Chapter 9

Sampling Distributions

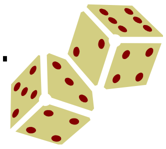
Sampling Distributions...

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.

For example, consider the roll of one and two dice…

Sampling Distribution of the Mean…



A fair **die** is thrown infinitely many times,
with the random variable X = # of spots on any throw.

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

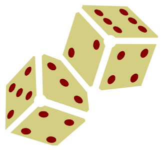
…and the mean and variance are calculated as well:

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

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

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

Sampling Distribution of Two Dice

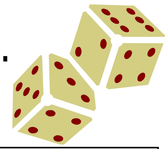


A sampling distribution is created by looking at
all samples of size n=2 (i.e. two dice) and their means…

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

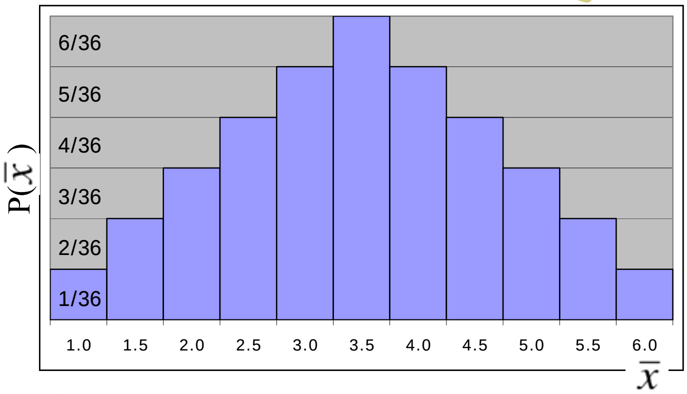
While there are 36 possible samples of size 2, there are only 11 values for $\bar{x}$, and some (e.g $\bar{x}$ =3.5) occur more frequently than others (e.g. $\bar{x}$ =1).

Sampling Distribution of Two Dice…



The ***sampling distribution*** of $\bar{x}$ is 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



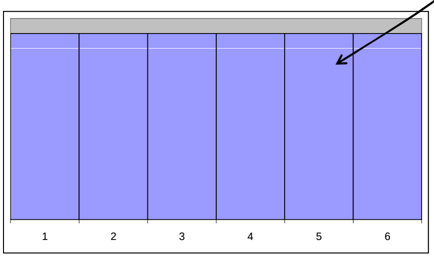
$$\mu_{\bar{x}} = \sum \bar{x}P(\bar{x}) = 1.0(\tfrac{1}{36}) + 1.5(\tfrac{2}{36}) + \ldots + 6.0(\tfrac{1}{36}) = 3.5$$

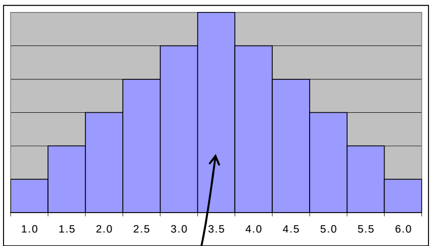
$$\sigma^2_{\bar{x}} = \sum (\bar{x} - \mu_{\bar{x}})^2 P(\bar{x}) = (1.0 - 3.5)^2(\tfrac{1}{36}) + \ldots + (6.0 - 3.5)^2(\tfrac{1}{36}) = 1.46$$

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

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:

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

…to **n**-dice:

$$\mu_{\bar{x}} = \mu$$
$$\sigma^2_{\bar{x}} = \frac{\sigma^2}{n}$$

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

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

Central Limit Theorem...

The sampling distribution of the mean of a random sample drawn from any population is ***approximately normal*** for a ***sufficiently large sample size***.

The larger the sample size, the more closely the sampling distribution of $\bar{X}$ will resemble a normal distribution.

Central Limit Theorem...

If the population is normal, then $\bar{X}$ is normally distributed for all values of n.

If the population is non-normal, then $\bar{X}$ is approximately normal only for larger values of n.

In most practical situations, a sample size of 30 may be sufficiently large to allow us to use the normal distribution as an approximation for the sampling distribution of $\bar{X}$.

Sampling Distribution of the Sample Mean

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

2. $\sigma_{\bar{x}}^2 = \sigma^2 / n \quad and \quad \sigma_{\bar{x}} = \sigma / \sqrt{n}$

3. If X is normal, $\bar{X}$ is normal. If X is nonnormal, $\bar{X}$ is *approximately* normal for sufficiently large sample sizes.

Note: the definition of "sufficiently large" depends on the extent of nonnormality of x (e.g. heavily skewed; multimodal)

Sampling Distribution of the Sample Mean

We can express the sampling distribution of the mean simple as

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

Sampling Distribution of the Sample Mean

The summaries above assume that the population is infinitely large. However if the population is finite the standard error is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N - n}{N - 1}}$$

where N is the population size and

$$\sqrt{\frac{N - n}{N - 1}}$$

is the ***finite population correction factor***.

Sampling Distribution of the Sample Mean

If the population size is large relative to the sample size the finite population correction factor is close to 1 and can be ignored.

We will treat any population that is at least 20 times larger than the sample size as large.

In practice most applications involve populations that qualify as large.

As a consequence the finite population correction factor is usually omitted.

Example 9.1(a)...

The foreman of a bottling plant has observed that the amount of soda in each "32-ounce" bottle is actually a normally distributed random variable, with a mean of 32.2 ounces and a standard deviation of .3 ounce.

If a customer buys one bottle, what is the probability that the bottle will contain more than 32 ounces?

Example 9.1(a)...

We want to find $P(X > 32)$, where X is normally distributed and $\mu = 32.2$ and $\sigma = .3$

$$P(X > 32) = P\left(\frac{X-\mu}{\sigma} > \frac{32-32.2}{.3}\right) = P(Z > -.67) = 1 - .2514 = .7486$$

"there is about a 75% chance that a single bottle of soda contains more than 32oz."

Example 9.1(b)...

The foreman of a bottling plant has observed that the amount of soda in each "32-ounce" bottle is actually a normally distributed random variable, with a mean of 32.2 ounces and a standard deviation of .3 ounce.

If a customer buys a carton of **four** bottles, what is the probability that the ***mean amount of the four bottles*** will be greater than 32 ounces?

Example 9.1(b)...

We want to find P($\overline{X}$ > 32), where X is normally distributed With μ = 32.2 and σ =.3

Things we know:

1) X is normally distributed, therefore so will $\overline{X}$.

2) $\mu_{\bar{x}} = \mu$ = 32.2 oz.

3) $\sigma_{\bar{x}} = \sigma / \sqrt{n} = .3/\sqrt{4} = .15$

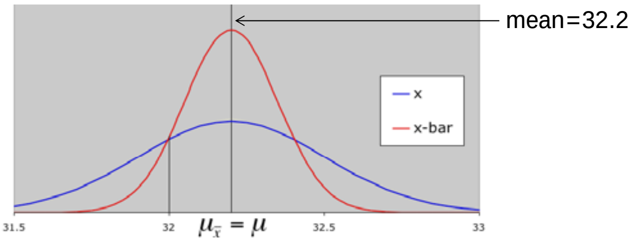
Example 9.1(b)...

If a customer buys a carton of **four** bottles, what is the probability that the ***mean amount of the four bottles*** will be greater than 32 ounces?

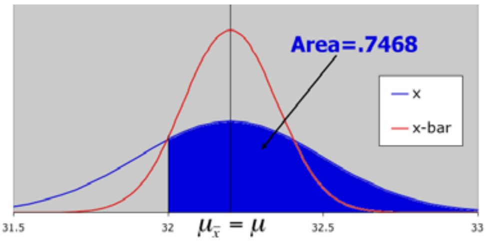
$$P(\overline{X} > 32) = P\left(\frac{\overline{X} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} > \frac{32 - 32.2}{.15}\right) = P(Z > -1.33) = .9082$$

"There is about a 91% chance the mean of the four bottles will exceed 32oz."

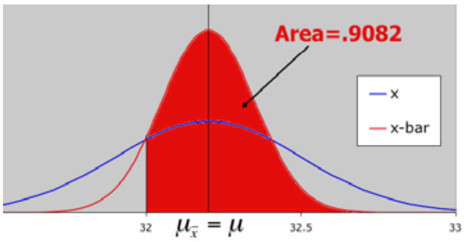
Graphically Speaking…



what is the probability that one bottle will contain more than 32 ounces?



what is the probability that the mean of four bottles will exceed 32 oz?



Chapter-Opening Example

Salaries of a Business School's Graduates

In the advertisements for a large university, the dean of the School of Business claims that the average salary of the school's graduates one year after graduation is \$800 per week with a standard deviation of \$100.

A second-year student in the business school who has just completed his statistics course would like to check whether the claim about the mean is correct.

Chapter-Opening Example

Salaries of a Business School's Graduates

He does a survey of 25 people who graduated one year ago and determines their weekly salary.

He discovers the sample mean to be $750.

To interpret his finding he needs to calculate the probability that a sample of 25 graduates would have a mean of $750 or less when the population mean is $800 and the standard deviation is $100.

After calculating the probability, he needs to draw some conclusions.

Chapter-Opening Example

We want to find the probability that the sample mean is less than $750. Thus, we seek

$$P(\overline{X} < 750)$$

The distribution of X, the weekly income, is likely to be positively skewed, but not sufficiently so to make the distribution of $\overline{X}$ nonnormal. As a result, we may assume that $\overline{X}$ is normal with mean

$$\mu_{\overline{x}} = \mu = 800$$

and standard deviation

$$\sigma_{\overline{x}} = \sigma / \sqrt{n} = 100 / \sqrt{25} = 20$$

Chapter-Opening Example

Thus,

$$
\begin{aligned}
&P(\overline{X} < 750) \\
&= P\left(\frac{\overline{X} - \mu_{\overline{x}}}{\sigma_{\overline{x}}} < \frac{750 - 800}{20}\right) \\
&= P(Z < -2.5) \\
&= .5 - .4938 \\
&= .0062
\end{aligned}
$$

The probability of observing a sample mean as low as $750 when the population mean is $800 is extremely small. Because this event is quite unlikely, we would have to conclude that the dean's claim is not justified.

Using the Sampling Distribution for Inference

Here's another way of expressing the probability calculated from a sampling distribution.

$$P(-1.96 < Z < 1.96) = .95$$

Substituting the formula for the sampling distribution

$$P(-1.96 < \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} < 1.96) = .95$$

With a little algebra

$$P(\mu - 1.96\frac{\sigma}{\sqrt{n}} < \overline{X} < \mu + 1.96\frac{\sigma}{\sqrt{n}}) = .95$$

Using the Sampling Distribution for Inference

Returning to the chapter-opening example where $\mu = 800$, $\sigma = 100$, and $n = 25$, we compute

$$P(800 - 1.96\frac{100}{\sqrt{25}} < \overline{X} < 800 + 1.96\frac{100}{\sqrt{25}}) = .95$$

or

$$P(760.8 < \overline{X} < 839.2) = .95$$

This tells us that there is a 95% probability that a sample mean will fall between 760.8 and 839.2. Because the sample mean was computed to be $750, we would have to conclude that the dean's claim is not supported by the statistic.

Using the Sampling Distribution for Inference

Changing the probability from .95 to .90 changes the probability statement to

$$P(\mu - 1.645\frac{\sigma}{\sqrt{n}} < \overline{X} < \mu + 1.645\frac{\sigma}{\sqrt{n}}) = .90$$

$\overline{X}$

Using the Sampling Distribution for Inference

We can also produce a general form of this statement

$$P(\mu - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \overline{X} < \mu + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

In this formula α (Greek letter *alpha*) is the probability that does not fall into the interval.

To apply this formula all we need do is substitute the values for μ, σ, n, and α.

Using the Sampling Distribution for Inference

For example, with $\mu = 800$, $\sigma = 100$, $n = 25$ and $\alpha = .01$, we produce

$$P(\mu - z_{.005}\frac{\sigma}{\sqrt{n}} < \overline{X} < \mu + z_{.005}\frac{\sigma}{\sqrt{n}}) = 1 - .01$$

$$P(800 - 2.575\frac{100}{\sqrt{25}} < \overline{X} < 800 + 2.575\frac{100}{\sqrt{25}}) = .99$$

$$P(748.5 < \overline{X} < 851.5) = .99$$

Sampling Distribution of a Proportion...

The estimator of a population proportion of successes is the ***sample proportion***. That is, we count the number of successes in a sample and compute:

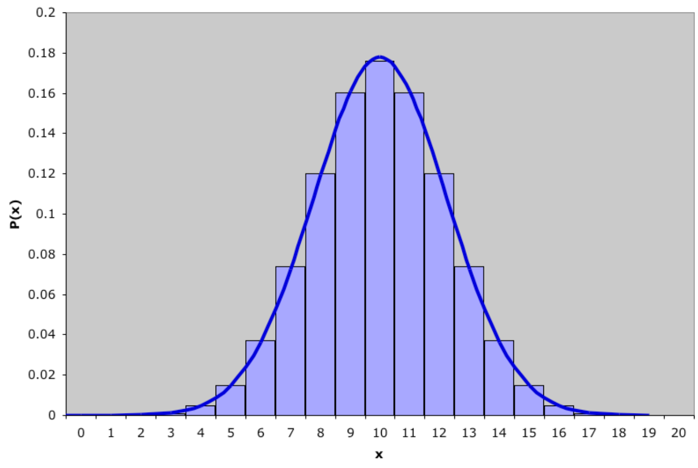
$$\hat{P} = \frac{X}{n}$$

(read this as "p-hat").

X is the number of successes, n is the sample size.

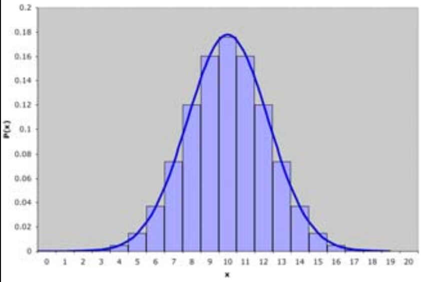
Normal Approximation to Binomial...

Binomial distribution with n=20 and p=.5 with a normal approximation superimposed (μ=10 and σ=2.24)



Normal Approximation to Binomial…

Binomial distribution with n=20 and p=.5 with a normal approximation superimposed (μ=10 and σ=2.24)



where did these values come from?!

From §7.6 we saw that:

$$\mu = np$$

$$\sigma^2 = np(1-p)$$

$$\sigma = \sqrt{np(1-p)}$$

Hence:

$$\mu = np = 20(.5) = 10 \quad \text{and} \quad \sigma = \sqrt{np(1-p)} = \sqrt{20(.5)(1-.5)} = 2.24$$

Normal Approximation to Binomial…

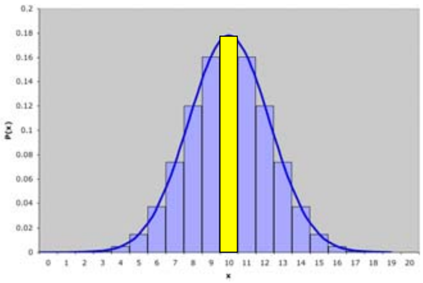
Normal approximation to the binomial works best when the number of experiments, n, (sample size) is large, and the probability of success, p, is close to 0.5

For the approximation to provide good results two conditions should be met:

1) np $\geq$ 5

2) n(1–p) $\geq$ 5

Normal Approximation to Binomial...

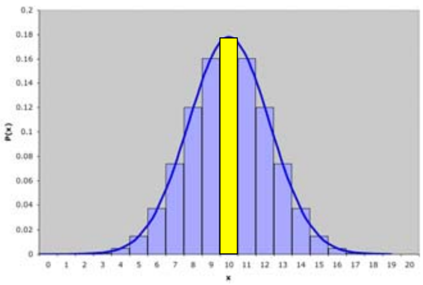


To calculate P(X=10) using the normal distribution, we can find the area under the normal curve between 9.5 & 10.5

$$P(X = 10) \approx P(9.5 < Y < 10.5)$$

where ***Y is a normal random variable*** approximating the ***binomial random variable X***

Normal Approximation to Binomial...



In fact:

P(X = 10) = .176

while

P(9.5 < Y < 10.5) = .1742

the approximation is quite good.

$$P(X = 10) \approx P(9.5 < Y < 10.5)$$

where ***Y is a normal random variable*** approximating the ***binomial random variable X***

Sampling Distribution of a Sample Proportion...

Using the laws of expected value and variance, we can determine the mean, variance, and standard deviation of $\hat{P}$. (The standard deviation of $\hat{P}$ is called the ***standard error of the proportion***.)

$$E(\hat{P}) = p$$

$$V(\hat{P}) = \sigma^2_{\hat{p}} = \frac{p(1-p)}{n}$$

$$\sigma_{\hat{p}} = \sqrt{p(1-p)/n}$$

Sample proportions can be standardized to a standard normal distribution using this formulation:

$$Z = \frac{\hat{P} - p}{\sqrt{p(1-p)/n}}$$

Example 9.2

In the last election a state representative received 52% of the votes cast.

One year after the election the representative organized a survey that asked a random sample of 300 people whether they would vote for him in the next election.

If we assume that his popularity has not changed what is the probability that more than half of the sample would vote for him?

Example 9.2

The number of respondents who would vote for the representative is a binomial random variable with n = 300 and p = .52.

We want to determine the probability that the sample proportion is greater than 50%. That is, we want to find

$$P(\hat{P} > .50)$$

We now know that the sample proportion $\hat{P}$ is approximately normally distributed with mean p = .52 and standard deviation

$$\sqrt{p(1-p)/n} = \sqrt{(.52)(1-.52)/300} = .0288$$

Example 9.2

Thus, we calculate

$$\begin{aligned} &P(\hat{P} > .50) \\ &= P\left(\frac{\hat{P} - p}{\sqrt{p(1-p)/n}} > \frac{.50 - .52}{.0288}\right) \\ &= P(Z > -.69) \\ &= .7549 \end{aligned}$$

If we assume that the level of support remains at 52%, the probability that more than half the sample of 300 people would vote for the representative is 75.49%.

Sampling Distribution: Difference of two means

The final sampling distribution introduced is that of the ***difference between two sample means***. This requires:

➔ ***independent*** random samples be drawn from each of **two *normal*** populations

If this condition is met, then the sampling distribution of the ***difference*** between the two sample means, i.e. $\bar{X}_1 - \bar{X}_2$ will be normally distributed.

(note: if the two populations are **not** both normally distributed, but the sample sizes are "large" (>30), the distribution of $\bar{X}_1 - \bar{X}_2$ is *approximately* normal)

Sampling Distribution: Difference of two means

The ***expected value*** and ***variance*** of the sampling distribution of $\bar{X}_1 - \bar{X}_2$ are given by:

mean: $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$

standard deviation: $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

(also called the standard error if the difference between two means)

Example 9.3...

Since the distribution of $\bar{X}_1 - \bar{X}_2$ is ***normal*** and has a

mean of $\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2$

and a standard deviation of $\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

We can compute Z (standard normal random variable) in this way:

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Example 9.3...

Starting salaries for MBA grads at two universities are normally distributed with the following means and standard deviations. Samples from each school are taken…

	University 1	University 2
Mean μ	62,000 $/yr	60,000 $/yr
Std. Dev. σ	14,500 $/yr	18,300 $/yr
sample size **n**	50	60

What is the probability that the sample mean starting salary of University #1 graduates will exceed that of the #2 grads?

Example 9.3...

"What is the probability that the ***sample mean*** starting salary of University #1 graduates will ***exceed*** that of the #2 grads?"

We are interested in determining $P(\bar{X}_1 > \bar{X}_2)$. Converting this to a difference of means, what is: $P(\bar{X}_1 - \bar{X}_2 > 0)$?

$$P(\bar{X}_1 - \bar{X}_2 > 0) = P\left(\underbrace{\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}}_{Z} > \frac{0 - 2000}{3128} \right) = P(Z > -.64) = .5 + .2389 = .7389$$

"there is about a 74% chance that the sample mean starting salary of U. #1 grads will exceed that of U. #2"

From Here to Inference

In Chapters 7 and 8 we introduced probability distributions, which allowed us to make probability statements about values of the random variable.

A prerequisite of this calculation is knowledge of the distribution and the relevant parameters.

From Here to Inference

In Example 7.9, we needed to know that the probability that Pat Statsdud guesses the correct answer is 20% (p = .2) and that the number of correct answers (successes) in 10 questions (trials) is a binomial random variable.

We then could compute the probability of any number of successes.

From Here to Inference

In Example 8.2, we needed to know that the return on investment is normally distributed with a mean of 10% and a standard deviation of 5%.

These three bits of information allowed us to calculate the probability of various values of the random variable.

From Here to Inference

The figure below symbolically represents the use of probability distributions.

Simply put, knowledge of the population and its parameter(s) allows us to use the probability distribution to make probability statements about individual members of the population.

Probability Distribution ----------→ Individual

From Here to Inference

In this chapter we developed the sampling distribution, wherein knowledge of the parameter(s) and some information about the distribution allow us to make probability statements about a sample statistic.

$$\frac{Population}{\&\ Parameter(s)} \quad \underline{Sampling \quad distribution} \text{ ------→ Statistic}$$

From Here to Inference

Statistical works by reversing the direction of the flow of knowledge in the previous figure. The next figure displays the character of statistical inference.

Starting in Chapter 10, we will assume that most population parameters are unknown. The statistics practitioner will sample from the population and compute the required statistic. The sampling distribution of that statistic will enable us to draw inferences about the parameter.

From Here to Inference

Statistic $\xrightarrow{\textit{Sampling distribution}}$ ------→ Parameter