Announcements

- Answer Keys have been posted, see “tests” link
- One Handout Today
- Previous Lecture’s handout up front
- Unclaimed exams & homework 1’s are up front
- Homework 2 will be available from your Teaching Assistants on Tuesday
- Homework 3 is due Wednesday
- Office Hours Today 2pm-4pm

Lecture 13 – Sampling Distributions OR the Distribution of All Possible Samples OR the population of Samples (Chapter 18 – Part 1)

- Recall **POPULATION, SAMPLE, PARAMETER and STATISTIC (review Chapter 12)**
- **Also recall: STATISTICAL INFERENCE**: a situation where the population parameters are unknown, and we draw conclusions from sample outcomes (those are statistics) to make statements about the actual value of the population parameters.
- When random samples are drawn from a population of interest to represent the whole population, they are generally unbiased and representative.

Sampling Distribution: Definition

- The sampling distribution is a theoretical /conceptual /ideal probability distribution of a statistic.
- A theoretical probability distribution is what the outcomes (i.e. statistics) of some random process (e.g. drawing a sample from population) would look like **if** you could repeat the random process over and over again and had information (that is statistics) from every possible sample.
- Note that a **sampling distribution** is the <u>theoretical</u> probability distribution of a statistic. The sampling distribution shows how a statistic varies from sample to sample and the pattern of possible values a statistic takes.
- We do not actually see sampling distributions in real life, they are simulated. They exist in theory.

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 ←μy (mu)
                        Largest       Std. Dev.      7.079283 ← σy (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

RULE 1: The mean of all possible sample means (all possible $\bar{y}$) is denoted $\mu_{\bar{Y}}$

- $\mu_{\bar{Y}}$ in theory should be equal to μ_Y (the true population mean).
- In other words, the mean of sample means ($\mu_{\bar{Y}}$) calculated from **<u>all possible samples</u>** (nearly infinite possibilities) of the same size from the same population should be equal to the true population mean.
- We can check this using a simulation.

A Simulated Sampling Distribution

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

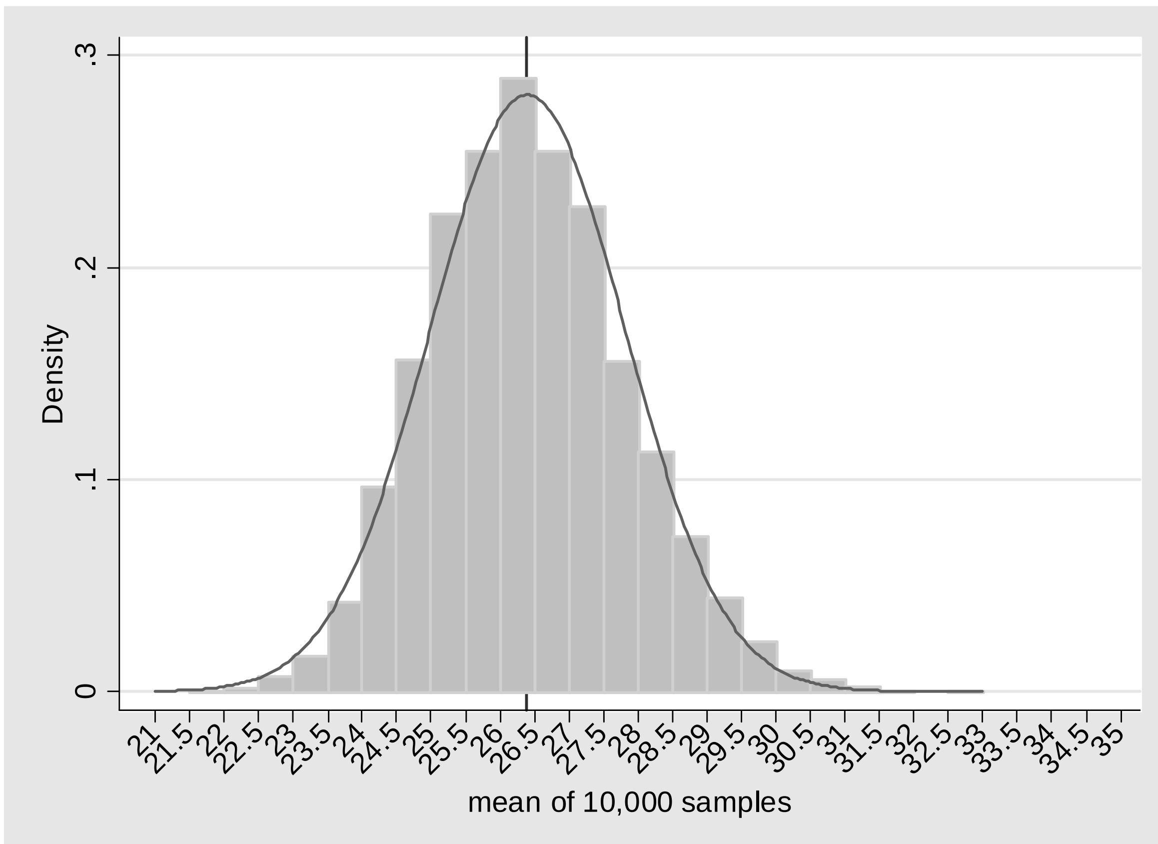
```
-------------------------------------------------------------
      Percentiles      Smallest
 1%         23.36         21.68
 5%          24.2         22.24
10%         24.64      22.29167       Obs               10000
25%          25.4         22.32       Sum of Wgt.       10000

50%         26.32                     Mean           26.39871  ←  μ_Ȳ
                        Largest       Std. Dev.      1.417049  ←  σ_Ȳ
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
```

$\mu_{\overline{Y}}$

$\sigma_{\overline{Y}}$

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



RULE 2. The theoretical standard deviation of all possible $\bar{y}$'s from all possible samples of size n is $\sigma_{\bar{Y}} = \frac{\sigma_y}{\sqrt{n}}$

- In our population data, σ_y is 7.079283
- so the theoretical standard deviation for a distribution of all possible sample means from samples of size 25 should be

$$\sigma_{\bar{Y}} = \frac{\sigma_y}{\sqrt{n}} = \frac{7.079283}{\sqrt{25}} = 1.4158566$$

- We can check whether this holds true: the standard deviation for our 10,000 sample means (from our samples of size 25) is 1.417049, again, very close
- This rule is approximately correct as long as your sample is no larger than 5% of your population.

Recap

- A sample has a mean $\bar{y}$ y-bar and it has a standard deviation s.
- A population has a mean μ_y and a standard deviation σ_y
- A sampling distribution or a distribution of all possible sample statistics, in this case -- sample mean -- also has a mean denoted $\mu_{\overline{Y}}$ and in theory it's equal to μ_y but with a standard deviation of

$$\sigma_{\overline{Y}} = \frac{\sigma_y}{\sqrt{n}}$$

Notes

- Your sample (or any real-life sample) is just one single realization from a population of samples.
- The standard deviation of all the SAMPLE MEANS $\sigma_{\bar{Y}} = \frac{\sigma_y}{\sqrt{n}}$ will be smaller than the SD for a single sample or the SD of the population.
- In other words, it is easier for us to predict the mean of many observations than it is to predict the value of a single observation (or to predict the average of small samples).
- Why? Examine the formula for the standard deviation of the sampling distribution, $\sigma_{\bar{Y}} = \frac{\sigma_y}{\sqrt{n}}$ note the effect of sample size on the standard deviation of all sample means. The bigger the sample size gets, the smaller $\sigma_{\bar{Y}}$ *becomes*.