Math 203

Topics for second exam

Statistics: the science of data

Chapter 5: Producing data

Statistics is all about drawing conclusions about the opinions/behavior/structure of large populations based on information about a relatively small part of that population, called a *sample*.

Samples can be gathered in a variety of ways:
Convenience sample: interview people as they walk by
Voluntary response sample (=call-in poll): the subject contacts the interviewer
Random sample: subjects are chosen 'randomly' from the target population

A sampling method is *biased* if it systematically favors one outcome over another (i.e., there are systematic reasons why the poll will not reflect the true behavior of the population).

E.g., comvenience poll: ask people at a mall 'Do you like malls?'

Random sample: every sample of the same size is as likely to be chosen as any other. Basic technique: assign every subject in the population a different number with the same number of digits, and then choose your sample by picking strings of numbers from a *table of random digits*.

Sampling variability: we would never expect two random samples to give the exact same result! There is always some variablility in results. However, this variability behaves in a very predictable way.

The individual sample results will be spread out around the 'true' value describing the opinions/structure of the whole population. But if our sample size is large, the values from our random samples will be more bunched together. This is described in terms of a 'margin of error' (MOE): 95% of our random samples will fall within the MOE of the true value, where the MOE is

5% for sample size 600, 4% for 1000, and 3% for 1500

Experiments: Unlike opinion polls, an experiment seeks to determine how one quantity will change when another is varied, e.g., how recovery time changes with the change in dosage for some new drug; does paying people more for the same job make them work harder, etc.

The basic problem in experimental design is making sure that only the thing we think we are varying (the causal variable), to see the change in the other quantity (the response variable), is actually changing!

Confounding: changes in two variables (usually, the one we're tracking and one we're not!) have similar effects on the response variable.

E.g., the placebo effect: giving patients something (even if it is 'nothing') will have some effect. Interviewer effects: might send subtle signals which will effect the subjects responses.

Solution: use a *control group*, to which you give a 'sham' treatment to. Both control group and experimental group are chosen randomly, to average out individual differences (randomized comparative experiment). For interviewer effects, use a double-blind experiment; neither interviewer nor subject know who is in control group.

When is a change in response variable really being caused by the change in causal variable? A change is *statistically significant* if it is unlikely to be the result of random chance. The idea is, even small changes can be significant if they occur to a large enough population.

Chapter 6: Describing data

Once you have generated your data, it's time to understand it! The idea is to look for patterns in the data, which tells you something about your population.

Typically, we try to disply the data graphically; the eye is very good at picking out patterns.

Dotplot: put one dot over the number 'N' for every sample that had value N. E.g., flip a coin 50 times, many times, record the number of heads. Typically, our dots will sort of clup together, with some *outliers* lying far away from the majority.

If our numbers can take on too many distinct values (e.g., 100 diffferent possible exam grades in a class of only 30), we might plot ranges of values (91-100, 81-90, etc.) recording the number of samples in each range as a rectangle with width = the range of values and height the number of samples. This is a *histogram* of the data.

With both we can try to describe the 'shape' of the data: any symmetry or lack of it (e.g., skewed (= stretched out) to the right or left; data has a 'tail'), where the center is, how spread out the data is. For this, there are two sets of numbers we can use:

Median/Quartiles: list the numbers in increasing order. The *median* M is the middle number on this list. The fist quartile Q_1 is the median of the bottom half (i.e., one-fourth of the way along the list), and the third quartile Q_3 is the median of the top half. The median describes the center; the quartiles describe the spread.

Mean/standard deviation: the *mean* μ is the 'average'; add all of the numbers up, and divide by how many numbers you have. This gives a (usually different) description of the center of the data.

The *standard deviation* σ describes the spread; it is the **square root** of the variance σ^2, which is calculated as (here $x_1, \ldots, x_n$ are our data points)

$$\sigma^2 = \frac{(\mu - x_1)^2 + (\mu - x_2)^2 + \cdots + (\mu - x_n)^2)}{(n-1)}$$

The larger σ is, the more spread out our data is.

The distinction between these two way of measuring center and spread is basically that μ and σ are much more sensitive to outliers, while M and the quartiles are almost completely ignorant of them.

In the case of M and the quartiles, we can add the highest and lowest values to create a *boxplot*; draw a horizontal line at M, a rectangle on top whose top and bottom are the quartiles, and finally a vertical line on top of this whose top and bottom are the highest and lowest values in the data set. This gives a very quick representation of the center, spread, and outliers in the data, especially for making comparisions between different data sets.

With data from experiments, on the other hand, our data comes in pairs; e.g., pairs of dosages and recovery times. To visualize this sort of data, we can use a *scatterplot*. Plot each pair as a point on a rectangular grid (with scale for the first number (causal variable)along the bottom, and for the second (response variable) along the left side), where the point lies above the first number and to the right of the second number.

With this we can look for patterns; form (clumped together or spread out?), direction (is there a trend in the data, does it follow a line/curve?), and strength (how cloe do all the data points stay to that line/curve?)

Can we predict values of the response variable for values of the causal variable that we didn't check? Use a *regression line*:

Can fit a line to the data 'by eye'; **slope** of the line (the 'a' of y=ax+b) tells us how the data trends. *Correlation* describes how well the line fits the data (the 'strength', above).

The least squares regression line goes through the point decribing the mean of both variables, and has slope equal to (where $(x_1, y_1), \ldots, (x_n, y_n)$ are our data points, and Σ means add up all of the quantities from 1 to n)

$$b = \frac{n\Sigma(x_i y_i) - (\Sigma x_i)(\Sigma y_i)}{n\Sigma(x_i^2) - (\Sigma x_i)^2}$$

Chapter 7: Probability: the mathematics of chance

Probability is the study of the long-term behavior of random phenomena, where *random*, basically, means that knowledge of what the phenomenon has done before will not let you decide what it will do <u>next</u>. Such phenomena include flipping coins, rolling dice, the Dow Jones Industrial Average, etc. The basic idea is that while the object's short term behavior is impossible to predict, its long term (average) behavior can be predicted with great accuracy!

Each observation of the object is a *trail* (e.g., the flip of a coin; and each possible outcome of the trial is an *event*. The *probability* of each event predicts how many times the event will occur, in a large number of trials.

We can express these things in a *probability model*. It consists of two things:

1. A *sample space* S = the collection of all possible outcomes for our trial
2. A *probability* (= a number between 0 and 1) for each outcome.

The idea is that the probability describes the fraction of times we would <u>expect</u> our outcome to occur in a very large number of trials.

The individual probablilities must add up to one, because: If we let an *event* mean, more generally, some collection of outcomes, then the probability of the event should be the sum of the individual probabilities of each event. Consequently, the sum od all the probabilities should be the probability that some one of the outcomes occurs in each trial, i.e., the fraction of the time that something happens! Since <u>something</u> always happens, this probability is one.

Ex: flipping a (fair!) coin; the sample space is {heads,tails}, and each has a probability of .5 .
Ex: rolling a pair of dice: there are 36 possible outcomes (if we keep track of which die is which), each having a probability of 1/36 .
These probability models describe *equally likely outcomes*; each event has the same probability.

Ex: the list of birthdays in a classroom of 20 students; there are 366^{20} possible outcomes; is each equally likely?
Ex: Roll a pair of dice, but don't keep track of which is which; there are then only 21 possible outcomes, and they are not equally likely!

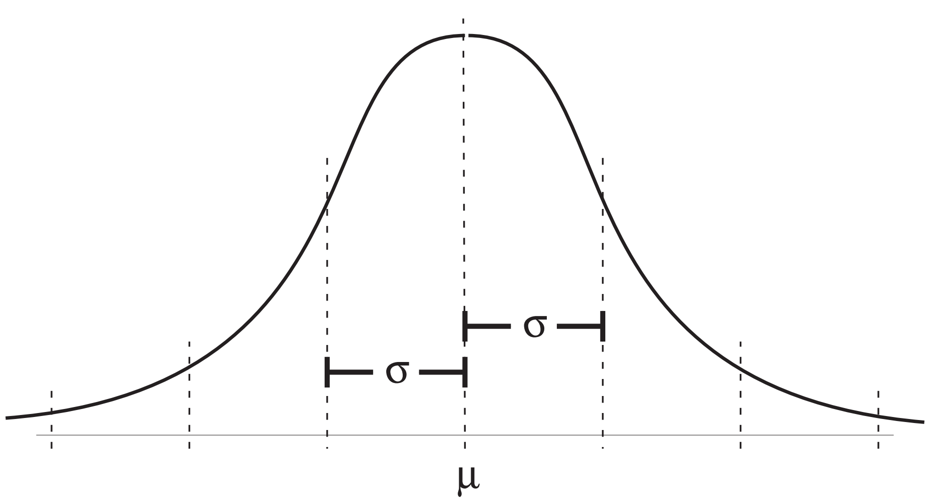
Just as with dotplots, is probability model has a *mean* and a *standard deviation* (and they mean much the same thing!):
For a probability model with sample space $S = \{x_1, \ldots, x_k\}$, where each outcome x_i occurs with probability p_i, then:
The mean μ of the model is the sum of the numbers $p_i x_i$ (think: for a large number N of trials, each x_i will occur approximately Np_i times; average these numbers!)
The standard deviation σ of the model is the square root of the sum of the numbers $p_i(\mu - x_i)^2$ (again, this is what we would expect from just thinking of a large number of trials).

The significance of these numbers stem from two important results:

The Law of Large Numbers: If we observe a random phenomenon that obeys a certain probability model (e.g., flip coins or roll dice) for larger and larger numbers of trials, then:
1. The fraction of times a particular outcome occurs will get closer and closer to the probability of that outcome, and
2. The average of all of the outcomes wil get closer and closer to the mean μ .

The Central Limit Theorem: under the same conditions, if we repeatedly average the oucomes of n trials, and plot the resulting averages, then these averages will have distribution that is (aprroximately) bell curve shaped (see below!); and the bell curve will have mean μ and standard deviation $\sigma/\sqrt{n}$.

The idea behind the Central Limit Theorem is that taking a survey of a population has alot in common with gambling; there is an underlying pattern (the fraction of the population that feels a certain way), but the individual answers we get will vary in a random way. If we plot the fraction from a large collection of samples of the same size, they will typically form a bell curve or *normal ditribution*. And the basic idea is that everything about a bell curve is determined by two numbers: its center (= mean = μ) and its spread (= standard deviation = σ). The standard deviation can be seen as the distance from the center to the point on the curve where the curve starts to 'level out'.



Every normal distribution has the same properties:
The median is μ (= the mean); the first and third quartiles are $Q_1 = \mu - 2\sigma/3$ and $Q_3 = \mu + 2\sigma/3$ (so half of the data from your trials will lie between these two numbers).
The 68-95-99.7 Rule:
68% of the data points will lie between $\mu - \sigma$ and $\mu + \sigma$.

95% of the data points will lie between $\mu - 2\sigma$ and $\mu + 2\sigma$.
99.7% of the data points will lie between $\mu - 3\sigma$ and $\mu + 3\sigma$.

The significance of the Central Limit Theorem is that the spread of the distribution for a larger number trials is always smaller (there is that $\sqrt{n}$ in the denominator of the standard deviation...). So these rules say that the fractions we get from large trials mostly fall <u>very</u> <u>close</u> to the mean.

This is why casinos make money. The mean of the probability model for all of their games is slightly negative (they get your money). If you play only a few times, the spread of possible outcomes for you is very large. This is what makes gambling exciting. But for the house, their spread is very small, since they average wins and losses over a very large number of plays. In fact, they are virtually guaranteed to make money every day, and that amount is nearly always very close to the mean (times the number of players!).

Chapter 8: Statistical inference

Here we return to something we introduced back in Chapter 5; the idea of the margin of error in a poll. The basic idea is that this is the same as the '95' in the 68-95-99.7 rule.

The results of opinion poll are usually given as a percent; '35% of those polled said.....' . Here we mostly think of this as a fraction; 35% <u>means</u> 35 out of 100, or .35 .

In general, we call such a statement a sample proportion. The idea is that it is supposed to represent a number close to the proportion of the entire population who would say whatever it is they would say. In general, any number obtained from a sample is called a *statistic*; the number for the whole population that it is intended to estimate is called a *parameter*. In the case of proportions, the smaple proportion is usually denoted $\hat{p}$, and the corresponding parameter is called p.

We know that different samples will yield different proportions; there is sampling variability. But the Central Limit Theorem (CLT) tells us what this distribution looks like! We can think of pick someone from the population as a random phenomenon, where they have a probability of p of saying whatever it is we reported (call this event '1'), and a probability of $1 - p$ of not saying it (call this event '0'). We can then calculate that the mean μ of this model is p, and its standard deviation σ is (unfortunately) $\sqrt{p(1-p)}$. *So*, if we carry out a number of trials (= ask n people), and look at the average of the n trials (= the sample proportion $\hat{p}$, and do this many times, the CLT says these will form a bell curve with mean μ and standard deviation $\sigma_{\hat{p}} = \sqrt{p(1-p)}/\sqrt{n}$.

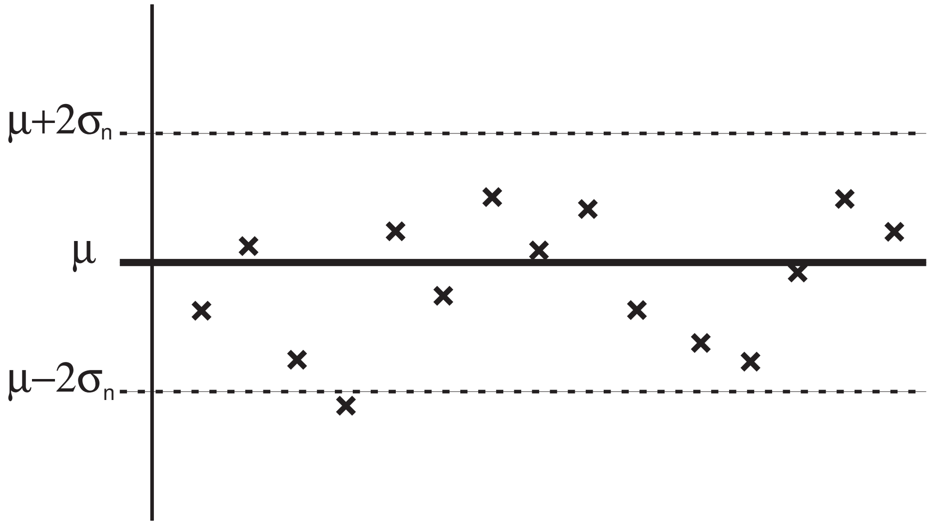
We can then use the 68-95-99.7 rule to tell us how often we expect the sample proportion $\hat{p}$ to fall within certain multiples of μ. In part, 95% of the time, $\hat{p}$ will be within $2\sqrt{p(1-p)}/\sqrt{n}$ of p.

But this, in turn, says that 95% of the time, p will be within $2\sigma_{\hat{p}} = 2\sqrt{p(1-p)}/\sqrt{n}$ of $\hat{p}$! Which is what we really want to know; it tells us how often our stastistic is close to the parameter we're trying to measure (and how close it would be). Unfortunately, in order to know what $2\sigma_{\hat{p}}$ is, we really need to know p (which we don't). But really, all we really want to know is that $2\sigma_{\hat{p}}$ is small, and what we can show is that no matter what p is, $2\sigma_{\hat{p}}$ is smaller than $1/\sqrt{n}$. Which tells us that if n is large (i.e., our sample is large), then most of the time our sample proportion $\hat{p}$ is close to the true proportion p.

We can say much the same thing for sample means: if a quantity varies, but its distribution is bell curve sahaped with mean μ and standard deviation σ, then if we take smaples of size n and average them, the distribution of these averages will also be a bell curve, with the same mean but <u>smaller</u> standard deviation; it will be $\sigma/\sqrt{n}$. So again we can use the 68-95-99.7 rule to describe how often the sample mean will be within certain values on either side of the mean.

One situation where these idea are put to use in practice is in *statistical process control*. Here the idea is to try to understand, by sampling the product coming off a production line or out of a machine, whether or not the machine is working properly. We basically think of some property of the objects coming off the line (their weight?) as being normally distributed, with known values for the mean μ and standard deviation σ. Any one item falling far from the mean <u>might</u> be due to simple chance; do we want to stop the process and inspect the machine because of it?

The idea is instead to use a measurement that we know should exhibit much less variation, namely the average value of some number n of the objects. What we do is build a *control chart*; each hour we randomly pick n (9, say) objects and measure them, and graph their average against the hour of the measurement.



If we draw in the lines at the mean and two standard deviations above and below, then we know that our average values should be more or less evenly distributed on either side of this line, and 95% of them should fall between the upper and lower lines. But since we have taken an average those upper and lower lines are much (in our case n=9, three times) closer to the mean than they would be with one sample. So even small changes in the average value can be significant.

In practice a halt is called and the maichine is inspected if one of our averages falls more than two (or if you want to be more sure the machine needs fixing, three) standard deviations away from the mean, or if eight averages in a row fall on the same side of the mean. Either of these indicates that it is very likely that the average object that the machine is really putting is different than the average value we designed the machine for. out is

One final observation; it is possible for completely correct statistics to lead us to make the wrong conclusion. For example, in a small town it is discovered that in a recent month the banks have approved 40% (18 of 45) of the loan request made by men, and 50% (9 0f 18) made by women. Does this represent discrimination by the banks? This information can be represented by a two-way table:

	applied	approved
Men	45	18
Women	18	9

	applied	approved
Men	5	3
Women	10	6

	applied	approved
Men	40	15
Women	8	3

In point of fact, however, there are two banks in town. Each actually grants the same fraction of loans to men and women. But since far more men applied at the bank that had the lower approval rate, when we add the results from the two banks together, we see an apparent difference between the two. This demonstrates that confounding variables (the difference in sizes of the two banks, in this example) can be very subtle, indeed!