**fe1. Definitions again**

Review the definitions of **POPULATION, SAMPLE, PARAMETER and STATISTIC**.
**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 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.

The key to understanding why samples behave this way is a difficult concept**: THE SAMPLING DISTRIBUTION.** 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 theoretical 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.

## 2. Sampling Distributions for Means

Let's suppose that the 1,428 or so people this example are a population.

And here is the mean $\mu_y$ (mu) and standard deviation $\sigma_y$ (sigma) of our population:

```
                            age
-------------------------------------------------------------
      Percentiles      Smallest
 1%           19             18
 5%           22             18
10%           25             18       Obs                1425
25%           32             18       Sum of Wgt.        1425

50%           42                      Mean           45.42035
                        Largest       Std. Dev.      17.11534
75%           56             89
90%           72             89       Variance       292.9348
95%           79             89       Skewness       .5865022
99%           87             89       Kurtosis       2.504332
```

Suppose we draw a simple random sample of size n from a large population. Call the observed values $Y_1$, $Y_2$, ..., $Y_n$. An example: draw a simple random sample (SRS) of 25 from the 1,425 persons with measured age.

Measure the average age from the sample of size 25 and compare it to the population average.

```
    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |        25       42.68    14.25868         21         71
```

A statistic: The mean of the sample of 25, 42.68 is just the old mean (from Chapter 5), here are the ages of the 25 people who were sampled:

```
71  60  55  55  43  41  25  30  24  43 24 50 36 66 57 32 29 21 41 43  26 58 43 55 39
```

We define the mean of a single sample as $\bar{y} = \frac{y_1 + y_2 + ... + y_n}{n}$ this is from chapter 5, and we define the Standard deviation of a single sample as $S_y = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}$ also from chapter 5

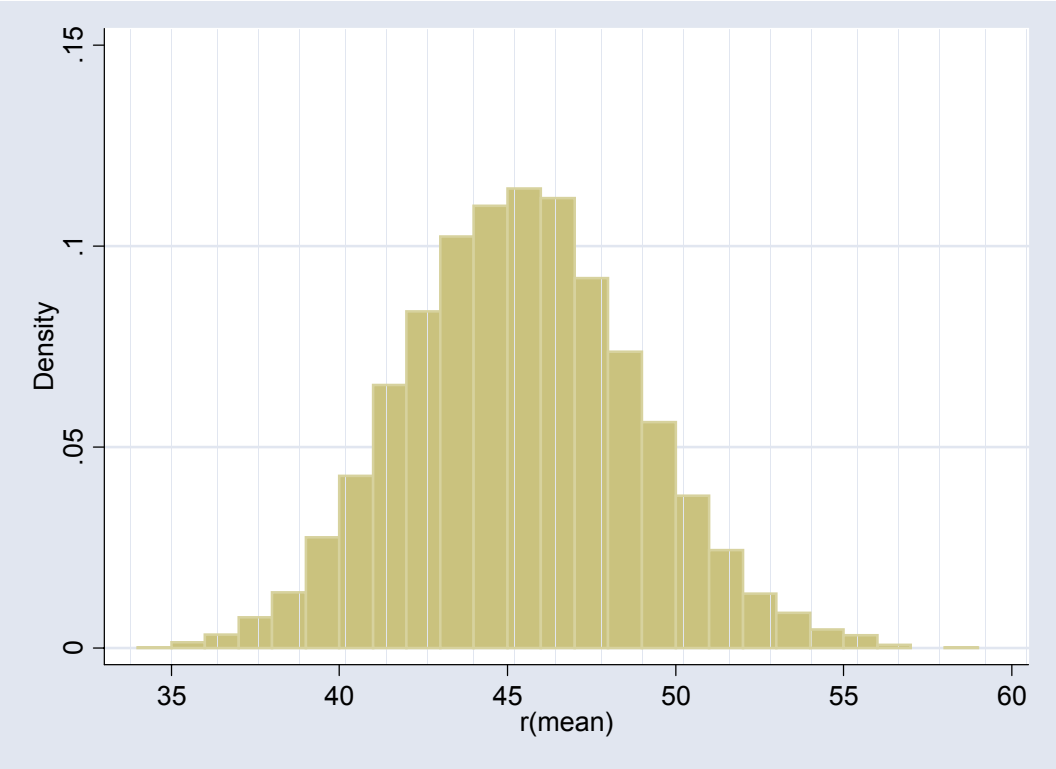
$\bar{y}$ can be thought of as the mean of a single sample of size 25 selected at random from all possible samples of size 25 that could have been generated from the population.

**RULE 1: The mean of all possible sample means (all possible $\bar{y}$ ) is denoted $\mu_{\bar{Y}}$** which in theory should be equal to $\mu_y$ (the true population mean). In other words, the mean of sample means ( $\mu_{\bar{Y}}$ ) calculated from all possible samples of the same size from the same population should be equal to the true population mean $\mu_y$. We can check this using a simulation. If I were to draw 10,000 samples of size 25 (with replacement) from our population of 1,428 (with mean age of 45.42035 years) the mean of all 10,000 sample means will be equal to, in theory, our true population mean.

```
                          r(mean)
-------------------------------------------------------------
      Percentiles      Smallest
 1%         37.72           34.8
 5%         39.88          35.12
10%         41.04          35.16       Obs                10000
25%            43          35.24       Sum of Wgt.        10000

50%         45.36                      Mean            45.39324
                        Largest        Std. Dev.        3.43353
75%         47.68           56.6
90%         49.84          56.64       Variance        11.78913
95%         51.16          58.16       Skewness        .1262797
99%         53.76          58.48       Kurtosis        2.951363
```

This is the overall average of 10,000 sample means from samples of size 25 drawn with replacement from our original population of 1,428. We got 45.39324 as the mean of the 10,000 sample means of all of our samples of size 25, this is very close to the true $\mu_{\bar{Y}}$ (or $\mu_y$ ) of 45.42035 Here's graph of the 10,000 sample means from our 10,000 samples of size 25:




Look familiar?

The mean of all sample means $\mu_{\bar{Y}}$ is considered an unbiased estimator of $\mu_y$(the true population mean) when it comes from a random sample. If your samples are not random, this relationship will not hold. For our first sample of 25 people, the mean of the sample is *42.68* but the mean of all 10,000 of the sample means is 45.39 and it's not too different from the true population mean of 45.42

**RULE 2. The theoretical standard deviation of all possible $\bar{y}$ 's from all possible samples of size n is**

$$\sigma_{\bar{Y}} = \frac{\sigma_y}{\sqrt{N}}$$

where $\sigma_y$ is the standard deviation of the population. In our population data, $\sigma_y$ is 17.11534 so the theoretical standard deviation for a distribution of all possible sample means from samples of size 25 should be

$$\sigma_{\bar{Y}} = \frac{\sigma_y}{\sqrt{N}} = \frac{17.11534}{\sqrt{25}} = 3.423068$$

We can check whether this holds true or not by examining the results of a simulation from the output above, the standard deviation for our 10,000 sample means (from our samples of size 25) is 3.43353, again, very close to what we get from the theory (3.423068).

This rule is approximately correct as long as your sample is no larger than 5% of your population. So please make a note of this:

- A sample has a mean $\bar{y}$ and it has a standard deviation s.
- A population has a mean $\mu_y$ and a standard deviation $\sigma_y$
- A sampling distribution or a distribution of all possible sample statistics, in this case
- the sample mean, also has a mean denoted $\mu_{\bar{Y}}$ and in theory it's equal to $\mu_y$ but with a standard deviation of $\sigma_{\bar{Y}} = \frac{\sigma_y}{\sqrt{N}}$.

Your sample (or any real-life sample) is just one single realization of all possible samples from a population of samples..

The standard deviation $\sigma_{\bar{Y}} = \frac{\sigma_y}{\sqrt{N}}$ of all the SAMPLE MEANS will be smaller than the standard deviation for a single sample. In other words, it is easier 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). What is causing this? Examine the formula for the standard deviation of the sampling distribution, note the effect of sample size on the standard deviation of all sample means. The bigger the sample size gets, the smaller $\sigma_{\bar{Y}} = \frac{\sigma_y}{\sqrt{N}}$ *becomes.*

***Some things to consider***

How close is $\bar{y}$ to $\mu_{\bar{Y}}$ **or** in other words, how accurate will our samples be? In order to do this, you will need to know the standard deviation of the population $\sigma_y$ and the sample size N

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.

## 3. RULES 3 & RULE 4: Normal Distributions and The Central Limit Theorem

Given a simple random sample of size n from a population having mean $\mu_y$ and standard deviation $\sigma_y$, the sample mean $\bar{y}$ will come from a sampling distribution of all possible sample means with mean $\mu_{\bar{Y}}$ and standard deviation $\sigma_{\bar{Y}} = \frac{\sigma_y}{\sqrt{N}}$

.

### A. Basic Distributional Result

If the original population had a normal distribution, then the distribution of the sample mean will also be normally distributed. This is good, because it means we can use the normal table to make inferences about a particular sample with a statement of probability or chance.

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

### B. The Central Limit Theorem (p. 343)

No matter what the distribution of the original population (recall our original one is left skewed), if the sample size is "sufficient", the distribution of the possible sample means will be close to the normal distribution. It is a very powerful theorem and it is the reason why the normal distribution is so well studied.

## C. Summary

Take a simple random sample from a population with mean $\mu_y$ and standard deviation $\sigma_y$. Let $\bar{y}$ be the average of the samples taken from the population. If either

**the original population is normally distributed OR the sample size n is sufficiently large,**

then all the $\bar{y}$ will be normally distributed with mean $\mu_{\bar{Y}} = \mu_y$ and standard deviation $\sigma_{\bar{Y}} = \frac{\sigma_y}{\sqrt{N}}$

If the histogram for the population follows a normal curve, or if the sample size is large enough each time, then the histogram for the possible values for $\bar{y}$ will follow a normal curve that has a mean of $\mu_y$ and a standard deviation of $\sigma_{\bar{Y}} = \frac{\sigma_y}{\sqrt{N}}$

.

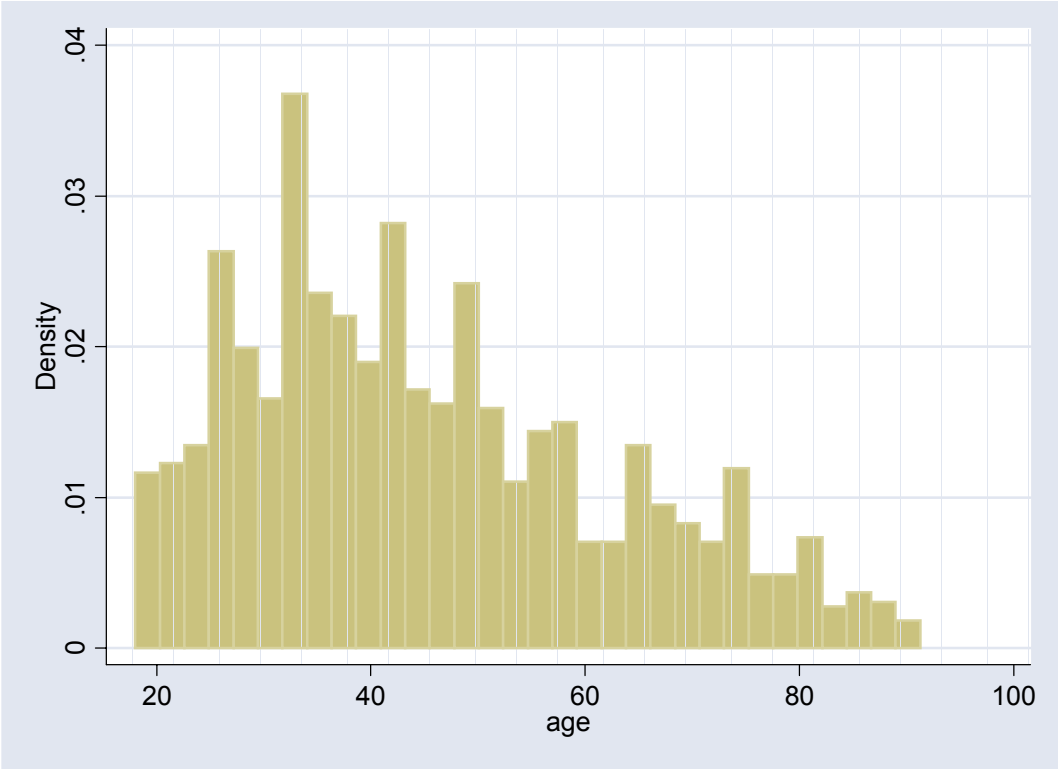
Thus, about 68% of the $\bar{y}$ will be within one standard deviation of the true population mean about 95% of the $\bar{y}$ will be within two standard deviations and 99.7% of the $\bar{y}$ will be within 3 SEs

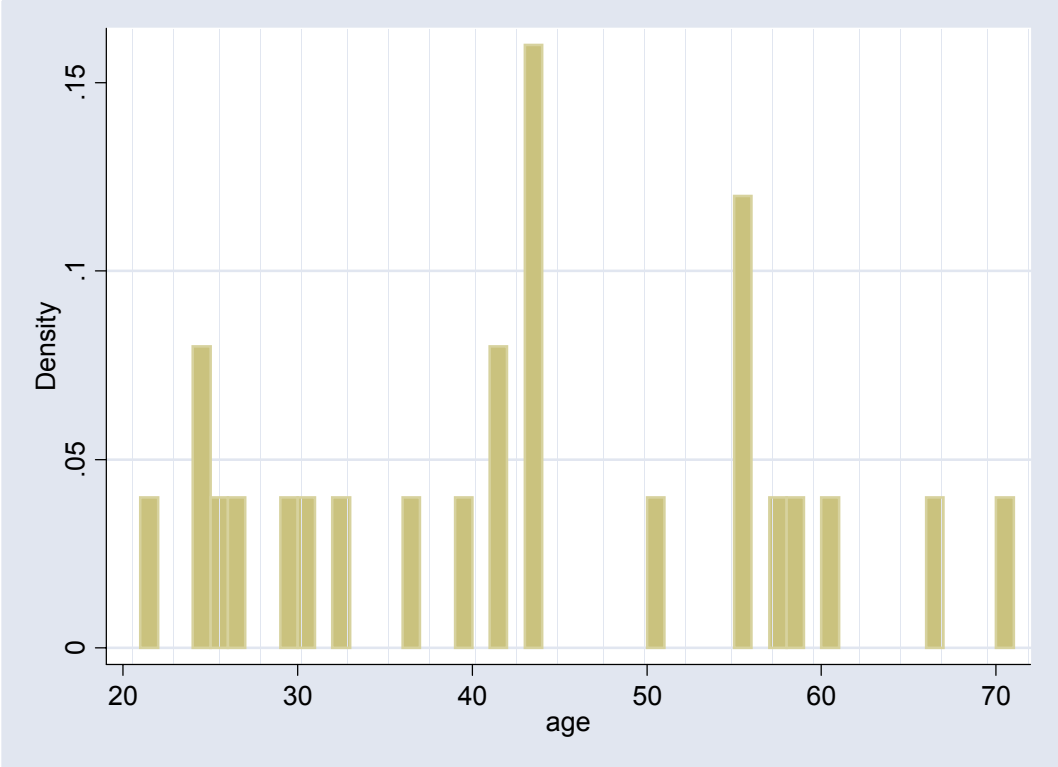
Let's go back to our first sample of 25 with its mean of 42.68. The chance of getting a mean that low or lower is: (1) calculate $Z = \frac{42.68 - 45.42}{\frac{17.11534}{\sqrt{25}}} = -.80$

. .

$Z$ about -.80, then (2) do a look-up from standard normal table and you get .2119 in the area beyond Z. So the chance (probability) of drawing a sample of size 25 with an average of 42.68 or lower when you were expecting the average to be 45.42 was about 21.19% Your interpretation is that about 21% of time you would get a sample average as low as the one got. This suggests that it's not too unusual to be this far from the true average even though you have done everything correctly (e.g. random sample).

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 scores in either the sample or in the population. For example, here is a graph of our age variable for the population followed by a graph of our sample of size 25 (from the beginning of this lecture)




Notice: neither one are normal, but we can use the normal curve to help us make statements of chance and accuracy because of the sampling distribution (it's normal as long as the sample size is sufficiently large)

## 4. A special case of means: The proportion

A proportion could be thought of as the mean of a special kind of population. The population only has values of 1 or 0. If a population has that feature, the population mean is

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

And the population standard deviation is

$\sigma_p = \sqrt{p * q}$ where q is the value of (1.0 – proportion of p's)

For example:

```
    clinton |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        379       43.36       43.36
          1 |        495       56.64      100.00
------------+-----------------------------------
      Total |        874      100.00
```

```
                          clinton
-------------------------------------------------------------
      Percentiles      Smallest
 1%             0             0
 5%             0             0
10%             0             0       Obs                 874
25%             0             0       Sum of Wgt.         874

50%             1                     Mean           .5663616
                        Largest       Std. Dev.      .4958603
75%             1             1
90%             1             1       Variance       .2458775
95%             1             1       Skewness      -.2678155
99%             1             1       Kurtosis       1.071725
```

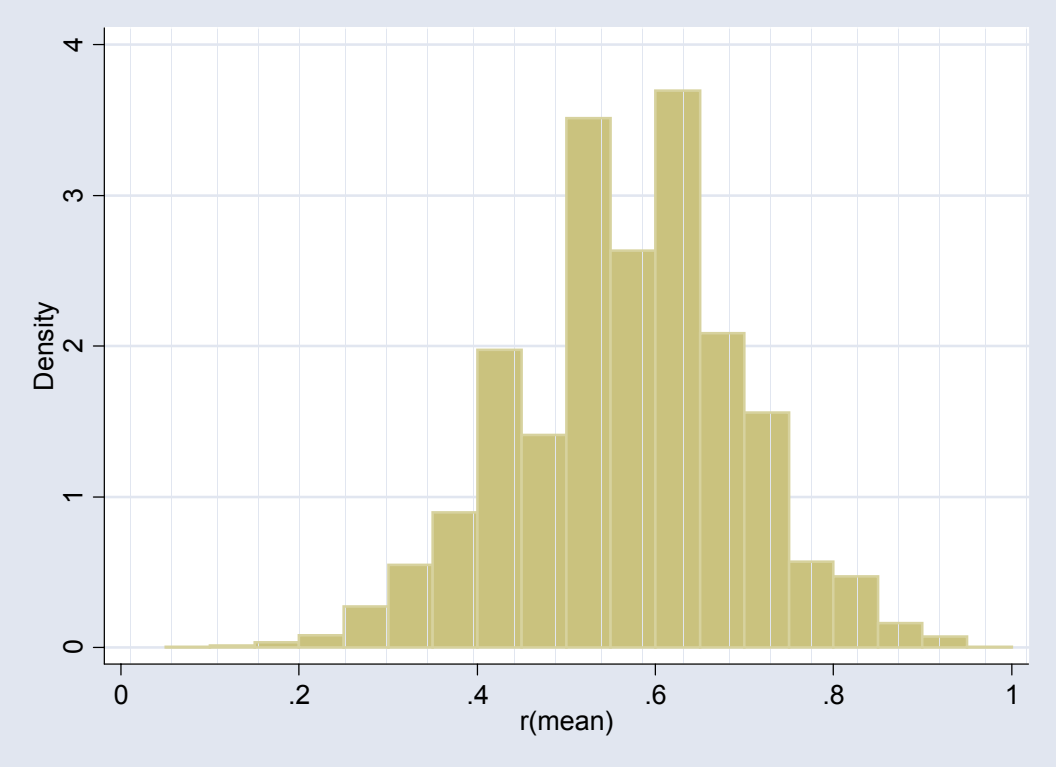
Proportions also have a sampling distribution, it's a distribution of sample proportions and this distribution has a mean of p and a standard deviation of

$$\sigma_p = \sqrt{\frac{pq}{n}}$$

And if I were to run a simulation of samples of size 25 for 10000 samples

```
                          clinton
-------------------------------------------------------------
      Percentiles      Smallest
 1%     .2666667       .0833333
 5%     .3529412       .0833333
10%           .4       .0909091       Obs               10000
25%     .4705882             .1       Sum of Wgt.       10000

50%     .5714286                     Mean           .5657526
                        Largest       Std. Dev.      .1289778
75%     .6470588       .9411765
90%     .7333333       .9444444       Variance       .0166353
95%     .7692308              1       Skewness      -.0970473
99%     .8571429              1       Kurtosis       2.948515
```




We can see that proportions behave like the mean, in theory it wants to center on the value of p (the true population proportion) and have a standard deviation of

$$\sigma_p = \sqrt{\frac{pq}{n}}$$