Statistics and Data Analysis in MATLAB

Kendrick Kay, kendrick.kay@wustl.edu

# Lecture 1: Probability distributions and error bars

## 1. Exploring a simple dataset: one variable, one condition

- Let's start with the simplest possible dataset. Suppose we measure a single quantity for a single condition. For example, suppose we measure the heights of male adults. What can we do with the data?
- The *histogram* provides a useful summary of a set of data—it shows the *distribution* of the data. A histogram is constructed by binning values and counting the number of observations in each bin.
- The *mean* and *standard deviation* are simple summaries of a set of data. They are *parametric statistics*, as they make implicit assumptions about the form of the data. The mean is designed to quantify the central tendency of a set of data, while the standard deviation is designed to quantify the spread of a set of data.

$$\text{mean}(x) = \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\text{std}(x) = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$$

In these equations, $x_i$ is the *i*th data point and *n* is the total number of data points.
- The *median* and *interquartile range (IQR)* also summarize data. They are *nonparametric statistics*, as they make minimal assumptions about the form of the data. The *X*th *percentile* is the value below which *X%* of the data points lie. The median is the 50th percentile. The IQR is the difference between the 75th and 25th percentiles.
- Mean and standard deviation are appropriate when the data are roughly Gaussian. When the data are not Gaussian (e.g. skewed, heavy-tailed, outliers present), the mean and standard deviation may be misleading and the median and IQR may be preferable.

## 2. Probability distributions

- A *probability distribution (or probability density function)* is a mathematical function of one or more variables that describes the likelihood of observing any specific set of values for the variables. Distributions can be *univariate* (pertaining to one variable) or *multivariate* (pertaining to more than one variable); we will stick with the univariate case for now. The integral of a probability density function necessarily equals one.
- The *Gaussian (or normal) distribution* is a very useful probability distribution. It is parametric in the sense that it places certain constraints on the distribution of the data (the distribution must be unimodal, symmetric, etc.). The Gaussian distribution has two parameters, the mean ($\mu$) and the standard deviation ($\sigma$), and is given by the following equation:

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

For any given value $x$, this equation specifies how to compute $p(x)$, the likelihood of that value. When points are drawn from a Gaussian distribution, 68% and 95% of the points will be within 1 and 2 standard deviations from the mean, respectively.
- Given a set of data, the Gaussian distribution that best describes the data (i.e. maximizes the likelihood of the data) is the one whose mean and standard deviation are matched to the mean and standard deviation of the data. Thus, when computing the mean and standard deviation of a set of data, you are in a sense fitting a Gaussian distribution to the data.
- An advantage of the Gaussian distribution is that it is simple and may be a reasonable approximation for many types of data. But what if the data are not Gaussian? If there is a suitable parametric probability distribution for the data (e.g. the Poisson distribution), we could choose to use it. Alternatively, we can adopt nonparametric techniques that take a more flexible approach, allowing the data themselves to determine the form of the probability distribution. Such techniques include histograms, bootstrapping, and kernel density estimation, and are covered later in this lecture.

3. Error bars
- When measuring some quantity, we may find that the measurement is different each time it is performed. We attribute this variability to *noise*, i.e. any factor that contributes to variability in the measurement.
- Statistically speaking, the measurements we make constitute a *sample* from the *population*, i.e. the underlying probability distribution that describes the measurement process. The problem is that we are interested in characteristics of the population but all we can observe is our finite sample from the population.
- A *statistic* (e.g. mean) computed on a random sample is subject to variability and is not the same as the statistic computed on the whole population (technically known as the *parameter*). Thus, we need to distrust, to some degree, the statistic computed on the sample. To indicate uncertainty on the statistic, it is useful to plot *error bars* indicating the *standard error*.
- To understand standard error, let's consider a simple example. Suppose we randomly draw $n$ points from a Gaussian distribution with standard deviation $\sigma$ and compute the mean of these points. Then suppose we repeat this process many more times. The distribution of the resulting means will have a standard deviation equal to $\frac{\sigma}{\sqrt{n}}$. This is the standard error, i.e. the standard deviation of the sampling distribution of the statistic. Thus, given a single sample of $n$ data points, the mean of the sample may be offset from the true population mean, and the standard error indicates about how far away the true population mean may be. (Note that when computing standard error on actual data, the standard deviation of the population is unknown, so we use the standard deviation of the sample as an estimate.)
- *Confidence intervals* are intimately related to standard error. Assuming that the sampling distribution is Gaussian, +/– 1 standard error gives the 68% confidence interval and +/– 2 standard errors gives the 95% confidence interval. Technically, the interpretation of confidence intervals is that with repeated experiments, we can expect that *X%* of the time, the true population parameter will be contained within the *X%* confidence interval. More loosely, we can use confidence intervals as indicators of our uncertainty in our estimates.
- *Test-retest* refers to the idea of collecting a set of data, performing some analyses on those data, and then repeating the whole process on a fresh set of data. Variation between the first set of results and the second set of results tells us something about the *reliability (or replicability or*

*reproducibility)* of the results. We can construe test-retest as a simple procedure for estimating error bars that involves drawing two points from a distribution.

4. Nonparametric approaches to error bars

- In the previous example describing standard error, we assumed that the underlying population distribution is Gaussian and we assumed that we want to estimate the mean. Let's drop these parametric assumptions—let's use bootstrapping to bypass the Gaussian assumption and let's compute the median instead of the mean.
- The procedure is simple: Given a set of *n* data points, draw *n* points with replacement from the data points and compute the median of the drawn points. Repeat the procedure many times, e.g. 10,000 times. Finally, summarize the resulting distribution using the median and the 68% confidence interval. (The reason for choosing the 68% confidence interval is that the range spanned by the 68% confidence interval is analogous to the range spanned by +/– 1 standard error in the case of Gaussian error bars.)
- Why does bootstrapping work? Recall that in the parametric approach to error bars, we assume a parametric probability distribution for the data and calculate error bars based on theoretical (analytical) considerations of what happens if we sample from that distribution. Bootstrapping adheres to exactly the same logic, but has two deviations: (1) instead of assuming a parametric probability distribution, the data themselves are used as an approximation of the underlying probability distribution, and (2) instead of calculating error bars analytically, brute computational force is used (random samples are drawn and the sampling distribution of the statistic is directly constructed).
- Note that in the limit, results from bootstrapping will approximately match those based on analytic assumptions. For example, if the data are truly Gaussian-distributed, then as the sample size grows, results from bootstrapping will be very similar to results based on Gaussian assumptions. (However, if the data are not Gaussian-distributed, then of course all bets are off.)
- Why is the same sample size (*n*) used when drawing a bootstrap sample? The reason is that the sample size determines the uncertainty in the estimate. If we artificially increased the number of samples drawn, we would be artificially (and incorrectly) decreasing the uncertainty.
- Why is sampling with replacement used when drawing a bootstrap sample? One reason is that we are simulating independent samples from a distribution; if the samples were drawn without replacement, then there would be dependencies across samples. A more mundane reason is that if we draw without replacement, all samples would be identical (which is clearly absurd).

5. Nonparametric approaches to probability distributions

- *Bootstrapping* can be cast as a nonparametric technique for characterizing probability distributions. The bootstrap estimate of the probability distribution that generated a set of data is simply a probability distribution consisting of a "spike" at each observed data point. This distribution looks a bit funny, but it is nonetheless a valid distribution—we can draw random samples from the distribution, just like any other probability distribution.
- The *histogram* is another nonparametric technique for characterizing probability distributions. To obtain an estimate of the probability distribution that generated a set of data, we simply construct a histogram of the data and then modify the scale of the *y*-axis so that the total area of the bars is equal to one.
- *Kernel density estimation (KDE)* is yet another nonparametric technique for characterizing probability distributions. In KDE, a probability distribution is constructed by placing a *kernel* at

each observed data point and averaging across kernels. Each kernel is itself a probability distribution.

- KDE bridges the gap between bootstrapping and the histogram. On the one hand, KDE can be seen as a smooth version of the bootstrap: instead of placing a sharp spike at each data point, we use a smooth bump. On the other hand, KDE can be seen as a more sophisticated version of the histogram: KDE performs the same basic function as the histogram but avoids the awkward discreteness of histogram distributions through slightly fancier mathematics.
- One issue in using the histogram is what bin size to use. (An analogous issue for KDE is what kernel size to use.) If the bin size is small, the resulting density estimate may be noisy (i.e. unstable across repeated measurements) but has the potential to reveal small-scale features in the data. If the bin size is large, the resulting density estimate will be less noisy (i.e. more stable) but may obscure small-scale features in the data. If the goal of the analysis is simply visualization and data exploration, a reasonable strategy is to try a variety of bin sizes, paying attention to the number of observations that fall into each bin (small numbers suggest an unstable and therefore untrustworthy estimate).