

210A Week 2 Notes

Distributional graphs

Graphs are useful for understanding the basic shape of the data. A lot of statistical concepts make the most sense in the context of a graph. For instance, standard deviation is defined as $\sqrt{\sum (y_i - \bar{y})^2 / n}$, but its real meaning is best understood with a graph. Later, we'll use graphs not just for data but more abstract issues, like standard error.

With categorical data pretty much all you can show is relative frequency by category. You can also show a continuous variable by category (eg, distribution of income for men vs women), but that's a different issue. Basically all you can do are frequency graphs like bar graphs (what Excel "column graphs") or pie charts. Note that you don't want to use a line graph with nominal data as that implies continuous or at least ordinal data.

A histogram is a special case of a frequency graph where the variable is continuous.¹ The way it works is that it divides the continuous data into "bins" and then plots the bins as a tightly grouped column graph. Bins can be tricky, especially if the sample is relatively small, and it's a good idea to experiment with different numbers of bins. For example, we can use the auto dataset to draw some histograms of car mileage.

```
sysuse auto, clear
histogram mpg, bin(5)
histogram mpg, bin(8)
histogram mpg, bin(12)
```

Stata treats the number of bins as an option, and as usual if you provide no specific option it makes a reasonable guess. In this case it doesn't really matter how we bin it, any way you draw it the basic shape is the same – a peak near but not all the way to the left and a downward slope to the right. However in other cases you can get weird stuff based on the number of bins.

Kernel density plots are like histograms but instead of binning they use a smoothed line to approximate the distribution. The smoothing algorithms are somewhat theory-laden and work better with some distributions than with others. You can draw a kernel density plot either by itself or super-imposed on a histogram. One nice thing about kernel density plots is that it is easier to use them to compare two similar distributions in a single graph than to do the same with histograms.

```
kdensity mpg
histogram mpg, bin(12) kdens
```

Regardless of whether you're using kernel density plot or a histogram, what you're looking for is the shape, which has implications for how you analyze the data. The first thing you're looking for is symmetry.

A symmetrical distribution with the mode (i.e., peak) in the middle and tapering off to the sides is a normal distribution or roughly normal distribution. Normal distributions have a lot of very convenient properties. Among these is that in true normal distribution, mean,

¹Stem and leaf plots are a special case of histograms where the histogram is drawn using digits. Textbooks like to talk about them but you rarely see them used in practice, in part because they don't scale well to large datasets.

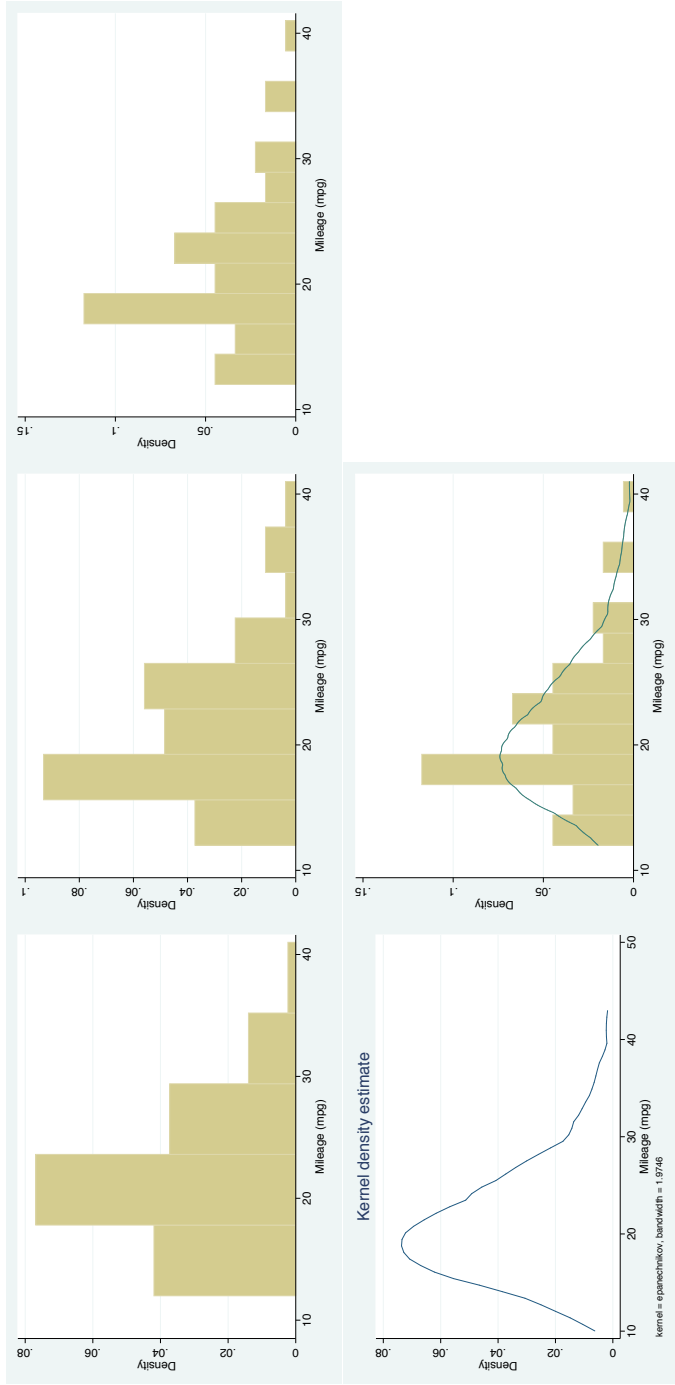


Figure 1: MPG Distribution

median, and mode will all be identical. That is, “average” has a common sense interpretation because all of its meaning are aligned.

A distribution that has the bulk of the area to one side and a tail pointing off to the other is “skewed” with the direction of the skew being the side with the tail.² When the distribution is extremely right-skewed it makes sense to just call it a count. The Poisson (eg, our MPG graph) is a mildly right-skewed graph. Such a distribution is similar to an exponentiated normal so you can force it to be normal by “logging” it (i.e., take the natural logarithm of x). Logging is a compressing transformation so you can think of it as scrunching the right-tail back towards the rest of the distribution. (Conversely, exponentiating, or taking e^x , stretches out the right part of the distribution). If you have Poisson data and are using a technique that requires normal data, you’ll need to log it. However it is increasingly common to find techniques optimized for Poisson data or negative binomial data. (The Poisson is a special case of the negative binomial).

```
gen ln_mpg=log(mpg)
kdensity ln_mpg
```

Note that logging can’t handle values less than or equal to zero so you’ll have to handle such cases (often by adding a constant sufficient to make all cases positive).

When you have skewed data, “average” is less obviously meaningful than it is with normal distribution because extreme values in the tail will bring up the mean much more than the median or mode. With skewed distributions median is usually considered a more meaningful figure than mean for getting a sense of the distribution. For instance, income is right-skewed and so it is much more common to talk about median income than mean income as the latter figure is sensitive to whether Warren Buffett and Bill Gates had a good quarter. Note though that for many purposes mean is still important, even in a count. For instance, Nassim Taleb’s argument is basically that investors are too attentive to the median return and not enough to the mean return. Likewise, most advanced stats are built on mean.

Some distributions are more complicated, for instance, a bimodal (two-peaked) distribution often indicates that you are conflating two populations that for analytical purposes might better be distinguished. For instance, consider a simulation of height for men and women, with each gender having a (separate) normal distribution for height.

```
clear
set obs 10000
gen male=0
replace male=1 in 1/5000
lab def male 0 "Women" 1 "Men"
lab val male male
*the rnormal function generates a random number from the
* normal distribution. also see rpoisson, runiform, etc
gen height=rnormal()*2+70 if male==1
replace height=rnormal()*1.5+64 if male==0
kdensity height
graphexportpdf w2_height
```

²Textbooks like to talk about both right- and left-skewed distributions. But left-skew is rare in practice. The only case I can think of is the grade distribution in my undergraduate class. However if you think of it not as points above zero earned, but as points below 100 docked, then it’s really a right-skewed distributions.

```

twoway (kdensity height if male==0, lwidth(thick)) /*
      */(kdensity height if male==1, lwidth(thick)), /*
      */ legend(order(1 "Women" 2 "Men"))
graphexportpdf w2_height_gender

```

When you look at height for both men and women together the distribution is bimodal and only approximately normal. However, when you break it apart by gender you see that there are really two populations, each of which is normally distributed. Note that in this example not only do women tend to be shorter than men but to have a narrower range of heights, which would be another issue with conflating them. Also note that there is some overlap between the two populations. In a few weeks we'll learn how to use a *t*-test to determine how much overlap is enough to say that two populations are similar as to a trait.

One more important type of distributional graph is the box plot, or box and whisker plot. It is similar to a histogram but more compact, making it well suited for comparing distributions between categories (eg, height of me vs height of women).

```

graph box height, by(male)

```

A boxplot shows a box encompassing the central half of the distribution (ie, the 25th percentile through the 75th percentile), which is also known as the “inter-quartile range.” The median is shown as a line inside this box. Extending from this box are two whiskers going to what are basically the minimum and maximum. If there are a few outliers (extreme values) these are shown as dots beyond the whiskers. Just like a histogram, you can see symmetry in a box-plot. A normal distribution will have about equal distance between minimum and 25th percentile, 25th and 75th, and 75th and maximum. In contrast, a right-skewed distribution will show an especially large distance between 75th and maximum with a few outliers beyond the maximum whisker.³

Some basic summary statistics

There are three meanings of “average” (or “central tendency”): mean, median, and mode. In a normal distribution they are equal, but in other distributions they can diverge. Mean is usually expressed as a horizontal bar atop the variable, like this \bar{x} .⁴ Of course $\bar{x} = \sum x_i/n$. The mean is highly sensitive to outliers when you are using skewed data.

The median is simply the middle value in a sort. Median is also known as the 50th percentile, other important percentiles are the quartile thresholds (p25, p50, and p75) and the decile thresholds (p10, p20, ... p90). Median is robust to outliers so it has a more intuitive meaning with skewed data.

The mode is simply the most common value. In a histogram or kernel density graph it's the peak. Technically, the mode is the single highest peak but we still refer to a graph with two distinct peaks as bimodal, even if one peak is lower than the other.

Range is (minimum, maximum). Unlike in a boxplot these are the true min and max, inclusive of outliers.

³Note that boxplots are usually drawn vertically not horizontally so “right-skewed” looks “top-skewed,” but we still call it “right-skewed.”

⁴Technically \bar{x} is the sample mean. The population mean is noted as μ_x , however this is more of interest to mathematicians than to practitioners.

Standard deviation is noted as σ (sigma) and defined as $\sqrt{\sum (y_i - \bar{y})^2 / n}$.⁵ Standard deviation has an interesting interpretation on a normal distribution if you imagine starting at the mean/mode/median and then going out so many sigmas and seeing how many cases you have encompassed. With a normal distribution, $\bar{x} \pm .67\sigma$ encompasses 50% of the cases and is equivalent to the interquartile range. $\bar{x} \pm \sigma$ gets 68% of the cases. $\bar{x} \pm 2\sigma$ encompasses 95% of the cases, which as you will learn is an important threshold for statistical significance (at least by convention, there's nothing intrinsically special about it). $\bar{x} \pm 3\sigma$ encompasses over 99% of the cases. These densities are given in Table A of your book (which is conveniently the inside back cover). Note that table A gives the density in one tail, so to see what's in the middle, double the tail and subtract this number from one. These figures are for a true normal. Most other distributions will have lower densities, with the limits being established by Chebyshev's Inequality.

Note that for the same reasons that we like to use median instead of mean to summarize skewed data we usually talk about quartiles or deciles instead of standard deviation.

Probability

A lot of statistics courses, particularly those taught by statisticians, start from probability theory. This is because statistical proofs are grounded in probability theory but it has little application for the practical researcher. For us, probability is mostly relevant so as to have a sense of what the null hypothesis means. That is, when we say something is significant, this is compared to what. For example the χ^2 asks, given the distribution of each variable, how unlikely are the observed cell frequencies?

In learning probability it is easiest to think of discrete probabilities, like coin tosses, because this lets us keep everything as arithmetic and algebra instead of calculus. However in practice we usually use continuous probability statistics. In fact, even when it looks like we're using binary outcomes, at a deep level we're often really using continuous probability distributions. Also note that probability is often statistics in reverse. That is what is stated as an assumption in probability theory (e.g., "if a is statistically independent of b , then $p(a \cap b) = p(a)p(b)$ ") is usually an empirical question in statistical practice (e.g., "given the observed relationship of a and b , how likely is it that they are statistically independent of one another"?). Here are a few important principles of probability.

Probabilities sum to one

$$\sum p_i = 1$$

More precisely, the probabilities of the full set of mutually exclusive possibilities sums to one. Also note that it's absurd to talk about something that occurs "less often than never" so $p_i \geq 0$.

The probability of something not happening is equal to 1 minus the probability of it happening

$$p(\text{not} - a) = 1 - p(a)$$

Note that in the notation of set theory, $\text{not} - a$ would be called the "complement of a " or a^c . This is purely an issue of nomenclature.

⁵Technically, σ is the population standard deviation and s is the sample standard deviation but we tend to use them interchangeably.

Independence obtains when the probability of co-occurrence is equal to the probability of one times the other

if independent, then $p(a \cap b) = p(a)p(b)$ ⁶

Another way to put this is to turn it around and say *if* that equation is true, then a and b are independent. This is exactly what a χ^2 does.

Also note that you don't add the probabilities, you multiply them. People have weird intuitions about probability. They often want to add the probabilities, but logically $p(a \cap b) < p(a)$. The failure to appreciate independence is the source of the "gambler's fallacy" that "I'm due for a win" or, conversely, that "I'm on a hot streak." (This is a fallacy because gambling equipment is designed to have each round be statistically independent, in many natural situations such inferences might be reasonable).

As a more extended example, consider page 303 of *The Other Greeks*, one of several books by Victor Hanson that changed the way classicists understood warfare in classical Greece:

To take a theoretical example: if an infantry force during this era fought one out of every two years ... a young farmer, twenty-one years of age, could potentially be called up twenty times before his sixtieth birthday. If he experienced an equal number of winning and losing battles, there was roughly a ten percent chance he would lose his life in any encounter. By age forty a hoplite could expect a theoretical certainty of death in battle.

Hanson is making the simplifying assumption that the chance of death was the same over the lifetime (which it wasn't, as in the ancient world it was common to put hardened old veterans in the back ranks to keep the antsy young men from retreating), but he understands this. The real problem is that he doesn't understand probability as well as he understands Greek warfare. Hanson's model seems to be that if we call chance of death in any given battle $p(\text{death}_i)$ then,

$$p(\text{deathbyage40}) = \sum p(\text{death}_i) = 10(p(\text{death}_i)) = 10(.1) = 1 \text{ [WRONG]}$$

Why this is wrong is more obvious if you extrapolate all the way to age 60 (and we know that there were late-middle-age hoplites).

$$p(\text{deathbyage60}) = \sum p(\text{death}_i) = 20(p(\text{death}_i)) = 20(.1) = 2 \text{ [WRONG]}$$

Of course it's impossible for any outcome to have a probability greater than one.

Here is the correct way to calculate it. First, consider the complement of death, which is survival. Then *multiply* the probabilities of ten or twenty consecutive survivals.

$$p(\text{deathbyage40}) = 1 - \prod (1 - p(\text{death}_i)) = 1 - (1 - p(\text{death}_i))^{10} = 1 - .9^{10} = .6513$$

$$p(\text{deathbyage60}) = 1 - \prod (1 - p(\text{death}_i)) = 1 - (1 - p(\text{death}_i))^{20} = 1 - .9^{20} = .8784$$

Thus a hoplite had a p of .65 of making it to 40 and a p of .88 of dying in battle before retirement from militia duty. This is still pretty grim but unlike a 200% probability of death by age 60, it's neither certain death nor logically impossible. Also note that this is a good example of how you get a count distribution, in this case the survival function for hoplites, with the areas of interest being "greater than or equal to ten battles" and "greater than or equal to 20 battles."

Expected value is the sum of outcomes times likelihoods

$$E(x) = \sum p_i x_i$$

⁶ \cap stands for "intersection" in set theory. If you are using L^AT_EX it is typeset as "\cap" but otherwise you can just type "&" as it means basically the same thing.

Expected value is usually used in the contexts of choices to evaluate the value of different options. The expected value of choice x is defined as the sum of the likelihood times the value of each outcome (all of which is assuming that choice x has been made, otherwise it goes into $E(not-x)$). This is useful for things like gambling. (When the expected value of an action is equal to its cost, we call this a “fair bet”).

The most famous example of expected value is Pascal’s Wager, a practical (as compared to epistemic) proof for believing in God. It goes that if you worship God and he exists, you go to heaven, but if he doesn’t, you lose nothing. In contrast if you don’t worship and he does exist, you go to hell, but if he doesn’t exist, you gain nothing. Since the afterlife lasts for eternity, assign infinite values to heavenly bliss and infernal torment. Therefore the expected values of worship and not worship are:

$$E(worship) = p_{God}\infty + (1 - p_{God})0$$

$$E(not - worship) = p_{God}(-\infty) + (1 - p_{God})0$$

Due to the weird properties of infinity, it turns out that as long as p_{God} is greater than zero, even if it is infinitesimally unlikely, the expected value is still greater for worshipping God than for not worshipping God. (If we force the outcomes to have finite values then the size of p_{God} does matter).

Bayes Theorem

Bayes Theorem is fundamentally about understanding cases where the assumption of independence is violated and so it makes sense to talk about conditional probabilities, which we express as $p(b|a)$.

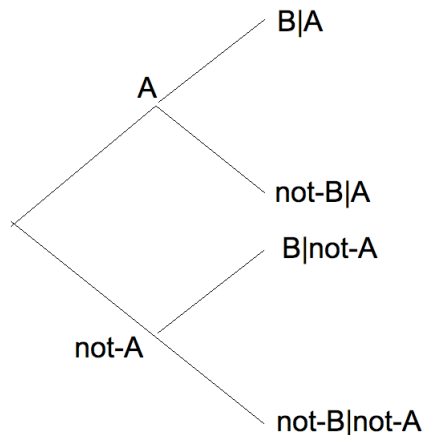
Note that people sometimes confuse “joint” and “conditional” probabilities. For instance, suppose there is a woman named Susan.

1. What is the probability that Susan wears glasses? $p(a)$
2. What is the probability that Susan is a librarian? $p(b)$
3. What is the probability that Susan is a librarian who wears glasses? $p(a \cap b)$

Since the number of librarians with perfect vision is either zero or positive, logically the third probability has to be the least but many people will say glasses & librarian is more likely than just librarian. The mistake seems to be that people are really thinking of a conditional probability, $p(glasses|librarian)$, which may indeed be higher than just $p(glasses)$. On the other hand $p(glasses)$ is *necessarily* higher than the joint probability $p(glasses \cap librarian)$.

Bayes theorem is useful for conditional probabilities when you do *not* have statistical independence.⁷ It’s easiest thought of as a tree diagram where each fork in the tree has probabilities. The probabilities are related such that the probability of what may happen is different depending on what has already happened. You distinguish between what *may* happen and what *has* happened with a vertical line so that $p(b|a)$ means what is the probability of b given that we know a has already occurred. It’s usually a good idea to draw a tree showing the different nested probabilities. For instance, note that all of the possibilities that branch off from a are shown as having the prior condition of a .

⁷We can view the independent model of joint probability as a special case of Bayes Theorem where $p(b|a) = p(b|not - a) = p(b)$.



The probabilities sum to one within prior conditions (but not across prior conditions). For example, $p(b|a) + p(not-b|a) = 1$. However $p(b|a) + p(not-b|a) + p(b|not-a) + p(not-b|not-a) \neq 1$. The easy way to think of it is that you sum all the tines of the specific fork.

You can also express the grand probability of a contingent condition across all contingencies as the sum of contingencies times priors. Such a probability will be less than one. For instance, $p(b) = p(a)p(b|a) + p(not-a)p(b|not-a) < 1$.

The usefulness of Bayes Theorem is if you know the base rate or $p(a)$ and you know $p(b|a)$ you can estimate $p(a|b)$. This is useful if you think of b as a crude proxy for a (itself unobservable in any one case) and you want to know how likely it is that an observed b is really an a . One useful application is to imagine that b is a test for a where you know the population prevalence of a or $p(a)$, the false positive rate or $p(b|not-a)$, and the false negative rate or $p(not-b|a)$. Now imagine that you want to know the probability that a positive test (“b”) is a true case of the disease.

$$p(a|b) = \frac{p(b|a)*p(a)}{p(b|a)*p(a)+p(b|not-a)*p(not-a)}$$

For instance, if we assume that the rate of HIV infection is 1%, and the test is 99% accurate, what is the probability that your friend with a positive test is actually healthy. It works out to:

$$\begin{aligned} p(HIV|positivetest) &= \frac{p(positivetest|HIV)*p(HIV)}{p(positivetest|HIV)*p(HIV)+p(positivetest|not-HIV)*p(not-HIV)} \\ p(HIV|positivetest) &= \frac{p(.99)*p(.01)}{p(.99)*p(.01)+p(.01)*p(.99)} \\ p(HIV|positivetest) &= .5 \end{aligned}$$

One implication of this is that if a disease is very rare then any given positive test is likely to be a false positive even if the test is known to be pretty accurate. The reason is that if $p(a)$ is low then $p(not-a)$ will be high and this will make for a large denominator.

Another implication (often called “Bayesian inference” or “statistical discrimination”) is that if you have information that base rates differ by groups then this information can be very useful. Surprisingly, this means that the same test result implies different things for different people. So imagine two people, a heroin addicted prostitute and a nun, each of whom gets a positive HIV test. The first person is probably actually sick and the second person actually healthy.

Bayes theorem is uncontroversial as far as we have discussed it, but some statisticians and philosophers have extrapolated from it a radical reconception of statistics called “Bayesianism.” We won’t cover these models but they are used by sociologists like Bruce Western and Scott Lynch and embodied in software like WinBUGS and the R package BRugs.