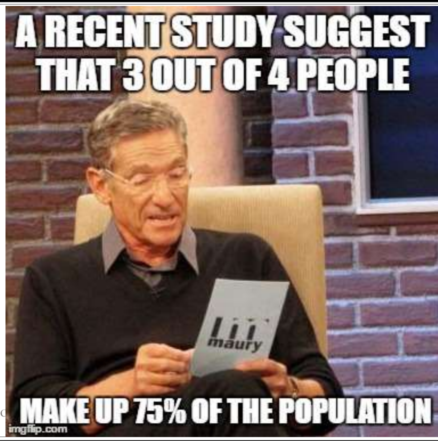


1

Keeping Terms Straight
(see Agresti and Finlay, p. 96-99)

- The **population distribution** refers to the distribution of a variable in the population;
- The **sample distribution** refers to the distribution of the values of a variable in the sample;
- The **sampling distribution** refers to the distribution of sample statistics (such as means or proportions) for all possible samples drawn from the population.

5 SOC424 @ CSUN - Ellis Godard

SOC424 – Statistics w/ Dr. Ellis Godard

CENTRAL LIMIT THEOREM

Main idea today, part I:

Means of repeated samples center around the population mean

- **Examples in Agresti & Finlay reading**
- **Population-of-three example last time**
- **Lab9 & HW3, across the class**
- **Two illustrations in lecture next time**

6 SOC424 @ CSUN - Ellis Godard

Outline for Today...

- **Basic Idea of the CLT**
- **Six Implications**
- **Sampling Distributions & Z-scores**
- **An example**
- **A lab**
- **Brilliance & wit interspersed throughout**

4 SOC424 w/ Dr. Ellis Godard -- Slide

Sampling distribution is more dispersed when sample size is small

means for recent MUSIC lab…

Ns of 10

		Frequency	Percent
Valid	3.00	4	28.6
	5.73	1	7.1
	6.25	1	7.1
	6.75	2	14.3
	7.30	1	7.1
	7.33	1	7.1
	7.33	1	7.1
	7.50	1	7.1
	7.66	1	7.1
	7.67	1	7.1
	Total	14	100.0

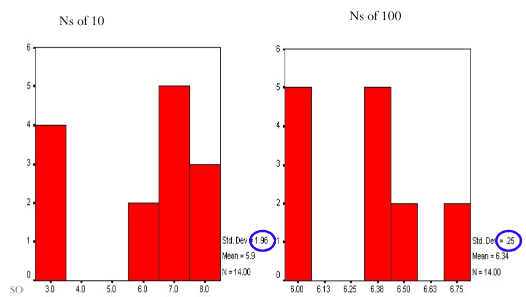
Ns of 100

		Frequency	Percent
Valid	6.02	1	7.1
	6.05	3	21.4
	6.05	1	7.1
	6.40	1	7.1
	6.43	1	7.1
	6.43	3	21.4
	6.48	2	14.3
	6.73	2	14.3
	Total	14	100.0

7 SOC424 @ CSUN - Ellis Godard

Sampling distribution more dispersed when sample size small

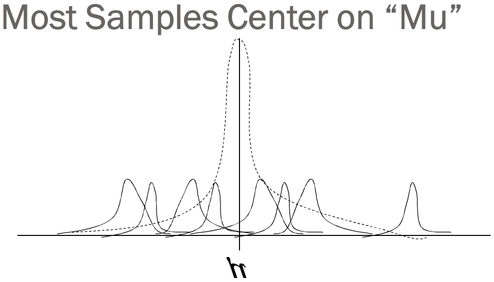
Compare the ranges or standard deviations…



Eventually, it would be normal

- That's only 14 samples.
  - With 1606 GSS cases, there are (1606!/1506!) possible samples of n=100, =(1606x1605x1604…)/(1506x1505x…), a number with over 312 digits!
- The more samples we add to that histogram the closer to normal it would look
  - And its mean would be …. what? ☺

Most Samples Center on "Mu"



- **This is one implication of the Central Limit Theorem. Five more, all today…**

CLT Overview

Statistics involves above all using sample statistics to make inferences about population parameters.

The central limit theorem implies that *if the population mean and population standard deviation are known,* it is possible to find the likelihood of obtaining a certain sample mean given that the sample is drawn from the population.

CLT (as stated in Agresti & Finlay)

Consider a random sample of n measurements from a population distribution having mean μ and standard deviation σ .

Then, if n is sufficiently large, the distribution of sample means (i.e. the sampling distribution) is an approximately normal distribution with mean μ and standard error $\sigma_{\bar{y}} = \sigma/\sqrt{n}$

CLT (in a few more words)

If a sample is taken <u>random</u>ly and if it is sufficiently <u>large</u>, the *sampling* distribution is <u>normal</u>, even if the *data* distribution (that is, the sample or population data) is not normal, i.e.

- An *estimated* 68.26% *of sample means* fall within 1 standard deviation of the *sampling* distribution – that is, within one standard error of the population mean
- An *estimated* 95.44% fall within 2 standard errors…
- An *estimated* 95% fall within 1.96 standard errors
- An *estimated* 99.77% fall within 2.57 standard errors

Note that these are *estimates* because we use the *sample* standard deviation when we don't know the population parameter (which is often)

3 Main Implications of Theorem:

- **sampling distribution's mean is approximately equal to the pop. mean**
  - Samples cluster around pop. Mean
  - That's above, and in last 2 set of notes
- **Standard error decreases as the sample size increases**
  - By formula, since n is denominator (see last 2 lects)
  - By logic: larger samples are closer to the pop. mean, but smaller samples could be weirder and further from it
- **Regardless of the shape of population distribution, the sampling distribution will be normal**

1st Implication: $\overline{Y}_{\overline{Y}} = \mu$

- **Imagine two simulations, repeatedly drawing samples from a larger set of cases:**
  - I drew 200 samples and got a mean of 500.4
  - I drew 2000 samples and got mean of 499.9.
  - Both are close to the population mean of 500.
- **The more cases there are, the closer the samples are *on* average to the pop parameter**
  - As the number of samples increased, we would expect the mean of the sampling distribution to become closer and closer to 500.

Example

$$if \quad \sigma = 100 \quad and \quad n = 1000, \quad then$$

$$\sigma_{\overline{Y}} = \sigma / \sqrt{n}$$

$$\sigma_{\overline{Y}} = 100/\sqrt{1000} = 100/31.6228 = 3.16$$

In the above simulations (see last lecture), the computed standard deviations were 3.13 and 3.22. Pretty close!

2nd Implication: n versus se

- **As n increases, the standard error decreases. Since the square root of n is in the denominator, a larger sample size will decrease the standard error**
- **See figure 4.14 on p.103 in A&F**

From above example, if ...

$$n=10, \quad then \quad \sigma_{\overline{Y}} = 100/\sqrt{10} = 31.6$$

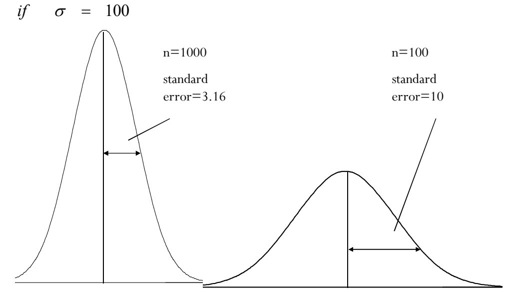
$$n=100, \quad then \quad \sigma_{\overline{Y}} = 100/\sqrt{100} = 10$$

$$n=1000, \quad then \quad \sigma_{\overline{Y}} = 100/\sqrt{1000} = 3.16$$

$$n=10000, \quad then \quad \sigma_{\overline{Y}} = 100/\sqrt{10000} = 1$$

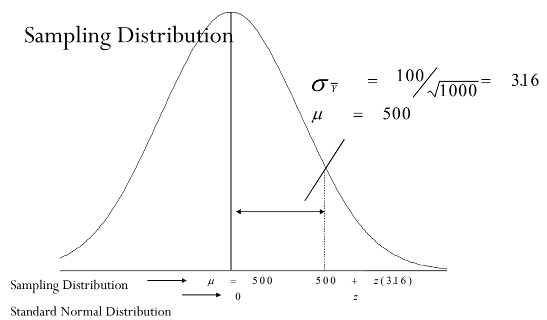
A Graphical Illustration

$$if \quad \sigma = 100$$



3rd Implication: Normality

- **The fact that the sampling distribution is approximately normal has nothing to do with the shape of the population distribution**
- ***Sampling* distribution is always normal**
- **See Figure 4.12 on page 101 of A&F**



Rhetorical Example I: Sports

- **What is the *expected* distribution of "number of years in the major league" for a baseball player?**
- **What is the shape of the sampling distribution?**

RE II: Political Sociology

- **What is the *expected* shape of the distribution of "presidential support" in a population of highly partisan Congress?**
- **What is the expected shape of the sampling distribution of sample means if the samples are "sufficiently large"?**
  - *Note: Next week, we'll define "sufficiently"*

RE III: Status and Orgs

- **What is the expected distribution of "academic reputation" for a population of 200 national universities?**
- **If we selected many samples of sufficient size (25 to 30 is usually enough according to Agresti and Finlay) and computed mean academic reputation for each sample, what is the shape of the distribution of these means? Approximately what is the mean?**

3 Main Implications of CLT:

- **sampling distribution's mean approximately equal to population mean**
  - Samples cluster around pop. mean
- **standard error decreases as the sample size increases**
  - Larger samples closer to pop. mean
- **regardless of shape of population distribution, sampling distribution normal**

4th Implication: Extreme Cases

1. What is the distribution of the sample mean if n = 1?

   The probability distribution for one randomly selected observation is the same as the population distribution of the variable.

2. What is the distribution of the sample mean if n=N (sample size equal population size)?

   The sample means always equals the population mean; the distribution is concentrated at the population mean μ

5th Implication: Parameters Matter

- **The larger the standard deviation of a variable in the population, the larger the standard error of the sampling distribution. This again follows from the definition of a standard error:**

$$\sigma_{\bar{Y}} = \sigma / \sqrt{n}$$

6th Implication: Size Matters

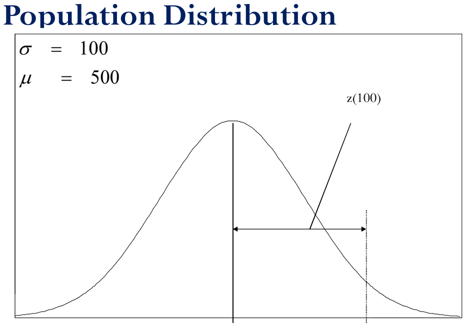
- **If the sample size n is small (less than 25-30), the shape of the sampling distribution is approximately normal *only if* the population distribution itself is approximately normal**
- **See Figure 4.15 on p. 104 of A&F**

Relation to z-scores

Recall that a z-score corresponding to a particular value is the number of standard deviations that this value is from the mean of the distribution.

$$z = \frac{Y - \mu}{\sigma}$$

Population Distribution



Normal Distribution → μ = 500 500 + z(100)

Standard Normal Distribution → 0 z

Z-Scores for External Values

- **Another sample's value**
  - How different from Family A is a family without a television?
    - $Z_0 = (0 - 10) / 6 = (-10)/6 = 1.66$
  - The central tendency of Family A is 1.66 standard deviations from a family with no television viewing.
- **An imaginary value**
  - What would be the z-score for someone who watched 22 hrs/week?
    - $Z_{22} = (22 - 10) / 6 = (12)/6 = 2$
  - They would be 2 standard deviations from this family's mean
- **An established standard**
  - If research calls more than 8 hrs unhealthy, how is Family A doing?
    - $Z_8 = (8 - 10) / 6 = (-2)/6 = 0.33$
  - This family exceeds the guidelines by 0.33 standard deviations

Distance by Standard Errors

- **I've illustrated z-scores w/ comparisons we might make (e.g. to an imaginary value, an external standard, or another sample's mean)**
- **But we make such comparisons in terms not of standard deviations but of *standard errors*:**

$$stderror = \frac{stdev}{\sqrt{n}}$$

- **For example, for Family A, the standard error is (6 / sqrt of 6), or 2.449**

***Specified to a sampling distribution,* the z-score of a particular sample mean is the number of standard errors that this sample mean is from the mean of the sampling distribution (which is equal to the population mean).**

$$z = \frac{\overline{Y} - \mu}{\sigma_{\overline{Y}}}$$

Estimating something new:

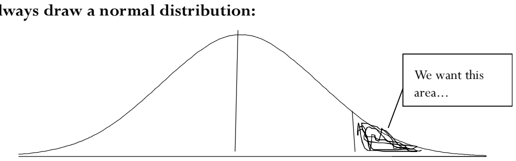
- **Had estimated likelihood of Y_i's**
  - Used a dataset
    - With a mean & stdev.
  - Calculated distance from that mean
    - Calculated zs for a Y_i
    - $Z = (\bar{y} - Y_i)$ / std.dev
  - Used z's/P's to estimate likelihood of getting a Y_i more different from the mean that the one we had
- **Now estimating likelihood of $\bar{y}$'s**
  - Using a sample
    - With a mean & stdev.
  - Calculated distance from that mean
    - Calculate zs for a $\bar{y}$
    - $Z = (\bar{y} - Y_i)$ / std.error
  - Use z's/P's to estimate likelihood of getting a $\bar{y}$ more different from the *population* mean that the one we had

Example from A&F (p89, # 15, 2nd edition)

Suppose the distribution of yearly incomes of all migrant workers in California has a mean of $8,000 and a standard deviation of $3,200.

a. Assuming that the distribution is normal, what is the probability that a particular migrant worker has an income over $9,000?

1. Always draw a normal distribution:



μ = $8,000 $\quad Y_i$ = $9,000

2. Compute the z-score and use Table A to find area:

$$z = \frac{Y - \mu}{\sigma} = \frac{9{,}000 - 8{,}000}{3{,}200} = 0.31$$

$$P(z > 0.31) = 0.3783$$

3. State conclusion in full sentence(s): The probability that a particular migrant worker has an income above $9,000 is 0.38 or 38 percent. That is, given the assumptions of the problem, we conclude that 38 percent of migrant workers in California have incomes above $9,000.

Second part…

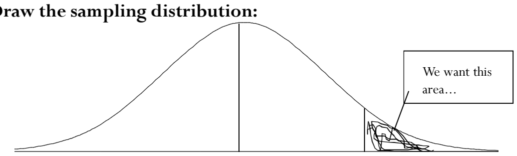
(b) If we plan on taking a random sample of 64 migrant workers, what is the sampling distribution of the sample mean income? Find the probability that the sample mean exceeds $9,000.

Solution:

1. Compute the standard error:

$$\sigma_{\bar{Y}} = \sigma / \sqrt{n} = 3{,}200 / \sqrt{64} = 400$$

2. Draw the sampling distribution:



$\mu = \$8{,}000$ $\bar{Y} = \$9{,}000$

3. Compute the z-score and use Table A to find area:

$$z = \frac{\bar{Y} - \mu}{\sigma_{\bar{Y}}} = \frac{9{,}000 - 8{,}000}{400} = 2.50$$

$$P(z > 2.50) = .0062$$

4. State conclusion in full sentence(s): The probability that the mean income of a sample of 64 migrant worker is above $9,000 is equal to 0.006 or 0.6 percent.

For your next lab…

- **Optional extra credit lab**
  - Count as an *additional* lab
  - Does *not* replace a skipped/missed lab
- **Use the world95.sav dataset**
- **Note two uses of "population"!!**
  - For a *population* of 109 countries
    - Not a sample, but all of them
  - For each case, its *population* is measured.
    - the number of people in that country
- **Group assignment**
  - 3-5 for secretary bonus
  - There's a form on the website.. Use it! ☺