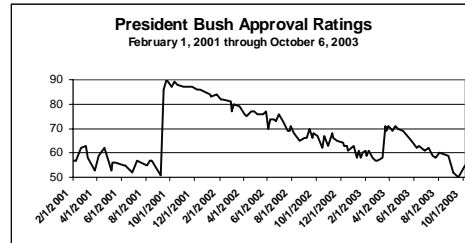


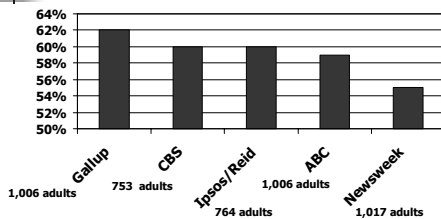
Sampling Distribution for the Mean

Dr Tom Ilvento
FREC 408

How is the President Doing?

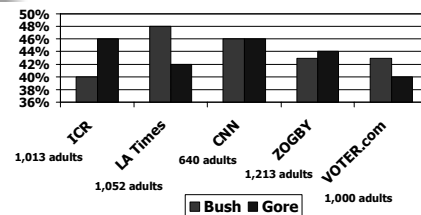


Bush Approval Ratings, Week of July 7, 2003



We expect variability from sample to sample – we call it sampling error

Presidential Poll Results Week of September 20-27, 2000



We expect variability from sample to sample – we call it sampling error (Def1.19 p48)

Now we move toward inference

- Remember we noted that
 - A **parameter** is a numerical descriptive measure of the population (Def3.15 p178)
 - We use Greek terms to represent it
 - It is hardly ever known
 - A **sample statistic** is a numerical descriptive measure from a sample (Def6.4 p311)
 - Based on the observations in the sample
 - We want the sample to be derived from a **random** process

Let's set up a small 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

A Priori we have the following expectation

X	1	2	3	4	5	6
P(x)	.167	.167	.167	.167	.167	.167

$$E(x) = 1(.1667) + 2(.1667) + 3(.1667) + 4(.1667) + 5(.1667) + 6(.1667)$$

$$E(x) = 3.500$$

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

$$E(x-\mu)^2 = 2.916667$$

$$\sigma = 1.7078$$

As an experiment

- Roll 3 times **5 4 1**
- Roll 3 times **4 4 3**
- Roll 3 times **5 5 2**
- Roll 3 times **6 1 1**
- Roll 3 times **6 4 2**
- Roll 3 times **3 3 2**

	Mean	Median
Roll 3 times 5 4 1	3.33	4
Roll 3 times 4 4 3	3.67	4
Roll 3 times 5 5 2	4.00	5
Roll 3 times 6 1 1	2.67	1
Roll 3 times 6 4 2	4.00	4
Roll 3 times 3 3 2	2.67	3

Results of 217 samples of size 3

	Mean	Median
Mean	3.47	3.47
Standard Error	0.07	0.09
Median	3.33	3.00
Mode	2.67	3.00
Standard Deviation	0.99	1.37
Sample Variance	0.98	1.88
Kurtosis	-0.41	-0.80
Skewness	0.01	0.07
Range	4.67	5.00
Minimum	1.00	1.00
Maximum	5.67	6.00
Sum	752	752
Count	217	217

Note: I worked out all possible outcomes

- 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 Sample Mean \bar{X} for rolling 3 die

Descriptives	
Mean	3.50
Standard Error	0.07
Median	3.50
Mode	3.33
Standard Deviation	0.99
Sample Variance	0.98
Kurtosis	-0.40
Skewness	0.00
Range	5.00
Minimum	1.00
Maximum	6.00
Sum	756
Count	216

We said that the standard deviation for rolling a die was: $\sigma = 1.7078$

Divide this figure by the Square root of 3 (sample Size),

Sampling Distribution of Sample Mean \bar{X} for rolling 3 die

Descriptives	
Mean	3.50
Standard Error	0.07
Median	3.50
Mode	3.33
Standard Deviation	0.99
Sample Variance	0.98
Kurtosis	-0.40
Skewness	0.00
Range	5.00
Minimum	1.00
Maximum	6.00
Sum	756
Count	216

We said that the standard deviation for rolling a die was: $\sigma = 1.7078$

Divide this figure by the Square root of 3 (sample Size), we get the Standard Deviation of \bar{X} , which is 0.99

This is also called the Standard Error of \bar{X} (p317)

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., blood pressure.
 - We think of the population (let's use the Population of adult males in Delaware age 18 to 85 in 2002).
- We believe there is an average blood pressure of this population, designated as μ . We want to take a sample to estimate μ .

Inferences from a sample

- Our **sample estimator** to population mean is:

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

- The **variance of our sample estimate** is given as:

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

- 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).
- I.e., what if we took a whole bunch of samples, and recorded the mean of each sample

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?

Inferences from a sample

- Sampling theory specifies the variance of the sampling distribution of a mean as:

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

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad \text{This is called the Standard Error of the mean (p317)}$$

Inferences from a sample

- The **standard error** of the mean is the standard deviation of a sampling distribution of means with population 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.

Here's Our 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 our observed sample as one of many possible samples, of size n , we could have drawn from the population

Here's our 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 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 for our estimator.

Properties of the Sampling Distribution of \bar{x} (p317)

- If \bar{x} is the mean of a random sample of size n from a population with mean μ and standard deviation σ , then:
 - The sampling distribution of \bar{x} has a mean equal to the population mean μ .
 - If we use $\mu_{\bar{x}}$ denote the mean of \bar{x} , then
 - $\mu_{\bar{x}} = \mu$

Properties of the Sampling Distribution of \bar{x} (cont.)

- And the sampling distribution of \bar{x} 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 . If we use $\sigma_{\bar{x}}$ denote the standard deviation of \bar{x} , then:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Properties of the Sampling Distribution of \bar{x} (cont.)

- And the sampling distribution of \bar{x} 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 . If we use don't know σ , then we use the sample standard deviation to estimate it

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

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**

For variables that are distributed normally

- If repeated samples of a variable Y of size **n** are drawn from a normal distribution, with mean μ and variance σ^2 ,
- the sampling distribution will be normal with mean μ and variance σ^2/n .

For variables that are distributed normally

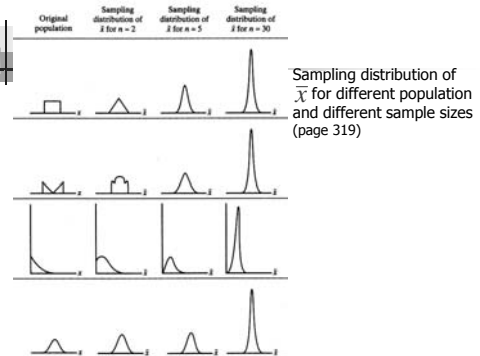
- What this means:
 - If we could repeatedly take random samples of size **n** from a normal distribution.
 - And then take the mean of each sample
 - We would expect the mean of the sample means to equal μ
 - And the variance of the sample means would equal σ^2/n

Central Limit Theorem

- If repeated sample of Y of size **n** are drawn from **any population** (regardless of its distribution as normal or otherwise) having a mean μ and variance σ^2 ,
- the sampling distribution of the sample means approaches normality, with μ and variance σ^2/n .
- As long as the sample size is **sufficiently large** (p317)

Central Limit Theorem

- The **Central Limit Theorem** is a very powerful theorem.
- It relaxes the assumption of the distribution of the population variable
- Note: this is based on the notion that our samples are drawn on a probability basis. That is, each element of the population has an equal chance of being selected



Inferences from a Sample

Comparison of the Characteristics of the Population, Sample, and the Sampling Distribution for the Mean

	Population	Our Sample	Sampling Distribution
Referred to as	Parameters	Sample Statistics	Statistics
How it is viewed	Assumed real	Observed	Theoretical
Mean	$\mu = \sum X/N$	$\bar{x} = \sum x/n$	$\mu = \sum \bar{x}/N$
Variance	$\sigma^2 = \sum (X - \mu)^2/N$	$s^2 = \sum (x - \bar{x})^2/(n-1)$	$\sigma_{\bar{x}}^2 = \sigma^2/n$
Standard Deviation	σ	$s = (s^2)^{.5}$	$\sigma_{\bar{x}} = \sigma/\sqrt{n}$

Note: N and X are for the Population, n and x are for the sample

So how do we use this information?

- We draw a random sample
- We think of our sample as one of many possible samples of size n from a population with parameters μ and σ .
- **If the variable is distributed normally**, we can use information about the sampling distribution of the mean to make inferences from the sample to the population.
- **Even if the variable is not distributed normally**, if our sample size (n) is large enough, we can assume the sampling distribution of sample mean \bar{x} is distributed normally (**Central Limit Theorem** p317)

Rare Event Approach

- Most inferences will be made using a rare event approach
- We will take a sample
- And compare it to a hypothesized population
- And see how close or far away our sample estimate is from the sampling distribution

Automobile Batteries

- The manufacturer claims that life of his automobile batteries is 54 months on average, with a standard deviation of 6 months.
- **We are involved in a consumer group and we decide to take a sample of 100 batteries and test the claim.**
 - We select **100 batteries at random**
 - Test them over time and record the length of the battery life

Our Sample Data

- The mean battery life for our sample is:
 - Mean = 52 months
 - Std Dev = 4.5 months
- So what do we do next?
- Our batteries didn't last as long on average as the manufacturer said, but it is just a sample.
- How can we test to see if the claim is bogus?

Our Testing Strategy

- If the world works as the manufacturer says
- And I would have taken repeated samples of size 100
- The **sampling distribution** would be a normal distribution
- And have a mean equal to the population mean for battery life, i.e., 54 months
- And a standard deviation of σ divided by the square root of n

$$\sigma_{\bar{x}} = \frac{6}{\sqrt{100}} = .60$$

Our Testing Strategy

- We want to look at our sample as being part of the theoretical Sampling Distribution. That is,
 - $\bar{x} \sim N(\mu_{\bar{x}}, \sigma_{\bar{x}})$
 - In this case, $\bar{x} \sim N(54, 0.6)$
- And see how likely it is that our sample came from that distribution
- **In other words, how likely is it to get a sample mean of 52 from a random sample of 100 batteries when the true population mean is 54 months**

How do I do this?

- I hypothesize that the true mean is 54
- I calculate a z-score based on my sample value (52.0) and the hypothesized mean and standard error (of the sampling distribution)
- I look up the probability of finding a z-score equal to or less than our calculated value

The Test

- Calculate my z-score

$$z = \frac{(52 - 54)}{.6} = -3.33$$

- Look up the value in the standard normal table

Draw it out!

$z = -3.33$ corresponds to a probability of .4990 up to that point

And $p = .0010$ after that point



This is really a rare event given the claim of the manufacturer – that the batteries really last 54 months on average

What is our sample size was 30?

- The standard error of the sampling distribution would change

$$\sigma_{\bar{x}} = \frac{6}{\sqrt{30}} = 1.0954$$

$$z = \frac{(52 - 54)}{1.0954} = -1.83$$

Draw it out!

$z = -1.83$ corresponds to a probability of .4664 up to that point

And $p = .0336$ after that point



While this is still a relatively rare event, I'm not as confident that I can refute the claim of the manufacturer

Example Problem

- The average live weight of a farmer's steers prior to slaughter in past years was 380 pounds. This year, 50 randomly chosen steers were fed on a new diet. Test to see if the mean weight for the sample on the new diet is larger than 380.
- Sample statistics
 - $n=50$
 - Mean = 390
 - $s = 35.2$
 - **NOTE: s is an unbiased estimator of σ . We use s when σ is unknown**

Strategy for the Problem

- Pretend that our sample really does come from a population with a mean = 380
- Compare our sample to the theoretical **sampling distribution** for $n = 50$ with
 - $\mu = 380$
 - $\sigma = 35.2/(50)^{.5} = 4.98$

Strategy for the Problem

- Next we see how likely it would be that our sample would come from a population so described,
- i.e., we see where our sample would lie in the sampling distribution if the population mean really was 380.
 - The book calls this a **rare-event approach (p215-16)**
 - This comparison is done using a z-score test statistic
 - Difference of our sample from the population parameter, divided by the standard deviation of the sampling statistic

Steer Problem

- Calculate Test Statistic
 - $z = (390 - 380)/[35.2/(50)^{.5}]$
 - $z = 10/4.98 = 2.01$

Steer Problem

- Look up the Test Statistic in the book
 - $z = 2.01$ relates to a probability of .4778 on the right side of the curve
 - Which means we would expect to find a value out there or further only at
 - $.5 - .4778 = .022$
 - Or 2.2% of the time

Steer Problem : conclusion

- While it is possible that a sample with a mean of 390 could come from a population with a mean of 380
- The probability is very small, .022
- If I was betting on this one I would have good evidence to suggest that the new diet resulted in increased weight
- I.e., that the sample did not come from a population with $\mu = 380$

I could also point a Bound of Error or Confidence Interval on my sample estimate

- I specify a level of confidence that I want
 - 95% C.I.
- Find the z value associated with this level
 - I take the 95% level
 - Divide it by 2 for each side **.475**
 - And search for the z-value associated with this probability level **1.96**

Confidence Interval

- I would calculate the confidence interval as
 - $390 \pm 1.96[35.2/(50)^{.5}]$
 - $390 \pm 1.96(4.98)$
 - 390 ± 9.76
 - 380.24 to 399.76

What does the Confidence Interval Mean?

- In repeated samples
- 95% of the samples drawn
- Would have a confidence interval that contains the true population parameter
- **It does not mean that the population parameter or the true value is between those two numbers I calculated for my sample— we don't know for sure.**
- We just know that in repeated samples, 95% of the samples would generate a confidence interval that contains μ

Summary

- We can make inferences from a sample to a population if:
 - The sample is drawn randomly
 - We use an estimator to generate sample statistics
 - We know something about the Sampling Distribution of our estimator - **This helps us generate the Standard Error of our estimator**

Summary

- We can use a interval estimate via a **Confidence Interval**
- Or a point estimate in a **Hypothesis Test**
- To make inferences to the population parameter.