Advanced Topics in Computer Systems, CS262B

Prof. Eric A. Brewer

Statistics Intro

January 20, 2004

One key difference between 262A and B is that this semester we will expect PhD level data analysis and presentation. This includes experimental design, data collection and measurement, and presentation with confidence intervals. In this lecture, we cover the basics, which will not be sufficient to deliver the above goals, but is a start. The difference will be made up as we go through the class projects.

I. Basics

A *distribution* is the set of outcomes of a system with the probability for each outcome. It can be discrete, like a coin or die, or continuous, like a measurement of time. A (fair) die has six possible outcomes with a probability of 1/6th for each.

Let *f(x)* be the probability of outcome *x*, $0 \le f(x) \le 1$. The sum of *f(x)* over all possible *x* must be 1. For the discrete case, we let *R* be the set of possible values for *x*, e.g. *R* ={Heads, Tails} for a coin. That is:

$$\sum_{x \in R} f(x) = 1 \qquad \int_{-\infty}^{\infty} f(x)dx = 1 \tag{1}$$

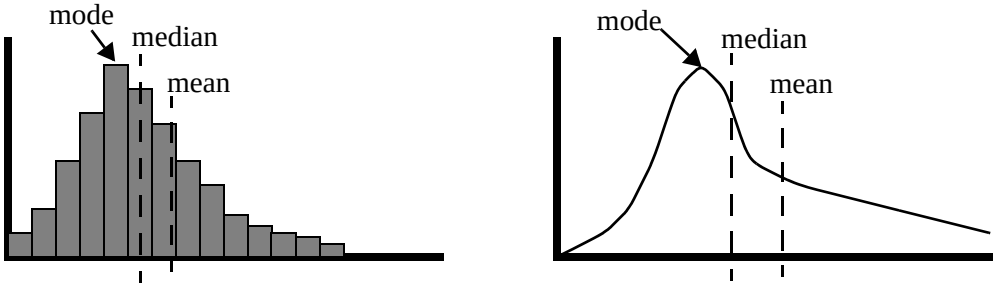
This function, *f(x)*, is called the *probability distribution function*, or *PDF*. For a continuous PDF, the probability of any single outcome is infinitely small, but we can find the probability of a narrow range by integrating over the range:

$$\text{Prob}\{a \le x \le b\} = \int_{a}^{b} f(x)dx \tag{2}$$

The *expected value* is just the sum of the values weighted by their probabilities. We use capital *X* to mean a random variable with a particular PDF, and lower case to be a particular outcome.

$$E[X] \equiv \sum_{x \in R} xf(x) \qquad E[X] \equiv \int_{-\infty}^{\infty} xf(x)dx \tag{3}$$

Note that this is different than an average, which refers to actual measurements rather than the expected outcome. There are three kinds of "average" that are regularly used. The *mean* is the average in the tradi-



tional sense: it is analogous to expected value; we use μ to represent the true mean and $\bar{x}$ to represent the

sample mean (the one that is computed from samples). The *mode* is the value that occurs the most, or for a continuous distribution, the value of the peak. The distributions shown here are *unimodal* (single peak), *bimodal* distributions have two peaks. There is only one mode even for a bimodal distribution, although it is not clear what happens if there is a tie; in general, the mode is only used for unimodal distributions. The *median* is the value with half the values above it and half below it. In a continuous curve the (expected) median has half the area of the PDF is on each side. (A measured median is not a continuous curve, since it is just a collection of samples.) These two plots are *skewed* to the right, which means that they have a long tail to the right. With long tails, we expect mode < median < mean. A left-skewed distribution would have mean < median < mode. Any order is possible, since the mode can be on either side of the median, and since you can "stretch" the tail on one side or the other to move the mean without moving the mode or median.

The *variance* measures the amount of variation in a distribution, in some sense, its "width". It measures the average of the square of the distance from the mean for each value. Think of the "square" part as a way to remove the effect of whether the difference is positive or negative.

$$Variance[X] \equiv E[(X-\mu)^2] = \sum_{x \in R} (x-\mu)^2 f(x) \qquad \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx \tag{4}$$

More useful and intuitive than the variance is the *standard deviation,* σ, which is just the square root of the variance, so $Variance[x] \equiv \sigma^2$. Unlike variance, standard deviation has an intuitive definition: it is the average distance from the mean of the samples. Keep in mind that variance is based of the squares of the distances, and stddev is the square root of the variance, making it essentially the average of the distances. It is also useful to realize that the standard deviation is measured in the same units as the random variable; i.e. if the outcomes are measured in seconds, then so are the mean and standard deviation. This is why we can talk about confidence intervals (below) that are based on the mean and stddev.

It is also useful to know that the expected value of a linear combination of variables is just the linear combination of the expected values, and there is a similar relationship for variance (but not stddev):

$$\text{if } Y = a_1X_1 + a_2X_2 + \ldots + a_nX_n \text{ then:}$$
$$E[Y] = \sum_i a_i E[X_i] \qquad Var[Y] = \sum_i a_i^2 Var[X_i] \tag{5}$$

II. Cumulative Distribution Functions

Before we get to confidence intervals, we should define *cumulative distribution functions,* or *CDFs*. Every PDF has a corresponding CDF which is the cumulative sum of the probabilities up to the current point, which is the same as the probability that x is less than the current point. Using *c(x)* to mean the CDF corresponding to the PDF *f(x)*:

$$c(a) \equiv Prob[x \le a] \equiv \int_{-\infty}^{a} f(x)dx \qquad c(+\infty) \equiv 1 \qquad c(-\infty) \equiv 0 \qquad 0 \le c(x) \le 1 \tag{6}$$

In practice, CDFs are only used for continuous PDFs, but they can be well defined for the discrete case if the values of the random variable are ordered. The main use of a CDF is that it can convert an integration

over the PDF into a subtraction of two CDF values. For example, to find the probability that the outcome in between *a* and *b*:

$$Prob[a \le x \le b] \equiv \int_a^b f(x)dx = c(b) - c(a) \tag{7}$$

This is a really simple but important equation. Intuitively, think of it as "the probability that $a \le x \le b$ equals the probability that $x \le b$ minus the probability that $x \le a$." In practice, we will look up the two CDF values in a table. Note also that:

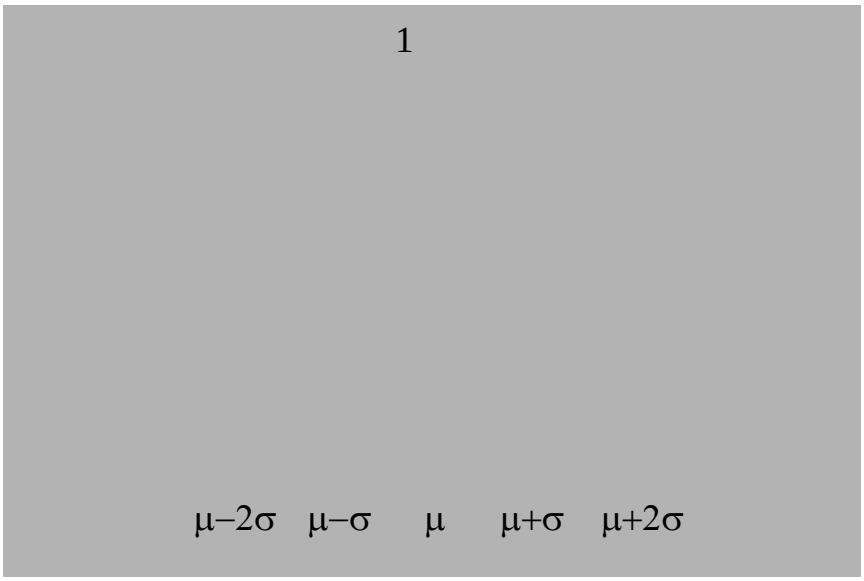
$$Prob[x > a] = 1 - c(a) \tag{8}$$

III. Normal Distribution

The normal distribution is the most important one, in part because it is a key tool for confidence intervals. We define $N(\mu,\sigma)$ as the normal distribution with mean μ and variance σ^2:

$$N(\mu, \sigma) \equiv \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{9}$$

This is a nasty looking integral, but fortunately the CDF table means we almost never have to deal with it. If you are bored, you can prove that $E[N] = \mu$, $Var[N] = \sigma^2$, and that the area under the curve is 1. We call $N(0, 1)$ the *standard* normal distribution, and its CDF is $\Phi(x)$, which is the function we actually look up in a table. The graph below shows N and its CDF plotted against μ and σ:



Since, $\Phi(x)$ is the CDF for $N(0,1)$, we need a way to get the CDF for $N(\mu, \sigma)$. Fortunately, this is just a linear translation:

$$Prob[a < x \le b] = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right) \tag{10}$$

Thus, given a random variable with distribution N(μ, σ), we have a way to compute the probability of any range using just two values from a standardized table.

IV. Sampling

So far we have only talked about distributions that are known a priori. In practice, we don't really know the distribution, or at least not its parameters, and we have to deduce them from samples. The collection of samples is called the *empirical distribution*.

The *sample mean* is average of the samples, typically represented as $\bar{x}$, called "x-bar". From (5), we deduce:

$$\bar{x} \equiv \sum \frac{1}{n} x_i$$

$$E[\bar{x}] = n \cdot \frac{1}{n} E[X] = \mu \qquad Var[\bar{x}] = n \cdot \left(\frac{1}{n}\right)^2 Var[X] = \frac{Var[X]}{n} \tag{11}$$

Since $E[\bar{x}] = \mu$, we say that the sample mean is an *unbiased estimator* of μ. This assumes that the samples are independent and all taken from the same distribution, or *iid* for "independent and identically distributed" (pronounced "eye-eye-dee"). This is a non-trivial assumption that is often wrong! For example, if you sample without replacement from a finite set, the distribution changes as you remove items. Steady-state systems typically have different distributions then the same system when it is warming up.

Perhaps more interesting than the expected value is the variance. The variance of the sample mean decreases linearly as we add samples! This implies that the standard deviation of the sample mean decreases with the square root of the number of samples. This is why adding samples increases the accuracy of the sample mean. This means, among other things, that if X is N(μ, σ), then $\bar{x}$ has the distribution N(μ, $\frac{\sigma}{\sqrt{n}}$).

At this point, it is worth mentioning that you can compute the sample mean and variance as you go, rather than waiting until you have all of the samples. There is an alternate (equivalent) definition of variance that is more useful for this purpose, as it is easy to use as part of a running sum:

$$Variance[X] = E[X^2] - \mu^2 = E[X^2] - (E[X])^2 = \frac{1}{n}\sum_i {x_i}^2 - \left(\frac{1}{n}\sum_i x_i\right)^2 \tag{12}$$

Using this definition, if we keep a running sum of n, x and x^2 then we can at any point calculate the sample mean and the variance for the samples so far. This is great for online systems, or systems looking for only good approximations, since they can wait until the variance drops below some threshold.

V. Central Limit Theorem

So far, we can't say anything about the distribution of $\bar{x}$, other than in the case where X is known a priori to be normally distributed. This is not very useful. Fortunately, there is an amazing theorem that saves us, the *Central Limit Theorem*:

> If $\bar{x}$ is the sample mean of a random variable, X, with an *unknown* distribution with mean μ and variance σ^2, then the distribution of $\bar{x}$ approaches a distribution that is $N(\mu, \frac{\sigma}{\sqrt{n}})$ as *n* becomes large.

This means that (magically) we can *ignore* the underlying distribution of X, as long as we have enough samples (and they are iid). The distribution of the sample mean always approaches the normal distribution.

VI. Confidence Intervals

The goal of a confidence interval is to capture how accurate we believe our sample mean to be. We know that as we take more samples, the variance of the mean declines and we rightfully expect that our measured mean is more accurate (to the real mean). In the rest of this discussion, we will assume that the sample mean is normally distributed, which is a good approximation given the central limit theorem. (Note the we are not assuming anything about the underlying real distribution.)

The basic idea is that if we want to be 90% sure that our sample mean is within *k* units of the real mean, then we pick *k* such that:

$$Prob[\mu - k < \bar{x} < \mu + k] = 0.9 \tag{13}$$

Since we know that $\bar{x}$ is $N(\mu, \frac{\sigma}{\sqrt{n}})$, we can use equation (10) to get:

$$Prob[\mu - k < \bar{x} < \mu + k] = \Phi\left(\frac{(\mu + k) - \mu}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{(\mu - k) - \mu}{\sigma/\sqrt{n}}\right) = \Phi\left(\frac{k\sqrt{n}}{\sigma}\right) - \Phi\left(\frac{-k\sqrt{n}}{\sigma}\right) \tag{14}$$

Since the standard normal distribution is symmetric around 0, we know that:

$$\Phi(a) = 1 - \Phi(-a) \tag{15}$$

We can thus simplify (14) to:

$$Prob[\mu - k < \bar{x} < \mu + k] = 1 - 2\Phi\left(\frac{-k\sqrt{n}}{\sigma}\right) = 0.9 \tag{16}$$

Solving for Φ, we get:

$$\Phi\left(\frac{-k\sqrt{n}}{\sigma}\right) = \frac{1}{2}(1 - 0.9) = 0.05 \tag{17}$$

This should make sense intuitively; we want 5% of the area on each side of the interval, so we find d such that $\Phi(d)$ is 0.05 and then we can solve for k. By scanning the table for Φ, we find that:[1]

$$\Phi(-1.65) = 0.050 \tag{18}$$

We can now solve for k:

$$\begin{aligned} \frac{-k\sqrt{n}}{\sigma} &= -1.65 \\ k &= 1.65\frac{\sigma}{\sqrt{n}} \end{aligned} \tag{19}$$

Thus, we say that the "90% confidence interval for $\bar{x}$" is:

$$\bar{x} \pm 1.65\frac{\sigma}{\sqrt{n}} \tag{20}$$

To save you from scanning the tables, here is a handy guide, using $s = \sigma/(\sqrt{n})$: as the standard deviation of the sample mean:

Confidence Interval	d	Result
90%	1.65	$\bar{x} \pm 1.65s$
95%	1.96	$\bar{x} \pm 1.96s$
99%	2.58	$\bar{x} \pm 2.58s$

A good rule of thumb to do confidence intervals quickly is to just multiply the standard deviation of the mean (s) by 2, which gives an approximate 95% confidence interval.

Finally, this discussion is based on the fact that you know σ exactly (the standard deviation of the *underlying* distribution). This is rarely true; instead we usually estimate it using equation (12). This increases the error of our confidence interval, so we must somehow make it wider, at least if we have a small number of samples. (With a large number of samples, the computed standard deviation becomes very close to the actual one and we can ignore this effect.) The table on the next page gives the modified values of d to use when you have a small number of samples, which is basically all the time.

1. Unfortunately, the tables in most books only have the right half of the Φ, which have values greater than one half. In this case, using equation (15), you look for $\Phi(-d) = 1 - 0.05$ (the mirrored part on the right), and find that $-d = 1.65$.

This table is based on the "Student's T Distribution", which allows you to adjust the width of the confidence interval when you have a small number of samples.[1]

Samples	90%	95%	99%	99.9%
1	6.314	12.71	63.66	637
2	2.920	4.303	9.925	31.6
3	2.353	3.182	5.841	12.92
4	2.132	2.776	4.604	8.610
5	2.015	2.571	4.032	6.869
6	1.943	2.447	3.707	5.959
7	1.895	2.365	3.499	5.408
8	1.860	2.306	3.355	5.041
9	1.833	2.262	3.250	4.781
10	1.812	2.228	3.169	4.587
11	1.796	2.201	3.106	4.437
12	1.782	2.179	3.055	4.318
13	1.771	2.160	3.012	4.221
14	1.761	2.145	2.977	4.140
15	1.753	2.131	2.947	4.073
16	1.746	2.120	2.921	4.015
17	1.740	2.110	2.898	3.965
18	1.734	2.101	2.878	3.922
19	1.729	2.093	2.861	3.883
20	1.725	2.086	2.845	3.850
21	1.721	2.080	2.831	3.819
22	1.717	2.074	2.819	3.792
23	1.714	2.069	2.807	3.768
24	1.711	2.064	2.797	3.745
25	1.708	2.060	2.787	3.725
26	1.706	2.056	2.779	3.707
28	1.701	2.048	2.763	3.674
30	1.697	2.042	2.750	3.646
32	1.694	2.037	2.738	3.622
34	1.691	2.032	2.728	3.601
36	1.688	2.028	2.719	3.582
38	1.686	2.024	2.712	3.566
40	1.684	2.021	2.704	3.551
42	1.682	2.018	2.698	3.538
44	1.680	2.015	2.692	3.526
46	1.679	2.013	2.687	3.515
48	1.677	2.011	2.682	3.505
50	1.676	2.009	2.678	3.496
55	1.673	2.004	2.668	3.476
60	1.671	2.000	2.660	3.460
65	1.669	1.997	2.654	3.447
70	1.667	1.994	2.648	3.435
80	1.664	1.990	2.639	3.416
100	1.660	1.984	2.626	3.390
150	1.655	1.976	2.609	3.357
200	1.653	1.972	2.601	3.340

1. This table was calculated by APL programs written by William Knight and is available online at http://www.math.unb.ca/~knight/utility/t-table.htm