

Section 5.2 The Sampling Distribution of a Sample Mean

Mean

- Most common statistic for quantitative data
- Averages are less variable than the individual observations
- Averages are more normal than individual observations

Mean and Standard Deviation of \bar{X}

Let X_i be an individual selected at random from population with mean, μ , and standard deviation σ . Then the sample mean is

$$\bar{X} = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$$

So

$$\mu_{\bar{X}} = \frac{1}{n}(\mu + \mu + \dots + \mu) = \mu \text{ and } \sigma_{\bar{X}}^2 = \frac{1}{n^2}(\sigma^2 + \sigma^2 + \dots + \sigma^2) = \frac{\sigma^2}{n}$$

Now, \bar{X} is the point estimate of population mean μ . The sample statistic \bar{X} is a sufficient and unbiased estimate of μ . Note that \bar{X} either equals μ or it does not equal μ , could we provide more information to make this estimation more meaningful? Let's look at what we know.

We measure the variability in the sample means by using the **standard error** of the sample means $\frac{\sigma}{\sqrt{n}}$.

The standard error is simply the standard deviation of the sample means. **Notice as the sample size increases, the variability in the sample means decreases.**

If the sample size is large, the sample contains more information about the population than if the sample size is small. Therefore, the sample means won't change as much from sample to sample. Since accuracy means repeatability of results, we know that sample means based on larger samples are more accurate (or repeatable) than sample means that are based on smaller samples.

If a census were conducted, the sample would be the entire population; the resulting mean would have no variability at all. In all other cases, we expect the means to vary by a certain amount.

How does the variability in the population affect the sample means?

As the population standard deviation increases, the variability among the sample means increases as well. This lowers the accuracy of our sample mean.

If we take samples of size n from a population where the original observations are quite homogeneous (don't have much variability), the resulting sample means will not have much variability. They will not change much from sample to sample.

If we take samples of size n from a population where the original observations are quite heterogeneous (a wide variety of possible values), then the sample means will reflect that heterogeneity. The sample means will have more variability from sample to sample.

Sampling Distribution of the Sample Mean

If population has a normal distribution with mean μ and standard deviation σ , then the sample mean \bar{X} of n independent observations has a normal distribution with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

Central Limit Theorem

Draw a SRS of size n from any population with mean μ and standard deviation σ . When n is large, the sampling distribution of \bar{X} is approximately normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

Recall our example of rolling a die from Chapter 3 notes:

The probability distribution for rolling a die is

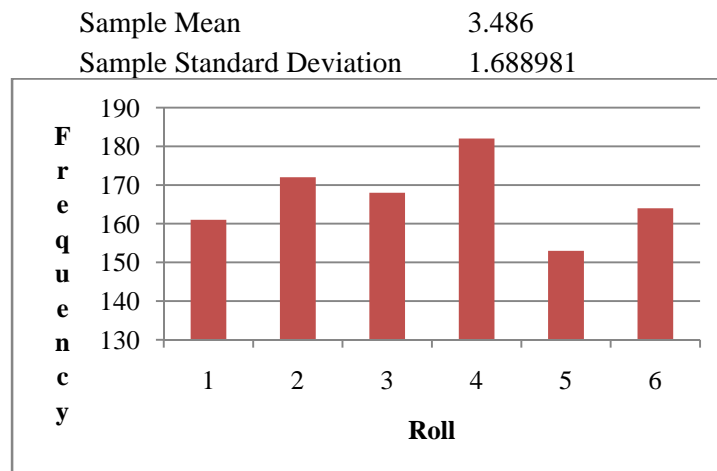
X	1	2	3	4	5	6
$p(x)$	1/6	1/6	1/6	1/6	1/6	1/6

The mean is $\mu = 1\left(\frac{1}{6}\right) + 2\left(\frac{1}{6}\right) + 3\left(\frac{1}{6}\right) + 4\left(\frac{1}{6}\right) + 5\left(\frac{1}{6}\right) + 6\left(\frac{1}{6}\right) = 3.5$ and the standard deviation is

$$\sigma = \sqrt{(1-3.5)^2\left(\frac{1}{6}\right) + (2-3.5)^2\left(\frac{1}{6}\right) + (3-3.5)^2\left(\frac{1}{6}\right) + (4-3.5)^2\left(\frac{1}{6}\right) + (5-3.5)^2\left(\frac{1}{6}\right) + (6-3.5)^2\left(\frac{1}{6}\right)}$$

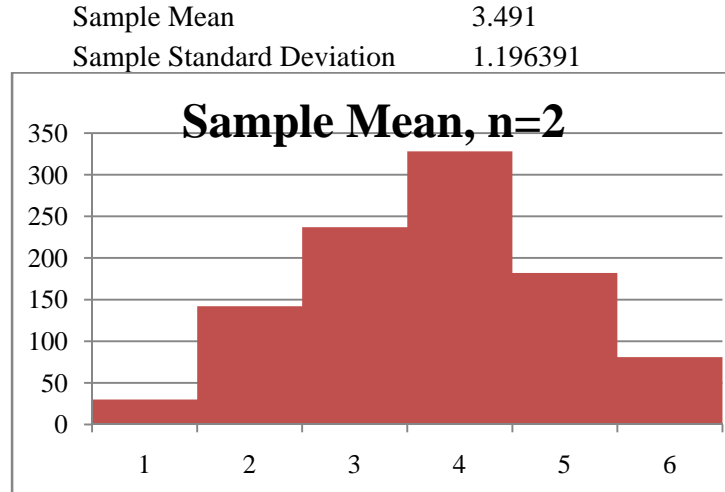
$$= 1.708$$

Look at the result if we roll a die 1000 times:



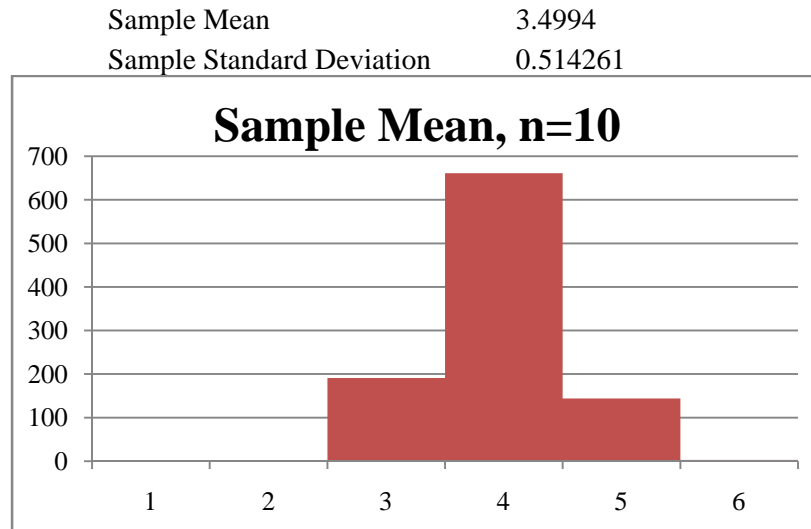
If the outcomes are all equally likely, why are the frequencies different for each?

Now consider rolling a die twice and finding the mean outcome for the two rolls and we will repeat this process 1000 times.



Note the bars on the histogram are $1 \leq \bar{x} < 2$, $2 \leq \bar{x} < 3, \dots, 5 \leq \bar{x} < 6$. The bars are no longer approximately equal why?

Now consider rolling a die ten times and finding the mean outcome for the ten rolls and we will repeat this process 1000 times.



The frequency in for the $1 \leq \bar{x} < 2$ groups is zero, in $2 \leq \bar{x} < 3$ is one, and in the $5 \leq \bar{x} < 6$ is three, why are these so low? (Note the minimum is 2 and maximum is 5.3.)

From the CLT, what does the 68-95-99.7% rule tell us?

Example: (#5.66, p 350)

Total SAT scores of high school seniors in a recent year had a mean $\mu = 1026$ and standard deviation $\sigma = 209$. The distribution of SAT scores is roughly normal.

- a. Julie scored 1110. If scores have a normal distribution, what percentile of the distribution is Julie's score?
- b. Now consider the mean \bar{x} of the scores of 80 randomly chosen students. If $\bar{x} = 1110$, what percentile of the sampling distribution of \bar{x} is this?
- c. Which of your calculations (a) or (b) is less accurate because SAT scores do not have an exact normal distribution? Explain your answer.

A few more facts:

- Normal approximation for sample proportions and counts is an example of the CLT
- Any linear combination of independent normal random variables is also normally distributed.
- More general versions of the CLT say that the distribution of a sum or average of many small random quantities is close to normal.
 - o Any variable that is the sum of many small influences will have an approximate normal distribution.