Unit 3
Sampling Distributions

Probabilities for Continuous Variables

In considering the sum when rolling two dice, or counting the number of tails in three tosses of a quarter, we are dealing with **discrete random variables** – ones that can take only certain values. But note that a sample space need not be finite. For example, if we are interested in the time it takes for a light bulb to burn out, the sample space is

$$S = \{\text{all values of } x \text{ such that } x \geq 0\}$$

Probabilities for Continuous Variables

In this case of a **continuous random variable**, how do we assign probabilities to the outcomes in S if there are infinitely many possible outcomes?

We accomplish this by assigning probabilities to intervals of values rather than to individual outcomes. And so the areas under density curves are actually representative of the **probability** of observing an outcome whose value falls in that interval.

Probabilities for Continuous Variables

A density curve is a model that assigns probabilities to a set of outcomes for a variable believed to be accurately described by the distribution.

We saw in the last unit that, instead of speaking of proportions under the normal curve, we can speak of probabilities.

A **random variable** is one whose value is a numerical outcome of a random phenomenon.

Example

Suppose it is known that pulse rates of adult females follow a normal distribution with mean 74 beats per minute and standard deviation 12 beats per minute.

What is the probability that a randomly selected adult female has a pulse rate above 57 beats per minute?

$$\textit{Answer}:\ P(X > 57) = P\left(Z > \frac{57-74}{12}\right)$$

$$= P(Z > -1.42) = 1 - P(Z < -1.42)$$

$$= 1 - 0.0778 = 0.9222.$$

Distribution of the Sample Mean

Suppose that, rather being interested in probability calculations for some variable on a **single individual**, we are instead interested in taking a random sample of size n and calculating the probability that the **sample mean** falls within some range of values.

To calculate such probabilities, we must examine the **distribution** of the sample mean.

Distribution of the Sample Mean

To find this distribution, we ask:

What would happen if we **repeatedly took samples of the same size $\boldsymbol{n}$ from the population and calculated $\bar{x}$?**

Distribution of the Sample Mean

To answer this, we take a random sample of size *n* from a normal distribution with mean μ and standard deviation σ, and record the value of the sample mean. Take another sample of the same size from the same distribution and again record the value of the sample mean.

Keep doing this, and after you have a large number of sample means, plot their values.

Distribution of the Sample Mean

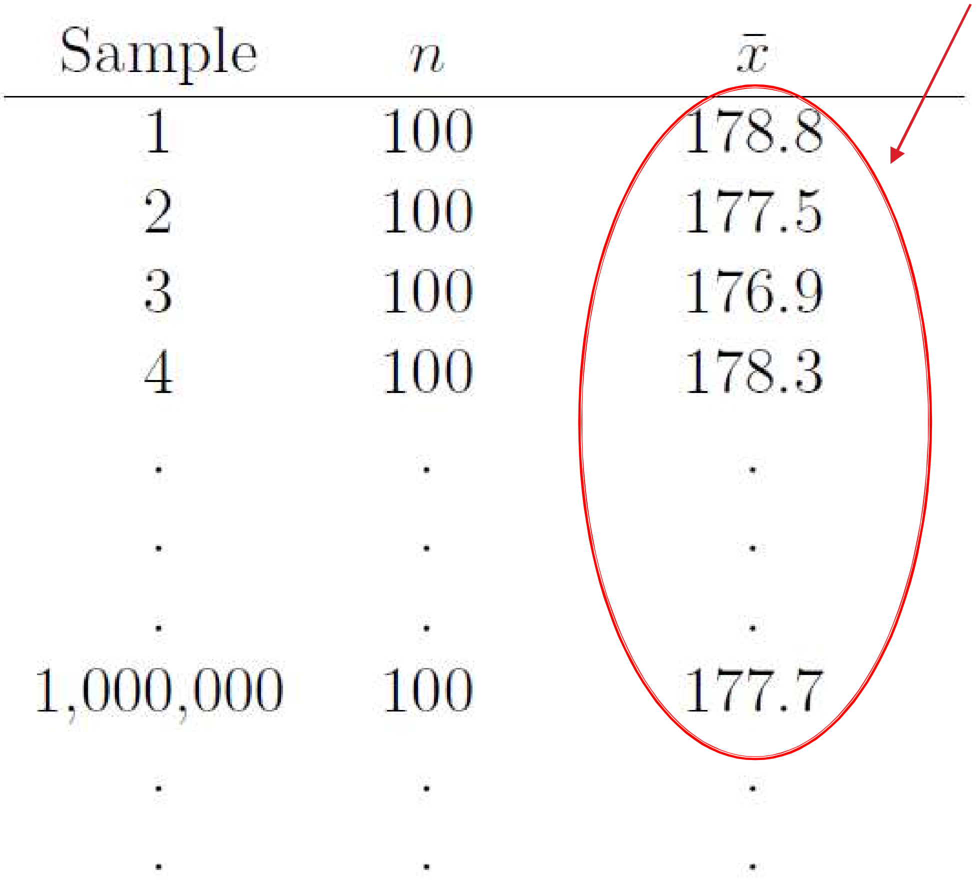
Recall our example about the heights X of adult Canadian males, where

$$X \sim N(178,\, 6)$$

Suppose we take a very large number of samples of 100 males and for each sample we calculate the value of $\bar{x}$. We then create a histogram of the values of $\bar{x}$.

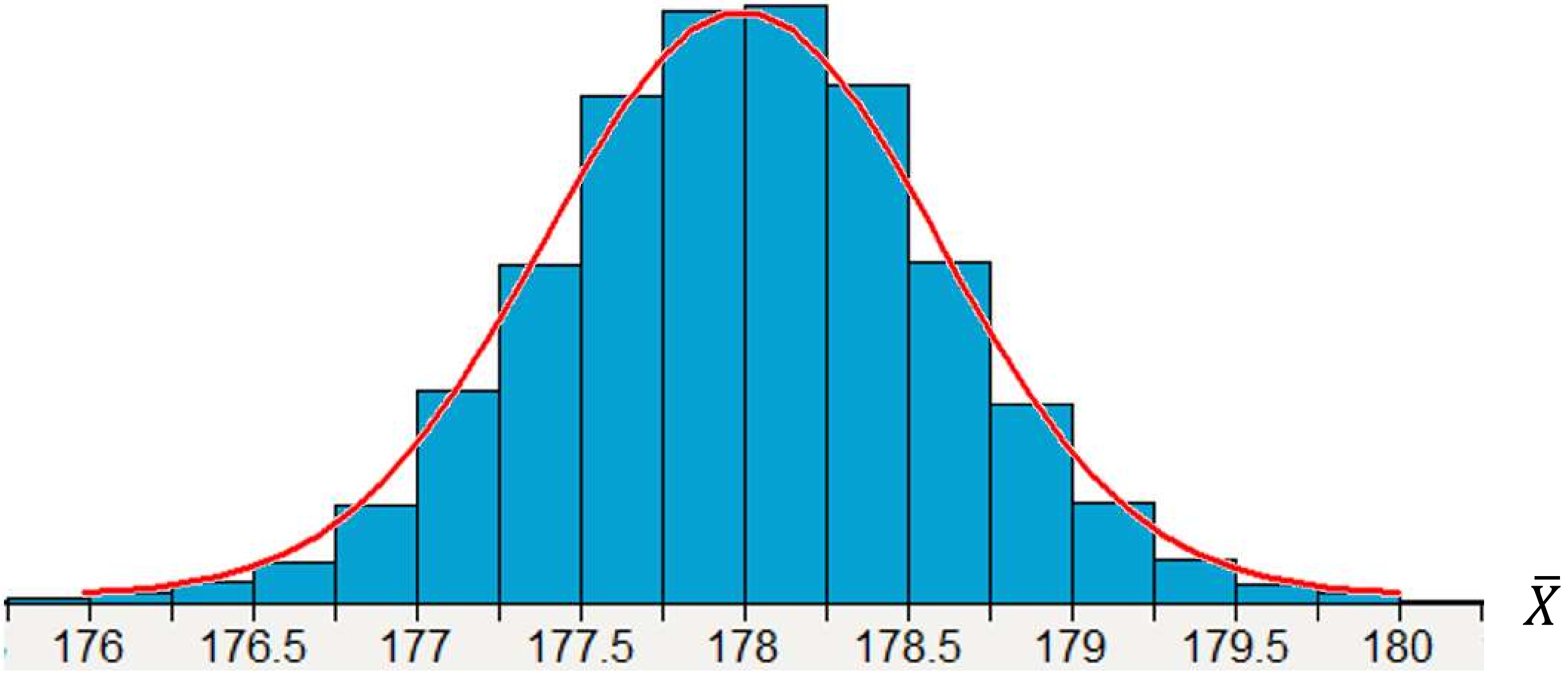
Distribution of the Sample Mean

Now make a histogram of these millions of values:



Sample	n	$\bar{x}$
1	100	178.8
2	100	177.5
3	100	176.9
4	100	178.3
.	.	.
.	.	.
.	.	.
1,000,000	100	177.7
.	.	.
.	.	.

Distribution of the Sample Mean



Mean = 178

Std. Dev. = 0.6

Distribution of the Sample Mean

We notice three things:

- The distribution of $\bar{X}$ is still normal.
- The mean of the distribution of $\bar{X}$ is the same as the mean of the population distribution of X, $\mu = 178$.
- The standard deviation of $\bar{X}$ is equal to

$$\frac{\sigma}{\sqrt{n}} = \frac{6}{\sqrt{100}} = 0.6$$

Distribution of the Sample Mean

Suppose that $\bar{X}$ is the mean of a random sample of size n drawn from a large population with mean μ and standard deviation σ. Then the mean of the sampling distribution of $\bar{X}$ is μ and its standard deviation is $\frac{\sigma}{\sqrt{n}}$.

Sampling Distribution

In particular, if $X \sim N(\mu, \sigma)$, then

$$\bar{X} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

This is known as the **sampling distribution** of the sample mean $\bar{X}$. The sampling distribution of a statistic is the distribution of values taken by the statistic in all possible samples of the same size from the same population.

Sampling Distribution

Note that the mean of the sampling distribution of $\bar{X}$ is equal to the mean of the population distribution of X, but the standard deviation is **lower**. This is due to the fact that **averages are less variable than individual observations**.

Example

If you randomly select one male, what is the probability that his height is greater than 180 cm?

$$P(X > 180) = P\left(Z > \frac{180-178}{6}\right) = P(Z > 0.33)$$

$$= 1 - P(Z < 0.33) = 1 - 0.6293 = 0.3707$$

Example

If you take a random sample of ten males, what is the probability that their **average** height is greater than 180 cm?

$$P(\overline{X} > 180) = P\left(\frac{\overline{X} - \mu}{\sigma / \sqrt{n}} > \frac{180 - \mu}{\sigma / \sqrt{n}}\right) = P\left(Z > \frac{180 - 178}{6 / \sqrt{10}}\right)$$

$$= P(Z > 1.05) = 1 - P(Z < 1.05) = 1 - 0.8531 = 0.1469$$

R Code

```
> pnorm(180, 178, 6/sqrt(10), lower.tail = FALSE)
[1] 0.1459203

> 1 - pnorm(180, 178, 6/sqrt(10))
[1] 0.1459203
```

Unbiased Estimators

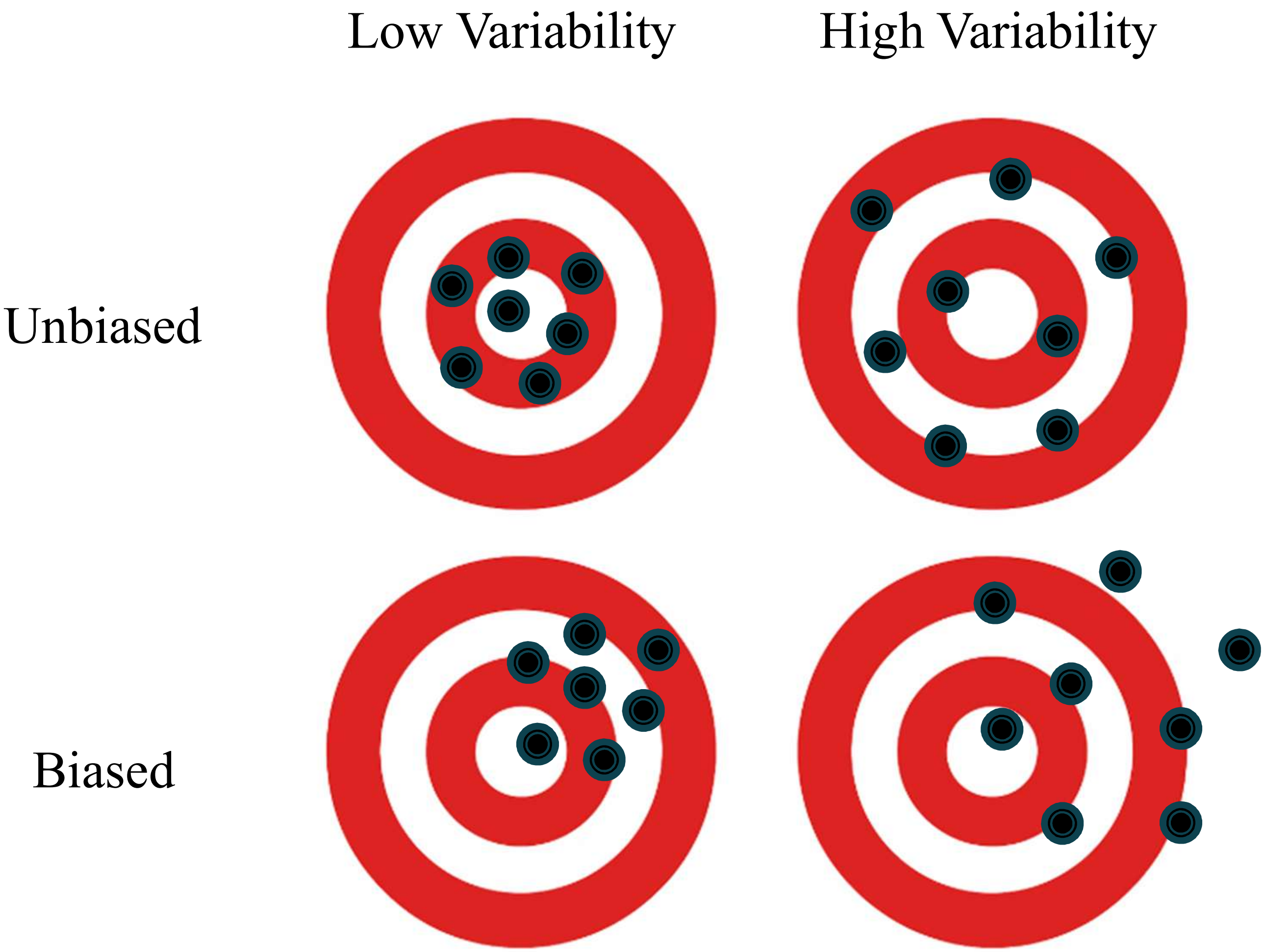
A statistic used to estimate a parameter is said to be **unbiased** if the mean of its sampling distribution is equal to the true value of the parameter.

Since the mean of the distribution of $\bar{X}$ is μ, the sample mean is said to be an unbiased estimator for the parameter μ.

The sample variance S^2 is also an unbiased estimator for the population variance σ^2.

A statistic is said to be **biased** if it systematically overestimates or underestimates the value of a parameter.

Unbiased Statistics & Variability



Low Variability

High Variability

Unbiased

Biased

Example

The pulse rates of adult women follow a normal distribution with mean 74 beats per minute and standard deviation 12 beats per minute. What is the probability that a random sample of 25 women has an average pulse rate between 68 and 80 beats per minute?

$$P(68 < \overline{X} < 80) = P\left(\frac{68-74}{12/\sqrt{25}} < Z < \frac{80-74}{12/\sqrt{25}}\right) = P(-2.50 < Z < 2.50)$$

$$= P(Z < 2.50) - P(Z < -2.50) = 0.9938 - 0.0062 = 0.9876$$

R Code

```
> pnorm(80, 74, 12/sqrt(25)) - pnorm(68, 74, 12/sqrt(25))
[1] 0.9875807
```

Random Samples

Question: How do we select our samples?

Answer: We would like our samples to be **representative** of the populations; that is, in a way that allows us to draw reliable conclusions about the population of interest. The fairest way to select a sample is to do so randomly. Intuitively, we would like every individual in the population to have the same chance to be in our sample.

Simple Random Samples

The type of sample we select is called a **simple random sample**.

A **simple random sample** (SRS) of size n consists of n individuals from the population chosen in such a way that every group of n individuals has an equal chance to be the sample actually selected.

It follows that each individual has an equal chance of being selected into the sample.

Simple Random Sampling

For example, suppose a teacher wants to select a random sample of five of the 40 students in her class to participate in a demonstration.

To do this, she could write the names of her students on 40 pieces of paper, place them in a hat, and randomly choose five of them.

Alternatively, she could use computer software to randomly select the sample. (For large populations, writing names on pieces of paper and placing them all in a hat becomes unrealistic.)

Simple Random Sampling

The teacher numbers her students as follows:

01 – Adams
02 – Alvarez
03 – Becker
04 – Bouchard
05 – Brown
06 – Campbell
07 – Chow
08 – Davis
09 – Evans
10 – Farzad

11 – Friesen
12 – Gamage
13 – Garcia
14 – Hossain
15 – Ibrahim
16 – Joshi
17 – Khan
18 – Kim
19 – Knight
20 – Lee

21 – Levesque
22 – Lewis
23 – Liu
24 – Martin
25 – Miller
26 – Nguyen
27 – Patel
28 – Pereira
29 – Roberts
30 – Shevchuk

31 – Singh
32 – Suzuki
33 – Taylor
34 – Tran
35 - Tremblay
36 – Usman
37 – Walker
38 – Weber
39 – Williams
40 – Zhang

Simple Random Sampling

Using R, we select our simple random sample of five students using the following code:

```
> class<-c(1:40)
> sample(class,5, replace=“FALSE”)
[1] 8 24 17 3 35
```

In this case, R has selected the students Davis (8), Martin (24), Khan (17), Becker (3) and Tremblay (35) to be in our sample.

In selecting this sample, R has simulated placing the 40 names in a hat and randomly selecting five of them.

Simple Random Sampling

The `replace="FALSE"` tells R to select a **simple random sample without replacement**. In this type of sample, once an individual has been chosen, they cannot be selected again.

If we were to `replace="TRUE"`, R would select a simple random sample with replacement, which would allow individuals to be selected more than once (i.e., after they have been chosen, their name "goes back in the hat").

In this course, we will exclusively use simple random samples **without** replacement.

Example: Sample Total

Weights of bricks used in construction follow a normal distribution with mean 8 pounds and standard deviation 0.2 pounds. What is the probability that a random sample of 4 bricks has a total weight greater than 33.2 pounds?

$$P(Total > 33.2) = P\left(\bar{X} > \frac{33.2}{4}\right) = P(\bar{X} > 8.3)$$

$$= P\left(Z > \frac{8.3 - 8.0}{0.2/\sqrt{4}}\right) = P(Z > 3.00) = 1 - P(Z < 3.00)$$

$$= 1 - 0.9987 = 0.0013$$

R Code

```
> pnorm(8.3, 8.0, 0.2/sqrt(4), lower.tail = FALSE)
[1] 0.001349898

> 1 - pnorm(8.3, 8.0, 0.2/sqrt(4))
[1] 0.001349898
```

Practice Question

The weights of eggs produced by a certain breed of hen are normally distributed with mean 65 grams and standard deviation 5 grams. If cartons of such eggs can be considered to be simple random samples of 12 eggs, what is the probability that the average weight of eggs in a carton is greater than 67 grams?

(A) 0.0655
(B) 0.0823
(C) 0.0000
(D) 0.0951
(E) 0.0722

Practice Question

The weights of eggs produced by a certain breed of hen are normally distributed with mean 65 grams and standard deviation 5 grams. If cartons of such eggs can be considered to be simple random samples of 12 eggs, what is the probability that the average weight of eggs in a carton is exactly 66 grams?

(A) 0.2451

(B) 0.6928

(C) 0.0000

(D) 0.7549

(E) 0.3072

Practice Question

The weights of eggs produced by a certain breed of hen are normally distributed with mean 65 grams and standard deviation 5 grams. If cartons of such eggs can be considered to be simple random samples of 12 eggs, what is the probability that the total weight of eggs in a carton less than 816 grams?

(A) 0.9049
(B) 0.9265
(C) 0.9474
(D) 0.9686
(E) 0.9812

Practice Question

The weights of certain breed of dog follow a normal distribution with mean 50 pounds and standard deviation 10 pounds. If we take a random sample of 25 dogs, there is an approximate 95% chance that their average weight is between:

(A) 42 and 58 pounds
(B) 30 and 70 pounds
(C) 55 and 65 pounds
(D) 46 and 54 pounds
(E) 40 and 60 pounds

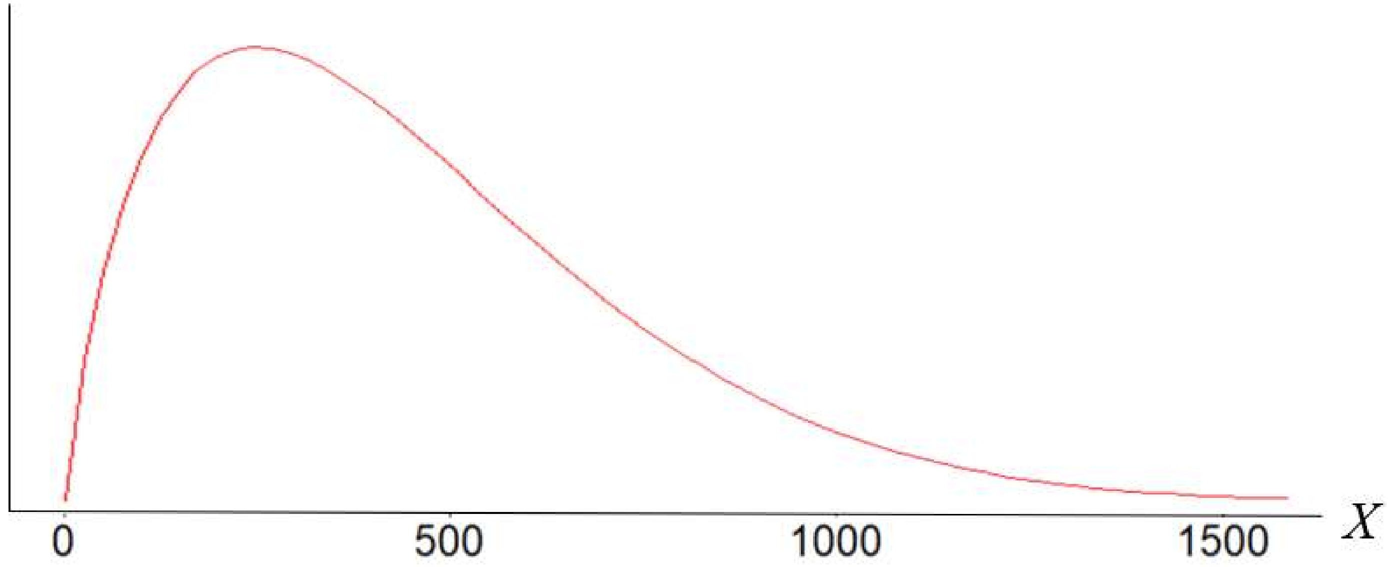
Distribution of the Sample Mean

Let us consider the general situation of a random variable following **any** distribution.

What happens to the distribution of the sample mean as we increase the sample size?

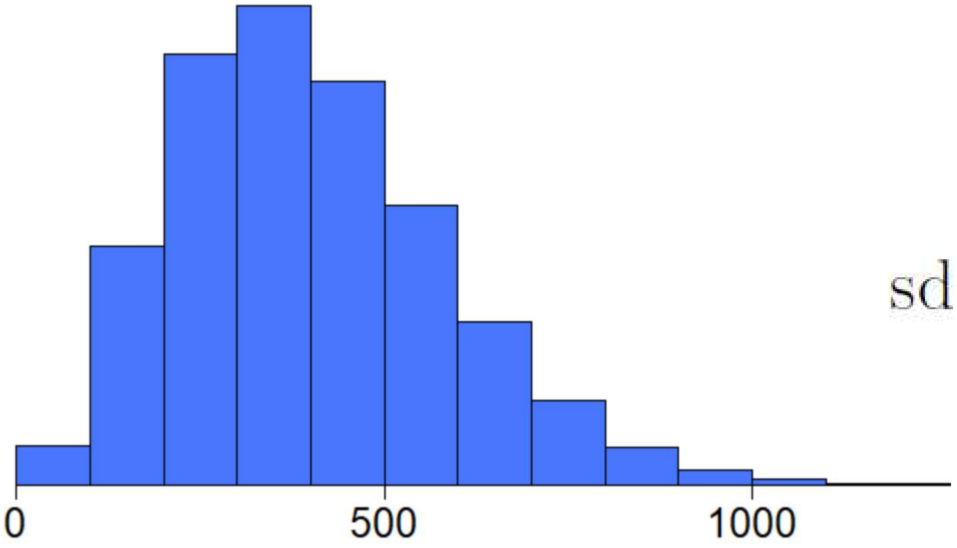
Distribution of the Sample Mean

The lifetime of a light bulb (i.e., the time until the bulb burns out) is known to follow a right-skewed distribution with mean 400 hours and standard deviation 250 hours.



Distribution of the Sample Mean

Now we take an SRS of n = 2 light bulbs and calculate the mean lifetime $\bar{x}$. Take another sample of size 2 and do the same thing. We repeat this many times and plot our values of $\bar{x}$ on a histogram:

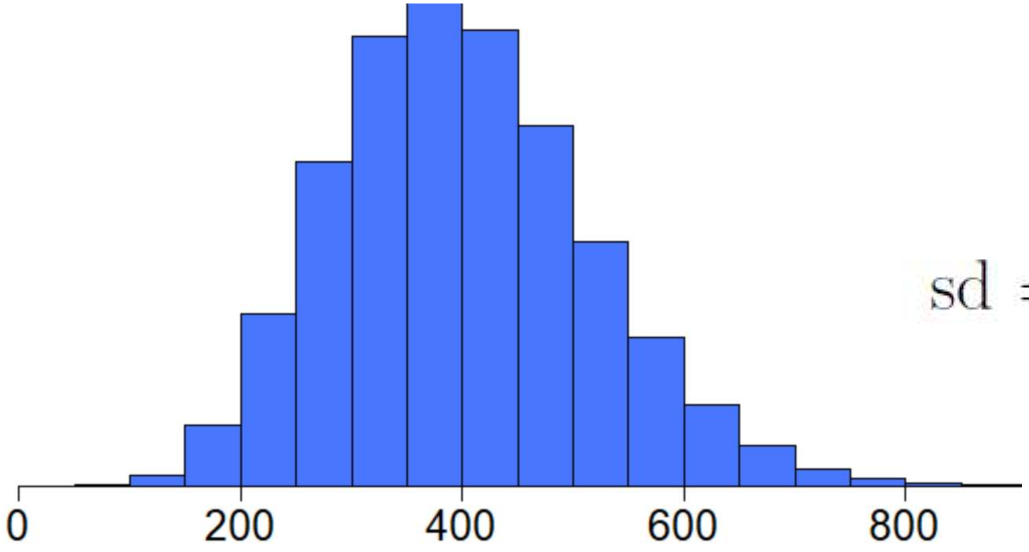


$$\text{mean} = \mu = 400$$

$$\text{sd} = \frac{\sigma}{\sqrt{n}} = \frac{250}{\sqrt{2}} = 176.78$$

Distribution of the Sample Mean

Now we take an SRS of $n = 5$ light bulbs and calculate the mean lifetime $\bar{x}$. Take another sample of size 5 and do the same thing. We repeat this many times and plot our values of $\bar{x}$ on a histogram:

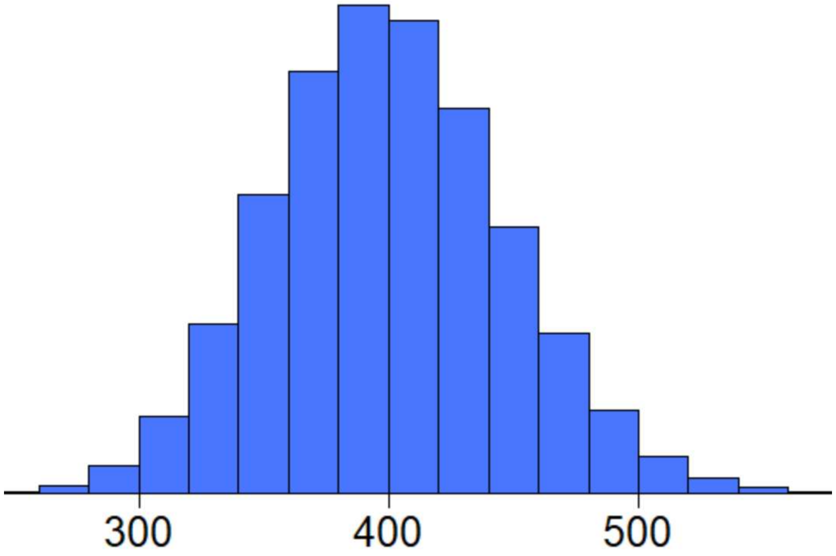


$$\text{mean} = \mu = 400$$

$$\text{sd} = \frac{\sigma}{\sqrt{n}} = \frac{250}{\sqrt{5}} = 111.8$$

Distribution of the Sample Mean

Now we take an SRS of n = 30 light bulbs and calculate the mean lifetime $\bar{x}$. Take another sample of size 30 and do the same thing. We repeat this many times and plot our values of $\bar{x}$ on a histogram:



$$\text{mean} = \mu = 400$$

$$\text{sd} = \frac{\sigma}{\sqrt{n}} = \frac{250}{\sqrt{30}} = 45.64$$

Distribution of the Sample Mean

As the sample size increases, the sampling distribution of $\bar{X}$ approaches a normal distribution!

Remember, **the form of the population distribution of X doesn't matter**. The sample mean will **always** be approximately normally distributed when the sample size is sufficiently large.

Not only that, but we know that the distribution will have mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

Central Limit Theorem

This result leads us to one of the fundamental theorems in statistics:

The Central Limit Theorem

Take an SRS of size *n* from **any** population with mean μ and standard deviation σ. **When *n* is large**, the sampling distribution of the sample mean is approximately normal:

“approximately follows”

$$\bar{X} \;\dot{\sim}\; N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

Central Limit Theorem

The sample size required for a sampling distribution to be approximately normal depends on the original distribution.

Sampling distributions of the sample mean for symmetric distributions approach a normal distribution much more quickly (i.e., for lower values of n) than those for strongly skewed distributions.

For the purpose of this course, we will assume it is safe to apply the Central Limit Theorem when $n \geq 30$.

Example

What is the probability that a random sample of 40 light bulbs has a mean lifetime greater than 450 hours?

Even though the distribution of lifetimes is not normal, the sample size $n = 40$ is high enough that we can use the Central Limit Theorem to calculate this probability, since the distribution of $\bar{X}$ **is** approximately normal.

Example

The probability that the average lifetime of the 40 bulbs exceeds 450 hours is approximately

$$P(\bar{X} > 450) \approx P\left(Z > \frac{450 - 400}{250/\sqrt{40}}\right)$$

$$= P(Z > 1.26) = 1 - P(Z \leq 1.26) = 1 - 0.8962 = 0.1038$$

R Code

```
> pnorm(450, 400, 250/sqrt(40), lower.tail = FALSE)
[1] 0.1029516

> 1 - pnorm(450, 400, 250/sqrt(40))
[1] 0.1029516
```

Example

A factory is producing metal bolts. The mean diameter of all bolts produced is known to be $\mu = 1.25$ cm and the standard deviation is $\sigma = 0.05$ cm. A random sample of 100 bolts is selected. What is the probability that the mean diameter of the bolts is between 1.24 and 1.26 cm?

Even though we don’t know the form of the distribution of diameters, since the sample size is high, we know that the distribution of $\bar{X}$ is approximately normal.

Example

It follows that

$$P(1.24 < \overline{X} < 1.26) \approx P\left(\frac{1.24 - 1.25}{0.05 \big/ \sqrt{100}} < Z < \frac{1.26 - 1.25}{0.05 \big/ \sqrt{100}} \right)$$

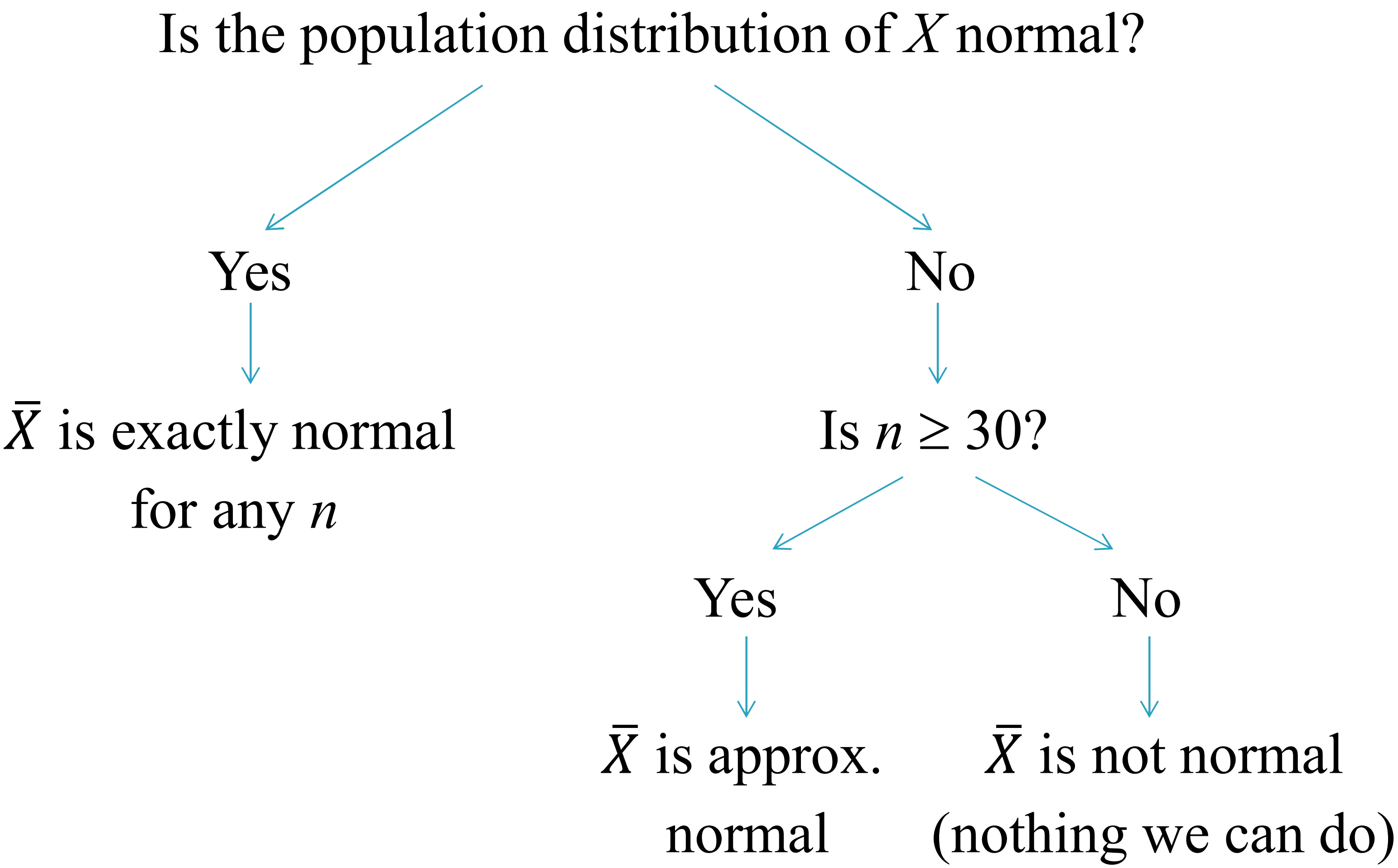
$$= P(-2.00 < Z < 2.00) = P(Z < 2.00) - P(Z < -2.00)$$

$$= 0.9772 - 0.0228 = 0.9544$$

Example

What is the probability that five randomly selected bolts have a mean diameter between 1.24 cm and 1.26 cm?

We can’t calculate this probability! We don’t know the form of the distribution of X, and the sample size is not high enough to use the Central Limit Theorem.



Is the population distribution of X normal?

Yes

$\bar{X}$ is exactly normal for any n

No

Is $n \geq 30$?

Yes

$\bar{X}$ is approx. normal

No

$\bar{X}$ is not normal (nothing we can do)

Practice Question

The amount of daily sales at a local coffee shop is known to follow a normal distribution with mean $2,500 and standard deviation $500. What is the probability that the average sales for a random sample of two days is greater than $3,000?

(A) 0.0329
(B) 0.0594
(C) 0.0793
(D) 0.0968
(E) impossible to calculate because of the low sample

Practice Question

Speeds of vehicles on a highway follow some right-skewed distribution with mean 105 km/h and standard deviation 20 km/h. What is the probability that a random sample of four cars on this highway have a mean speed less than 100 km/h?

(A) 0.6915

(B) 0.0228

(C) 0.1867

(D) 0.3085

(E) impossible to calculate with the information given.

Distribution of the Sample Mean

In summary:

- If the mean and the standard deviation of a random variable X are μ and σ, respectively, then, regardless of the distribution of X, the mean of $\bar{X}$ is μ and standard deviation of $\bar{X}$ is $\frac{\sigma}{\sqrt{n}}$.

Distribution of the Sample Mean

In summary:

- If X follows a normal distribution with mean μ and standard deviation σ, then the sampling distribution of the sample mean $\bar{X}$ is normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$, regardless of the sample size n.

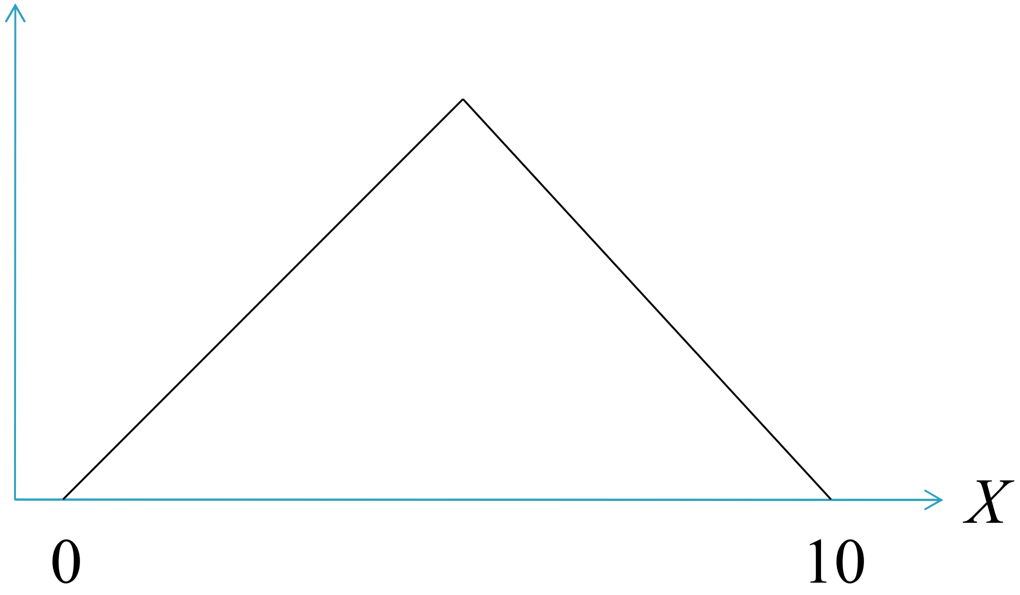
Distribution of the Sample Mean

In summary:

- Central Limit Theorem – If X follows any distribution with mean μ and standard deviation σ, then the sampling distribution of the sample mean $\bar{X}$ is approximately normal with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$, provided that the sample size is large.

Practice Question

A variable X follows a triangular distribution, as shown below, with mean 5 and standard deviation 2.



Practice Question

Suppose that you take a random sample of 400 observations from the population above and make a histogram. You expect the histogram to be:

(A) approximately normal with mean close to 5 and standard deviation close to 2.

(B) triangular with mean close to 5 and standard deviation close to 0.1.

(C) approximately normal with mean close to 5 and standard deviation close to 0.005.

(D) triangular with mean close to 5 and standard deviation close to 2.

(E) approximately normal with mean close to 5 and standard deviation close to 0.1.

Practice Question

We select a simple random sample of size 400 from the above distribution and calculate the sample mean $\overline{X}$. The sampling distribution of $\overline{X}$ is:

(A) approximately normal with mean 5 and standard deviation 2.

(B) triangular with mean 5 and standard deviation 0.1.

(C) approximately normal with mean 5 and standard deviation 0.005.

(D) triangular with mean 5 and standard deviation 2.

(E) approximately normal with mean 5 and standard deviation 0.1.

Practice Question

Lengths of movies shown at a certain theater follow a normal distribution with mean 110 minutes and standard deviation 20 minutes. What is the probability that the average length of a random sample of ten movies is less than 123 minutes?

(A) 0.9564
(B) 0.9656
(C) 0.9761
(D) 0.9803
(E) impossible to calculate with the information given.

Practice Question

Lengths of movies shown at a certain theater follow a normal distribution with mean 110 minutes and standard deviation 20 minutes. What is the probability that the total length of a random sample of seven movies is greater than 700 minutes?

(A) 0.9066
(B) 0.9115
(C) 0.9279
(D) 0.9357
(E) 0.9474

Sample Proportions

So far in this unit, we’ve studied the sampling distribution of the sample mean $\bar{X}$. Now suppose that instead of being interested in the mean of some variable, we are interested in the proportion of individuals who possess some characteristic:

- the proportion of people with brown eyes
- the proportion of defective items produced in a large factory
- the proportion of tails in 100 flips of a quarter
- the proportion of voters who support the NDP

Sample Proportions

We denote the true **population proportion** by ***p***, and the **sample proportion** by $\hat{\boldsymbol{p}}$. We take a sample of size n and count the number of individuals X who possess some characteristic. (We call X the number of "successes".) Then the sample proportion is

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

Note that p is a **parameter**, whereas $\hat{p}$ is a **statistic** and an **estimator** for p (just as the statistic $\bar{x}$ is an estimator for the parameter μ).

Distribution of a Sample Proportion

Let $\hat{p}$ be the sample proportion of successes in a simple random sample drawn from a large population having population proportion p of successes.

It can be shown that the mean and standard deviation of $\hat{p}$ are

$$\mu_{\hat{p}} = p$$

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

Distribution of a Sample Proportion

For example, the true proportion of Canadians who live in Manitoba is $p = 0.036$. Suppose we didn’t know this true proportion and we wanted to estimate it. (Our populations of interest are often very large, so we often won’t know the true value of p). We could take a sample of $n = 1000$ Canadians and count the number of people X in our sample who live in Manitoba. The sample proportion of Manitobans in the sample is $\hat{p} = X/1000$.

Distribution of a Sample Proportion

Now suppose we calculated the sample proportion $\hat{p}$ for every possible sample of 1000 Canadians. Then the mean of the sample proportions would be

$$\mu_{\hat{p}} = p = 0.036$$

and the standard deviation would be

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.036(0.964)}{1000}} = 0.00589$$

Distribution of a Sample Proportion

The Central Limit Theorem says that if a variable represents a sample mean, then the sampling distribution of the variable is approximately normal when *n* is high. Specifically,

$$Z = \frac{variable - mean}{std.\ dev.} \;\dot{\sim}\; N(0,1)$$

But we can think of $\hat{p}$ as a kind of sample mean, because we are adding up the number of successes, and dividing by the sample size *n*.

Distribution of a Sample Proportion

So when the sample size *n* is high,

$$\hat{p} \,\dot{\sim}\, N\left(p, \sqrt{\frac{p(1-p)}{n}}\right) \quad \Rightarrow \quad Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \,\dot{\sim}\, N(0,1)$$

We can safely use this approximation provided that

$$np \geq 10 \quad \text{and} \quad n(1-p) \geq 10$$

and that the population is very large compared to the sample.

Comparing Means of Independent Samples

An illustration of what we mean by "the sampling distribution of $\hat{p}$"…

Imagine taking millions of random samples of 1000 Canadians and calculating the sample proportion of Manitobans $\hat{p}$ for each sample:

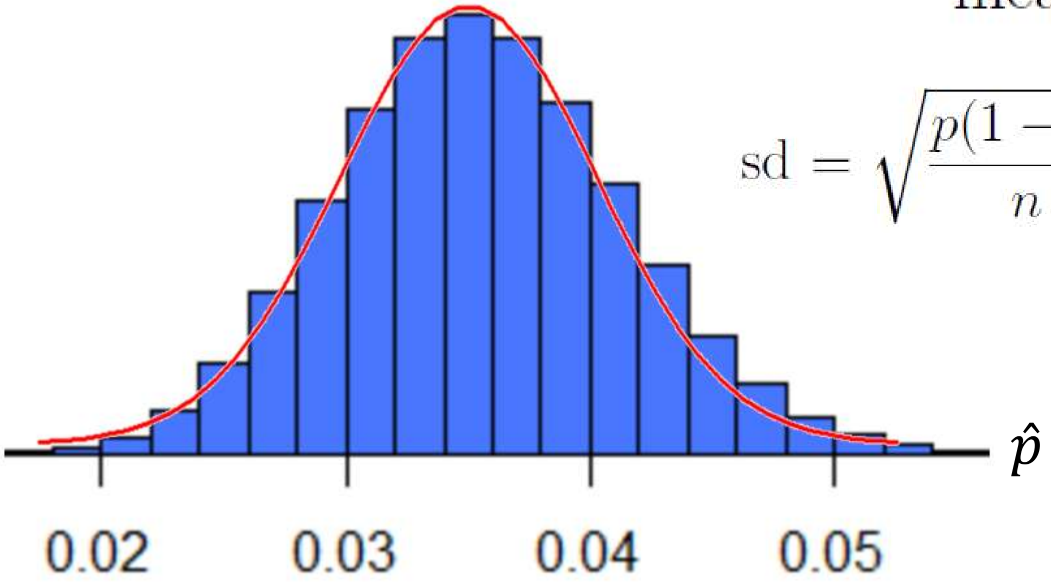
Comparing Means of Independent Samples

Now make a histogram of these millions of values:

Sample	n	Proportion of Manitobans $\hat{p}$
1	1000	0.042
2	1000	0.033
3	1000	0.029
4	1000	0.038
.	.	.
.	.	.
.	.	.
1,000,000	1000	0.035
.	.	.
.	.	.

Comparing Means of Independent Samples

And we get the sampling distribution of $\hat{p}$:



$$\text{mean} = p = 0.036$$

$$\text{sd} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.036(0.964)}{1000}} = 0.0589$$

Example

Suppose we randomly select 200 U of M students and ask them whether they are left- or right-handed. Assuming 10% of all people are left-handed, what is the approximate probability that at least 25 (12.5%) of the sampled students are left-handed?

Let $\hat{p}$ be the proportion of left-handed students in the sample. Then for a sample of size 200, the mean and standard deviation of $\hat{p}$ are

$$\mu_{\hat{p}} = p = 0.10$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.10(0.90)}{200}} = 0.02121$$

Example

We calculate $np = 200(0.10) = 20 > 10$ and $n(1 - p) = 200(0.90) = 180 > 10$. The population is large, so we can use the normal distribution.

The approximate probability at least 12.5% of the sampled students are left-handed is

$$P(\hat{p} \geq 0.125) \approx P\left(\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \geq \frac{0.125 - p}{\sqrt{\frac{p(1-p)}{n}}} \right) = P\left(Z \geq \frac{0.125 - 0.10}{0.02121} \right)$$

$$= P(Z \geq 1.18) = 1 - P(Z < 1.18) = 1 - 0.8810 = 0.1190.$$

Example

Suppose it is known that 22% of Canadians speak French. If we take a random sample of 500 Canadians, what is the approximate probability that less than 20% of them speak French?

Let $\hat{p}$ be the proportion of people in the sample who speak French. Then for a sample of size 500, the mean and standard deviation of $\hat{p}$ are

$$\mu_{\hat{p}} = p = 0.22$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.22(0.78)}{500}} = 0.01853$$

Example

We calculate $np = 500(0.22) = 110 > 10$ and $n(1 - p) = 500(0.78) = 390 > 10$. The population is large, so we can use the normal distribution.

The approximate probability that less than 20% of the Canadians in the sample speak French is

$$P(\hat{p} < 0.20) \approx P\left(Z < \frac{0.20 - 0.22}{0.01853}\right)$$

$$= P(Z < -1.08) = 0.1401$$

Practice Question

It is known that 20% of a certain type of lottery tickets are winners. If you buy 100 lottery tickets, what is the approximate probability that at least 25 of them are winners?

(A) 0.1056
(B) 0.1539
(C) 0.2061
(D) 0.2578
(E) 0.3085

Example

A slot machine wins on 17% of all spins. If you spin the slot machine 400 times, what is the approximate probability you win between 60 and 80 times?

We are looking for

$$P(60 < X < 80) = P\left(\frac{60}{400} < \hat{p} < \frac{80}{400}\right) = P(0.15 < \hat{p} < 0.20)$$

Example

Let $\hat{p}$ be the proportion of spins that are winners. The mean and standard deviation of $\hat{p}$ are

$$\mu_{\hat{p}} = p = 0.17$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.17(0.83)}{400}} = 0.0188$$

We calculate $np = 400(0.17) = 68 > 10$ and $n(1 - p) = 400(0.83) = 332 > 10$, and the population is large, so we can safely use the normal approximation.

Example

Therefore, the approximate probability that you win between 60 and 80 times is

$$P(0.15 < \hat{p} < 0.20) \approx P\left(\frac{0.15 - 0.17}{0.0188} < Z < \frac{0.20 - 0.17}{0.0188}\right)$$

$$= P(-1.06 < Z < 1.60) = P(Z < 1.60) - P(Z < -1.06)$$

$$= 0.9452 - 0.1446 = 0.8006$$

Practice Question

In a large city, 37% of households have a pet dog. We take a random sample of 275 households in the city. What is the approximate probability that less than 40% of them have a dog?

(A) 0.9082

(B) 0.7881

(C) 0.8485

(D) 0.9656

(E) 0.7224

Example

What if the sample size isn’t large enough to satisfy our conditions to use the normal distribution approximation for proportions? For example, if you flip a fair coin ten times, what is the probability you observe Heads exactly 7 times?

We won’t be able to calculate such probabilities in this course (so we wouldn’t ask you a question like this on an exam), but if you take either STAT 2150 or STAT 2400 in the future, you will learn to calculate exact probabilities such as this using something known as the binomial distribution.