Chapter 14: Sampling Distribution Models

A **population parameter** is a numerical summary of population data.
- unknown

A **sample statistic** is a numerical summary of sample data. use to estimate
- known after sample selected

Given a population of size N and a variable of interest, we consider three important distributions:

a) **Population distribution:** distribution of population data values

- population parameter, such as
  - population mean μ and population standard deviation σ. quantitative data
  - population proportion p. → categorical data

b) **Sample distribution**: distribution of a sample of data values (Data)
(sample size n)

- sample statistic, such as
  - sample mean $\bar{y}$ and sample standard deviation s.
  - sample proportion $\hat{p}$.

c) **Sampling distribution**: distribution of a sample statistic
(sample size n)

- for the distribution of $\bar{y}$, mean $\mu_{\bar{y}} = \mu$ and standard deviation $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$
- for the distribution of $\hat{p}$, mean $\mu_{\hat{p}} = p$ and standard deviation

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

$\bar{y}$: $n \geq 30$ $\quad$ $\hat{p}$: $np \geq 10$, $n(1-p) \geq 10$

usually

Central Limit Theorem: For a sufficiently large n, the sampling distribution of a sample statistic ($\bar{y}$ or $\hat{p}$) will be approximately normal (even if the population distribution is not).

Sampling Distributions

The value of a statistic depends on the specific sample selected from a population and it changes from sample to sample. This is known as **sampling variability** or **sampling error**.

For a given population and a fixed sample size n, a statistic is a **random variable** whose values are determined by taking all possible samples of size n from the population and computing the statistic for each one.

As such, a statistic (for a fixed sample size) has

Shape ?

- a probability distribution, called its **sampling distribution**
- a mean Centre
- a variance and a standard deviation Spread
  ↳ standard error

Sampling Distribution of a Proportion

For a given population, we often want to know what proportion of the population has a specific characteristic (categorical variable).
↳ Size N

Those who have the characteristic are called **successes** and those who do not are called **failures**.
S
F
- binary categorical variable

Notation:

$p =$ proportion of the population that has a specific characteristic.

$$= \frac{\#\ S \text{ in population}}{N}$$

$\hat{p} =$ the proportion of a random sample of size n that has a specific characteristic.

$$= \frac{\#\ S \text{ in sample}}{n}$$

ex. 73 people in a sample of 152 like hockey

$$\hat{p} = \frac{73}{152} \simeq 0.48$$

For a binary categorical variable with population proportion p and for a fixed sample size n,

- the average of all possible values of $\hat{p}$ is the mean $\mu_{\hat{p}}$ of the sampling distribution of $\hat{p}$ with sample size n. It is given by:

$\mu(\hat{p})$

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

- the standard deviation of all possible values of $\hat{p}$ is the standard deviation $\sigma_{\hat{p}}$ of the sampling distribution of $\hat{p}$ with sample size n. It is given by:

$\sigma(\hat{p})$, $SD(\hat{p})$

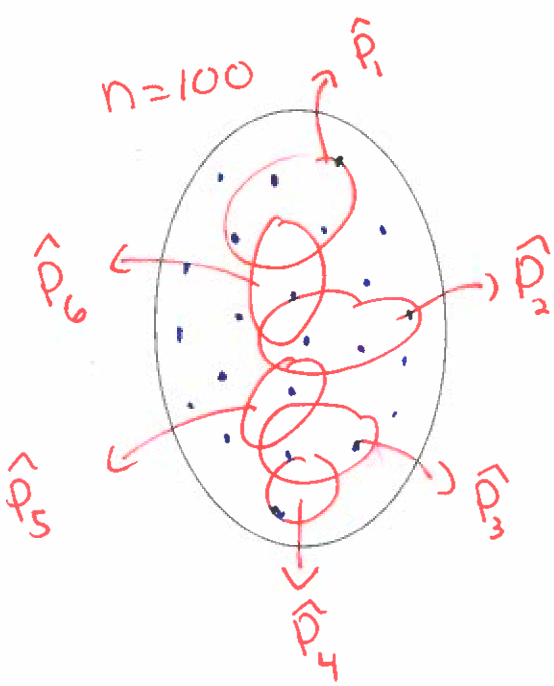
$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

$\uparrow n$ $\quad$ $\sigma_{\hat{p}} \downarrow$

Variability goes down

Example: One of the ways that people deal with stress is to eat sweets. Suppose the proportion of Canadians that eat sweets when stressed is $p = 0.46$. Find the mean and standard deviation of the sampling distribution of $\hat{p}$ with sample size $n = 100$.

$p = 0.46$



mean of all possible $\hat{p}$ s

$$\mu_{\hat{p}} = p = 0.46$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.46(0.54)}{100}}$$

$$\simeq 0.0498$$

standard deviation of all possible $\hat{p}$ s

Assumptions and Conditions for use of Normal Model:

Sample Proportion

Assumptions:

a) **Independence Assumption:** values in sample must be independent of each other.

b) **Sample Size Assumption:** sample size must be sufficiently large.

Conditions:

a) **Randomization Condition:** a simple random sample is selected from the population. → independence

b) **Success/Failure Condition:** there should be at least ten successes and ten failures in the sample, that is,

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

# of individuals in sample with characteristic.

c) **10% Condition:** sample size should be less than 10% of population size. → when taken without replacement

↳ more affects independence.

Central Limit Theorem for Sample Proportions

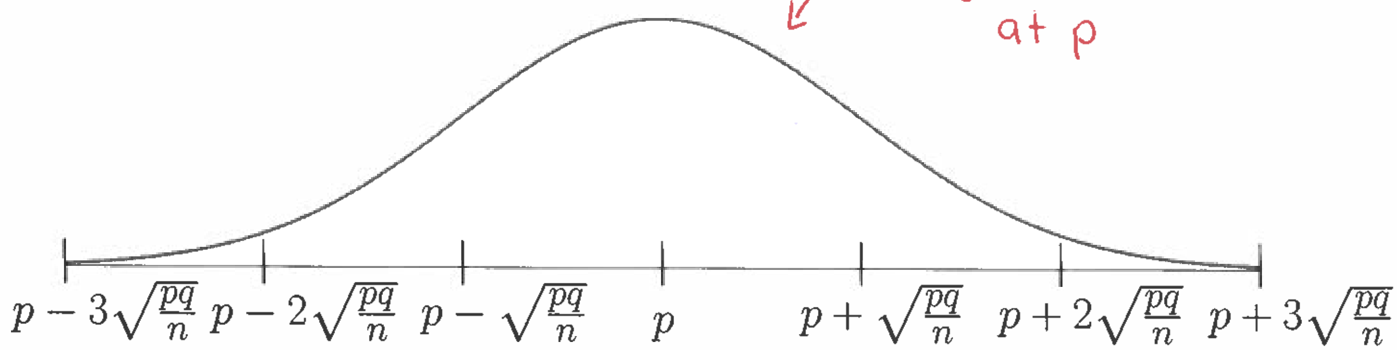
Provide that the sampled values are independent and that the sample size n is sufficiently large, the sampling distribution of $\hat{p}$ is approximately normal, that is, it can be described by a Normal model

$np \geq 10$
$n(1-p) \geq 10$

$$N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$$

with mean $\mu_{\hat{p}} = p$ and standard deviation $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$.

histogram centred at p



Note: The larger the sample size n and the closer p is to 0.5, the better the approximation. The closer p is to either 0 or 1, the larger n must be for the approximation to be reasonable.

→ Empirical Rule
(68 - 95 - 99.7 Rule)

→ Standardize to Z.

Example: $p = 0.3$

Suppose that 30% of all students at the U of A wear contact lenses.

a) If we randomly select a sample of $n = 20$ students, can we approximate the sampling distribution of $\hat{p}$ with a Normal model?

$np = 20(0.3) = 6 < 10$

No!

Now suppose that we select a random sample of $n = 100$ students.

b) What can you say about the sampling distribution of $\hat{p}$? • random sample

• $\left.\begin{cases} np = 100(0.3) = 30 \geq 10 \\ n(1-p) = 100(0.7) = 70 \geq 10 \end{cases}\right\}$ Success/Failure condition met

• $n = 100 <$ 10% of student population

$\therefore$ Sampling distribution of $\hat{p}$ with $n = 100$ is approx. normal (CLT)

c) Find the mean and standard deviation of the sampling distribution of $\hat{p}$. with $n = 100$

$\mu_{\hat{p}} = p = 0.3$, $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.3)(0.7)}{100}}$

$\simeq 0.0458$

$\hat{p} \sim N(0.3, 0.0458)$

d) What is the probability that less than a third of the students in this sample wear contacts?

$\frac{1}{3} \simeq 0.3333$

$P(\hat{p} < \frac{1}{3})$

$= P\left(z < \frac{0.3333 - 0.3}{0.0458}\right)$

$= P(z < 0.73)$

$= 0.7673$

Example: Suppose that a cable company includes the Shopping Channel in its basic cable package and that 20% of their customers watch it at least once a week. The cable company is trying to decide if it wants to continue to offer the Shopping Channel in its basic package or remove it. The company randomly selects a sample of 100 customers. The company will continue to offer the Shopping Channel if at least a quarter of those selected indicate that they watch it at least once a week. $\frac{1}{4} = 0.25$

a) Find the mean and standard deviation of the sampling distribution of $\hat{p}$.

$p = 0.2$, $n = 100$

$\mu_{\hat{p}} = p = 0.2$, $\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{(0.2)(0.8)}{100}}$

$= 0.04$

b) What can you say about the sampling distribution of $\hat{p}$?

$np = 100(0.2) = 20 \geq 10$

$n(1-p) = 100(0.8) = 80 \geq 10$

$\Rightarrow$ Sampling distribution of $\hat{p}$ is approx normal: $\hat{p} \sim N(0.2, 0.04)$

c) What is the probability that the company will keep the Shopping Channel in its basic package?

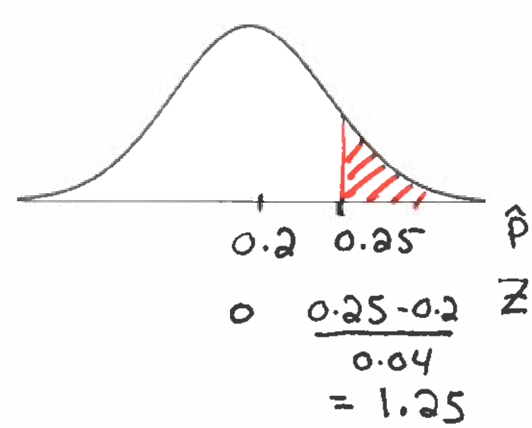
$P(\hat{p} \geq 0.25)$

$= P\left(z \geq \frac{0.25 - 0.2}{0.4}\right)$

$= P(z \geq 1.25)$

$= P(z \leq -1.25)$

$= 0.1056$



Sampling Distribution of a Mean

For a quantitative variable with population mean μ and population standard deviation σ, and for a fixed random sample size n,

- the average of all possible values of $\bar{y}$ is the mean $\mu_{\bar{y}}$ of the sampling distribution of $\bar{y}$ with sample size n. It is given by:

total = n(mean)

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

$\mu_{y_{total}} = n\mu$

- the standard deviation of all possible values of $\bar{y}$ is the standard deviation $\sigma_{\bar{y}}$ of the sampling distribution of $\bar{y}$ with sample size n. It is given by:

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

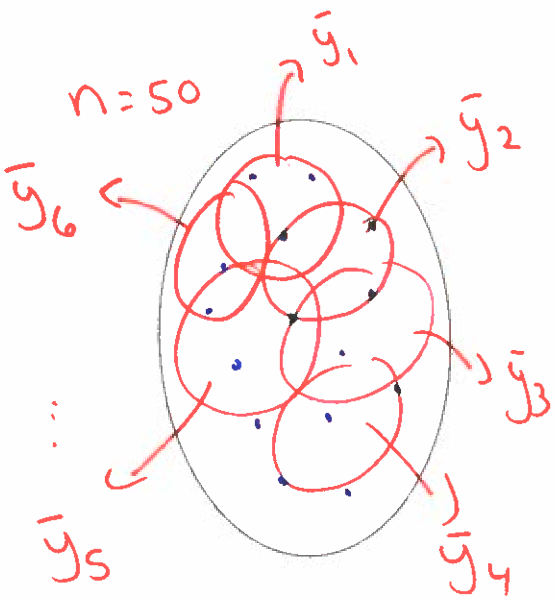
$\sigma_{y_{total}} = \sqrt{n}\,\sigma$

$\uparrow n \quad \sigma_{\bar{y}} \downarrow$

variability goes down.

Example: Suppose the length of the western diamondback rattlesnake has mean $\mu = 107$ cm and standard deviation $\sigma = 5.2$ cm. Find the mean and standard deviation of the sampling distribution of $\bar{y}$ with sample size $n = 50$.

$\mu = 107$, $\sigma = 5.2$



mean of all possible $\bar{y}$s

$\mu_{\bar{y}} = \mu = 107$

$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{5.2}{\sqrt{50}} \simeq 0.735$

standard deviation of all possible $\bar{y}$s

Assumptions and Conditions for use of Normal Model: Sample Means

Fact: If the population has a **Normal** distribution, then the sampling distribution of $\bar{y}$ with sample size n will be exactly Normally distributed, regardless of the sample size n.

Assumptions/Conditions for CLT:

a) **Independence Assumption:** values in sample must be independent of each other.

b) **Sample Size Assumption:** sample size must be sufficiently large, usually

$$n \geq 30$$

c) **Randomization Condition:** data values are randomly sampled.

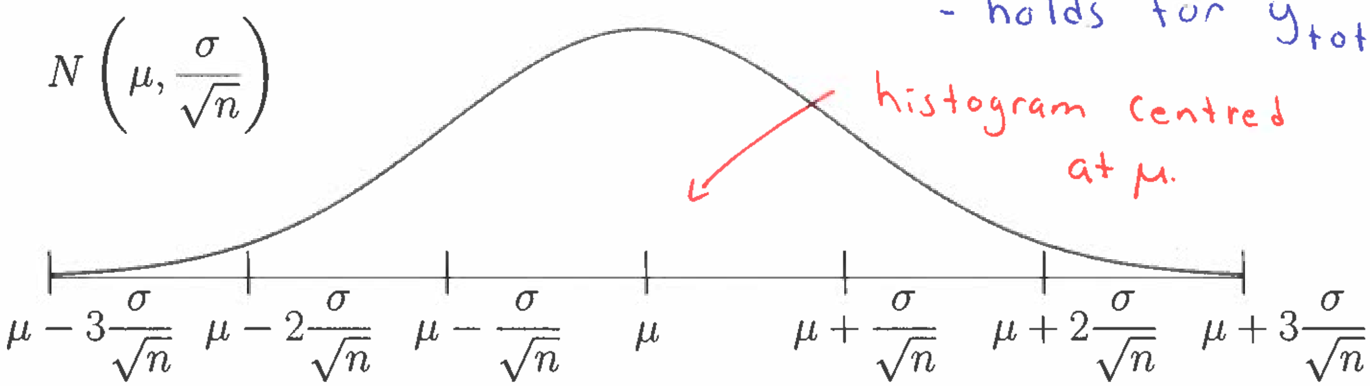
↳ independence

Central Limit Theorem (CLT)

If random samples of size n are selected from a population with mean μ and standard deviation σ, then when n is sufficiently large, the sampling distribution of $\bar{y}$ is approximately normally distributed with mean $\mu_{\bar{y}} = \mu$ and standard deviation $\sigma_{\bar{y}} = \dfrac{\sigma}{\sqrt{n}}$.

$n \geq 30$

- holds for y_{total}



Note: The CLT holds regardless of the distribution of the population. The approximation becomes better and better with increasing sample size.

Example: A company sells water-softener salt. Suppose that the bags contain an average of 40 lb of salt with a standard deviation of 1.5 lb and that the weights are normally distributed.

$\mu = 40$, $\sigma = 1.5$ pop. normally distributed

a) What is the probability that a randomly selected bag of water-softener salt will be 39 lb or less?

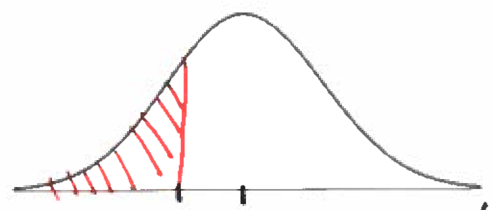
y = weight of bag

$P(y \le 39)$

$= P\left(z \le \frac{39-40}{1.5}\right)$

$= P(z \le -0.67)$

$= 0.2514$



b) What is the probability that the mean weight of 10 randomly selected bags of water-softener salt will be 39 lb or less? Sampling distribution

$\bar{y}$ = average weight of 10 bags

The Sampling distribution for $\bar{y}$ with $n = 10$ has

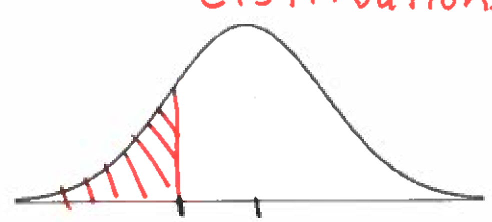
- $\mu_{\bar{y}} = \mu = 40$
- $\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}} = \frac{1.5}{\sqrt{10}} \approx 0.4743$
- normal distribution since population is normal

$P(\bar{y} < 39)$

$= P\left(z < \frac{39-40}{0.4743}\right)$

$= P(z < -2.11)$

$= 0.0174$



Example: Suppose that the mean value of the interpupillary distance for all adult males is 65 mm and the population standard deviation is 5 mm. What is the probability that a random sample of 100 males has a mean between 64.7 mm and 66.2 mm?

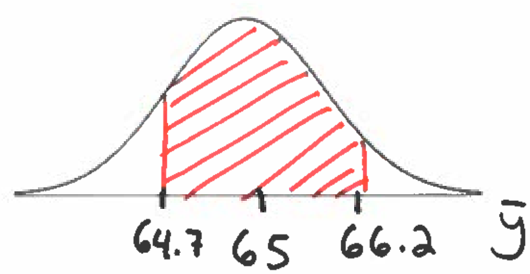
↳ Sampling distribution

$\mu = 65$, $\sigma = 5$, $n = 100 \geq 30$

$\bar{y}$ = mean interpupillary distance in a sample of 100 males

The sampling distribution for $\bar{y}$ with $n = 100$ has:

- $\mu_{\bar{y}} = \mu = 65$ $\qquad \bar{y} \sim N(65, 0.5)$
- $\sigma_{\bar{y}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{5}{\sqrt{100}} = 0.5$
- normal distribution, by CLT since $n = 100 \geq 30$



$\dfrac{64.7-65}{0.5} = -0.6 \qquad 0 \qquad \dfrac{66.2-65}{0.5} = 2.4 \qquad z$

$P(64.7 < \bar{y} < 66.2)$

$= P\left(\dfrac{64.7-65}{0.5} < z < \dfrac{66.2-65}{0.5}\right)$

$= P(-0.6 < z < 2.4)$

$= P(z < 2.4) - P(z < -0.6)$

$= 0.9918 - 0.2743$

$= 0.7175$

Example: The number of complaints per day received by a cell phone company has a mean of 1.1 and a standard deviation of 1.136. What is the probability that the company will receive more than 105 complaints in 90 days?

↳ in total

$n = 90$, $\mu = 1.1$, $\sigma = 1.136$, population not normally distributed (discrete variable)

y = # complaints per day

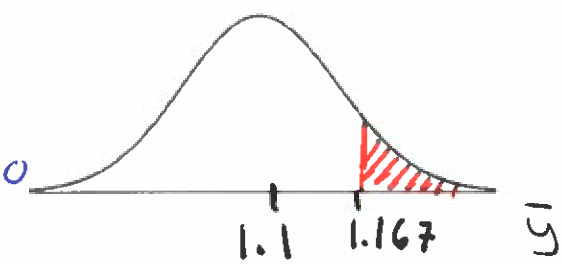
y_{total} = # complaints in total in 90 days

$\bar{y}$ = mean # complaints per day during 90 days

The sampling distribution of $\bar{y}$ with $n = 90$ has:

- $\mu_{\bar{y}} = \mu = 1.1$
- $\sigma_{\bar{y}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{1.136}{\sqrt{90}} = 0.1197$
- normal distribution, by CLT since $n = 90 \geq 30$

$\bar{y} \sim N(1.1, 0.1197)$



$0 \qquad \dfrac{1.167 - 1.1}{0.1197} = 0.56 \qquad z$

$P(y_{total} > 105)$

$= P\left(\dfrac{y_{total}}{90} > \dfrac{105}{90}\right)$

$= P(\bar{y} > 1.167)$

$= P\left(z > \dfrac{1.167 - 1.1}{0.1197}\right)$

$= P(z > 0.56) = P(z < -0.56) = 0.2877$