Lecture 6

Last time

We learned how to design experiments to test ideas like what temperature to bake a cake at or if a new medicine for arthritis works. The big theme in the experiments and in sampling is randomization. Why is randomization so important?

Randomness

In this section we are going back to the ideas of probability, statistical inference and independence.

How can probability quantify randomness?

There's an essential component that satisticians rely on to try to avoid bias in designing experiments and in sampling observational studies. This is **randomness** - randomly assigning subjects to treatments or randomly selecting people for a sample. As kids, we all employed randomization in games we played when we rolled dice, spun a wheel or flipped a coin. These tools make the game fair. These are just simple ways to represent the randomness that occurs with randomized experiments. For instance, the "head" or "tail" outcome of a fair coin flip can represent "drug" and "placebo" when a medical study randomly assigns a subject to receive one of two treatments.

For a *small* number of observations, outcomes of random phenomena may look quite different from what you expect. For instance, you may expect to see random pattern in the outcomes; but instead, exactly the same outcome may happen a few times in a row. That's not necessarily a surprise, as unpredictability for any given observation is the essence of randomness. We'll see, however, that with a *large* number of observations, our summary statistics (means, standard deviations,...) become surprisingly nonrandom. For instance, as we make more and more observations, the proportion of times that a particular outcome occurs gets closer and closer to a certain number. This long-run proportion provides the basis for the definition of probability.

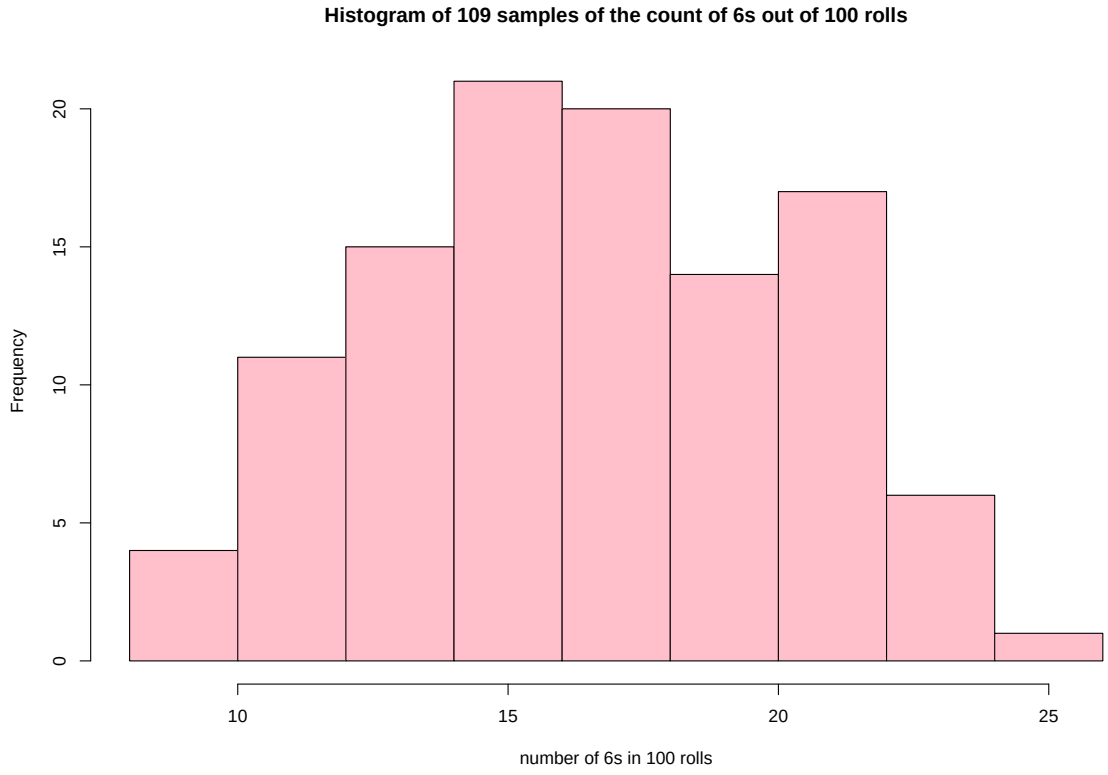
Example:

The board game you've been playing has a die that determines the number of spaces moved on the board. After you've rolled the die 100 times, the number 6 has appeared 23 times, more frequently than each of the other 5 numbers, 1 through 5. At one stage, it turns up three times in a row, resulting in you winning a game. Your opponent then complains that the die favors the number 6 and is not a fair die.

If a fair die is rolled 100 times, how many 6s do you expect?

Would it be unusual to get 23 6s out of 100 rolls? How can I prove this?

Well, we could each roll a die 100 times and count how many 6s we each get. Then we have 109 different counts of 6s out of 100 rolls. This is like a sample of 109 different possible 100 rolls of a die. We can then graph these counts on a histogram to see the results.



So is it so unusual?

Independent Trials

With random phenomena, many believe that when some outcome has not happened in a while, it is due to happen - meaning that its probability of happening goes up until it happens. In many rolls of a fair die, if a particular value (like 5) has not occurred in a long time, some think it's due and that the chance of a 5 on the next roll is greater than 1/6. If a family has four girls in a row and is expecting another child, are they due to get a boy? Does the next child have more than a 1/2 chance of being a boy?

With many random phenomena, such as the outcomes of rolling a die or having children, what happens on a previous trial does not affect the trial that's about to occur. The trials are **independent** of each other.

Samples & Populations

Example:

A major marketing firm decides to randomly sample 100,000 individuals across the country and ask their ages. They determine that the mean age for those 100,000 people is 42.3. They will use this figure to estimate the mean age of all U.S. citizens. This information helps them target their advertising. Hypothetically speaking, we happen to know that the true mean age of all U.S. citizens is 45.6. This number is practically impossible to determine.

In this example, which number would we call the true mean age? Does it change?

Which number would we call the sample mean age? Can the sample mean change?

How do we distinguish between the two?

A **parameter** is a number used to describe the entire population which is rarely known because of the difficulties encountered in examining the entire population.

Example:

the mean of a population = μ

the standard deviation of a population = σ

A **statistic** is a number computed from a sample (taken from the population) without using any unknown parameters. It's a function only of the data collected.

Example:

the mean of a sample = $\bar{X}$

the standard deviation of a sample = s

Distributions & Samples & Populations

We use statistics like $\bar{X}$ in order to infer something about the parameter μ. We call $\bar{X}$ a **random variable** since it changes with each new random sample chosen.

Things to notice about how $\bar{X}$ relates to :

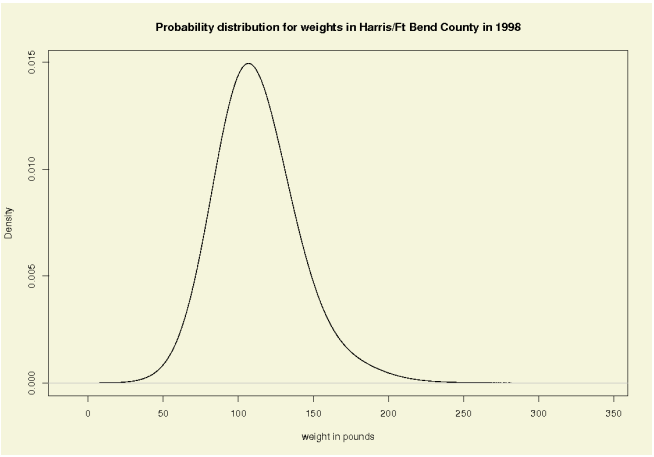
$\bar{X}$ is the statistic that estimates μ

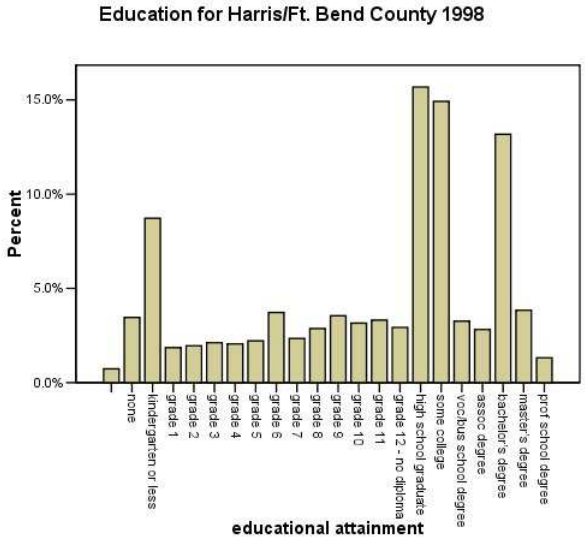
μ, in practice, is fixed but rarely known

If a different SRS is taken, we'd have a slightly different $\bar{X}$. $\bar{X}$varies from sample to sample

The larger the sample size, the closer $\bar{X}$ is to μ

Recall that density curves and distributions show the frequency or proportion of an observation from a sample.





Another way to look at a **(probability) distribution** of a random variable, X, is to show the values of X and their probabilities:

X value	**Probability**
x_1	p_1
x_2	p_2
x_3	p_3
.	.
.	.
.	.
x_k	p_k

Note that these probabilities must satisfy two requirements:
1. all probabilities, p_i, must be between 0 and 1
2. all the probabilities add up to one: $p_1 + p_2 + p_3 + ... + p_k = 1$

There will always be a distribution for both your population and your sample.

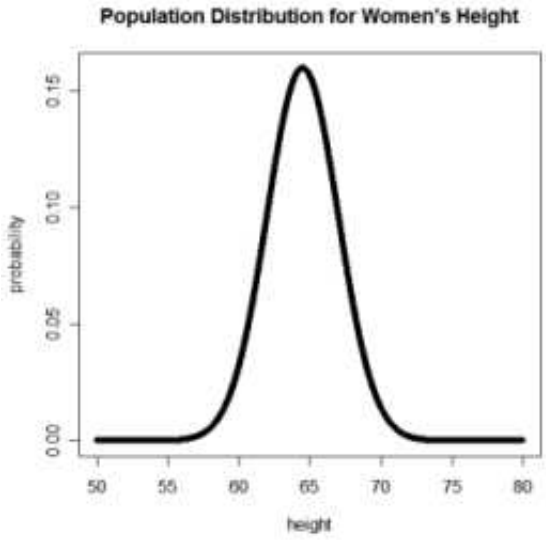
A **population distribution** is the probability distribution of a variable for all members of the population

A **sampling distribution** is the probability distribution of values of a variable in all possible samples of the same size from the same population.

Our goal is to get our sampling distribution to be as close as possible to the population distribution.

Example:

The heights of women between the ages of 18 to 24 is approximately normally distributed with mean 64.5 inches and standard deviation 2.5 inches. We call N(64.5, 2.52) the population distribution because we are talking about all women, not just a sample.



Given this information about the population distribution, computer software can do simulations of sampling by taking many samples of size ten and calculating $\bar{X}$=average height of the sample for each of them. (It would not be practical to do this by hand.) See below is a histogram of these means of 1000 samples of size ten. This is the sampling distribution.