CHAPTER FOURTEEN

The Central Limit Theorem

Suppose you have a small computer business with five hourly employees. You have an employee value system that goes from 0 to 10.

Name	Value
Al	9
Bob	7
Carl	8
Dot	7
Ely	9

In an attempt to ascertain the population (all 5 workers) mean, you sample 2 workers, find their mean, and use that as an estimate of the population mean.

Logically, it is unlikely that a sample mean would be identical to the population mean. The difference between the sample statistic and the population parameter is called the sampling error, which is due to chance.

For instance, suppose our sample contained Carl and Ely:

$\bar{X} = \frac{8+9}{2} = 8.5$ while $\mu = \frac{9+7+8+7+9}{5} = 8$

Due to this sampling error, how do we trust our statistic as an estimate of a parameter?

If we organized the means of all possible samples of size 2 into a probability distribution, we would logically obtain the sampling distribution of the means.

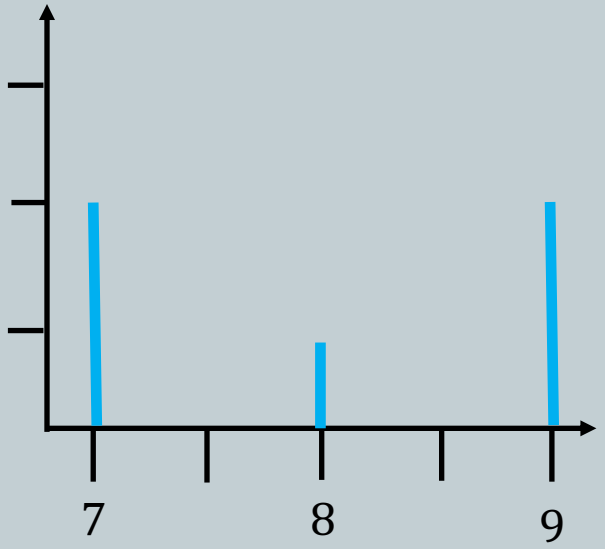
$$\binom{5}{2} = 10$$

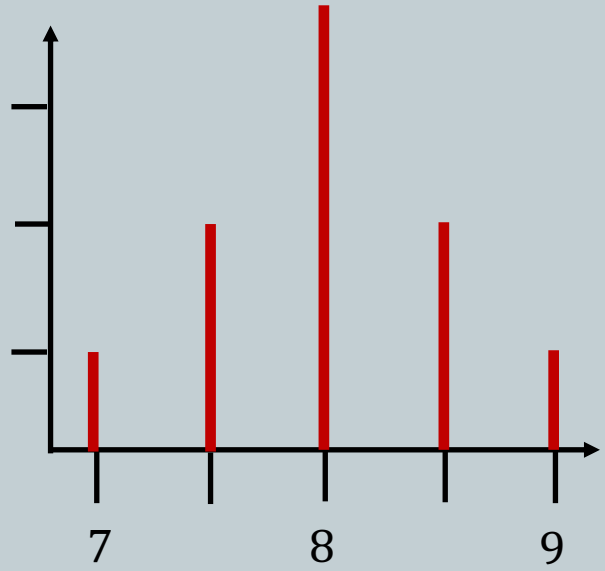
Name	Rank
Al	9
Bob	7
Carl	8
Dot	7
Ely	9

$\mu = 8$

Sample	mean
A, B	8
A, C	8.5
A, D	8
A, E	9
B, C	7.5
B, D	7
B, E	8
C, D	7.5
C, E	8.5
D, E	8

$\mu_{\bar{x}} = 8$





Of what importance is this sampling distribution of the means?

The previous example was intentionally kept small to emphasize that the mean of the sample means is exactly equal to the mean of the population, and to emphasize that the shape of the distribution of sample means is not necessarily the same as that of the population.

Why is the tendency toward the normal distribution so important?

The Central Limit Theorem: For a population with a mean of μ and a population standard deviation σ, the sampling distribution of the means of all possible samples of size n generated from the population will be approximately normally distributed – with the mean of the sampling distribution equal to μ and the standard deviation equal to $\frac{\sigma}{\sqrt{n}}$ – assuming that the sample size is sufficiently large.

BTW – there is no common agreement as to what constitutes a "sufficiently large" sample size. Some statisticians say 30; others go as low as 12, the example concerning employee value worked quite well with a sample size of 2.

Important facts about the CLT:

1] If the sample size n is sufficiently large, the sampling distribution of the means will be approximately normal, regardless of parent distribution.

2] The mean of the population, μ, and the mean of all possible samples, $\mu_{\bar{X}}$, are equal. If the population is large and a large number of samples are selected from that population, then the mean of the sample means will be close to the population mean.

3] The standard deviation of the distribution of sample means is determine by $\frac{\sigma}{\sqrt{n}}$

Let X_1, X_2, X_3, X_4, X_5 be random variables denoting the result of rolling a fair six-sided die five times. The sample mean of these five random variables is:

$$\bar{X} = \frac{X_1 + X_2 + X_3 + X_4 + X_5}{5}$$

$\bar{X}$ itself is also a random variable:

$\bar{X} = 3.5$ is a possible outcome

$\bar{X} = 2$ is another possible outcome, etc.

Now imagine that we rolled a die 5 times and obtained the series of values: $4, 6, 1, 5, 4$

The sample mean for these 5 values is given by: $\bar{x} = \frac{4+6+1+5+4}{5} = 4$

The sample mean $\bar{x}$ of these 5 particular numbers is not a random variable.

$\bar{x}$ is a single number.

An event occurred where $X_1 = 4,\ X_2 = 6,\ X_3 = 1, X_4 = 5$, and $X_5 = 4$.

Lowercase letters are often numbers, Uppercase letters are often random variables.

Let's perform a mind-experiment:

Imagine a heard of buffalo.

You want to determine the average weight of a buffalo from the herd (the population.) We call this mean the population mean, or μ.

However, we cannot possibly weigh each and every buffalo to discover μ.

Instead, we catch 5 buffalo and weigh them. we then find the mean weight of those 5 buffalo. Let's call this $\bar{X}_1$.

Now, return those 5 buffalo to the herd and catch a second sample of 5 buffalo. Find this samples mean weight and call it $\bar{X}_2$.

Continue this process a number of times to produce: $\bar{X}_1, \bar{X}_2, \bar{X}_3, \ldots, \bar{X}_n$

Consider the weights of the buffalo in a number of samples:

$$Smp_1 = [1000, 1100, 900, 800, 1050] \qquad \bar{X}_1 = 970, \quad S^2 = 65{,}100$$
$$Smp_2 = [850, 1200, 950, 700, 1000] \qquad \bar{X}_2 = 940, \quad S^2 = 141{,}000$$
$$Smp_3 = [900, 1000, 950, 900, 850] \qquad \bar{X}_3 = 920, \quad S^2 = 13{,}000$$
$$\vdots$$

The values of the sample mean, $\bar{X}$, and the sample variance, S^2, depend on the selected random sample.

That is to say, $\bar{X}$ and S^2, are continuous random variables in their own right. Therefore, they themselves should each have a particular:

probability distribution, mean, and standard deviation.

The variability seen in sample means would diminish if larger samples were taken.

The sampling distribution of the sample means will be approximately normal.

How normal? Well, how does one measure normality?

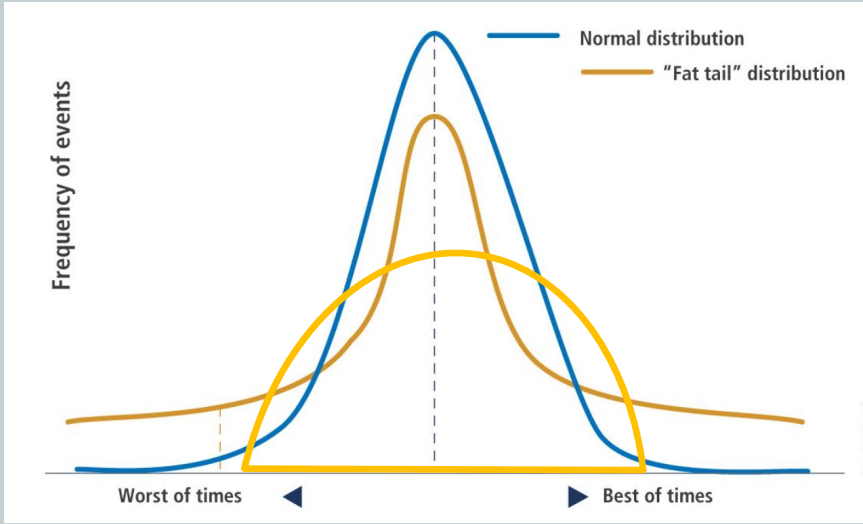
Skew:

right (positive) skew and left (negative) skew

Kurtosis:

positive implies fat tails and pointy peak

negative implies smaller tails and rounded top



The Central Limit Theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement, then the distribution of the sample means will be approximately normally distributed.

This holds true regardless of whether the source population is normal or skewed, provided the sample size is sufficiently large (usually $n \geq 30$).

This infers that we can use the normal probability model to quantify uncertainty when making inferences about a population mean, μ, based on the sample mean, $\bar{X}$.

For the random samples we take from the population, we can compute the mean of the sample means: $\mu_{\bar{X}}$

$$\mu_{\bar{X}} = \frac{\bar{X}_1 + \bar{X}_2 + \bar{X}_3 + \cdots + \bar{X}_n}{n} = \mu$$

The mean of the sampling distribution of the mean is the mean of the population from which the scores were sampled.

Therefore, if a population has a mean μ, then the mean of the sampling distribution of the mean is also μ.

We can also compute the variance of the sample means:

$$\sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \qquad \text{Therefore,} \quad \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

Why is the variance divided by n ?

Larger samples implies less variability

Visual on the process of sampling and the Central Limit Theorem:
https://onlinestatbook.com/stat_sim/sampling_dist/

The above link was developed by Rice university.

It is a great visualization of the CLT.

To reiterate:

The sampling distribution of the sample means approaches a normal distribution as the sample size gets larger.

No matter what the shape of the population distribution.

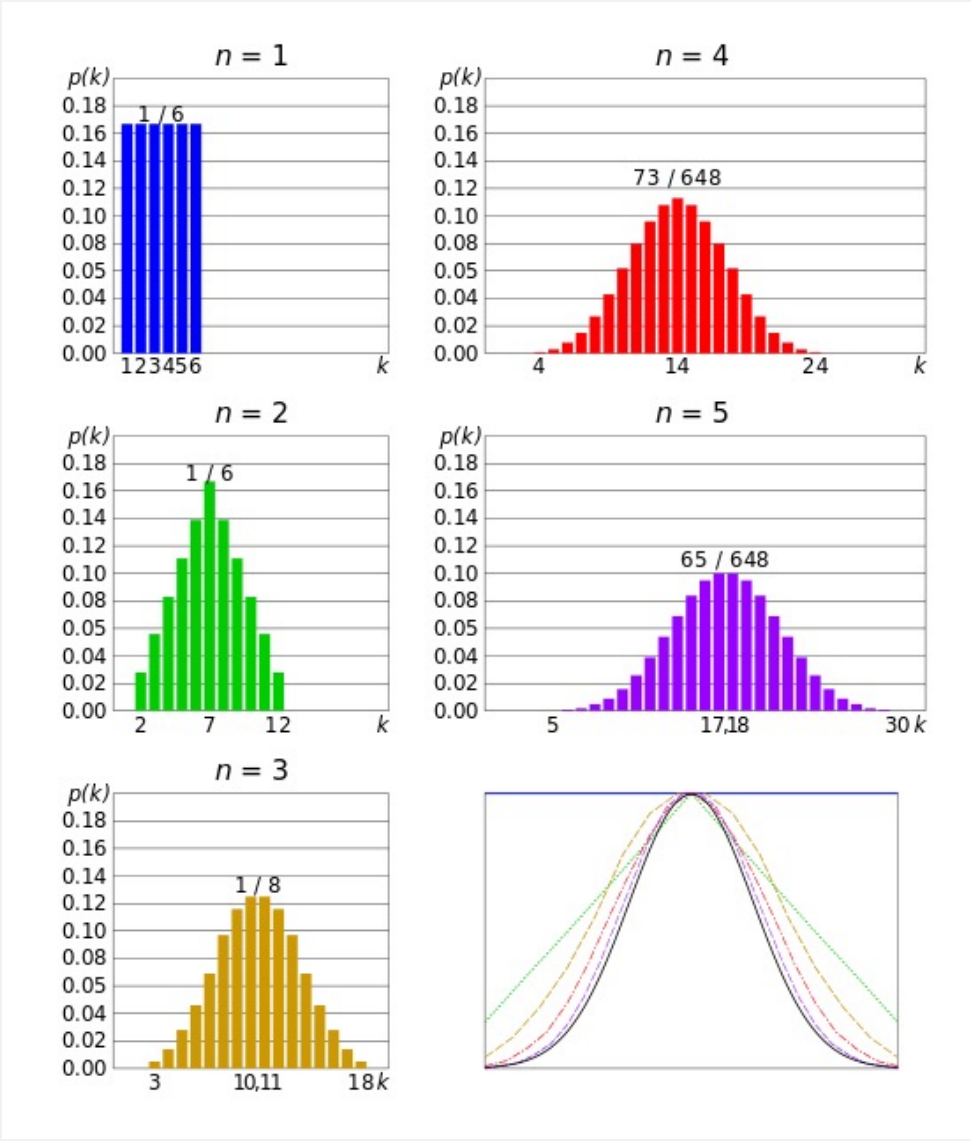
"This fact holds especially true for sample sizes over 30."

All this is saying is that as you take more samples, especially large ones, your graph of the sample means will look more like a normal distribution.

The CLT; Graphically

The various graphs show the simple experiment of rolling a fair die.

The more times you roll the die, the more likely the shape of the distribution of the means tends to look like a normal distribution graph.



VOCABULARY

Consider a population of N elements.

Population {4, 3, 2, 3, 3, 1, 4, 1, 4, 5}

A population has parameters (mean and variance): μ, σ^2 (3, 1.6)

You then sample n elements of that population.

Sample {1, 2, 3, 3, 4, 4}

A sample has statistics (mean and variance): $\bar{x}$, s^2 (2.83, 1.36)

What is the sample mean, $\bar{x}$?

What is the sample variance, s?

These are estimates of the population parameters.

BTW, `numpy.var(set_x, ddof = 1)`

K

Population standard deviation:

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$

Sample standard deviation:

$$s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

Random Samples

The random variables $X_1, X_2, X_3, \ldots, X_n$ form a random sample of size n, if

1] All X_k are independent

2] All X_k came from the same distribution; "identically distributed".

Summarizing our iid sample:

$\bar{X}$ is the sample mean estimator of our population mean μ.

$\hat{p}$ is the sample proportion

in the sample satisfying some characteristic of interest / # in sample

s^2 is the sample estimator for σ^2.

Suppose $X_1, X_2, X_3, \ldots, X_n \sim N(\mu, \sigma^2)$

Then $\bar{X} \sim N(\mu, {}^{\sigma^2}/_{n})$

Estimators and their distributions
Any estimator, including the sample mean, is a random variable
(since it is based on a random sample)

The sampling distribution depends on:
1] the population distribution
2] sample size, n
3] The method of sampling

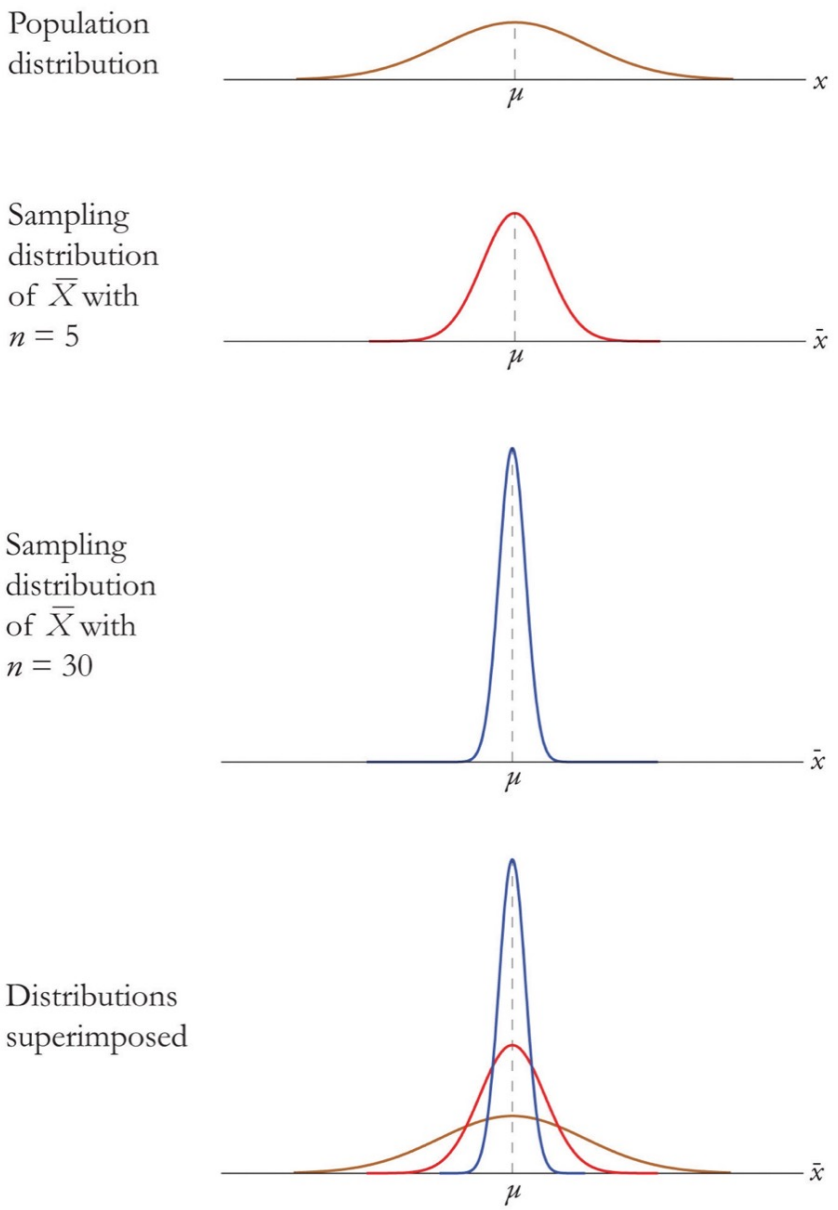


***Figure* 6.2.4**: *Distribution of Sample Means for a Normal Population*

$$Z = \frac{\bar{X} - \mu}{\sigma} \qquad \text{versus} \qquad Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

Which one should you use?

Technically you always divide by $\sqrt{n}$. However, occasionally the square root of n sometimes equals 1.

For example, if you are choosing one person and trying to figure out the probability their weight is under x pounds, then $n = 1$.
In other words, if you are calculating a z-score, you can always use $\sqrt{n}$.

$$Z = \frac{\bar{X} - \mu}{\sigma} \qquad \text{versus} \qquad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Why do we use $\sigma/\sqrt{n}$?

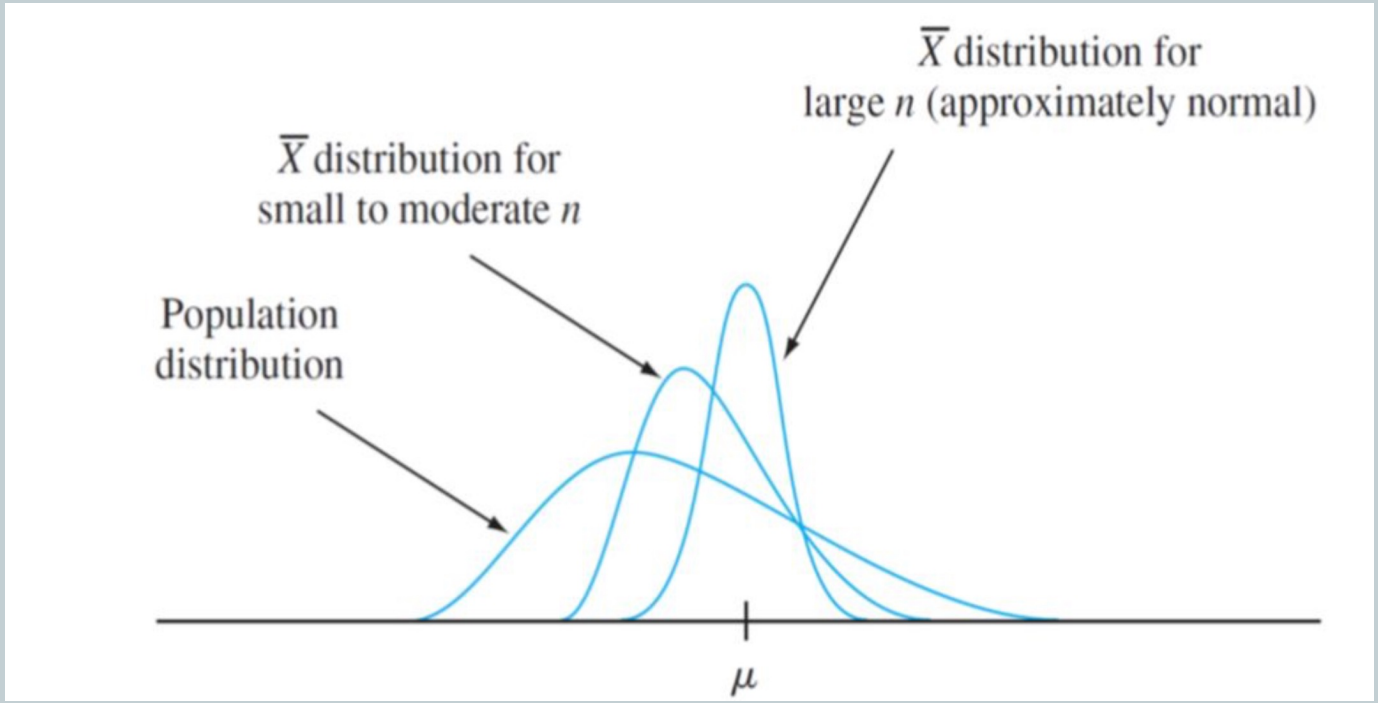
When you are estimating the standard error for the mean
(i.e., the standard deviation of the means of samples)
the larger your sample size, the smaller the standard deviation.

For example, if you took a sample of 200, you would be much more likely to get close to the true mean than if you took a sample of 2.

The larger your n, the smaller the standard deviation.

Pro: Larger samples produce more accurate results.

Con: Larger samples require more time and money.



Due to COVID a political policy was put in place to offset the unemployment situation. Qualifying recipients receive government funds in various amounts. The average is $110 per week with a standard deviation of $20.
In order to determine the proper transference of funds, a random sample of 25 people is taken. What is the probability their mean benefit will be greater than $120 per week?

Due to COVID a political policy was put in place to offset the unemployment situation. Qualifying recipients receive government funds in various amounts. The average is $110 per week with a standard deviation of $20.
In order to determine the proper transference of funds, a random sample of 25 people is taken. What is the probability their mean benefit will be greater than $120 per week?

ANS:

$\mu = 110 \quad \sigma = 20 \quad n = 25 \quad P(\bar{X} > 120) = \ ?$

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = 2.5$$

$$P(\bar{X} > 25) = 1 - P(\bar{X} \leq 2.5) = 0.0062$$

A population of 17-year-old fast-food workers has a mean salary of $\$29{,}321$ with a standard deviation of $\$2120$.

If a sample of 100 workers is taken, what is the probability that their mean salaries will be less than $\$29{,}000$?

A population of 17-year-old fast-food workers has a mean salary of $\$29{,}321$ with a standard deviation of $\$2120$.

If a sample of 100 workers is taken, what is the probability that their mean salaries will be less than $\$29{,}000$?

ANS:

$\mu = 29321 \quad \sigma = 2120 \quad n = 100 \quad P(\bar{X} < 29000) = \ ?$

$$Z = \frac{29000 - 29321}{2120/10} = -1.514$$

$$P(Z \leq -1.514) = 0.0655$$

There are 250 dogs at a dog show who weigh an average of 12 pounds, with a standard deviation of 8 pounds.
If 4 dogs are chosen at random, what is the probability they have an average weight of greater than 8 pounds and less than 25 pounds?

ANS:

$N = 250 \quad \mu = 12 \quad \sigma = 8 \quad n = 4 \quad P(8 < \bar{X} < 25) = \ ?$

$Z_8 = \frac{8 - 12}{8/2} = -1 \qquad Z_{25} = \frac{25 - 12}{8/2} = 3.25$

$P(Z < 3.25) = 0.9994 \qquad P(Z < -1) = 0.1587$

$0.9994 - 0.1587 = 0.8407$

Let $X_1, X_2, \ldots$ be any sequence of independent identically distributed random variables with finite positive variance.

Let μ be the expected value and σ^2 the variance of each X_i.

For $n \geq 1$, let $Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$

Then for any number a,

$$\lim_{n \to \infty} F_{Z_n}(a) = \Phi(a)$$

where Φ is the distribution function of the $N(0, 1)$ distribution.

Next time, confidence intervals!