Hypothesis Testing

Estimation

We use sample statistics to estimate population parameters. We want our sample estimates to be good in the sense that the sample numbers are close to the population values.

Q1: is the sample statistic equal to the population parameter on average?

If the answer to this question is "yes," the statistic is said to be an **unbiased** estimator of the parameter. If the answer is "no," the statistic is a **biased** estimator of the parameter.

Q2: what is the average distance of the statistic from the parameter? If the statistic is unbiased, this is the standard deviation of the statistic, also known as the **standard error**. Other things being, equal, we want the average statistic to equal the parameter, and we want the average distance from the parameter to the statistic to be small.

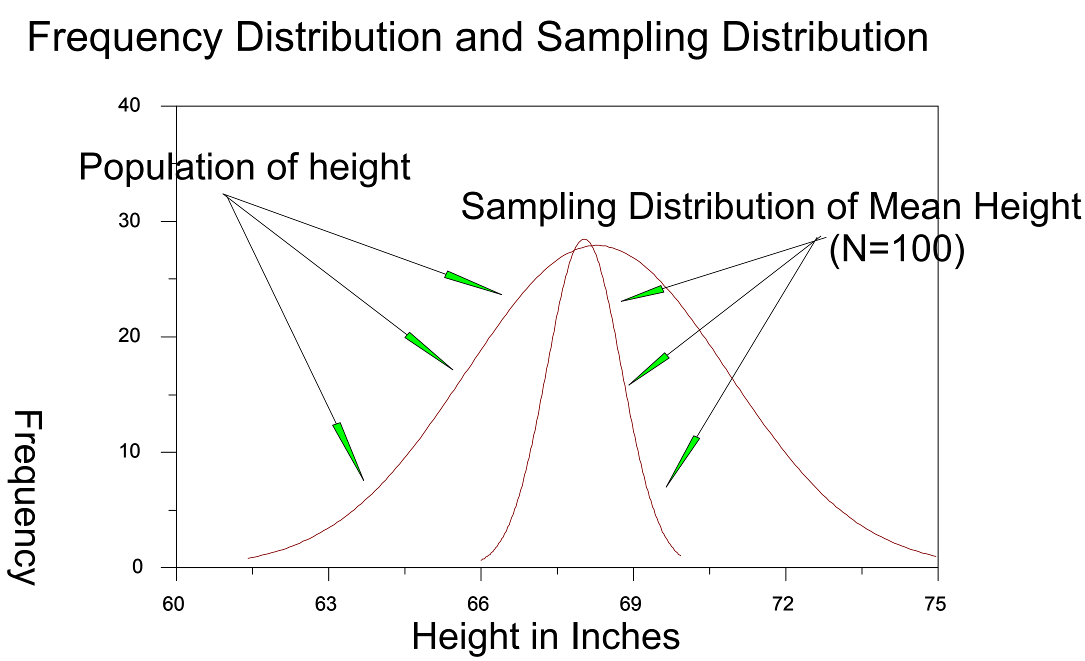
Sampling distribution

The sampling distribution is what we get when we do the following:

1. Take a sample of size N (a given number) from a population (with replacement).
2. Compute the statistic (parameter estimate, $\overline{X}$, *s, s^2, r*) and record it.
3. Repeat steps 1 and 2 a lot (infinitely).
4. The resulting distribution, that is, the distribution of the statistic that comes from the repeated samples is called a *sampling distribution*.

Estimating the Mean

Example: estimating mean height of USF students.



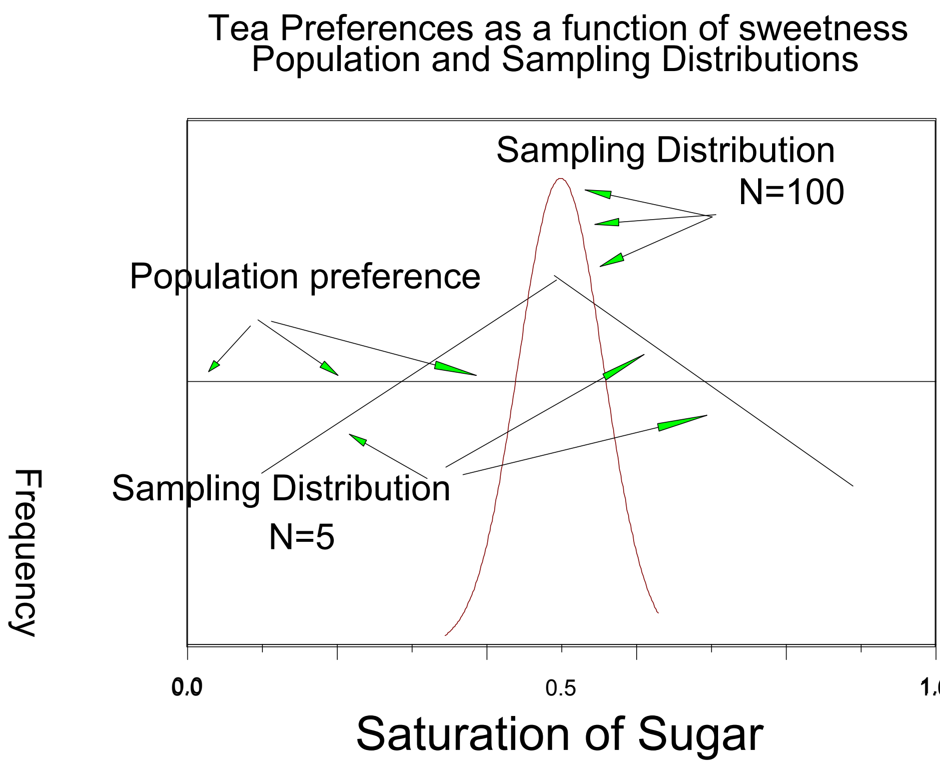
Points to notice:

1. The mean of the sampling distribution is close to the mean of the population.
2. The standard deviation of the sampling distribution is much smaller than the standard deviation of the population. The relation between the two can be expressed:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{N}}$$

where $\sigma_{\overline{X}}$ is the standard deviation of the sampling distribution of the mean, σ is the standard deviation of the population, and N is the sample size used to estimate the mean. This says that the standard deviation of the sampling distribution gets smaller as the sample size gets larger, by a factor of the square root of N. $\sigma_{\overline{X}}$ is called the standard error of the mean.

Example 2. Preferences for sweetness of tea.



Points to notice:

1. The sampling distribution means are equal to the population mean (all = .5).
2. The sampling distribution for N =100 has a smaller variance than the sampling distribution for N=5 (by a factor related to N).
3. The sampling distribution for N=100 has a more Normal shape than the distribution for N=5.
4. The shape of each of the distributions differs from each of the others. As the sample size increases, the sampling distribution of means becomes normal.

The sample mean is an unbiased estimator of the population mean.

Estimating the Variance and Standard Deviation

The sample variance and standard deviations are *biased estimators* of their population values. We can make the estimates of the variance and standard deviation unbiased by changing the denominator from N to N-1, like this:

$$\hat{\sigma}^2 = \frac{\sum (X - \overline{X})^2}{N-1}$$

Where $\hat{\sigma}^2$ is the estimated variance, and the rest of the symbols are as previously defined.

For the standard deviation, we just take the square root:

$$\hat{\sigma} = \sqrt{\frac{\sum (X - \overline{X})^2}{N-1}}$$

The *z* score and the *t* score

As the sample size (N) increases, the sampling distribution of means becomes Normal. The sampling distribution with large N has a *z* distribution if we subtract $\overline{X}$ and divide by $\sigma_{\overline{X}}$. This is cool because it lets us conduct statistical tests. For example, we can test whether Pepsi tastes better than Coke. With smaller N (which is usually the case), the distribution of our sample statistics instead follows the *t* distribution. The function of *t* is to boost the size of the *z* score for error in *x-bar* and *s* due to small sample size. As the sample size increases, *t* and *z* become closer and closer until they become the same number. There is little difference between *z* and *t* for N greater than 100.

Confidence Intervals

We use confidence intervals to bracket an estimate. Opinion polls are usually reported this way, e.g., 40 percent in favor plus or minus 2 percent. Because of the nature of the sampling distribution, we can choose the size of the confidence interval so that we are correct about the population mean a given percentage of the time. For example, we could choose the size of the interval so that we would be correct 95 percent of the time. If we choose the interval this way, we would be constructing a *95 percent confidence interval.* The actual calculation of a confidence interval is found by:

$$95\%\, CI = \overline{X} \pm t_{.05} s_{\overline{X}}$$

where $\overline{X}$ is the sample mean (estimate of the population mean), $t_{.05}$ is the value of the *t* distribution that cuts of 95 percent of the values of the distribution (like a *z* of 1.96) and $s_{\overline{X}}$ is our estimate of the standard error of $\overline{X}$ (the estimate of the standard deviation of the sampling distribution of the mean). This is just like the *z* score computations. The value of t such that it cuts off a certain percentage of scores (e.g., $t_{.05}$, $t_{.01}$) is called the critical value.

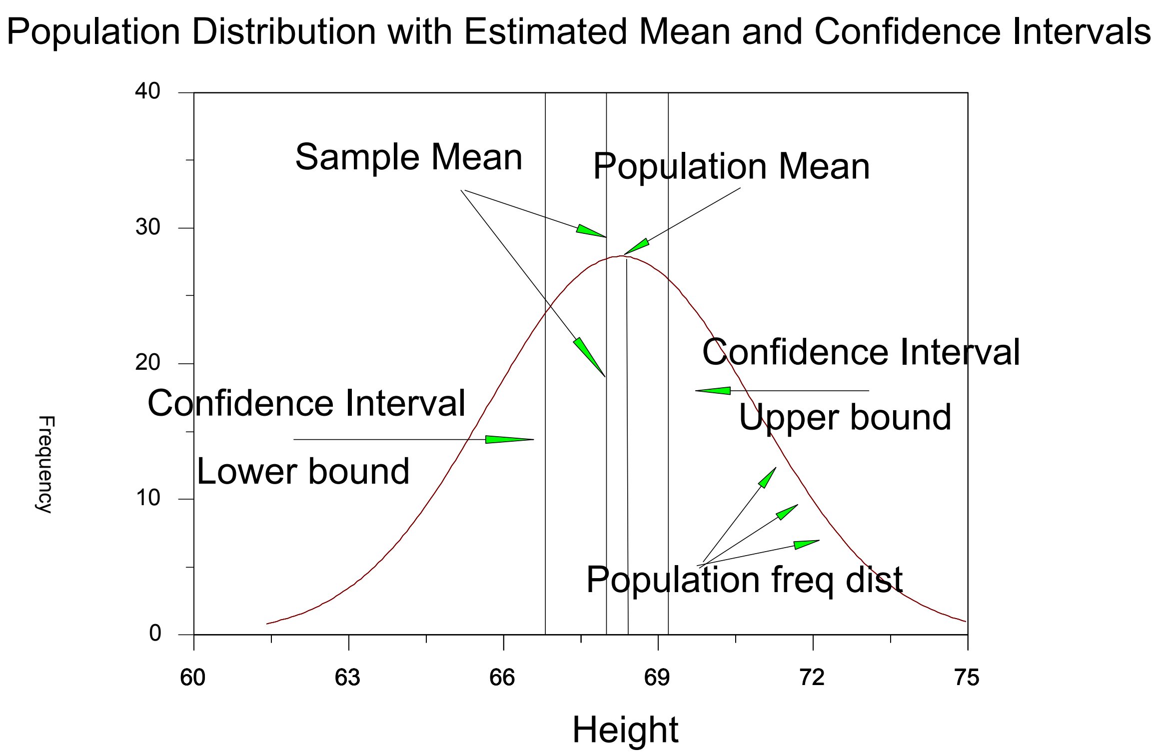
Example: let's say we sampled N=100 students. Their mean height was 68 inches and their standard deviation was 6 inches. Then we can construct a confidence interval like this:

95 % CI = 68±(1.98)[6/sqrt(100)]

=68±1.98*.6

=68±1.19 or the interval {66.81 to 69.19}

Lets say I know the actual population distribution of height and pick a sample of N=100 at random to get the above numbers. To illustrate the confidence interval graphically:



Points to notice:
1. The sample mean is close to the population mean, but does not equal the population mean.
2. The confidence interval is constructed about the sample mean (plus and minus 1.19 inches).
3. The confidence interval contains the population mean!!! This will be true 95 percent of the time we construct a 95 percent confidence interval, so we can be confident the mean is in there even when we don't know the actual population value as we do in this example.

Hypothesis Testing as a Decision Aid

The statistical approach to hypothesis testing is really a way of making decisions under uncertainty.

Decisions Right & Wrong.

DM jargon	State of the World

Our decision (study outcome)	Means different	Means same
Different	Correct rejection (of the null hypothesis)	Mistake (Type I error, alpha, α)
Same	Mistake (Type II error, beta, β) Power = $1-\beta$	Correct decision, but usually unwanted

Degrees of Freedom (*df*)

1. Degrees of freedom are the number of numbers in a distribution that are free to vary after a constraint is imposed.
2. Degrees of freedom are used up to estimate things.

In fact, *df* are not easy to figure, but they are always a simple function of N, like N-1 or N-2 or (N-1)*(N-2). We will only be doing a few tests but you must know **(memorize)** the proper *df* for each. There are only a few different numbers of *df*, so this is easy to do (*r*, *t*, and *F* are the main ones).