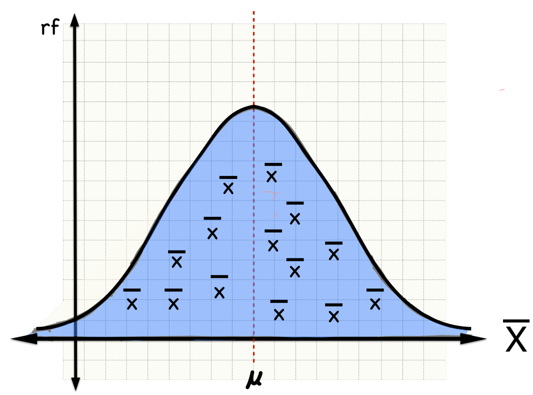
Central Limit Theorem

We put together the information we discovered through the **Simulation of the Sampling Distributions**. To be clear, the **Sample Distribution of Sample Means** tends to be a Normal Distribution, if the sample size is large enough. Even when the original distribution of data is not normal. The procedures, and concepts, discussed here present the foundation of very meaningful applications in Statistics.

When selecting a random sample form a Population (any Distribution) with mean μ and standard deviation σ we need to understand these principles:

- If $n > 30$, then the **sample means $\bar{x}$** have a distribution that can be **approximated** by a Normal Distribution with **mean μ** and **standard deviation $\frac{\sigma}{\sqrt{n}}$**.

Let $\bar{X}$ represents the **sample distribution of the means** and $\mu_{\bar{X}} = \mu$ with $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$



Sampling Distribution of Means

- If $n \leq 30$ and the original population has a Normal Distribution, then the sample means have a distribution that can be approximated by a Normal Distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

- If $n \leq 30$ and the original population is not from a Normal Distribution, then the Central Limit Theorem does not apply.

Definition of a Statistic (Estimator)

The **measured values** of a sample mean, sample proportion, sample variance, and a sample standard deviation, used to infer the value of a corresponding **population parameter**.

$$\bar{x}, \bar{p}, s^2, s$$

Definition of a Population Parameter

The **true value** of a mean, proportion, variance, and standard deviation from **Population**. The values associated with the taking of a Census.

$$\mu, p, \sigma^2, \sigma$$

Practical use of the Central Limit Theorem

When working with the mean from a sample, verify that the Normal Distribution can be used by **confirming that the original Population has a Normal Distribution or the sample size $n > 30$.**

Working with a Mean from a Sample (group) of Values

$z = \frac{x-\mu}{\sigma}$ becomes $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ for a group of values

The Rare Event Rule

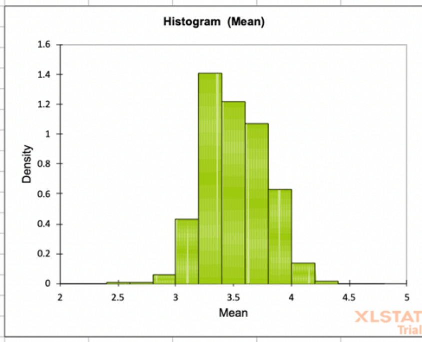
If under a given assumption, we observe the probability of a particular event is **exceptionally small**, we conclude that the original assumption is **probably not correct**. The **Central Limit Theorem** and the concept of a **rare event** is the **basis for Hypothesis Testing** in future topics of study in Statistics.

A closer look at $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

Consider the 500 simulations of Rolling a Die (**Uniform Distribution**) with **sample size 40**. We know the Population Mean is 3.5 and the Population Standard Deviation is 1.732. Please refer to the **Uniform Distribution Simulation** notes. The Central Limit Theorem applies as the sample size $n > 30$. The **sampling distribution of the means** reveals the following.

Summary statistics:

Variable	Observations	Obs. with missing	Obs. without	Minimum	Maximum	Mean	Std. deviation
Mean	500	0	500	2.500	4.350	3.499	0.265



Descriptive statistics for the intervals (Mean):

Lower bound	Upper bound	Frequency	Relative frequency	Density
2	2.2	0	0.000	0.000
2.2	2.4	0	0.000	0.000
2.4	2.6	1	0.002	0.010
2.6	2.8	1	0.002	0.010
2.8	3	6	0.012	0.060
3	3.2	43	0.086	0.430
3.2	3.4	141	0.282	1.410
3.4	3.6	122	0.244	1.220
3.6	3.8	107	0.214	1.070
3.8	4	63	0.126	0.630
4	4.2	14	0.028	0.140
4.2	4.4	2	0.004	0.020
4.4	4.6	0	0.000	0.000
4.6	4.8	0	0.000	0.000

$$\mu = 3.5$$
$$\sigma = 1.732$$

$$\mu_{\bar{X}} = 3.495$$
$$\sigma_{\bar{X}} = 0.265$$

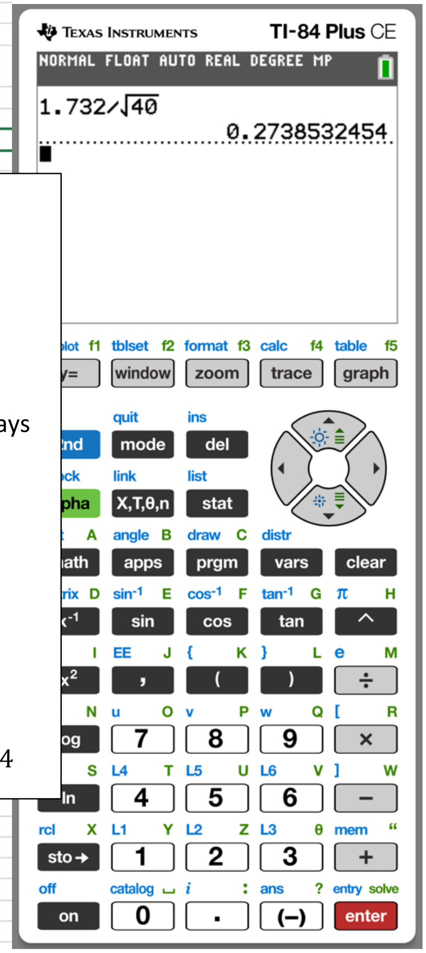
Central Limit Theorem Says

$$\mu_{\bar{X}} = \mu$$

$$3.495 \approx 3.5$$

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

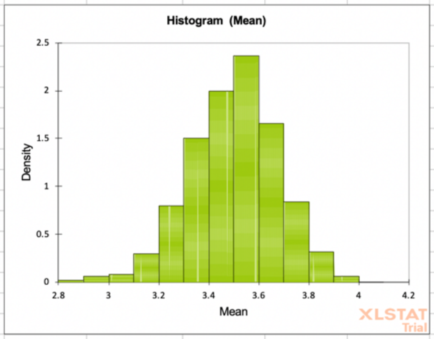
$$0.265 = \frac{1.732}{\sqrt{40}} \approx 0.274$$



Consider the 500 simulations of Rolling a Die with sample size 100. We know the Population Mean is 3.5 and the Population Standard Deviation is 1.732.

Summary statistics:

Variable	Observations	Obs. with missing	Obs. without	Minimum	Maximum	Mean	Std. deviation
Mean	500	0	500	2.870	3.990	3.495	0.174



Descriptive statistics for the intervals (Mean):

Lower bound	Upper bound	Frequency	Relative frequency	Density
2.8	2.9	1	0.002	0.020
2.9	3	3	0.006	0.060
3	3.1	4	0.008	0.080
3.1	3.2	15	0.030	0.300
3.2	3.3	40	0.080	0.800
3.3	3.4	75	0.150	1.500
3.4	3.5	100	0.200	2.000
3.5	3.6	118	0.236	2.360
3.6	3.7	83	0.166	1.660
3.7	3.8	42	0.084	0.840
3.8	3.9	16	0.032	0.320
3.9	4	3	0.006	0.060
4	4.1	0	0.000	0.000

$$\mu = 3.5$$
$$\sigma = 1.732$$

$$\mu_{\bar{X}} = 3.495$$
$$\sigma_{\bar{X}} = 0.265$$

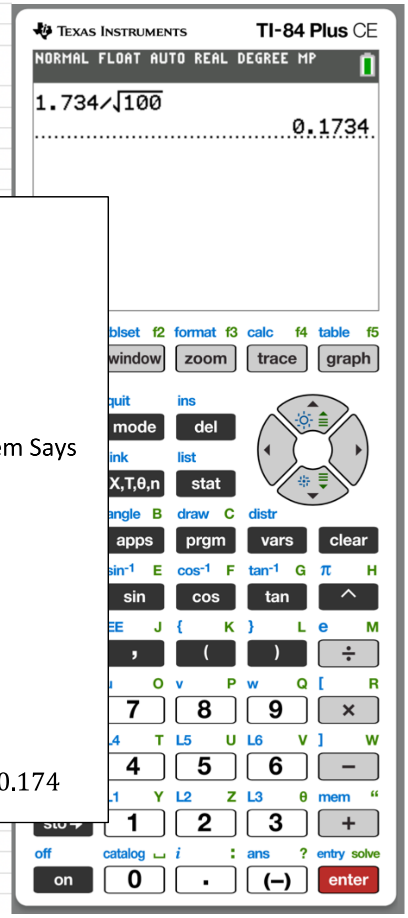
Central Limit Theorem Says

$$\mu_{\bar{X}} = \mu$$

$$3.495 \approx 3.5$$

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

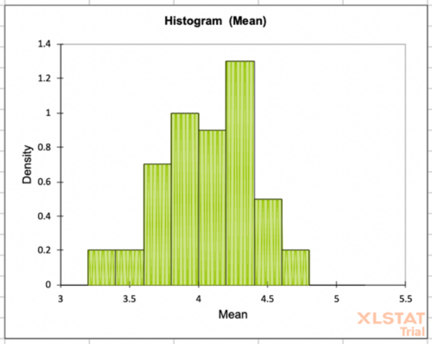
$$0.174 = \frac{1.734}{\sqrt{100}} \approx 0.174$$



Consider the 50 simulations of sample size 40 for the Poisson Distribution (student arrivals) with $\mu = 4$ and $\sigma = 2$.

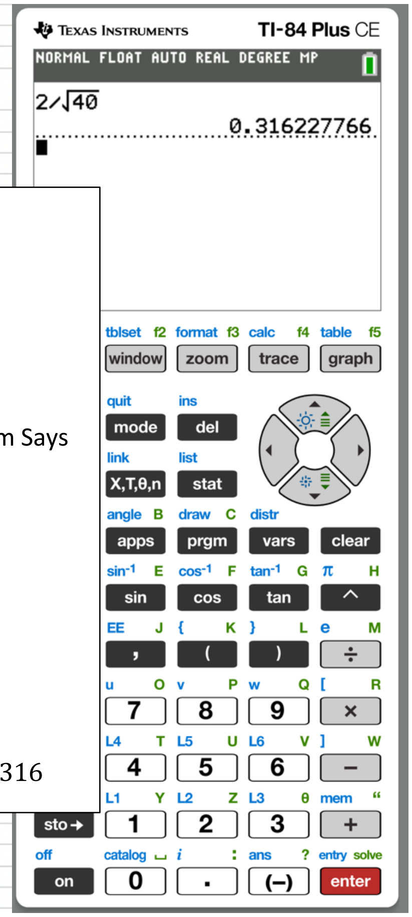
Summary statistics:

Variable	Observations	Obs. with missing	Obs. without	Minimum	Maximum	Mean	Std. deviation
Mean	50	0	50	3.275	4.725	4.054	0.321



Descriptive statistics for the intervals (Mean):

Lower bound	Upper bound	Frequency	Relative frequency	Density
3	3.2	0	0.000	0.000
3.2	3.4	2	0.040	0.200
3.4	3.6	2	0.040	0.200
3.6	3.8	7	0.140	0.700
3.8	4	10	0.200	1.000
4	4.2	9	0.180	0.900
4.2	4.4	13	0.260	1.300
4.4	4.6	5	0.100	0.500
4.6	4.8	2	0.040	0.200
4.8	5	0	0.000	0.000
5	5.2	0	0.000	0.000



$$\mu = 4$$
$$\sigma = 2$$

$$\mu_{\bar{X}} = 4.054$$
$$\sigma_{\bar{X}} = 0.321$$

Central Limit Theorem Says

$$\mu_{\bar{X}} = \mu$$

$$4.054 \approx 4$$

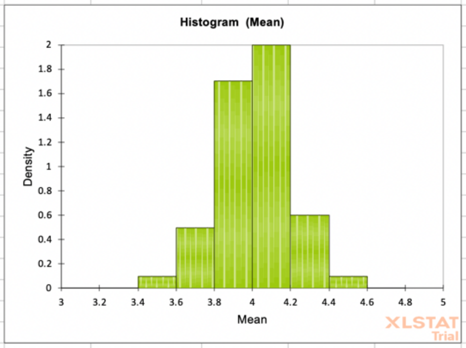
$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

$$0.321 = \frac{2}{\sqrt{40}} \approx 0.316$$

Consider the 50 simulations of sample size 100 for the Poisson Distribution (student arrivals) with $\mu = 4$ and $\sigma = 2$.

Summary statistics:

Variable	Observations	Obs. with missing data	Obs. without missing data	Minimum	Maximum	Mean	Std. deviation
Mean	50	0	50	3.460	4.440	4.003	0.194



Descriptive statistics for the intervals (Mean):

Lower bound	Upper bound	Frequency	Relative frequency	Density
3	3.2	0	0.000	0.000
3.2	3.4	0	0.000	0.000
3.4	3.6	1	0.020	0.100
3.6	3.8	5	0.100	0.500
3.8	4	17	0.340	1.700
4	4.2	20	0.400	2.000
4.2	4.4	6	0.120	0.600
4.4	4.6	1	0.020	0.100
4.6	4.8	0	0.000	0.000

$$\mu = 4$$
$$\sigma = 2$$

$$\mu_{\bar{X}} = 4.054$$
$$\sigma_{\bar{X}} = 0.321$$

Central Limit Theorem Says

$$\mu_{\bar{X}} = \mu$$

$$4.054 \approx 4$$

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

$$0.194 = \frac{2}{\sqrt{100}} \approx 0.200$$

Definition Standard Error of the Mean

$$SEM = \sigma_{\bar{X}}$$

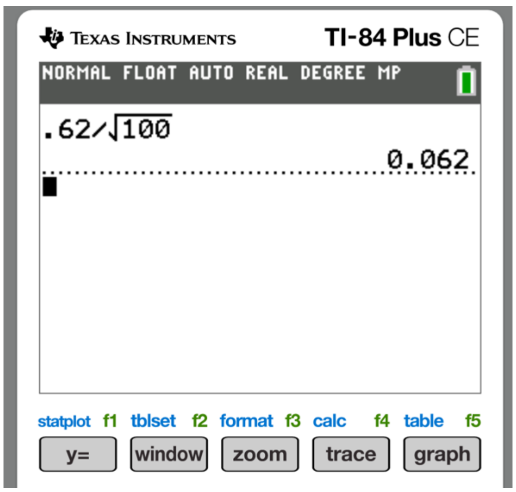
The Rare Event Rule Application for the Central Limit Theorem

If under a given assumption, we observe the probability of a particular event is **exceptionally small**, we conclude that the original assumption is **probably not correct**. The **Central Limit Theorem** and the concept of a **rare event** is the **basis for Hypothesis Testing** in future topics of study in Statistics.

It is widely believed that the body temperature of human beings is 98.6^{o}F.

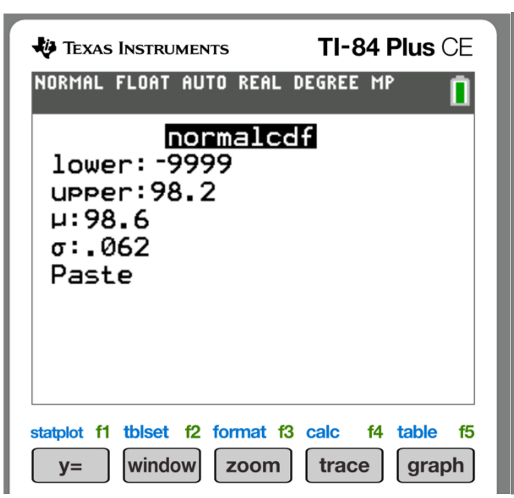
However, researchers have gathered a sample of 200 body temperatures and obtained a mean of 98.2^{o}F with a standard deviation of 0.62^{o}F. If we **assume the human body temperature is 98.6^{o}F**, what's the probability a sample size of 100 data values will be no more than 98.2^{o}?

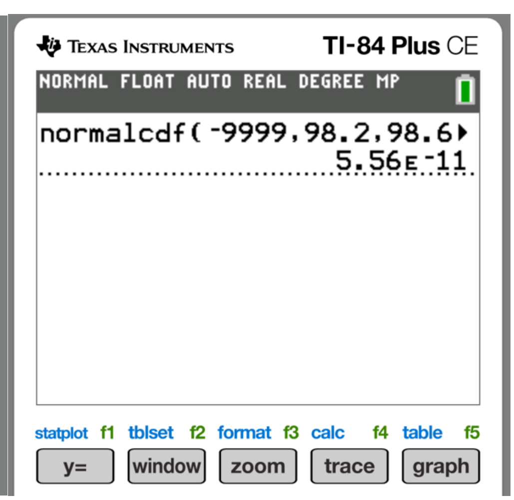
$$\mu_{\bar{x}} = \mathbf{98.6^{o}F} \text{ (assumption)}$$
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{0.62}{\sqrt{100}} \approx 0.062$$



$$p(x \leq 98.2) \approx 0.00000000006 \approx 0.000$$

Not likely, the original assumption is incorrect!





Race Based Life Expectancy

https://www.thebalance.com/the-racial-life-expectancy-gap-in-the-u-s-4588898

The life Expectancy in the United States is 78.9 years according to the CDC in 2014. However, different ethnicities have different Life Expectancy rates. They even differ by States in the Union. The causes of Death also vary from ethnicity to ethnicity and by State.

Native Americans live on average of 75.06 years
African Americans live on average of 75.54 years
White Americans live on average of 79.12 years
Hispanic Americans live on Average of 82.89 years
Asian Americans live on Average of 86.67 years

We will use the Central Limit Theorem to determine probabilities for a group of individuals from the same population using information from https://www.worldlifeexpectancy.com/usa-cause-of-death-by-age-and-gender

If the Life Expectancy of all races in the United States is Normally distributed with a mean of 78.86 years and a standard deviation of 15.9 years (this is an assumption):
https://www.thinkingaheadinstitute.org/en/News/Public/News/2018/12/A-basic-question-about-life-expectancy-that-even-actuaries-struggle-to-answer

Approximate Answers to the Nearest Ten Thousandths

1. What's the probability 5 randomly selected people living at least 90 years of age?
2. What's the probability 10 randomly selected people living less than 60 years of age?
3. What's the probability 20 randomly selected people living between 60 years and 90 years of age?
4. What's the probability 35 randomly selected people living more than 100 years of age?
5. What's the probability 50 randomly selected people living between 60 and 70 years of age?

If the Life Expectancy of Hispanic Americans in California is 83.17 years with a standard deviation of 15.9 years (this is an assumption): https://www.worldlifeexpectancy.com/usa/life-expectancy-hispanic

Approximate Answers to the Nearest Ten Thousandths

6. What's the probability of 5 randomly selected Hispanic Americans in California living no more than 65 years?
7. What's the probability of 10 randomly selected Hispanic Americans in California living more than 90 years?
8. What's the probability of 20 randomly selected Hispanic Americans in California living between 65 and 90 years of age?
9. What's the probability of 35 randomly selected Hispanic Americans in California living less than 45 years?
10. What's the probability of 50 randomly selected Hispanic Americans in California living between 65 and 85 years of age?

If the Life Expectancy of African America Male from Alabama is 69.01 years with a standard deviation of 16.5 years (this is an assumption): https://www.worldlifeexpectancy.com/usa/life-expectancy-hispanic
Approximate Answers to the Nearest Ten Thousandths

11. What's the probability of 5 randomly selected Asian America Females living at least 80 years of age?
12. What's the probability of 10 randomly selected Asian America Females living less than 50 years of age?
13. What's the probability of 20 randomly selected Asian America Females living between 50 and 80 years of age?
14. What's the probability of 35 randomly selected Asian America Females living more than 95 years?
15. What's the probability of 50 randomly selected Asian America Females living between 95 and 100 years of age?

If the Life Expectancy of African American Males living in Alabama is 88.89 years with a standard deviation of 18.2 years (this is an assumption): https://www.worldlifeexpectancy.com/usa/life-expectancy-hispanic
Approximate Answers to the Nearest Ten Thousandths

16. What's the probability of 5 randomly selected African America Male from Alabama living more than 60 years of age?
17. What's the probability of 10 randomly selected African America Males from Alabama living less than 70 years of age?
18. What's the probability of 15 randomly selected African America Male from Alabama living between 80 and 100 years of age?
19. What's the probability of 35 randomly selected African America Male from Alabama living less than 45 years of age?
20. What's the probability of 50 randomly selected African America Male from Alabama living no more than 75 years of age?