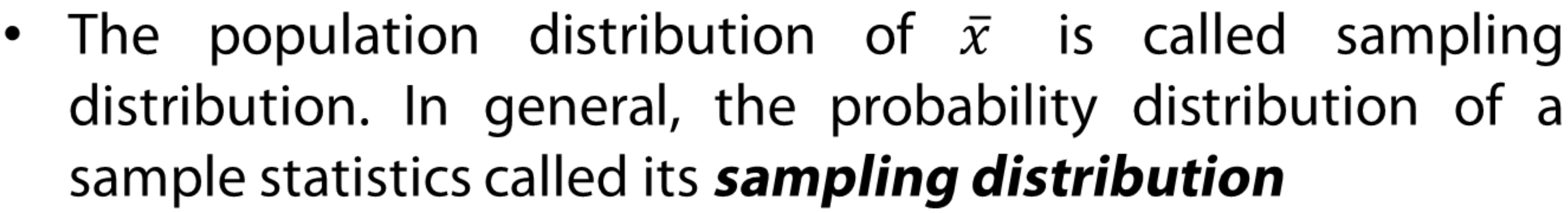
CHAPTER 6

Sampling & Sampling Distribution

Population Distribution

- The population distribution is the probability distribution of the population data

Sampling Distribution



- The population distribution of $\bar{x}$ is called sampling distribution. In general, the probability distribution of a sample statistics called its ***sampling distribution***

Sampling…

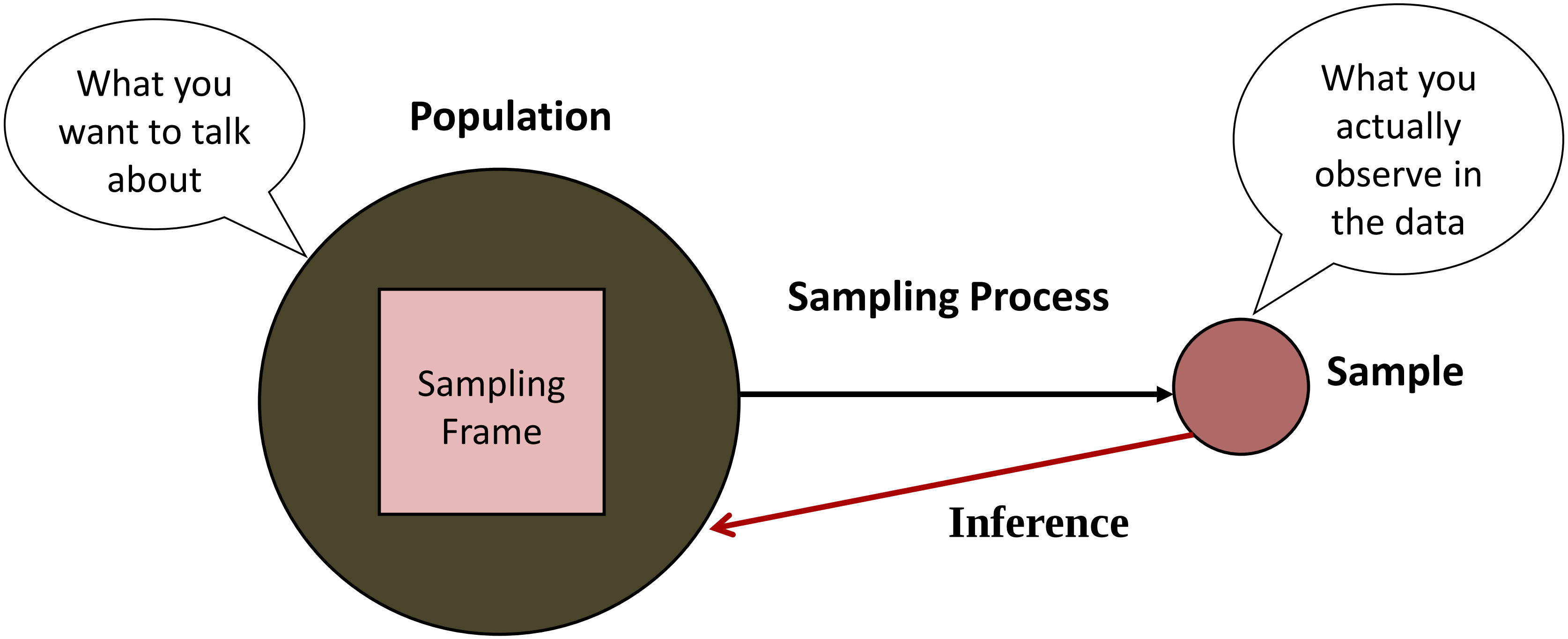
3 factors that influence sample representative-ness

- Sampling procedure
- Sample size
- Participation (response)

When might you sample the entire population?

- When your population is very small
- When you have extensive resources
- When you don’t expect a very high response

What is Sampling?



What you want to talk about

Population

Sampling Frame

Sampling Process

What you actually observe in the data

Sample

Inference

A **sampling frame** is the source material or device from which a sample is drawn. It is a list of all those within a population who can be sampled, and may include individuals, households or institutions.

Example

Suppose there are 5 students in an advanced statistics class and the midterm scores of these 5 students are

70 78 80 80 95

What is the population mean ? Consider all possible samples of three scores each that can be selected without replacement form that population?

Errors…

Two major types of error can arise when a sample of observations is taken from a population:

i. Sampling error
ii. Non - sampling error.

Sampling Error…

Sampling error refers to differences between the sample and the population that exist only because of the observations that happened to be selected for the sample.

$$\textbf{Sampling Error} = \bar{x} - \mu$$

- Sampling error occurs because researchers draw different subjects from the same population but still, the subjects have individual differences.

Increasing the sample size **will** reduce this type of error.

Non - sampling Error…

Nonsampling errors are more serious and are due to mistakes made in the acquisition of data or due to the sample observations being selected improperly. Three types of nonsampling errors:

i. Errors in data acquisition,

ii. Nonresponse errors, and

iii. Selection bias.

Increasing the sample size **will** reduce this type of error.

Example:

Suppose there are 5 students in an advanced statistics class and the midterm scores of these 5 students are

70 78 80 80 95

Find the sampling error?

Sampling Distribution of The Mean

A sampling distribution is created by, as the name suggests, ***sampling***.

The method we will employ on the ***rules of probability*** and the ***laws of expected value and variance*** to derive the sampling distribution;

- The mean of the sample means is equal to the mean of the population from which the samples were drawn.
- The variance of the distribution is s divided by the square root of n. (the standard error.)

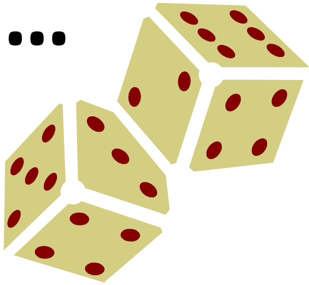
Mean if the Sampling Distribution of $\overline{\boldsymbol{x}}$: Is always equal to the mean of the population. Thus,

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

Standard Deviation of the Sampling Distribution of $\overline{\boldsymbol{x}}$:

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

Sampling Distribution of the Mean…



A fair **die** is thrown infinitely many times,

with the random variable X = # of spots on any throw.

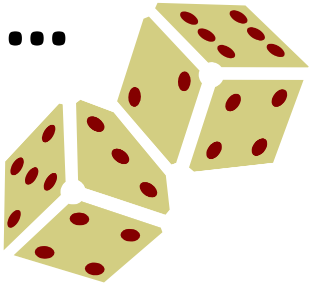
i. How many samples of size 2 are possible?

ii. List all possible samples of size 2, and compute the mean of each sample.

iii. Compute the mean of the sample means and the population mean. Compare the two values.

iv. Compare the dispersion in the population with that of the sample means.

Sampling Distribution of the Mean…



A fair **die** is thrown once,

The probability distribution of X is:

x	1	2	3	4	5	6
P(x)	1/6	1/6	1/6	1/6	1/6	1/6

…and the mean and variance are calculated as well:

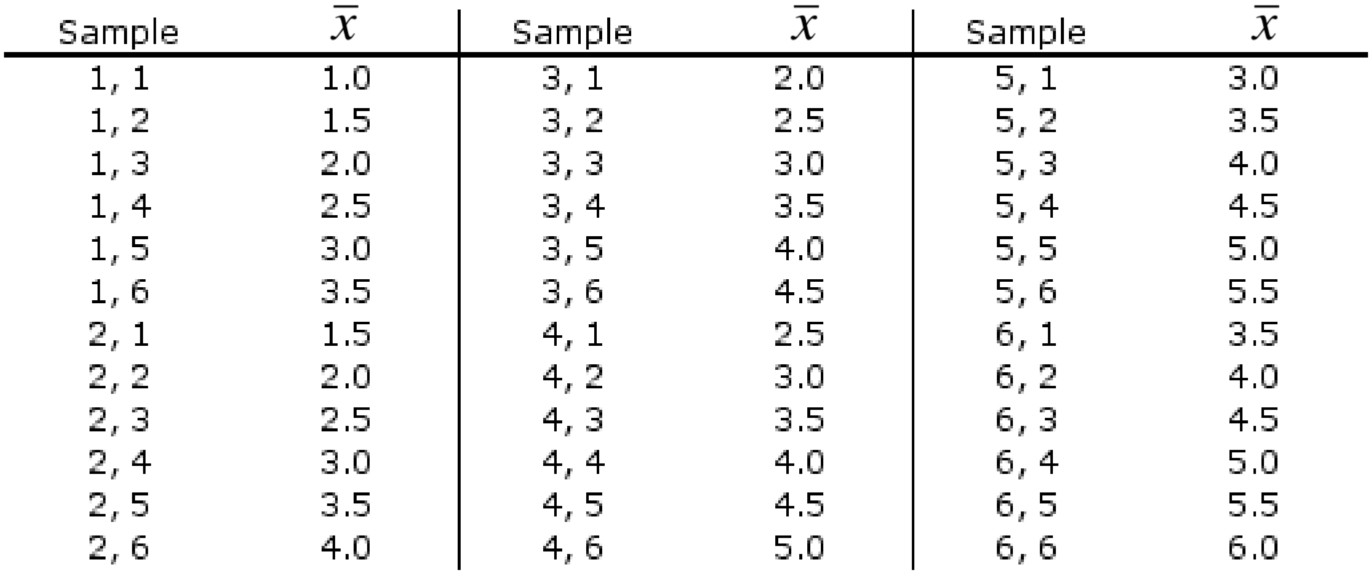
$$\mu = \sum xP(x) = 1(\tfrac{1}{6}) + 2(\tfrac{1}{6}) + \ldots + 6(\tfrac{1}{6}) = 3.5$$

$$\sigma^2 = \sum (x-\mu)^2 P(x) = (1-3.5)^2(\tfrac{1}{6}) + \ldots + (6-3.5)^2(\tfrac{1}{6}) = 2.92$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{2.92} = 1.71$$

Sampling Distribution of Two Dice

A sampling distribution is created by looking at all samples of size n=2 (i.e. two dice) and their means…



Sample	$\bar{x}$	Sample	$\bar{x}$	Sample	$\bar{x}$
1, 1	1.0	3, 1	2.0	5, 1	3.0
1, 2	1.5	3, 2	2.5	5, 2	3.5
1, 3	2.0	3, 3	3.0	5, 3	4.0
1, 4	2.5	3, 4	3.5	5, 4	4.5
1, 5	3.0	3, 5	4.0	5, 5	5.0
1, 6	3.5	3, 6	4.5	5, 6	5.5
2, 1	1.5	4, 1	2.5	6, 1	3.5
2, 2	2.0	4, 2	3.0	6, 2	4.0
2, 3	2.5	4, 3	3.5	6, 3	4.5
2, 4	3.0	4, 4	4.0	6, 4	5.0
2, 5	3.5	4, 5	4.5	6, 5	5.5
2, 6	4.0	4, 6	5.0	6, 6	6.0

While nly 11 values for , and some (e.g. =3.5) occur more frequently than others (e.g. =1).

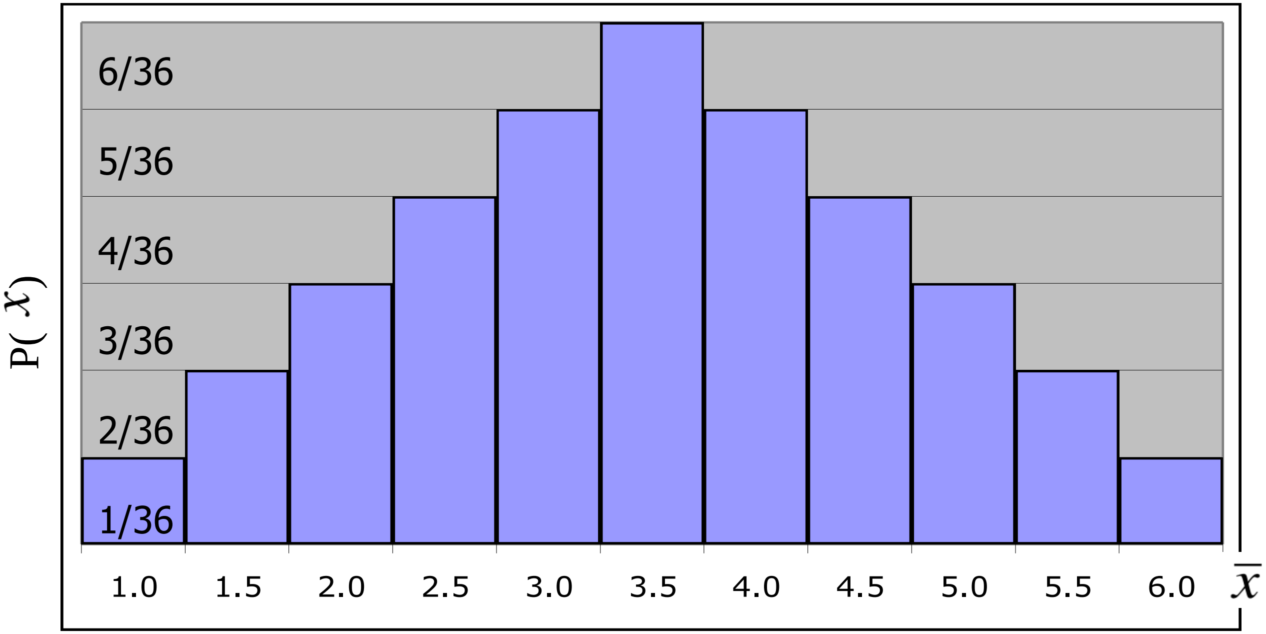
$\bar{x}$ $\bar{x}$

$\bar{x}$

Sampling Distribution of Two Dice…

The ***sampling distribution*** of $\bar{x}$ shown below:

$\bar{x}$	$P(\bar{x})$
1.0	1/36
1.5	2/36
2.0	3/36
2.5	4/36
3.0	5/36
3.5	6/36
4.0	5/36
4.5	4/36
5.0	3/36
5.5	2/36
6.0	1/36



$$\mu_{\bar{x}} = \sum \bar{x}P(\bar{x}) = 1.0(\tfrac{1}{36}) + 1.5(\tfrac{2}{36}) + \ldots + 6.0(\tfrac{1}{36}) = 3.5$$

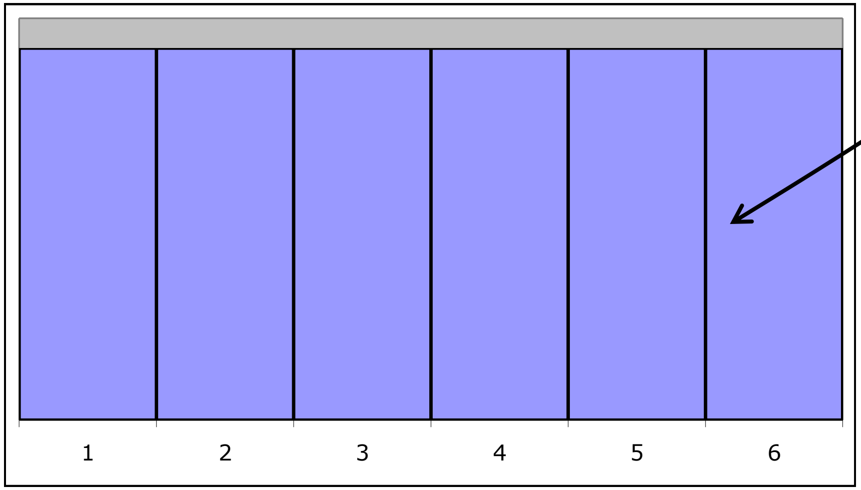
$$\sigma^2_{\bar{x}} = \sum (\bar{x} - \mu_{\bar{x}})^2 P(\bar{x}) = (1.0 - 3.5)^2(\tfrac{1}{36}) + \ldots + (6.0 - 3.5)^2(\tfrac{1}{36}) = 1.46$$

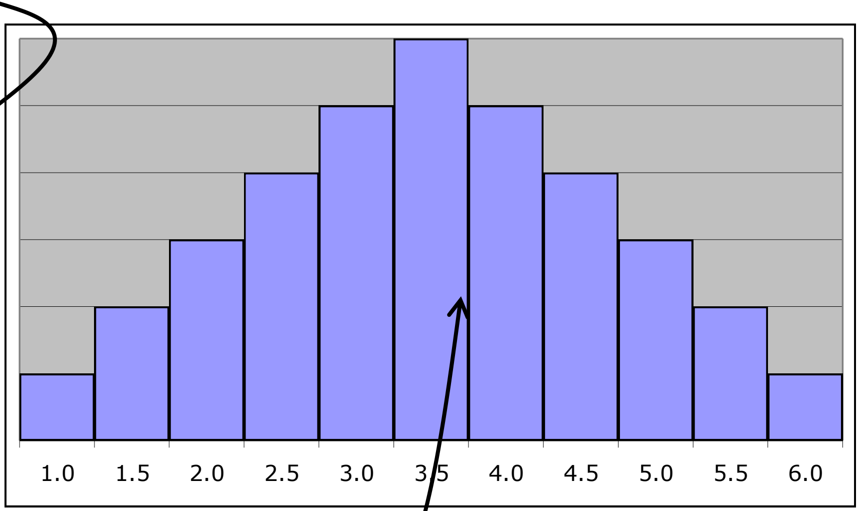
$$\sigma_{\bar{x}} = \sqrt{\sigma^2_{\bar{x}}} = \sqrt{1.46} = 1.21$$

Sampling Distribution of Two Dice…

Compare…

Compare the distribution of X…





…with the sampling distribution of $\bar{x}$

As well, note that:

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

$$\sigma^2_{\bar{x}} = \sigma^2 / 2$$

Generalize…

We can generalize the mean and variance of the sampling of two dice:

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

$$\sigma_{\bar{x}}^2 = \sigma^2 / 2$$

…to **n**-dice:

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

$$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$$

The standard deviation of the sampling distribution is called the ***standard error***:

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

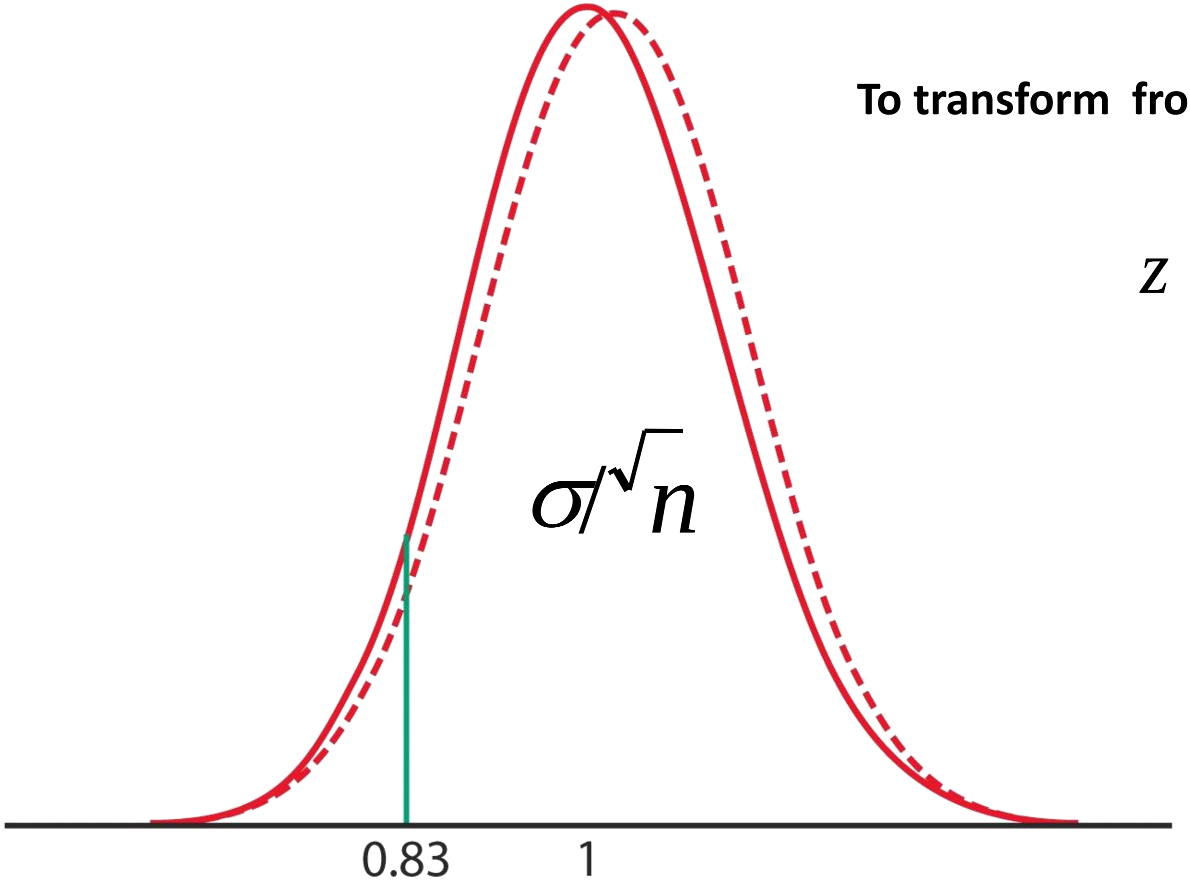
3.2 Central Limit Theorem

- Specifies a theoretical distribution
- Formulated by the selection of all possible random samples of a fixed size *n*
- A sample mean is calculated for each sample and the distribution of sample means is considered

How Large is Large?

- If the sample is **normal**, then the sampling distribution of $\bar{x}$ will also be normal, no matter what the sample size.
- When the sample population is approximately **symmetric**, the distribution becomes approximately normal for relatively small values of *n.*
- When the sample population is **skewed**, the sample size must be **at least 30** before the sampling distribution of $\bar{x}$ becomes approximately normal.

(CLT allows us to use Normal probability calculation to answer questions about the sample means)



To transform from $\bar{x}$ into *z*:

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

Example

A certain brand of tires has a mean life of 25,000 miles with a standard deviation of 1600 miles.

What is the probability that the mean life of 64 tires is less than 24,600 miles?

Example… SOLUTION

The sampling distribution of the means has a mean of 25,000 miles (the population mean)

$$\mu = 25000\,mi$$

and a standard deviation (i.e.. standard error) of:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1600}{\sqrt{64}} = 200$$

Convert 24,600 mi. to a z-score and use the normal table to determine the required probability.

$$z = \left(24600 - 25000\right) / \,200 = -2$$

$$P\left(z < -2\right) = 0.0228$$

or 2.28% of the sample means will be less than 24,600 mi.

Example

The average weekly income of graduates one year after graduation is $600. Suppose the distribution of weekly income has a standard deviation of $100. What is the probability that 25 randomly selected graduates have an average weekly income of less than $550?

Solution

$$P(\bar{x} < 550) = P(\frac{\bar{x} - \mu}{\sigma_{\bar{x}}} < \frac{550 - 600}{100 / \sqrt{25}})$$

$$= P(z < -2.5) = 0.0062$$

If a random sample of 25 graduates actually had an average weekly income of $550, what would you conclude about the validity of the claim that the average weekly income is 600?

Solution

- With m = 600 the probability to have a sample mean of 550 is very low (0.0062). The claim that the average weekly income $600 is probably unjustified.
- It will be more reasonable to assume that m is smaller than $600, because then a sample mean of $550 becomes more probable.