bar{x}$ and plot what you get, it will look like a Normal distribution. Note that x doesn't need to be distributed as a Normal for $\bar{x}$ to be Normal.

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Pareto: Fancy name for a power law, but one that is correctly normalized.

Gamma: A general distribution that has many well-known specific examples. It is a combination of a power law and an exponential. The general form is

$$p(x) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} x^{\alpha-1} e^{-\frac{x}{\beta}}$$

where Γ is the gamma function, a generalization of the notion of an integer factorial. The mean and variance are $\mu = = \beta\alpha$, and $\sigma^2 = \beta^2\alpha$. Reduces to the *exponential distribution* for $\alpha = 1$ and $\beta = \theta$, and to the *Chi-square distribution* when $\alpha = \mathrm{n}/2$ and $\beta = 2$.

2. Parametric

2.1. Techniques

Maximum Liklihood: (p90) Assume the data are drawn from some population, change the parameters (e.g., mean and variance) until the the probability that the data were drawn from the distribution with those specific parameters is maximized. Derived by maximizing the ln of the joint probability (a product, that turns into a sum because of the ln).

Bootstrap (paired): Take a subsample of the data and analyze it for a fit. Put the data back and resample as much as you like (p54, p157). Helpful for quantifying errors of parameters because it doesn't assume a particular distribution for the errors. The data resampling naturally take the distribution into account.

Jacknife: This is bootstrap with a sample of n−1.

2.2. Regression

Regression fits a function to the data.

Logical regression: The variable y is either 0 or 1

Poisson regression: The variable y is an integer based on a counting process

Methods for outliers: Trimmed, Huber, and Tukey bisquare are methods used to minimize the influence of outliers on the data. You can also go nonparametric with Thiel-Sen.

2.2.1. Linear

Linear Model: The model is y = β_0 + β_1x + ϵ, where β_0 and β_1 are parameters to be solved and ϵ is a combination of the measurement error, which has some distribution, and an intrinsic scatter, which also has some distribution (p151, p165). The more you know about how the errors are distributed the better you can account for their effects on the fit.

Ordinary Least Squares: Fits β_0 and β_1 by minimizing the sum of squared deviations from the fit (p154).

Weighted Fits: Pearson, Neyman, and 'Astronomer' weight the ordinary least squares in the denominator by the model fit, the observation, or the square of the error, respectively (p163). These are better than ordinary least squares because they make use of all the information in the data. When the errors are Poisson, Astronomer = Neyman. When the errors are not Poisson, 'Astronomer' is not distributed as a χ^2, necessarily, so hypothesis testing is problematic.

Errors in both x and y: Can be dealt with if you know the ratio of systematic to statistical errors (p167). or if you know the actual error distribution function, in which case you use maximum liklihood (p169).

2.2.2. Nonlinear and multivariate fits

Usually minimize sum of squared residuals. There is something called generalized linear modeling (p170). This type of fit is related to PCA.

Number of multivariate parameters: Fits are always better with more parameters. But there is an "adjusted R^2" that penalizes for too many parameters to avoid over-fitting the data (p173).

2.2.3. Principal Component Analysis (PCA)

The concept is that for n-dimensional data you find the axis where the data are most spread out, and that axis is the first eigenvector. Then, go perpendicular to that in n-1 dimensional space, and the direction that aligns with the greatest spread is the second eigenvector, and so on. Depending on the data, you might only need some subset p < n dimensions to adequately match the data, in which case you truncate the dimensionality. This dimensional reduction is a great power of PCA. PCA is sometimes used as an initial step for other methods for this

reason. Usage is typically either (i) some collection of n attributes of the data (e.g. weight, height, age), each of which creates a new dimension; (ii) a spectrum, where each wavelength in the digital data is a new dimension; and (iii) a sound recording or other time-varying quantity, where each time interval is a new dimension. Some properties:

Centering and Whitening: Typically one subtracts the mean from the data (an average spectrum) before processing. This is called centering the data. If the data are a bunch of points in n-space that have different dimensions and you don't want the variable with the largest numerical spread to dominate the component choice, then you can 'whiten' the data by normalizing out the variance for each variable so that each dimension is treated the same. With spectra one typically centers but does not whiten. For example, if height and weight of a collection of people are in millimeters and milligrams and their age in centuries, the age spread is going to be tiny compared with the weight and height spread unless the data are whitened.

Data cubes: Data cubes are typically a collection of spectra, each of which has an (x,y) label indicating its position in the image that is not used when determining the PCA vectors. The dimensionality is the number of wavelength bins in the spectrum.

Eigenvectors: Are by definition perpendicular to one-another. For the case of a collection of spectra, the eigenvectors are spectra and are chosen such that their dot-product is zero. This can make it hard to interpret the PCA output from a physical standpoint. In principal there are as many eigenvectors as their are dimensions in the data set (number of attributes, number of wavelengths or number of time steps depending on the data). In practice you never need all the eigenvectors to describe the data. The PCA eigenvectors are those of the covariance matrix of the data points. If you whiten the data first by subtracting out the means and normalizing the variances along each dimension, then the PCA eigenvectors are those of the correlation matrix of the data points.

Eigenvalues: The eigenvalues λ_i of PCA component i have the property that λ_i^2 is proportional to the percent of the total variance in the data that is accounted for by component i.

Projections along components: You can think of the eigenvectors as a new basis set of vectors, so each spectrum $\mathrm{s(t)} = \sum_{i=1}^{n} a_i e_i(\mathrm{t})$, where $e_i(\mathrm{t})$ are the eigenvectors and a_i the projections of the data point along that eigenvector. You can think of the $e_i(\mathrm{t})$ as ethnicities if you like, so if each spectrum is like a person, it has some fraction of German, Irish, Spanish, etc. Since the data have (x,y), one can plot the amount $a_i(\mathrm{x,y})$ of any eigenvector e_i as an image, and look for spatial clustering (Chinatown) or do Fourier analysis of the resulting image for each component (running out on the analogy here). Interesting structures can emerge in

such images, but they may be hard to interpret.

3. Nonparametric

Nonparametric methods do not depend on parametric assumptions or require that the data arise from some specific distribution.

3.1. See if a sample population comes from, say, a Poisson or Normal distribution

Test to see if the observed e.d.f. $\widehat{F}(x)$ matches a known c.d.f. F(x)

F(x) can be any known distribution, either from a standard distribution function like a Poisson or Normal, or from some other known c.d.f. Typically one tests a null hypothesis that the observed data are drawn from some distribution, and tries to reject the hypothesis at some confidence level using tabulated values. This is a rather 'rough' test in that it doesn't take much to make a significant difference, and it doesn't explain why the difference occurs.

K-S test: Kolmogorov-Smirnoff (p108). Works by looking at the maximum difference between $\widehat{F}(x)$ and F(x). Popular, works well. Not that sensitive near the tails of the distributions.

CvM test: Cramer-von Mises (p109). Same as K-S but uses sum of squared differences. Works better than K-S in some cases.

AD test: Anderson-Darling (p109). Same as K-S but designed to be more sensitive at the extreme ends of the distribution. Uses a funny-looking sum.

3.2. See if sample populations come from the same distribution

K-S test: There is a two-sample version (p113) that tests if the e.d.f. of two samples are the same or not.

Wilcoxian rank sum: a.k.a. the MWW test. Like a nonparametric version of the t-test for different means. Your data has two samples, and you rank the whole sample of both components and look at the sum of the ranks for the first sample. This tells you if the means are different for the two samples. The samples do not have to be the same size.

Kruskal-Wallis test: Same as Wilcoxian but when the data are divided into more than two samples.

3.3. Determine if there is a correlation between two variables that have unknown distributions

These types of tests refer to pairs of data. Often the pair is the (x_i,y_i) coordinate of point i.

Spearman's ρ rank test: An equation with a bunch of rank sums to test significance of correlation. There is a 'D^2+T' table, where D is the rank differences between x and y, and T takes care of ties.

Kendall's τ: Splits the data up into all possible pairs, notes if the slopes are either positive or negative and counts them, kind of like a sign test. Approaches normality faster than Spearman. Robust. The more 'concordant' pairs there are, the more likely there is a correlation. Can also be used to identify anticorrelations.

3.4. Fit a line but without knowing error distributions

Thiel-Sen: Calculates a median of all possible combinations of the slopes. Robust (p160).

3.5. Test for a shift between two populations

Hodges-Lehmann test: Determines if two populations are shifted relative to one-another with some confidence level.

3.6. Hypothesis testing that the mean or median will have some value

Sign test: The sign test (p112) is good for this. It just uses a binomial distribution to see how many points are above or below the value. Insensitive to the amount by which they are above or below.

3.7. Presence of a property among a population split into various pieces

Contingency tables: These compare the observed and expected frequency of the property throughout the table, and are distributed as χ^2 (p114).

4. Time-Series Analysis

A lot of astronomical data has data taken at uneven time intervals. Most of the techniques are for evenly-spaced intervals though.

4.1. Techniques and definitions

Differencing filter: Defining y(t) = x(t) − x(t−1) removes secular trends.

Correlogram: A plot of the autocorrelation function.

Smoothing: Sometimes you want to smooth the data to reduce high frequency noise

Lag-Scatter plots: Plot x(t) vs. x(t+k) for all t. A circle indicates periodic signal, linear implies stochastic autoregression.

4.2. Autoregression for equally-spaced data

Autocorrelation: To get an autocorrelation function (ACF) you multiply the curve by itself and sum it up. That gives the point at x=0. Then you shift by one pixel and multiply by the original and sum that up for the next pixel, and so on. The ACF should equal zero everywhere except for x=0 if there is no time signal.

Partial ACF: An ACF that removes effects of correlations at shorter lag times.

Models: Autoregression models look like x(t) = x(t−1) + ϵ(t), where ϵ is an error term. In other words the counts at time t equal those at time t−1, plus an error. If it depends on t−2 that is one more term to the model. The spectrum of the autoregressive coeficients may be diagnostic.

Number of Coefficients: You can use the Akaike information criterion (p300) to help define how many coefficients you should keep. The criterion kicks in when the value of the coefficients gets too small.

4.3. Unevenly-spaced data

Discrete Correlation Function: The DCF and the *slot autocorrelation* are similar, and search for all the pairs with given time intervals, and then bin the data to get something like an ACF (p303). The *structure function* might also be useful. Once you have this, then you can go on to typical ACF analysis.

Fourier Power spectrum f(ω): Is what its name implies. Mathematically it is related to the ACF (p306). White noise means ACF = 0 and f(ω) = constant. You can iteratively subtract peaks until no other significant peaks are left.

Tapering and Smoothing: Reduce ringing in Fourier data.

Finding periods in unevenly-spaced data: The two main methods are (i) Lomb-Scargle and (ii) DQSE, neither formula very intuitive-looking.

Nonparametric method: Choose a bunch of periods, connect the dots and look for *minimum string length.*

Significance of peaks: Lots of assumptions go into false alarm probabilities like white Gaussian noise with known sigma, so these are hard to interpret and shouldn't be taken too literally for real data.

Nonstationary time-series: Looks for things like change points where the time-behavior of the system changes, and the analysis can get rather involved.