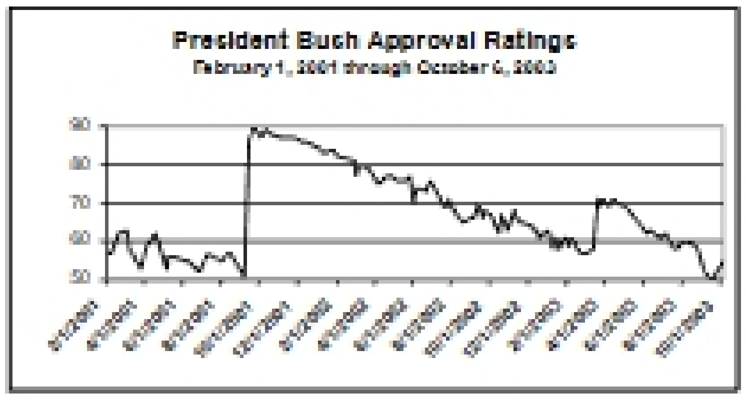
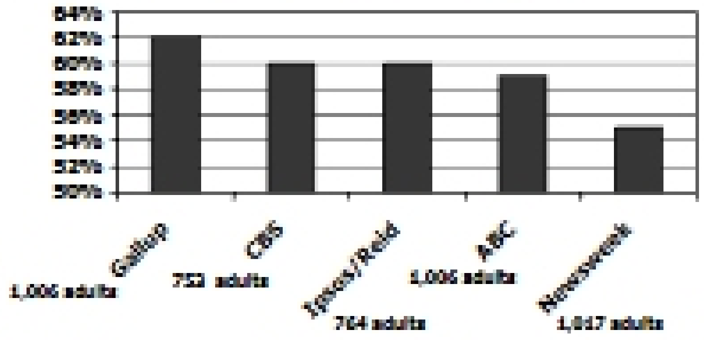
Sampling Distribution for the Mean

Dr Tom Ilvento
FREC 408

How is the President Doing?

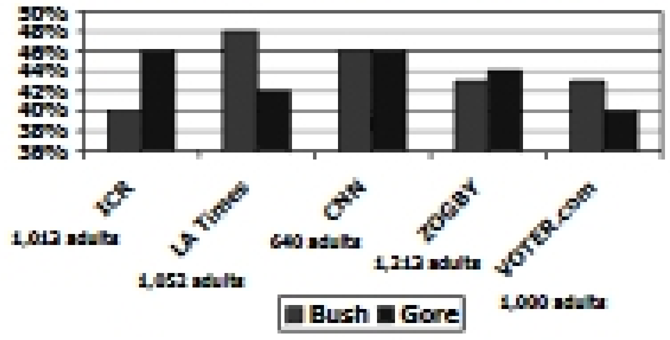


Bush Approval Ratings, Week of July 7, 2003



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

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

$\mathbf{E(x) = 3.500}$

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

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

$\boldsymbol{\sigma = 1.7078}$

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
3.33	4
3.67	4
4.00	5
2.67	1
4.00	4
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:
$\boldsymbol{\sigma = 1.7078}$

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:
$\boldsymbol{\sigma = 1.7078}$

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)

Stem and Leaf of 3-die of a Priori Sampling Distribution

1	0	1
1	3 3 3	3
1	7 7 7 7 7 7	6
2	0 0 0 0 0 0 0 0 0 0	10
2	3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	15
2	7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7	21
3	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	25
3	3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	27
3	7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7 7	27
4	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0	25
4	3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3	21
4	7 7 7 7 7 7 7 7 7 7 7 7 7 7 7	15
5	0 0 0 0 0 0 0 0 0 0	10
5	3 3 3 3 3 3	6
5	7 7 7	3
6	0	1

Stem is the ones; leaf is the first decimal place.

Compare the Mean and Median of our Sample Distribution

	Mean	*Median*
Mean	3.47	3.47
Standard Error	0.07	0.08
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

The mean has Minimum Variance

Sampling Distribution

- Sample statistics are **random variables**
- They have a probability distribution based on repeating the sampling experiment many times. We may get different sample statistics each time
- Repeating the experiment many times results in a **sampling distribution**
- **The sampling distribution of a sample statistic calculated from a sample of n measurements is the probability distribution of the statistic** (Def6.5 p311)

What do we want to see?

- If $\bar{x}$ is a good estimator of μ
- We would expect the values of $\bar{x}$ 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 $\bar{x}$'s around μ is tighter than the sampling distribution of some other possible estimator (minimum variance)

I asked a past class to help construct a sampling distribution for the mean

- It was repeated samples of size n = 50
- From x ~ N (75, 10)
- The "mean of the means" should equal the population parameter μ
- The standard deviation of this new distribution should be related to σ
  - But we might expect it to be smaller

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

$10/(50)^{.5} = 1.41$

AKA the Standard Error (p317)

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

Sampling Distribution for the Mean

71	1
72	8 7 7 8
73	3 8 7 8 8
74	[illegible]
75	[illegible]
76	[illegible]
77	1 8 8
78	3

Stem = 10's place Leaf = one's digit

mean $\bar{x}$ [illegible]	n [illegible]
Std Dev [illegible]	Median [illegible]

The exercise resulted in a Std Dev = **1.35**

This is close to the expected Std Error of

$10/(50)^{.5} = 1.41$