MATH 1070 Introductory Statistics
Lecture notes
Sampling theory, Experiments, and Probability

Objectives:

1. Learn the basic concepts of sampling theory
2. Learn what bias is and why it is bad
3. Learn the difference between
   - voluntary response,
   - convenience,
   - systematic, and
   - simple random samples
4. Create a simple random sample with a random number table
5. Issues with sampling (undercoverage, nonresponse)
6. Learn the concepts of experiments, including
   - subjects, factors, and treatments,
   - double blind, and
   - placebo effect
7. Identify the different types of experiments
8. Learn the basic concepts and vocabulary of probability
9. Identify and use the rules of probability

Why do we sample?

To do statistics one needs data. We cant use the population because usually its too large. We cant get our hands on it all, at least not cost effectively. So we use a sample, a smaller subset of the population. But we cant use just any sample, we must choose correctly.

Choosing correctly means avoiding bias. Since we are making decisions about the population from our sample, if the sample is biased, or unrepresentative, then the conclusions based upon it are flawed.

Sometimes we can collect our own data. This is called **primary data**. We have ultimate control over it. But sometimes we cant afford the time or money to collect data so we rely on someone else, like the government. This is called **secondary data**. We have to trust that the data were collected in an unbiased manner.

It would be nice to have access to all the data we need to make a decision or a prediction. The hard fact is, though, we don't always have all the data. Sometimes we can't hope to have all the data if the data set is quite large (like the US population) or constantly arriving (like a manufacturing process). In that case we must use a smaller portion of the whole data and generalize back to the population at large. This smaller portion is called a **sample**.

There are several ways to select samples. The easiest are sometimes the most prone to bias. Creating a sample is relatively easy, but there are rules to be followed if we want a useful sample and a useful conclusion.

What are we doing?

The data we are sampling can either be from an **observational study** or from an **experiment**. An observational study does not attempt to influence the phenomenon being observed. It can be a survey which only collects opinions and data. An experiment attempts to influence the outcome of the phenomenon being studied, usually to determine which of several approaches is the best solution (if any).

Various sample designs

What we want to avoid is **bias**. Biased samples produced biased conclusions, and this does not serve the purpose of sampling and generalizing of conclusions. It can, but then the outcome was already known and the data collected were just being used to support a predefined conclusion. We want a truly representative sample of the population so we know what is true in general.

Samples must be designed and are not actually arbitrary. Have you noticed that the elections are often predicted with a rather small (1,500) number of registered voters? This isn't just asking the next 1,500 people that walk in the door.

That is, however, a method of sampling. **Voluntary reponse** is a method whereby the sample "self selects". People choose to respond or participate based on a general appeal. This is what happens on call-in radio and television programs. This is a bad method, if cheap and easy, because the participants usually are those people with strong opinions and very often negative opinions to the question at hand. Why do you think this is?

Another cheap and easy sampling method is the **convenience sample**. It's name tells you about it. In this method, the sample is composed of those people easiest to reach. This is probably not a very good representation of the population in question, unless the pollster has a very large group of friends and colleagues.

This can be useful, though, in a test situation. When developing a new survey tool it must be actually taken to identify errors and confusing statements. It would be best if the group used to identify these items is not the real group being surveyed. Surveys cost money to deploy and collect, so you want to make sure that data being returned is good. A convenience sample can help in the "beta testing" of a survey tool. College students (and co-workers) make excellent survey testers.

The easiest (good) sample design is a **simple random sample**. This is a sample where every member of the population had an equal chance at being selected. This is true no matter how many samples are selected from the population. Repeated samples should each be equally representative of the population.

Using random numbers to generate a sample

How do we generate a simple random sample? Randomly, of course. We user random numbers to select who might be asked to participate in the survey or study. Random numbers come from a variety of sources such as a roll of a die, a flip of a coin, or a random number generator on a computer. A nontechnological method is to use a table of random numbers, one of which appears in every statistics textbook.

So how do we use a table of random numbers? First we assign each observation in the dataset a number. This number can start at zero or at one, or anywhere else for that matter. The number of digits in the number should be padded with leading zeros so that all the assigned numbers have the same length. For example, if there are ten to ninety-nine observations the

numbers assigned should be $01, 02, 03, 04, \ldots, 98, 99$. If there are one-hundred to nine-hundred ninety-nine then the assigned numbers should be three digits. And so forth.

In this example, the total number of observations would be 50 which would include one observation for each state. The assigned numbers should then be two digits in length.

01	7263	16347	Alabama
02	12441	28395	Alaska
03	9185	19017	Arizona
04	6271	14641	Arkansas
05	10729	21537	California
06	9552	21279	Colorado
⋮	⋮	⋮	⋮

Now, choose a row in the table of random numbers. Since any row should give as good a sample as any other, the choice of row does not matter. Starting at the left end of the row, choose numbers of the same size as your assigned labels. In the above example, if our table row was

$$07511889154126716853845697936732337033 16$$

then our two digit random numbers would be

$$07, 51, 18, 89, 15, 41, 26, 71, 68, 53, \ldots$$

Notice that not all the numbers selected from the row are usable. We don't have 89 states so there isn't an observation with 89 as it's assigned number. We discard that number from the row and continue. The same action is taken when we encounter a number we've already used. When we get to the end of the row we continue on the next row until we have enough observations in our sample.

How many is enough? It depends on the size of the population you are trying to generalize about. There are calculations depending upon the statistical test you want to run that will help you decide the right size for your sample. It isn't by chance that we use a sample of 1,500 registered voters to predict the national elections.

So if we were to want a sample of five states from the fifty states and District of Columbia (a population of 51) we could use the SRS methodology as follows:

1. Assign each observation (state) a two digit number beginning at 01 and ending with 51. (See above for an example.)
2. Choose a row in the random number table.

00	15544	80712	97742	21500	97081	42451	50623	56071	28882	28739
01	01011	21285	04729	39986	73150	31548	30168	76189	56996	19210

3. Beginning at the first digit on line 00, begin selecting two digit numbers. When a gap intervenes, consider the two digits on either side as one two digit number. The numbers from the first row are

15	54	48	07	12	97	74	22	15	00
97	08	14	24	51	50	62	35	60	71
28	88	22	87	39					

4. Only those numbers which fall within the range 01 to 51, inclusive, are useful in selecting our sample. Those numbers are

15	54	**48**	**07**	**12**	97	74	**22**	**15**	00
97	**08**	**14**	**24**	**51**	**50**	62	35	60	71
28	88	**22**	87	**39**					

This line of the table appears to be rich in values that meet our needs so we don't need to use any further rows. The values that we would select for our random sample of five would be

$$15, 48, 07, 12, 22$$

Note that the next number in the sequence was another 15. Since we already have that number in our sample we would discard the duplicate and continue on if we still needed numbers. So our sample consists of

15	Indiana	20535
48	Washington	20018
07	Connecticut	23149
12	Hawaii	22750
22	Massachusetts	21166

Other sampling methods

There are two other sampling methods that are routinely used. **Systematic sampling** selects every n^{th} observation, such as "every third car" or "every fifth student." Sometimes the choice of n is based upon the population size divided by the desired sample size ($\frac{N}{n}$). Rounding this value to the nearest whole number, most likely the largest integer not greater than, or $floor$, will give you the choice of which observations to select.

The other method is **stratified random sampling**. Sometimes the population naturally separates into groups, called **strata**. A simple example of this would be gender. Any population of people can be divided by gender into two relatively homogeneous group. If there are discernible groups within the population it is advisable that samples be taken from each of the subgroups to insure that the whole sample is representative of the population.

Another example of this would be voters by county. Especially in Georgia, a majority of the population resides in the urban areas. In this case, predominantly in the Atlanta MSA (roughly 4,768,685 out of 9,072,576 people in Georgia, 52.5%). If you take a random sample of voters in Georgia the majority of them will naturally reside in the Atlanta MSA. This may not give a truly representative sample and may not give an accurate response to your question. It would be better to group the voters by county and then take simple random samples from each of the groups.

Experiments in statistics

If we aren't doing an observational study, then we are doing an experiment where we hope to affect the outcome. What is the effect of adding 10g of fertilizer to the water for bean plants? What is the effect of giving a 110mg does of an antiviral versus 130mg? To make sure we get useful data from our experiment we must design it correctly.

There are some vocabulary words we need to learn regarding experiments.

Experimental units The items studied in an experiment.

Subjects The individuals studied in an experiment when they are people.

Factor The explanatory (independent) variables in an experiment.

Treatment Any specific experimental condition applied to the experimental units or subjects. If there are several factors, a treatment is a combination of specific value of each factor.

An example from Moore, if we are interested in the optimal length (30 or 90 seconds) and frequency (1, 3, or 5 times) of commercials to achieve maximum retention and potential sales, we could design an experiment to test this. In this experiment we have two factors, length and frequency, and the factors have different values assigned to them. The factor length has two levels, 30 seconds and 90 seconds. The factor frequency has three levels, 1, 3, or 5 times. When we combine these two together we get a total of six possible treatments (levels of factors):

- 1 time, 30 seconds
- 3 times, 30 seconds
- 5 times, 30 seconds
- 1 time, 90 seconds
- 3 times, 90 seconds
- 5 times, 90 seconds

In an experiment we hope to affect the outcome or observe some change. Sometimes this change can be affected simple because our subjects think there should be some change. If you tell someone they should feel better because a person in a white coat gave them a pill, some of them will feel better (or think they do) because of that. To control for this behavior we use a fake or sham treatment called a **placebo**. The **placebo effect** is the term used to describe this behavior.

Ways to Prevent Bias in Experiments

Significant Results

The goal of any experiment is to obtain results that are solid and defensible. Since we are using a sample to give us an estimate of what the population would do, we want to make sure our sample and experiment are without bias. We also want results so solid that they couldn't have just happened. They're real. Results that are too large to have happened by chance are said to be **statistically significant**. This is a term that is reported with the results of many studies. This phrase says that these results are really "the way it is" and we didn't just pick a lucky sample.

Basic Concepts of Probability

Things happen. Can we make a reasonable estimation as to how and when?

Probability is **empirical**, that is, it is based on observation. The probability that an event will occur is the relative frequency with which that event can be expected to occur. **Relative frequency** is the observed occurrences of an event. If we assume that the event is random then individual outcomes are uncertain, but there is nonetheless a regular distribution of outcomes in a large number of repetitions.

But what gives this any validity? The fact is, chance behavior is unpredictable in the short run, but has a regular and predictable pattern in the long run. So we can call a phenomenon random if the individual outcomes are uncertain, but a regular distribution appears with a large number of repetitions.

The **probability** of any outcome of a random phenomenon is the proportion of times the outcome would occur in a very long series of repetitions. Our repetitions must be **independent**, however. What happens in one trial cannot have an effect on another.

Vocabulary of Probability

Experiment Any process that yields a result or an observation.

Outcome A particular result of an experiment.

Sample Space The set of all possible outcomes of an experiment $\{S\}$. Each element in the set is a sample point.

Event Any subset of the sample space. Can be one or more outcomes of the same experiment.

Probability Model A mathematical description of a random phenomenon consisting of a sample space (S) and a way of assigning probabilities to events.

Complement The opposite of an event, or, the set of all sample points in the sample space that do not belong to event A. We denote this as "A-bar" $\bar{A}$.

Note: Outcomes must not overlap (mutually exclusive) and there cannot be any exceptions (inclusive).

We can represent the sample space in several ways: tree diagram, set of ordered tuples, a list, a grid, etc.

The outcomes can be affected by whether the experiment is conducted **with replacement** or **without replacement**. If you select an item and return it so that it may be selected again that is **with replacement**. If you do not return the item such that it can only be selected once, that is **without replacement**.

In a sample space containing sample points that are equally likely to occur, the probability $P(A)$ of an event A is the ratio of the number $n(A)$ of points that satisfy the definition of even A to the number $n(S)$ of sample points in the entire sample space:

$$P(A) = \frac{(\text{number of times } A \text{occurs in sample space})}{(\text{number of elements in } S)}$$

or said another way

$$P(A) = \frac{(\text{number of ways } A \text{can occur})}{(\text{total number of outcomes in the sample space})}$$

Rules of Probability

There are some rules to Probability:

1. All probabilities must be between 0 and 1, inclusive.

$$0 \leq P(A) \leq 1 \text{ for any event } A$$

2. All probabilities must add up to 1.
3. The probability that an event does not occur is 1 minus the probability that the event does occur $(1 - P(A))$.
4. If events are mutually exclusive, the probability that one of them happens is the sum of their individual probabilities. (Also known as the **addition rule for disjoint events**.

Random Variables and Probability Distributions

The outcome of any trial (or experiment) can take on any of the possible values in the sample space. In an experiment where we flip two coins and record the results, we could have any of the following: HH, HT, TH, or TT. When we roll a single fair die the results could be any number between 1 and 6, inclusive. Since the values can change from any given trial to any other this makes the results a variable. And since we don't know what the value might be that makes it random. The "box" that holds this random value is called a **random variable**. The value in this box depends on chance and can be any of the values possible. These values could be either integer (whole) values or values such as "heads" or "tails". In these cases we call these values **discrete** and further classify the variables as discrete random variables. If the values could be more precisely defined to a more granular degree (written in real, or floating point, notation, with a decimal point) we call them continuous random variables. For our purposes we will concentrate on the discrete random variables (**d.r.v.**) for now. Some examples of d.r.v. are

- The number of cars through the GA 400 toll booth in one day.
- Passengers through Hartsfield-Jackson Airport during Thanksgiving.
- The score on the math portion of the SAT or ACT.

Each value in a random variable has a portion of the total probability assigned to it. We know the total has to be 1 when we sum all the probabilities and each share of the total probability is between zero and one. The assignment of probability to each value is referred to as the **probability distribution**. This can be represented as either a function with output values or as a table.

> it Probability distribution: A listing of the possible values and corresponding probabilities of a discrete random variable; or a formula for the probabilities. (Weiss, p 291)

To construct a probability distribution first identify the values for the discrete random variable. Then compute the probabilities for each of those values. Then construct a table with the values for the d.r.v. and the corresponding probability.

For example, if X is the d.r.v. in an experiment involving rolling one fair die, the possible values for the d.r.v. are 1,2,3,4,5, and 6. Since the die is fair we know each value is equiprobable (has equal probability) so we can simply divide the total probability by six to compute the

amount of probability to assign to each value ($\frac{1}{6}$). Constructing our table we get something like this:

roll of die X	1	2	3	4	5	6
probability	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{1}{6}$

Of course, the probability does not have to be evenly distributed. If we look at students in high school, grades 9 through 12, and treat the grade level as a random variable (X) then we could construct a probability distribution that looks like this

grade level X	9	10	11	12
probability	0.300	0.262	0.230	0.208

Each of these probabilities is between zero and one and the sum is .300+.262+.230+.208 = 1.0. This is a valid probability distribution.

Law of Large Numbers

Empirical, or experimental, probability is the observed relative frequency with which an event occurs. The value assigned to the probability of event A as a result of experimentation is:

$P\prime(A)$ = (number of times A occurred) / (number of trials)

We use $P\prime$ (P prime) to denote empirical probability.

The more experiments we run the larger the set of results we collect. The larger the set of results the closer to the "true" probability of occurrence. This leads us to the law of large numbers which states:

> As the number of times an experiment is repeated increases, the ratio of the number of successful occurrences to the number of trials will tend to approach the theoretical probability of the outcome for an individual trial.

That is, the larger the number of experimental trials n, the closer the empirical (observed) probability $P\prime(A)$ is expected to be to the true or theoretical probability $P(A)$.

Compound events and probability

We have talked about instances where an item must belong to one, and only one, group. These are called **mutually exclusive**. Events, outcomes of an experiment, can also be mutually exclusive, that is, an event can be defined in such a way that the occurrence of one event precludes the occurrence of any of the other events. (JK) These can also be called **disjoint events**. (DM)

There are three possible compound event outcomes:

1. Either event A or event B will occur, $P(A \text{ or } B)$.
2. Both event A and event B will occur, $P(A \text{ and } B)$.
3. Event A will occur given that event B has occurred, $P(A \mid B)$.

Each has a different rule applied to it for various reasons.

An example: Roll two dice and look at the sum of the numbers. Let there be three outcomes of interest:

A The sum is seven.

B The sum is ten.

C The two numbers are the same.

Clearly events **B** and **C** can happen at the same time since ten can be the sum of five and five. But **A** and **C** cannot happen together since seven is not evenly divisible by two. (JK)

We have special rules for computing the probability of mutually exclusive events. These are the **General Addition Rule** and the **Special Addition Rule**.

The **General Addition Rule** states:

> Let A and B be two events defined in a sample space S. Then
>
> $$P(A \text{ or } B) = P(A) + P(B)P(A \text{ and } B)$$

The **Special Addition Rule** states:

> Let A and B be two events defined in a sample space S. If A and B are mutually exclusive events, then
>
> $$P(A \text{ or } B) = P(A) + P(B)$$

Its simple, really. If two (or more) events are mutually exclusive then there cannot be overlap. Double counting cannot occur. We can then add the probabilities together to determine the probability.

Addition is not always the answer, though.

Consider the experiment of flipping a coin. Its been shown that the probability of either side of a coin resulting from a flip is one-half (0.5). If we toss a coin twice, the probability does not change because we have flipped the coin previously. The probability of the second toss is still 0.5. The probability of the same face twice in a row would then be 0.25, or 0.5×0.5. This is because the two events are independent.

Two events A and B are independent events is and only if the occurrence (or nonoccurrence) of one does not affect the probability assigned to the occurrence of the other. If this can be proven or reasonably assumed, then we can use the Special Multiplication Rule to calculate probability, which is

$$P(A \text{ and } B) = P(A)P(B)$$

If one event can affect the outcome of another, the events are dependent. And example would be event A being "sum of 10" on a roll of two dice and event B being "doubles." These two events are dependent in nature since there is only one outcome that satisfies both, (5,5). (JK)

Another example of dependence is the color of a card dealt from a standard deck. There are 52 cards, 26 each of black and red. If the first card dealt is red, then the probability of the next card being red has been changed from 26/52 (0.5) to 25/51 (0.49).

If we want to know the probability of an event occurring given that a previous event occurred, then we have conditional probability which we write as $P(A \mid B)$, the probability that A will occur given B.

Conditional probability is computed by taking the number of elements in the intersection of the two events (A and B) and dividing by the number of elements in the contingency (B). That is,

$$P(A \mid B) = P(A \text{ and } B)/P(B)$$

If there is dependence between the events, the computation of the events in sequence is determined by the General Multiplicative Rule, which states

$$P(A \text{ and } B) = P(A) \times P(B \mid A)$$