Sampling Distribution of the Mean

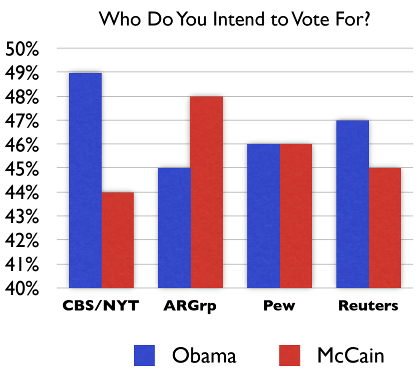
Dr Tom Ilvento
Department of Food and Resource Economics

Overview

- This is one of the most important topics in the course
- Sampling distributions are key to the logic of inference
- It is based on the notion of taking many, many samples….
- And making an estimate from each sample
- In reality we only take one sample, but we will think of it as one of many possible
- And our estimate as one of many estimates that follow a certain distribution……

Who Do you intend to vote for in the 2008 Presidential Election?

- All polls were taken in mid-September, 2008
- We expect variability from sample to sample
- we call it **sampling error**



Now we start toward inference

- Remember we noted that
- A **parameter** is a numerical descriptive measure of the population
  - We use Greek terms to represent it
  - It is hardly ever known
- A **sample statistic** is a numerical descriptive measure from a sample
  - Based on the observations in the sample
  - We want the sample to be derived from a **random process**

Help me do a little experiment

- Toss a die three times
  - Each time we toss the die three times we note and record the faces
  - Then calculate mean and median
- We can do this a number of times

Results of some trials of the experiment

Sample	Result	Mean	Median
1	5 4 1	3.33	4
2	4 4 3	3.67	4
3	5 5 2	4.00	5
4	6 1 1	3.67	1
5	6 4 2	4.00	4
6	3 3 2	2.67	3

A Priori we have the following expectation and variance

x	1	2	3	4	5	6
P(X)	0.167	0.167	0.167	0.167	0.167	0.167

- $E(x) = \mu = 1(.1667) + 2(.1667) + 3(.1667) + 4(.1667) + 5(.1667) + 6(.1667)$
  - **$E(x) = \mu = 3.500$**
- $E(x-\mu)^2 = \sigma^2 = (1-3.5)^2(.1667) + (2-3.5)^2(.1667) + (3- 3.5)^2(.1667) + (4-3.5)^2(.1667) + (5-3.5)^2(.1667) + (6-3.5)^2(.1667)$
  - **$E(x-\mu)^2 = \sigma^2 = 2.916667$**
  - **$\sigma = 1.7078$**

I used the following as a sample

- There are 6*6*6 = 216 different combinations of outcomes of rolling three die
- If I take the mean of each possible outcome
- And take the summary statistics (including the mean of the means)
- I get the following table (from Excel)

Sampling Distribution of the Sample Mean for rolling 3 die

- Notice the **mean of “the sample means” is 3.5**, the population value
- But the standard deviation is something less than the 1.7078 we expected
- But try this - divide 1.7078 by the SQRT(3) - the sample size (n)
- This value is **.986** - which is essentially the standard deviation of this variable
- And the distribution looks very much like a normal distribution

Comparing the Mean vs the Median for this Sampling Distribution

	Mean	*Median*
Mean	**3.50**	**3.50**
Standard Error	0.07	0.09
Median	**3.50**	**3.50**
Mode	3.33	3.00
Standard Deviation	**0.99**	**1.37**
Sample Variance	**0.98**	**1.89**
Kurtosis	-0.40	-0.81
Skewness	0.00	0.00
Range	5	5
Minimum	1	1
Maximum	6	6
Sum	756	756
Count	216	216

- I also calculated the statistics for the median of each sample
- If I compare the mean to the median, **the standard deviation for the mean is smaller**
- The mean as a measure of this sampling distribution has **minimum variance**

Sampling Distribution

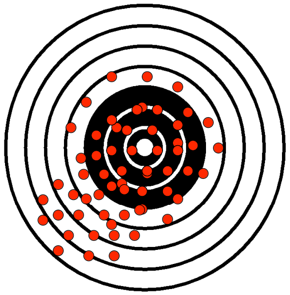
- Sample statistics are **random variables** - they vary from sample to sample
- We can look at their probability distribution based on repeating the sampling experiment many times - we will get different sample statistics each time
- If we do repeat the experiment many times results in a **sampling distribution**
- The sampling distribution of a sample statistic calculated from many samples of n measurements results in the **probability distribution of the statistic**.

What do we want from an estimator of the population parameter?

- If the sample statistic is a good estimator of μ
- We would expect the values of the sample means to cluster around μ
- We wouldn’t want to cluster to be at a point above or below μ (**not be biased**)
- And we might say our estimator is “good” if the cluster of the sample means around is tighter than the sampling distribution of some other possible estimator (**minimum variance**)

Think our our estimator in terms of a Bulls-Eye Target

- We want our estimates to center around the true value
- A tighter fit around the target – Minimum Variance – better and preferred
- A biased estimator may have a tight fit, but consistently misses the target in a discernible way
- This is the kind of pattern we would like to see – a tight fit around the population parameter



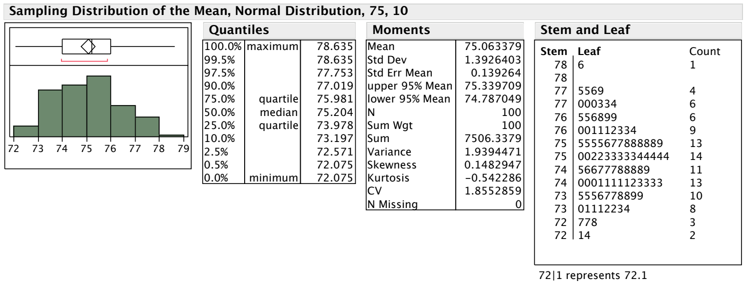
I did a simulation to help construct a sampling distribution for the mean

- We took 100 random samples of size **n = 50**
- From a population of **X ~ N (75, 10)**
- The "mean of the means" should equal the population parameter, **μ = 75**.
- The standard deviation of this new distribution should be related to **σ**, but we might expect it to be smaller

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad = 10/(50)^{.5} = 1.41$$

The standard deviation of this sampling distribution is called a standard error

Results of the Experiment based on repeated samples from a population with a mean of 75 and a standard deviation of 10 – 100 different random samples



Sampling Distribution of the Mean, Normal Distribution, 75, 10

Quantiles

100.0%	maximum	78.635
99.5%		78.635
97.5%		77.753
90.0%		77.019
75.0%	quartile	75.981
50.0%	median	75.204
25.0%	quartile	73.978
10.0%		73.197
2.5%		72.571
0.5%		72.075
0.0%	minimum	72.075

Moments

Mean	75.063379
Std Dev	1.3926403
Std Err Mean	0.139264
upper 95% Mean	75.339709
lower 95% Mean	74.787049
N	100
Sum Wgt	100
Sum	7506.3379
Variance	1.9394471
Skewness	0.1482947
Kurtosis	-0.542286
CV	1.8552859
N Missing	0

Stem and Leaf

Stem	Leaf	Count
78	6	1
78		
77	5569	4
77	000334	6
76	556899	6
76	001112334	9
75	5555677888889	13
75	00223333344444	14
74	56677788889	11
74	0001111123333	13
73	5556778899	10
73	01112234	8
72	778	3
72	14	2

72|1 represents 72.1

- The exercise resulted in a Mean of 75.06, very close to 75.0
- Std Dev = 1.39
- This is very close to the expected Std Error of $10/(50)^{.5}$ = 1.41

Sampling theory and sampling distributions help make inferences to a population

- Let's use the example of the mean to set up our discussion of a sampling distribution.
  - Suppose we are looking at a variable, e.g., average purchases of customers at an Internet sales company.
  - We think of the population (let's use the Population of adult customers of our Internet company in 2008).
- We believe there is an average purchases of this population, designated as μ.
- We want to take a random sample to estimate μ.

Inferences from a sample

- Our sample estimator of the population mean is:

$$\overline{x} = \frac{\sum x}{n}$$

- The variance of our sample estimate is given as:

$$s^2 = \frac{\sum (x - \overline{X})^2}{n - 1}$$

  - where n is equal to the sample size
  - s^2 is a unbiased estimator of population variance σ^2

Inferences from a sample

- The standard deviation represents the average deviation around the sample mean.
- But we only took one sample out of an infinite number of possible samples.
- A reasonable question would be what is the deviation around our estimator (i.e., the sample mean).
- That is, what if we took a whole bunch of samples, and recorded the mean of each sample – what it would look like?

Inferences from a sample

- If we could take an infinite number of samples, each sample would most likely yield a different sample mean.
- Yet, each one would be expressed as a reasonable estimate of the true population mean.
- So, if we were able to take repeated samples, each of sample size n, what would be the standard deviation of the sample estimates?
- Sampling theory specifies the variance of the sampling distribution of a mean as:

$$Var(\overline{x}) = Var\left(\frac{\sum \overline{x}}{n}\right) = \frac{\sigma^2}{n}$$

The square root is called the **Standard Error** of the mean

Inferences from a sample

- The standard error of the mean is the standard deviation of a sampling distribution of means from samples taken from a population with parameters equal to **μ** and **σ^2**.
- If we don't know **σ^2** we use the unbiased sample estimate of **s^2** to estimate the sampling variance of the mean.

$$S_{\overline{x}} = \frac{S}{\sqrt{n}}$$

This is the strategy

- We use the **theoretical sampling distribution** to make inferences from our sample to the population.
- The sampling distribution of an estimator is based on repeated samples of size n.
  - We may never actually take repeated samples
  - But we could think of this happening
  - And we think of our observed sample as one of many possible samples, of size n, we could have drawn from the population

This is the strategy

- We expect that the standard deviation of sampling distribution of the estimator (in this case the mean) will be smaller than that of the population or the samples themselves.
- We expect some variability across samples, but not as much as we would find in the population.
- **Thus the sampling error is smaller than the standard deviation for the population.**

The Standard Error

- The **Standard Error** depends upon:
  - The size of n (as n gets larger the SE gets smaller)
  - The variance of the population variable itself. We can think of this as homogeneity.
- The larger the sample size, and the more homogeneous the population, the smaller the standard error is for our estimator.

Properties of the Sampling Distribution for the Mean

- If a random sample of size **n** is drawn from a population with mean **μ** and standard deviation **σ**, then:
- The sampling distribution of the means has a mean equal to the population mean **μ**.
- And the sampling distribution of the mean has a standard deviation equal to the standard deviation of the population standard deviation **σ**, divided by the square root of the sample size **n**.
- And if we don't know **σ**, we use the sample standard deviation, **s.**

$$s_{\bar{x}} = \frac{s}{\sqrt{n}}$$

We use two theorems to help us make inferences

- In the case of the mean, we use two theorems concerning the normal distribution that help us make inferences
- One depends upon the variable being normally distributed
- The other does not - **Central Limit Theorem**

Summary

- We are on our journey toward inference
- A key part of this journey is understanding a sampling distribution
  - The sampling distribution is the theoretical distribution of making an estimate of a parameter from many, many samples
  - The sampling distribution of the mean (and proportion), will follow a normal distribution (under certain circumstances)
- Knowing the form of the sampling distribution will help in inference through confidence intervals and hypothesis tests