Announcements

- One Handout Today
- Previous Lecture's handout up front
- Unclaimed exams & homework 1's & 2's are up front
- Homework 3 is due TODAY

A Population

All 1,109 American deaths in Iraq as of 10/28/2004

Age

	Percentiles	Smallest		
1%	18	18		
5%	19	18		
10%	20	18	Obs	1109
25%	21	18	Sum of Wgt.	1109
50%	24		**Mean**	**26.39405** ← μ_y **(mu)**
		Largest	**Std. Dev.**	**7.079283** ← σ_y **(sigma)**
75%	30	52		
90%	37	54	Variance	50.11625
95%	41	55	Skewness	1.295218
99%	49	59	Kurtosis	4.393614

A Sample

A single sample of size 25 taken from the population

Age

	Percentiles	Smallest		
1%	18	18		
5%	19	19		
10%	21	21	Obs	25
25%	23	21	Sum of Wgt.	25
50%	26		**Mean**	**28.36** ← *(y-bar)*
		Largest	**Std. Dev.**	**7.169844** ← **s**
75%	35	38		
90%	39	39	Variance	51.40667
95%	41	41	Skewness	.4979002
99%	42	42	Kurtosis	2.004631

The actual sample:

24 27 23 36 25 26 26 19 34 18 28 39 35 21 29 23 36 28 41 38 42 21 24 21 25

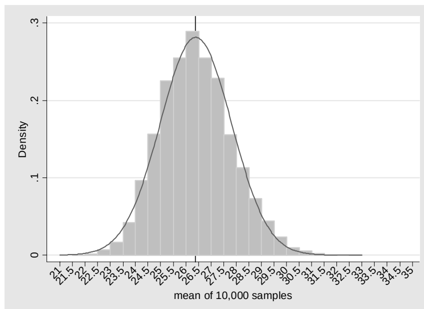
A Simulated Sampling Distribution

10,000 means of the variable age from 10,000 samples of size 25

	Percentiles	Smallest		
1%	23.36	21.68		
5%	24.2	22.24		
10%	24.64	22.29167	Obs	10000
25%	25.4	22.32	Sum of Wgt.	10000
50%	26.32		**Mean**	**26.39871** ← $\mu_{\overline{Y}}$
		Largest	**Std. Dev.**	**1.417049** ← $\sigma_{\overline{Y}}$
75%	27.32	31.44		
90%	28.28	31.5	Variance	2.008029
95%	28.84	31.8	Skewness	.2502371
99%	29.92	32.52	Kurtosis	3.016403

Histogram of the means of 10,000 samples of size 25

•If I were to draw 10,000 samples of size 25 (with replacement) from our population of 1,109 (with mean age of 26.39405 years) the mean of all 10,000 sample means will be equal to, in theory, our true population mean. We got 26.39871 for the 10,000 samples



If sampling distributions are theoretical, what's the point?

- We want to know how close is $\bar{y}$ to μ_y **(or $\mu_{\bar{Y}}$)** that is, how accurate will our samples be?
- In order to answer this, you will need to know the standard deviation of the population σ_y and the sample size **n and also that the sampling distribution is normal**
- Note how the standard deviation of the sampling distribution changes with sample size. For big samples, the standard deviation for the sample mean will be small and for small samples, the standard deviation for the sample mean is large. $\sigma_{\bar{Y}} = \frac{\sigma_y}{\sqrt{n}}$

Rule 3. Normal Populations and Normal Sampling Distributions

- Given a simple random sample (SRS) of size n from a population having mean μ_y and standard deviation σ_y, the sample mean will originate from a sampling distribution of all possible sample $\bar{y}$ means with mean $\mu_{\bar{Y}}$ and standard deviation $\sigma_{\bar{Y}} = \frac{\sigma_y}{\sqrt{n}}$
- If the original population had a normal distribution, then the sampling distribution will also be normally distributed. This is good, because it means we can use the standard normal table to make inferences about a particular sample which results in a statement of probability or chance.

Example of Rule 3

- Example. IQ scores are normally distributed with a mean of 100 and a standard deviation of 15. A sample of 25 persons is drawn.
- How likely is it to get a sample average of 108 or more? (0.38%)

$$Z = \frac{\bar{y} - \mu_Y}{\frac{\sigma_Y}{\sqrt{n}}} = \frac{108 - 100}{\frac{15}{\sqrt{25}}} = +2.67$$

How likely is it for the first score to be 108 or more? (Like Chapter 6 – 29.8%)

$$Z = \frac{\bar{y} - \mu_Y}{\frac{\sigma_Y}{\sqrt{n}}} = \frac{108 - 100}{\frac{15}{\sqrt{1}}} = +.53$$

Rule 4. **The Central Limit Theorem (p. 343)**

- No matter the distribution of the original population (recall our Iraqi Freedom age at death is right skewed), if the sample size is "sufficiently large" (> 50 in your text) and the sample is random, the distribution of the possible sample means will be close to the normal distribution even for skewed populations.
- It is a very powerful theorem and it is the reason why the normal distribution is so well studied.

C.L.T. more fully

- Take a simple random sample from a population with mean μ_y and standard deviation σ_y. Let $\bar{y}$ be the average of the random sample taken from the population. If either
  - **the original population is normally distributed**

  OR
  - **the sample size n is sufficiently large (>50),**
- then the population of all possible $\bar{y}$ will be normally distributed with $\mu_{\bar{Y}} = \mu_y$ and standard deviation $\sigma_{\bar{Y}} = \frac{\sigma_y}{\sqrt{n}}$ this was demonstrated last time.

C.L.T. Results

- Thus, about 68% of the $\bar{y}$ will be within one standard deviation of the true population mean
- about 95% of the $\bar{y}$ will be within two SDs and 99.7% of the $\bar{y}$ will be within 3 SDs
- Let's go back to our first sample of 25 with its mean of 28.36. The chance of getting a mean as high or higher is:

$$Z = \frac{\bar{y} - \mu_Y}{\frac{\sigma_Y}{\sqrt{n}}} = \frac{28.36 - 26.394}{\frac{7.079283}{\sqrt{25}}} = \frac{1.966}{1.4159} = +1.39$$

- Look-up +1.39 from standard normal table and you get .9177. What we want is the area beyond Z which is (1.0-.9177=.0823). So the chance (probability) of drawing a sample of size 25 with an average of 28.36 or higher when you were expecting the average to be 26.394 was about 8.23% Your interpretation is that about 8% of time you would get a sample average as high as the one you got. This suggests that while we were off by 2 years, it's not really an outlier and it wasn't impossible to be this far from the true average even though you have done everything correctly (e.g. random sample, large enough)

Remember…

- NOTE: The Central Limit Theorem only applies to the distribution of possible sample averages (i.e. the sampling distribution) it says nothing about the distribution of individual values in either the sample or in the population. (Look at the handout from the previous lecture)

A special case of means: The proportion

- A proportion is the mean of a special kind of population. This population only has values of 1 or 0. For this type of population, the mean is

  p which is the proportion of 1's in your population.

- And the population standard deviation is $\sigma_p = \sqrt{p * q} = \sqrt{p(1-p)}$ where q is the value of (1.0 – proportion of p's)

Example

```
For example: The Recall of Governor Davis in October 2003, Final Count
(population)

      Davis |      Freq.     Percent        Cum.
------------+-----------------------------------
         NO |  3,559,436       44.63       44.63       Let's treat the YES=1 and
        YES |  4,415,398       55.37      100.00       the NO=0
------------+-----------------------------------
      Total |  7,974,834      100.00

                            Davis
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            0              0       Obs             7974834
25%            0              0       Sum of Wgt.     7974834

50%            1                      Mean            .5536664
                        Largest       Std. Dev.       .4971116
75%            1              1
90%            1              1       Variance        .2471199
95%            1              1       Skewness       -.2159131
99%            1              1       Kurtosis        1.046618
```

Suppose we had a sample of size 49 likely voters

- One of the final polls before the 2003 recall showed that 51% wanted Davis recalled.
- It is like having a sample of 49 with 24 "No" (label them with zeroes) and 25 "Yes" (label them with ones)
- Proportions also have sampling distributions, it's a distribution of $\hat{p}$ (p-hat instead of y-bar) and the distribution has a theoretical mean of **p** (true population proportion) and a standard deviation of

$$\sigma_{\hat{p}} = SD(\hat{p}) = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}}$$

- The question: what was the chance of getting a sample with 51% (p-hat) in favor of a recall when the true proportion, p, was really 55.37%?

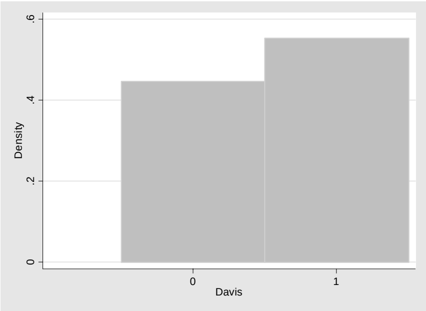
Results & Application of the Sampling Distribution for Proportions

- The chance of getting a sample proportion (p-hat) as low as 51% or lower from a single sample of 49 when the true proportion is 55.37% is

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{.51 - .5537}{\sqrt{\frac{.5537(.4463)}{49}}} = \frac{-.0437}{\sqrt{\frac{.2471}{49}}} \approx -.62$$

- Use a Z score because our one sample comes from a larger sampling distribution which is normal. The area to the left of -.62 is .2676 or they had a 26.76% chance of getting a sample proportion as low as or lower than 51% (when the true proportion was 55.37%)

Note: The Population is NOT NORMAL – It's Histogram



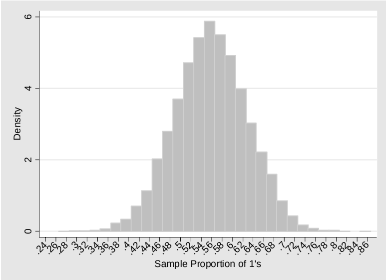
Simulation Results

BUT if I were to run a simulation of samples of size 49 for 25,000 different samples from our population of 55.37% recall (1's), 44.63% do not recall (0's):

```
Sample Proportion of "1's" for 25,000 different samples of size 49
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .3877551       .2653061
 5%     .4285714       .2857143
10%     .4693878       .2857143       Obs              25000
25%     .5102041       .2857143       Sum of Wgt.      25000

50%     .5510204                      Mean          .5541796
                        Largest       Std. Dev.     .0710398
75%     .5918368       .8163266
90%     .6530612       .8163266       Variance      .0050466
95%     .6734694       .8571429       Skewness     -.0111072
99%     .7142857       .8571429       Kurtosis      2.956921
```

Graph of 25,000 different samples of size 49



We can see that sample proportions behave like sample means, in theory they wants to center on the value of p (the true population proportion) and have a standard deviation of $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

Important Conditions

- To have a normal sampling distribution for a proportion (and then to apply Z to help you calculate chances)
  - np > 10 & nq > 10
  - The sample size can't be larger than 10% of the population
- Formulas: for proportions → $Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}}$

$$Z = \frac{\bar{y} - \mu_Y}{\frac{\sigma_Y}{\sqrt{n}}}$$ ← For Means