Review of Statistics
(SW Ch. 3)

Ercan Karadas

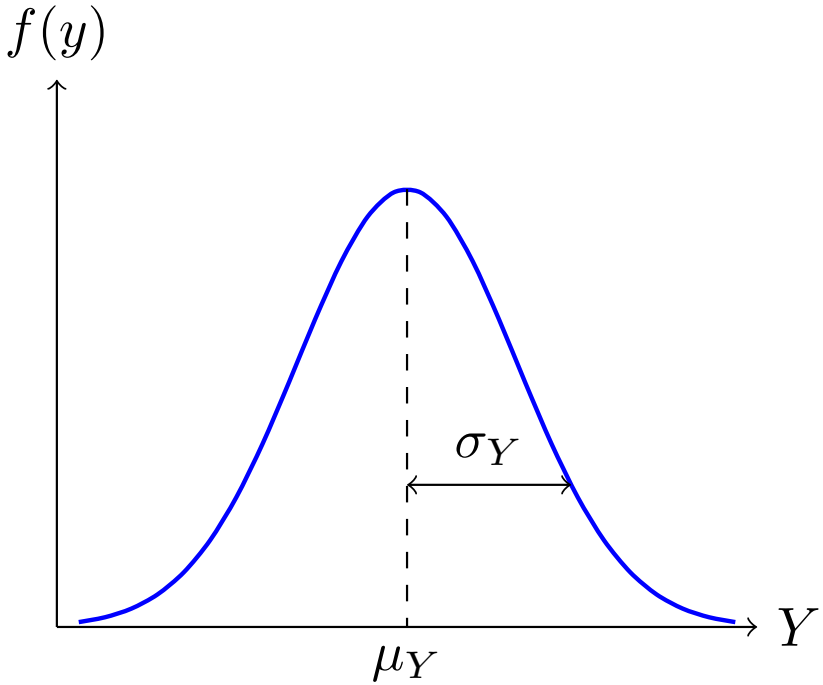
Econometrics (ECON 3112)

Belk College of Business, UNCC

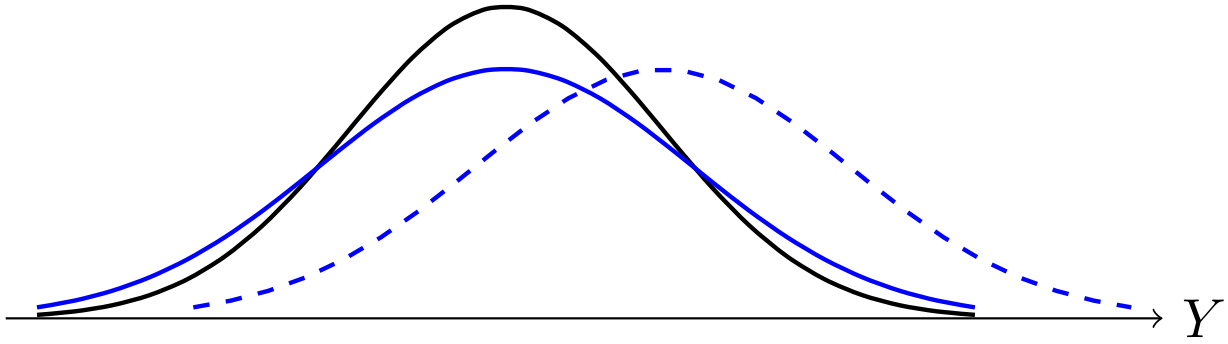
September 6, 2018

Outline

The Normal Distribution



- The most important distribution of all:
  - Closely approximates the p.d. of a wide range of phenomena
  - Distributions of sample means approaches a normal distribution as the sample size gets larger (more on this later)
- μ_Y determines the location,
- σ_Y determines the spread,
- The normal distribution is
  - Bell shaped
  - Symmetrical
  - Mean, Median and Mode are equal to μ_Y

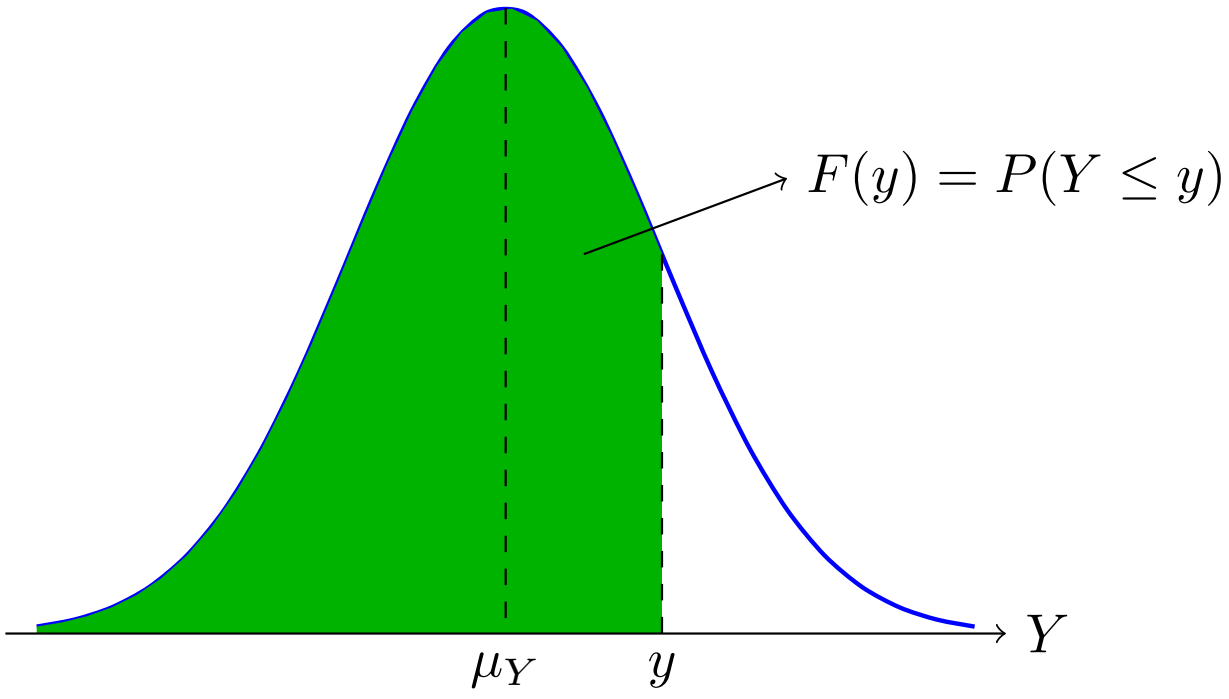


- By varying the parameters μ_Y and σ_Y, we obtain different normal distributions
  - Changing μ_Y shifts the distribution left or right
  - Changing σ_Y increases or decreases the spread,
  - The normal distribution is
- Given the mean μ_Y and variance σ_Y^2 we define the normal distribution using the notation

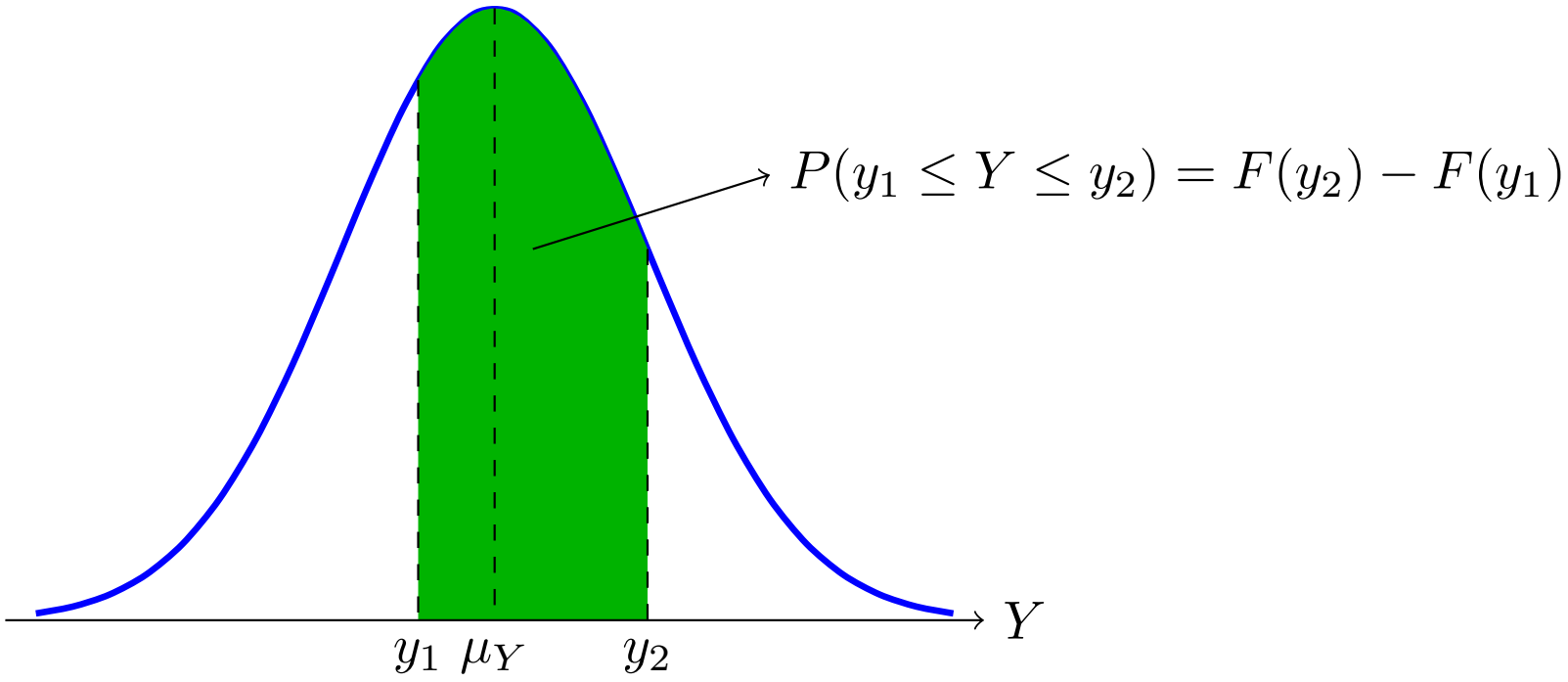
$$Y \sim N\left(\mu_Y, \sigma_Y^2\right)$$

Read: Y is normally distributed with mean μ_Y and variance σ_Y^2

Cumulative Normal Distribution



Finding Probabilities with The Normal Distribution



The Standard Normal Distribution

- Suppose we have a r.v. X

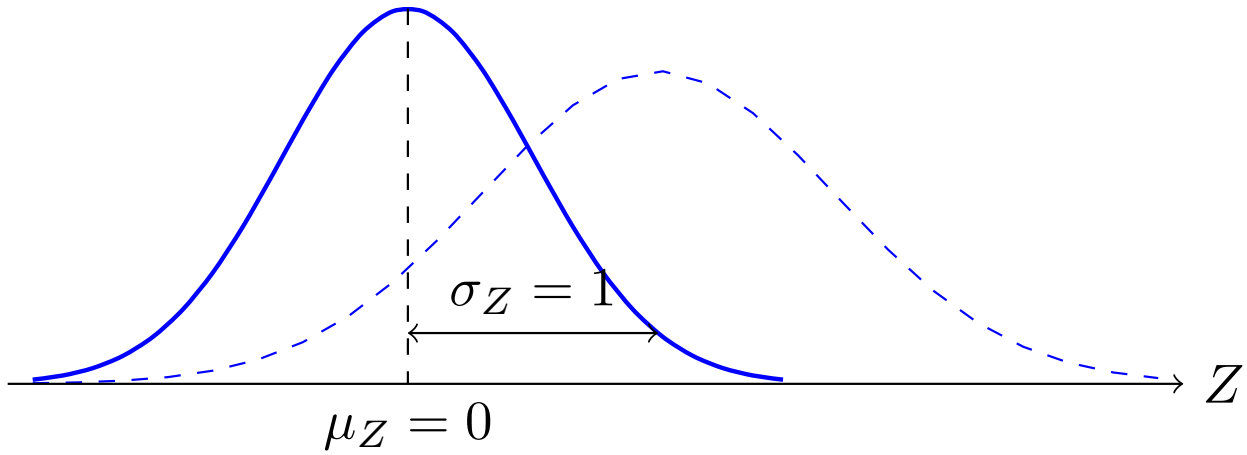
$$Y \sim N\left(\mu_Y, \sigma_Y^2\right)$$

Then

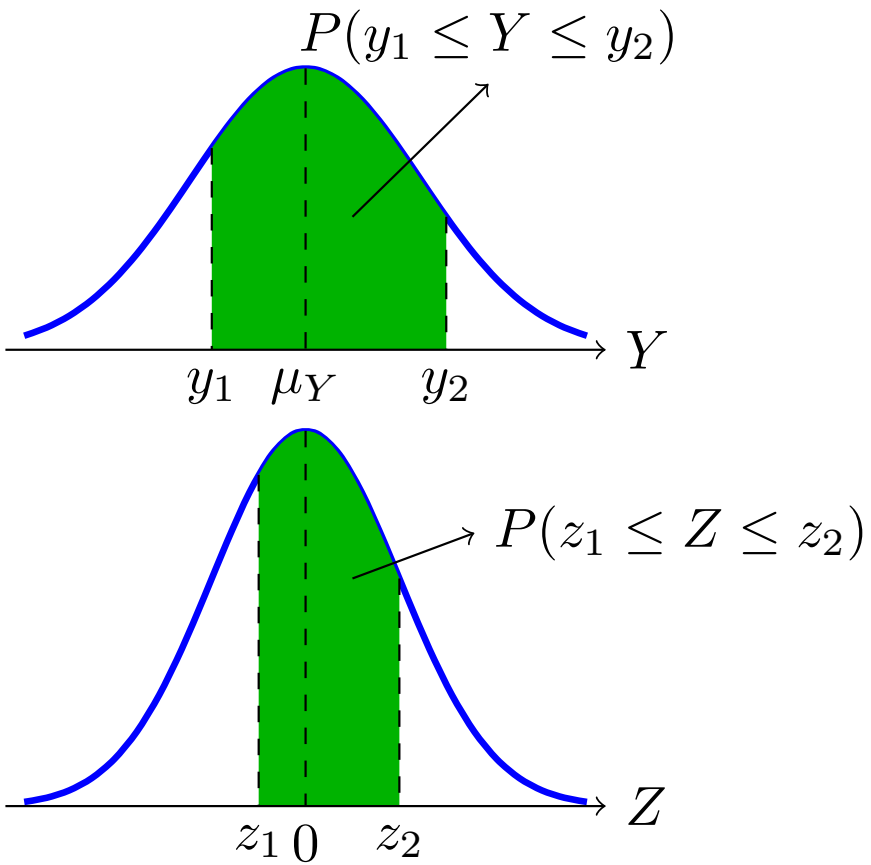
$$Z = \frac{Y - \mu_Y}{\sigma_Y} \sim N(0, 1)$$

i.e. Z transforms Y to a normal distribution with mean 0 and variance 1.

- The distribution $Z \sim N(0, 1)$ is called the **standard normal distribution**.



Finding Probabilities using The Standard Normal Distribution



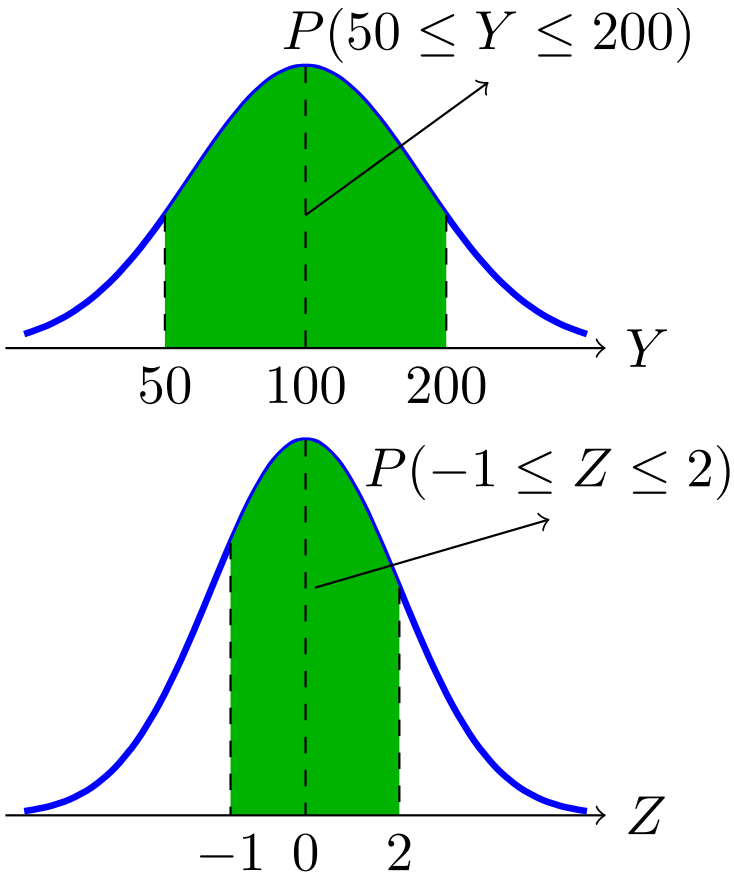
To find $P(y_1 \leq Y \leq y_2)$

- Draw the normal curve for the problem in terms of $Y \sim N(\mu_Y, \sigma_Y^2)$
- Translate Y-values to Z-values

$$z_1 = \frac{y_1 - \mu_Y}{\sigma_Y} \text{ and } z_2 = \frac{y_2 - \mu_Y}{\sigma_Y}$$

- Use the Standard Cumulative Normal Table

Example: $Y \sim N(100, 50^2) \implies P(50 \leq Y \leq 200) = ?$



To find $P(y_1 \leq Y \leq y_2)$

- Draw the normal curve in terms of $Y \sim N(100, 50^2)$
- Translate Y-values to Z-values

$$z_1 = \frac{50 - 100}{50} = -1.0 \text{ and}$$

$$z_2 = \frac{200 - 100}{50} = 2.0$$

- Use the Z-Table

$$\begin{aligned} &P(y_1 \leq Y \leq y_2) \\ &= P(z_1 \leq Z \leq z_2) \\ &= P(Z \leq 2) - P(Z \leq -1) \\ &= 0.9772 - 0.1587 = 0.8185 \end{aligned}$$

Linear Combinations of Two Random Variables

- For two continuous r.v.'s X and Y the following linear combination:

$$W = aX \pm bY$$

  defines a new r.v. just like in the discrete case.

- The **Expected Value** for the linear combination of two r.v.'s

$$E[W] = aE(X) \pm bE(Y) \quad \text{or}$$
$$\mu_W = a\mu_X \pm b\mu_Y$$

- The **Variance** for the linear combination of two r.v.'s

$$Var[W] = a^2 Var(X) + b^2 Var(Y) \pm 2abCov(X,Y) \quad \text{or}$$
$$\sigma_W^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 \pm 2abCov(X,Y)$$

Linear Combinations of Two Normal Random Variables

- In general, even if we manage to compute $E(W)$ and $Var(W)$, we can not say much about the shape of its distribution.
- However, when X and Y are normal r.v.s: $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_Y^2)$, then W is normally distributed as well and

$$W \sim N(\mu_W, \sigma_W^2)$$

where

$$\mu_W = a\mu_X \pm b\mu_Y$$
$$\sigma_W^2 = a^2\sigma_X^2 + b^2\sigma_Y^2 \pm 2abCov(X, Y)$$

- This is a very important property because it basically allows us to compute all kinds of probabilities that we might be interested in regarding W.

Example

- Two tasks, X and Y, must be performed by the same worker
  - The worker completes X in 20 minutes on average, with a standard deviation of 5.
  - The worker completes Y in 30 minutes on average, with a standard deviation of 8.
  - Suppose X and Y are normally distributed and independent.
- What is the probability that the worker will complete both tasks in less than 45 minutes?
  - Define $W = X + Y$
  - Then
$$\begin{aligned} \mu_W &= \mu_X + \mu_Y \\ &= 20 + 30 \\ &= 50 \end{aligned} \qquad \begin{aligned} \sigma_W^2 &= \sigma_X^2 + \sigma_Y^2 + 2Cov(X,Y) \\ &= 5^2 + 8^2 + 2 \times 0 \\ &= 89 \end{aligned}$$
  - Therefore, $W \sim N(50, 89)$, and finally
$$\begin{aligned} P(W < 45) &= P(Z < \frac{45 - 50}{\sqrt{89}}) \\ &= P(Z < -0.535) \approx 0.3 \end{aligned}$$

Example

- Consider two stocks, A and B.
  - The price of stock A is normally distributed with mean 12 and standard deviation of 4.
  - The price of stock B is normally distributed with mean 20 and standard deviation of 16.
  - Stock prices are positively correlated with correlation $\rho_{AB} = 0.5$.
- Suppose you own 10 shares of A and 30 shares of B.
- What is the probability that your portfolio value will be less than \$ 500?

What is the probability that your portfolio value will be less than $ 500?

- Define $W = 10A + 30B$
- Then,

$$\begin{aligned}
\mu_W &= 10\mu_A + 30\mu_B \\
&= (10)(12) + (30)(20) \\
&= 720 \\
\sigma_w^2 &= 10^2\sigma_A^2 + 30^2\sigma_B^2 + 2(10)(30)Cov(X,Y) \\
&= 10^2(4^2) + 30^2(16^2) + 2(10)(30)(0.5)(4)(16) \\
&= 251,200
\end{aligned}$$

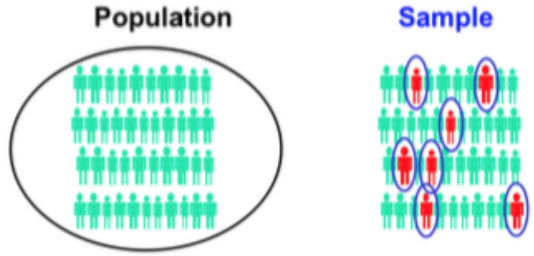
- Therefore, $W \sim N(720,\, 251\,200)$, and finally

$$\begin{aligned}
P(W < 500) &= P(Z < \frac{500 - 720}{\sqrt{251,200}}) \\
&= P(Z < -0.44) \approx 0.33
\end{aligned}$$

$\rightarrow$ So the probability is 0.33 that your portfolio value will be less than $500.

Sampling from a Population

- A **Population** is the set of all items, individuals or entities of interest.
  - All likely voters in the next election
  - All sales receipts for August
- A **Sample** is a subset of the population
  - 1000 voters selected at random for interview
  - Random 350 receipts selected for auditing



- Why Sampling?
  - Less time consuming and less costly to administer than a census
  - It is possible to obtain statistical results of a sufficiently high precision based on samples

Simple Random Sampling

- In a Simple Random Sample, n objects are drawn such that
  (i) Every object in the population has the same probability of being selected, i.e. **identically** drawn

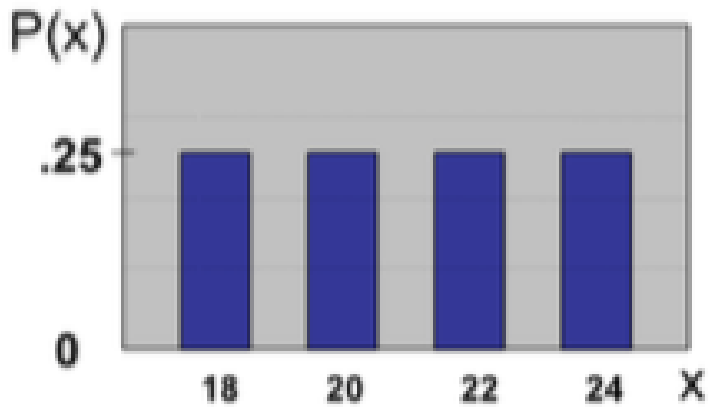
  (ii) Objects are selected **independently**
- When $Y_1, \ldots, Y_n$ are drawn from the same distribution and are independently distributed, they are said to be independently and identically distributed or simply we say Y_i's are **i.i.d.**
- A simple random sample is the ideal against which other sampling methods are compared
- Decide whether the sampling is random in the following cases:
  - Interview college students to predict the next election
  - Use ice cream sales in August to predict the average ice cream consumption
  - Test the quality of a river's water by taking samples of the water from multiple places in the river, on different dates and at times

Sampling Distribution of $\overline{Y}$: Examples

- A **sampling distribution** is a probability distribution of all the possible values of a statistic for a given size sample selected from a population
- For example, the mean, maximum, median are all statistics and we can talk about their sampling distributions
- But we are mostly be interested in the sampling distribution of sample means $\overline{Y}$

Example 1: Developing a Sampling Distribution from a Uniform Distribution

- Consider a population of college students with 4 different age groups: 18, 20, 22, 24
- Let r.v. X denote age of individuals, $X \in \{18, 20, 22, 24\}$
- And suppose the population distribution of X is uniform:



- Now consider all possible samples of size $n = 2$:

1st Obs	2nd Observation			
	18	20	22	24
18	18,18	18,20	18,22	18,24
20	20,18	20,20	20,22	20,24
22	22,18	22,20	22,22	22,24
24	24,18	24,20	24,22	24,24

16 possible samples (sampling with replacement)

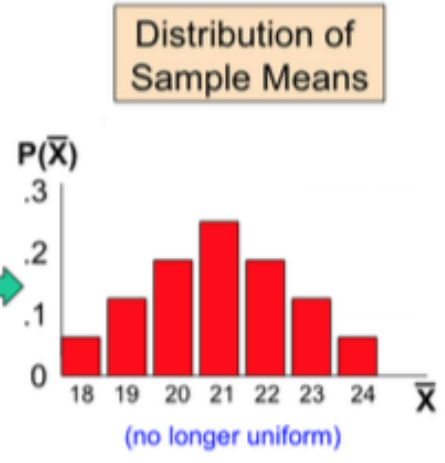
16 Sample Means

1st Obs	2nd Observation			
	18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24

- Sampling Distribution of All Sample Means:

16 Sample Means

1st Obs	2nd Observation			
	18	20	22	24
18	18	19	20	21
20	19	20	21	22
22	20	21	22	23
24	21	22	23	24



Example 2: Developing a Sampling Distribution from a Bernoulli Distribution

- Let Y be a r.v. with the following probability distribution (Bernoulli)

$$P(Y = 0) = 0.22 \text{ and } P(Y = 1) = 0.78$$

- Suppose we take two draws from this distribution, Y_1 and Y_2, and then compute

$$\overline{Y} = \frac{Y_1 + Y_2}{2}$$

- What is the distribution of $\overline{Y}$?

- What are the all possible values for $\overline{Y}$?

$$\overline{Y} \in \{0, 1/2, 1\}$$

- What are the corresponding probabilities?

$$P(\overline{Y} = 0) = .22^2 = .0484$$
$$P(\overline{Y} = 1/2) = 2 \times .22 \times .78 = .3432$$
$$P(\overline{Y} = 1) = .78^2 = .6084$$

$\overline{Y}$	0	1/2	1
$P(\overline{Y})$	.0484	.3432	.6084

- Compare $E(Y)$ and $E(\overline{Y})$:

$$E(Y) = .22 \times 0 + .78 \times 1 = .78$$
$$E(\overline{Y}) = .0484 \times 0 + .3432 \times 1/2 + .6084 \times 1 = .78$$

- Compare $V(Y)$ and $V(\overline{Y})$:

$$V(Y) = .22 \times (0 - .78)^2 + .78 \times (1 - .78)^2 = .1716$$
$$V(\overline{Y}) = .0484 \times (0 - .78)^2 + .3432 \times (1/2 - .78)^2 + .6084 \times (1 - .78)^2$$
$$= .0858 (= 0.1716/2)$$

- Suppose we now take five draws (n=5) from the same Bernoulli distribution, $Y_1, \ldots, Y_5$, and then compute

$$\overline{Y} = \frac{Y_1 + \ldots + Y_5}{5} = \frac{1}{5}\sum_{i=1}^{5} Y_i$$

- The same question: what is the distribution of $\overline{Y}$?
- What are the all possible values for $\overline{Y}$?

$$\overline{Y} \in \{0, 1/5, 2/5, 3/5, 4/5, 1\}$$

- What are the corresponding probabilities?

$\overline{Y}$	0	1/5	2/5	3/5	4/5	1
$P(\overline{Y})$	.0005	.0091	.0647	.2298	.4071	.2887

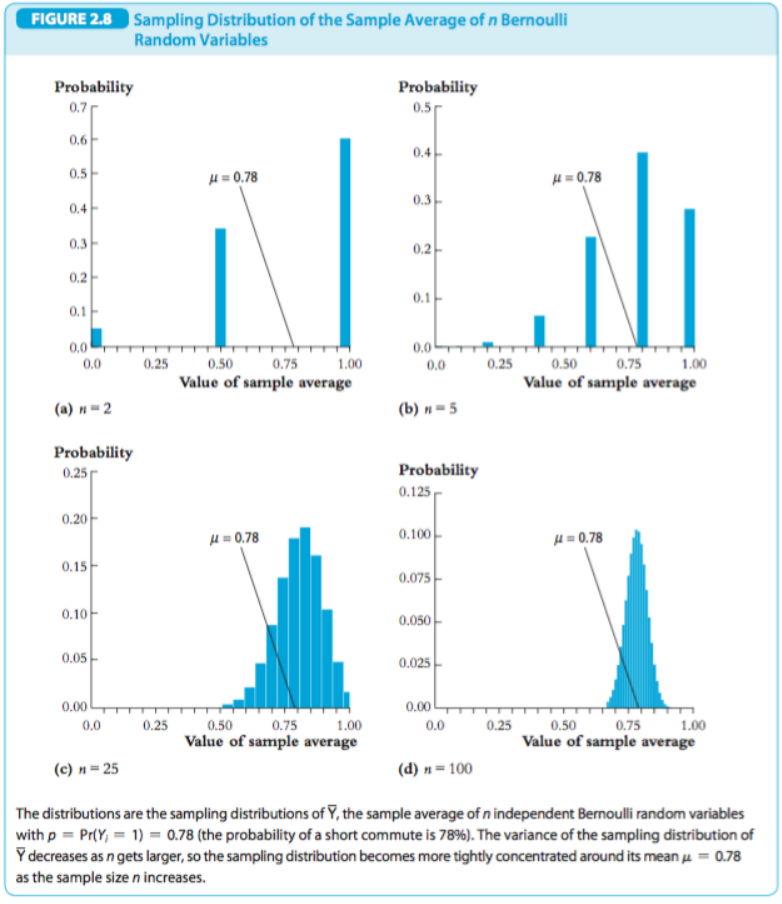
where we used $P(\overline{Y} = y/5) = C_y^5 (0.78)^y (0.22)^{5-y}$ to compute the probabilities.

- Compute $E(\overline{Y})$ and $V(\overline{Y})$:

$$E(\overline{Y}) = .78$$

$$V(\overline{Y}) = .03432 (= .1716/5)$$

The Sampling Distribution of $\overline{Y}$ for Different n's



- **Observation:** as the sample size (n) increases, the distribution of $\overline{Y}$ approaches to a Normal distribution!

Sampling Distribution of $\overline{Y}$: General Case

General Case: consider a sample of size n drawn identically and independently from a population distribution of Y with the mean μ_Y and variance σ^2_Y:

$$\{Y_1, Y_2, \ldots, Y_n\}$$

Want: Characterize the sampling distribution of $\overline{Y}$, i.e. $P(\overline{Y})$?

Properties of $\overline{Y}$

Let us start by studying the main statistical properties of $\overline{Y}$:

- $\overline{Y}$ is an unbiased estimator of the population mean:

$$E(\overline{Y}) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\left(\sum_{i=1}^{n} E(Y_i)\right)$$
$$= \frac{1}{n}\left(\sum_{i=1}^{n} \mu_Y\right) \qquad (Y_i's \text{ have identical means})$$
$$= \mu_Y$$

- $\overline{Y}$ is a consistent estimator of the population mean:

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \longrightarrow \mu_Y$$

We will not prove this result but, intuitively, the fraction of heads $(\overline{Y})$ approaches to $1/2$, which is equal to the population mean μ_Y, if we keep tossing an unbiased coin many many times.

- The variance of the sample mean:

$$
\begin{aligned}
V(\overline{Y}) \equiv \sigma^2_{\overline{Y}} &= V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2} V\left(\sum_{i=1}^{n} Y_i\right) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n} V(Y_i)\right) \qquad (Y_i's \text{ are independent}) \\
&= \frac{1}{n^2}\left(\sum_{i=1}^{n} \sigma^2_Y\right) \qquad (Y_i's \text{ have identical variances}) \\
&= \frac{\sigma^2_Y}{n}
\end{aligned}
$$

(!) We can always calculate $E(\overline{Y})$ and $V(\overline{Y})$ in terms of the population parameters, but, in general, these two are not enough to characterize the distribution of $\overline{Y}$, **except** when Y is normally distributed, which is the case we now turn to...

Sampling Distribution of $\overline{Y}$ when Y is Normal

Consider a sample of size n identically and independently drawn from a Normal distribution, $Y \sim N(\mu_Y, \sigma_Y^2)$:

$$\{Y_1, Y_2, \ldots, Y_n\}$$

Then

- As before, we have these properties:

$$E(\overline{Y}) = \mu_Y \quad \text{and} \quad V(\overline{Y}) = \frac{\sigma_Y^2}{n}$$

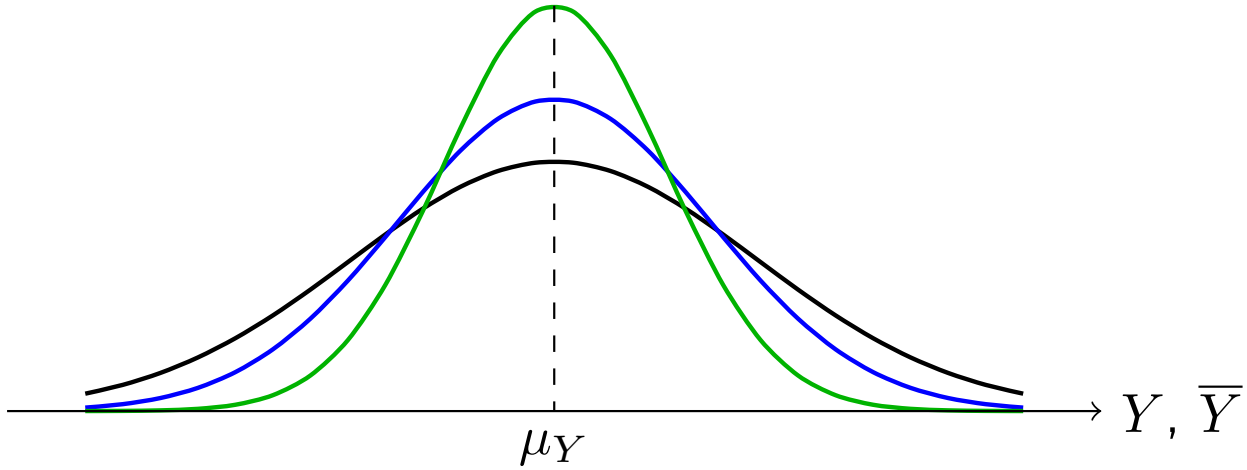
- And more importantly, we now also have

$$\overline{Y} \sim N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right)$$

  i.e. $\overline{Y}$ is normally distributed with mean μ_Y and variance $\frac{\sigma_Y^2}{n}$

- This result holds for any sample size n, i.e. no matter whether n is large or small

- The population distribution of Y and the sampling distributions of $\overline{Y}$ for samples of size n=2 and n=5...



- What about when Y is not Normally distributed?
  - In general, it is difficult to characterize the sampling distribution of $\overline{Y}$
  - However, **when n is large**, as we will see, the sampling distribution of $\overline{Y}$ is **approximately** Normal ...

Sampling Distribution of $\overline{Y}$ when the Sample Size is Large ($n \geq 30$)

Consider a sample of size $n \geq 30$ drawn identically and independently from a population distribution Y with the mean μ_Y and variance σ_Y^2.
We have **two fundamental results (theorems)**:

(1) As the sample size n increases, the distribution of $\overline{Y}$ becomes more and more tightly centered around μ_Y

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \longrightarrow \mu_Y \text{ as } n \uparrow$$

This result is known as **The Law of Large Numbers (LLN)**

(2) As the sample size n increases, the distribution of $\overline{Y}$ becomes approximately Normal

$$\overline{Y} \underset{approx.}{\sim} N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right) \qquad \text{when } n \geq 30$$

This result is known as **The Central Limit Theorem (CLT)**

The Law of Large Numbers (LLN)

If $\{Y_1, Y_2, \ldots, Y_n\}$ are i.i.d. and $0 < \sigma_Y^2 < \infty$, then

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \longrightarrow \mu_Y \text{ as } n \uparrow$$

- Here $\overline{Y}$ is said to be a **consistent** estimator of μ_Y.
- In general, an estimator is **consistent** if the probability that it falls within an interval of the true population value tends to one as the sample size $\uparrow$
- This is a desired property for an estimator because as the sample size increases we want our estimator to be more precise in locating the unknown population parameter
- So far we have seen that $\overline{Y}$ has two desired properties as a point estimator of μ_Y:
  - $E(\overline{Y}) = \mu_Y$, i.e. $\overline{Y}$ is an unbiased estimator of μ_Y
  - $\overline{Y} \longrightarrow \mu_Y$, i.e. $\overline{Y}$ is a consistent estimator of μ_Y

The Central Limit Theorem (CLT)

If $\{Y_1, Y_2, \ldots, Y_n\}$ are i.i.d. and $0 < \sigma_Y^2 < \infty$, then

$$\overline{Y} \underset{approx.}{\sim} N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right) \qquad \text{when } n \geq 30$$

- There is an alternative way of stating the CLT in terms of the standard normal distribution:

$$\frac{\sqrt{n}\left(\overline{Y} - \mu_Y\right)}{\sigma_Y} \underset{approx.}{\sim} N\left(0, 1\right) \qquad \text{when } n \geq 30$$

- The larger is n, the better is the approximation.
- Here the condition is a rule of thumb, recall that
  - in our uniform distribution example, the sampling distribution of $\overline{Y}$ was quite close to a normal distribution even for $n = 2$
  - on the other hand, in our Bernoulli example, we needed $n = 100$ to obtain a good approximation

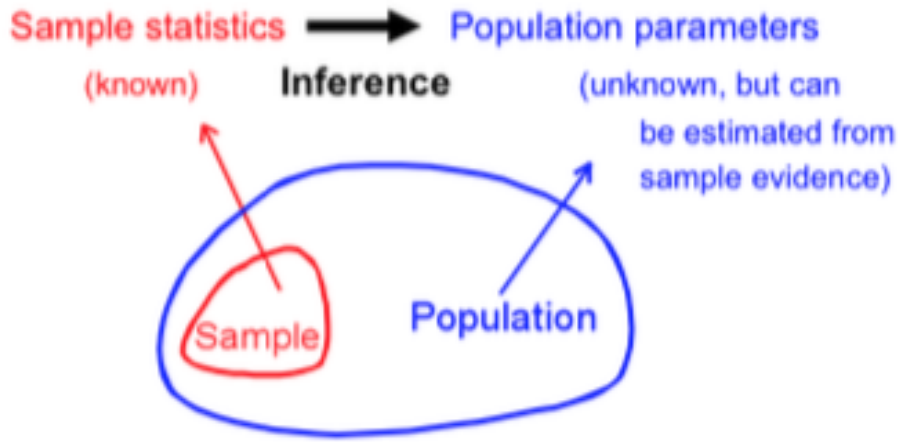
Summary: The Sampling Distribution of $\overline{Y}$

For $\{Y_1, Y_2, \ldots, Y_n\}$ i.i.d. with $0 < \sigma_Y^2 < \infty$, we have seen that

- The exact (finite sample) sampling distribution of $\overline{Y}$ has mean μ_Y and variance $\sigma_{\overline{Y}}^2 = \sigma_Y^2/n$
- Other than its mean and variance, the exact distribution of $\overline{Y}$ is complicated and depends on the distribution of Y (the population distribution)
- When n is large ($n \geq 30$), the sampling distribution simplifies:
  - $\overline{Y} \longrightarrow \mu_Y$ **(The Law of Large Numbers)**
  - $\overline{Y} \underset{approx.}{\sim} N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right)$ **(The Central Limit Theorem)**

Inferencial Statistics

- **Inferencial Statistics:** Making statements about a population by examining sample results



- Suppose you want to know the population mean weight. A natural way to make inferences about this mean is to compute the sample average
  - **Estimation:** e.g. estimate the population mean weight using the sample mean weight $\overline{Y}$. Here $\overline{Y}$ is called an **estimator** for the population mean μ_Y.
  - **Hypothesis Testing:** e.g. use sample evidence to test the claim that the population mean weight is 120 pounds

Estimators and Estimates

- An **estimator** is a function of a sample data to be drawn randomly from a population.
- An **estimate** is the numerical value of the estimator when it is actually computed using data from a specific sample.
- For instance, $\overline{Y}$ is an estimator (for the population mean), and for a particular sample suppose $\overline{Y} = 120$, this is an estimate.

Unbiasedness, Consistency and Efficiency

There are three desirable characteristics of an estimator

- **Unbisadness:** When you evaluate an estimator many times over repeated randomly drawn sample, if you obtain the population parameter on average, the estimator is called unbiased. Mathematically, if $E(\hat{\mu}) = \mu$, then $\hat{\mu}$ is said to be an unbiased estimator of μ.
- **Consistency:** If the estimator gets closer and closer to the population parameter as the sample size increases, then the estimator is said to be a consistent estimator.
- **Efficiency:** If we have two candidate estimators $\hat{\mu}$ and $\tilde{\mu}$, we pick the one with smaller variance.

Estimation of the Population Mean

- Suppose you want to know the mean value of Y (that is, μ_Y) in a population, such as the mean earnings of women recently graduated from college.
- A natural way to estimate this mean is to compute the sample average $\overline{Y}$ from a sample of n i.i.d. observations, $\{Y_1, \ldots, Y_n\}$.
- $\overline{Y}$ has the following properties as an estimator of μ_Y:
  - $\overline{Y}$ is an unbiased estimator of μ_Y, i.e. if you flip an unbiased coin repeatedly half of the time you would observe Heads on average
  - $\overline{Y}$ is a consistent estimator of μ_Y, i.e. if you start flipping an unbiased coin, the fraction Heads would get closer to $1/2$ as you keep flipping
  - $\overline{Y}$ is an efficient estimator of μ_Y in the sense that its variance is the lowest among all linear unbiased estimators

The Main Concepts of Hypothesis Testing

- A **hypothesis** is a claim/conjecture about a population parameter:
  - The mean monthly cell phone bill in NYC is \$52, i.e. $\mu = 52$
  - Standard deviation of SAT scores of NYU students is a least 25, i.e. $\sigma > 25$.
- **The Null Hypothesis,** H_0 states the claim/conjecture to be tested. It is always numeric.
  - In the examples above

$$H_0 : \mu = 52 \quad \text{and} \quad H_0 : \sigma > 25$$

  - It is always about a population parameter, not about a sample statistics

$$H_0 : \mu = 52, \quad \cancel{H_0 : \overline{X} = 52}$$

- **The Alternative Hypothesis,** H_1 is the opposite of the null hypothesis.

$$H_1 : \mu \neq 52 \quad \text{and} \quad H_1 : \sigma \leq 25$$

H_0 vs H_1

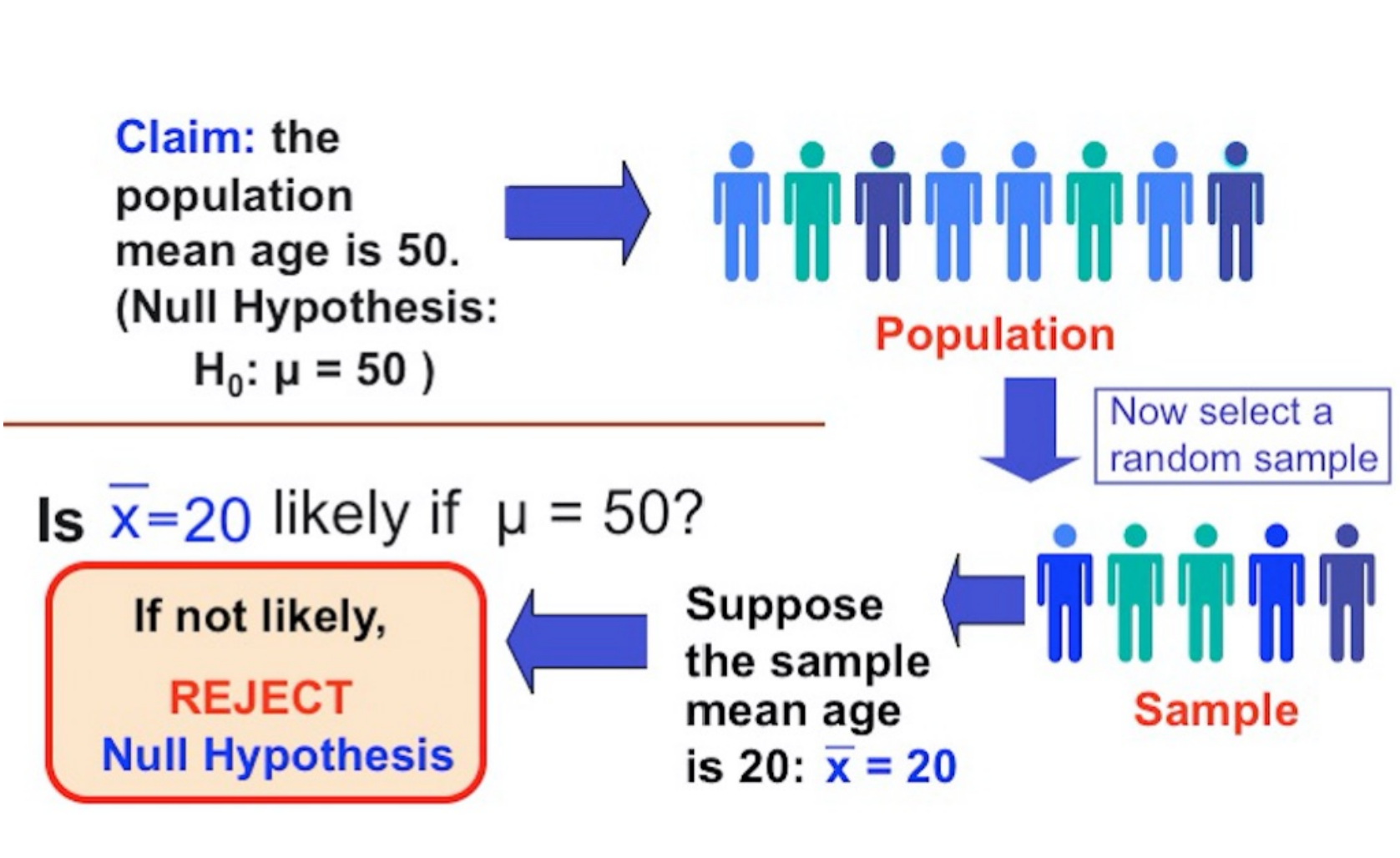
$\mathbf{H_0}$

- Begin with the assumption that the null hypothesis is true $\rightarrow$ Similar to the notion of innocent until proven guilty
- Refers to the status quo
- Always contains "$=$", "$\leq$", or "$\geq$" sign
- May or may not be rejected

$\mathbf{H_1}$

- Challenges the status quo
- Never contains the "$=$", "$\leq$", or "$\geq$" sign
- May or may not be supported
- Is generally the hypothesis that the researcher is trying to support!

Hypothesis Testing Process



The Main Idea

- Suppose you **claim/conjecture** that the population mean age is 50

$$H_0 : \mu = 50$$

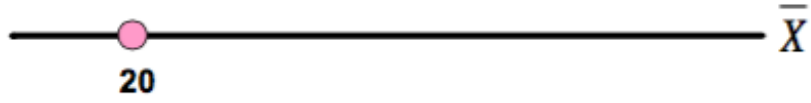
- Suppose the population is normally distributed with $\sigma = 18$
- You have a sample of 15 i.i.d. draws and you compute $\overline{X} = 20$
- Let us summarize, what we know:

$$X \sim N(??, 18^2), \quad \overline{X} = 20, \quad \sigma = 18, \quad n = 15$$

  where X is the age of an individual member of the population.
- Does the data support your initial claim? What do you think?

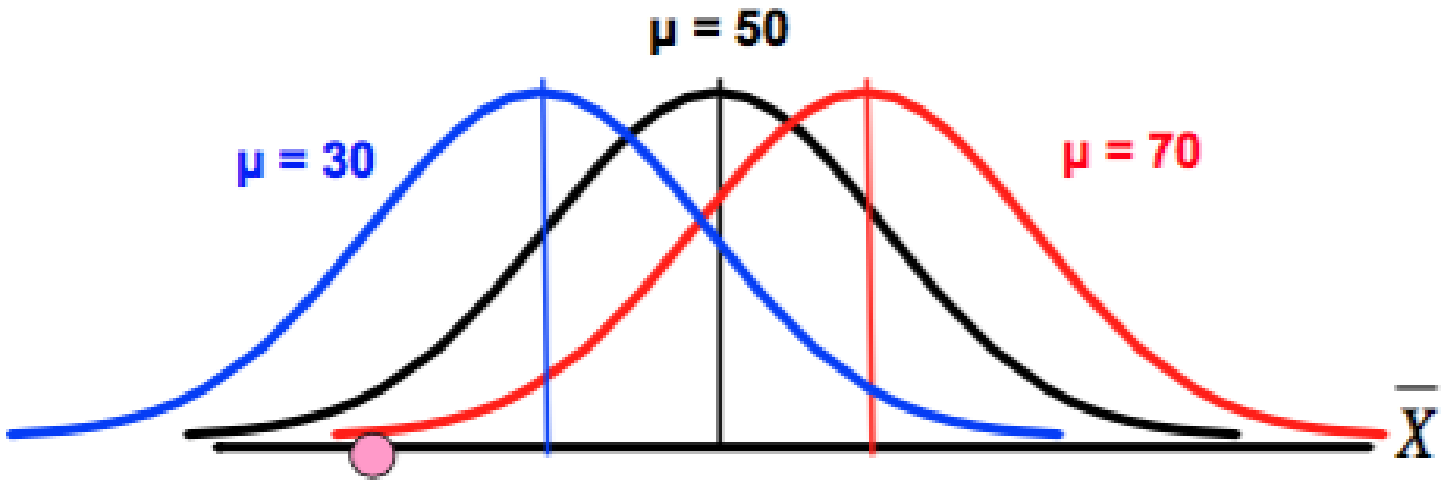
Some Questions

- Can we determine the distribution of $\overline{X}$?
- Yes! We know that
$$\overline{X} \sim N(\mu, 18^2/15)$$
- And since our sample generated $\overline{X} = 20$, this should also be somewhere on the support of the distribution of $\overline{X}$

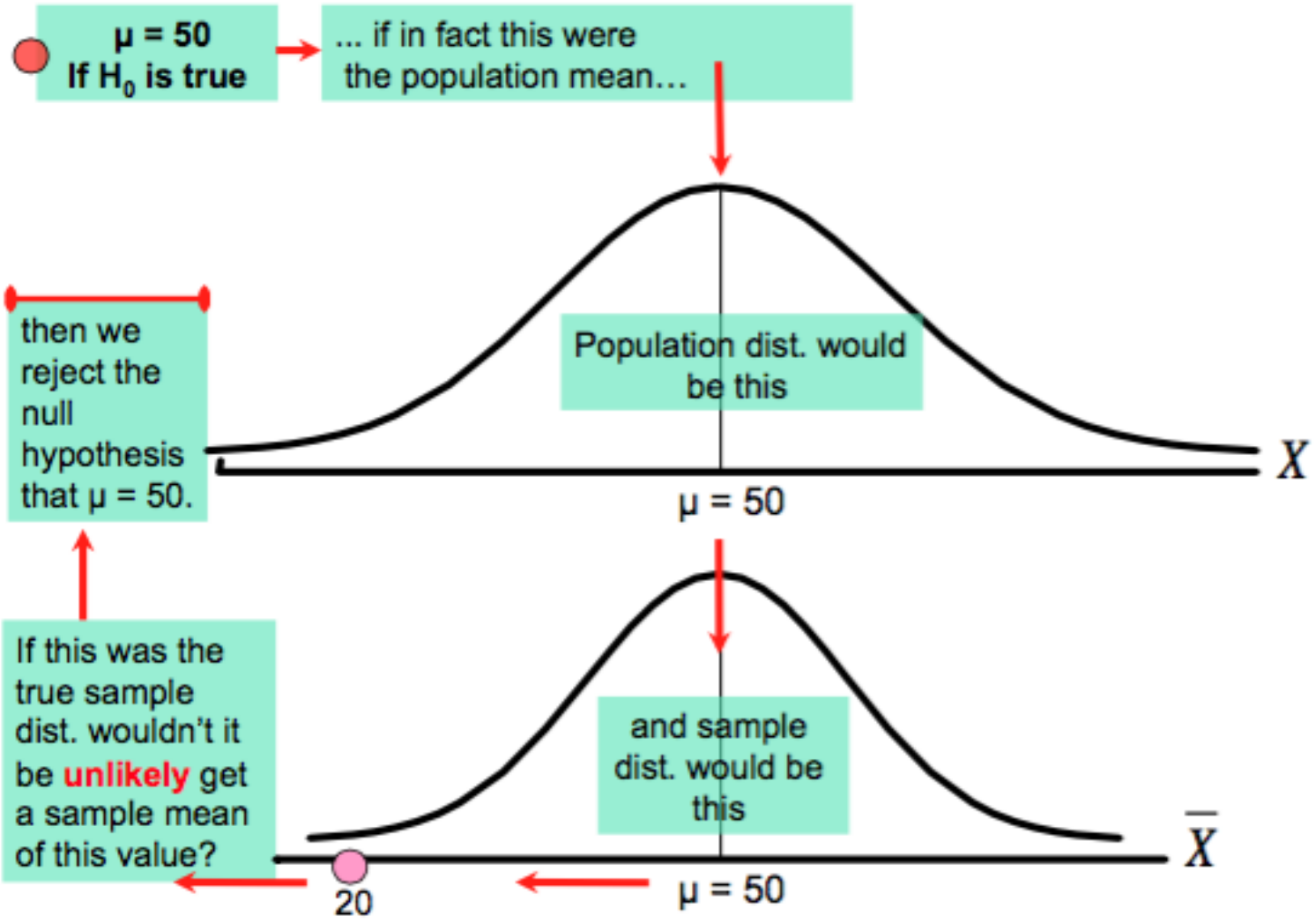


- But actually as long as we do not say something about μ we can not go any further!

- However, under different assumptions on μ we know exactly what the distribution of $\overline{X}$ would look like:



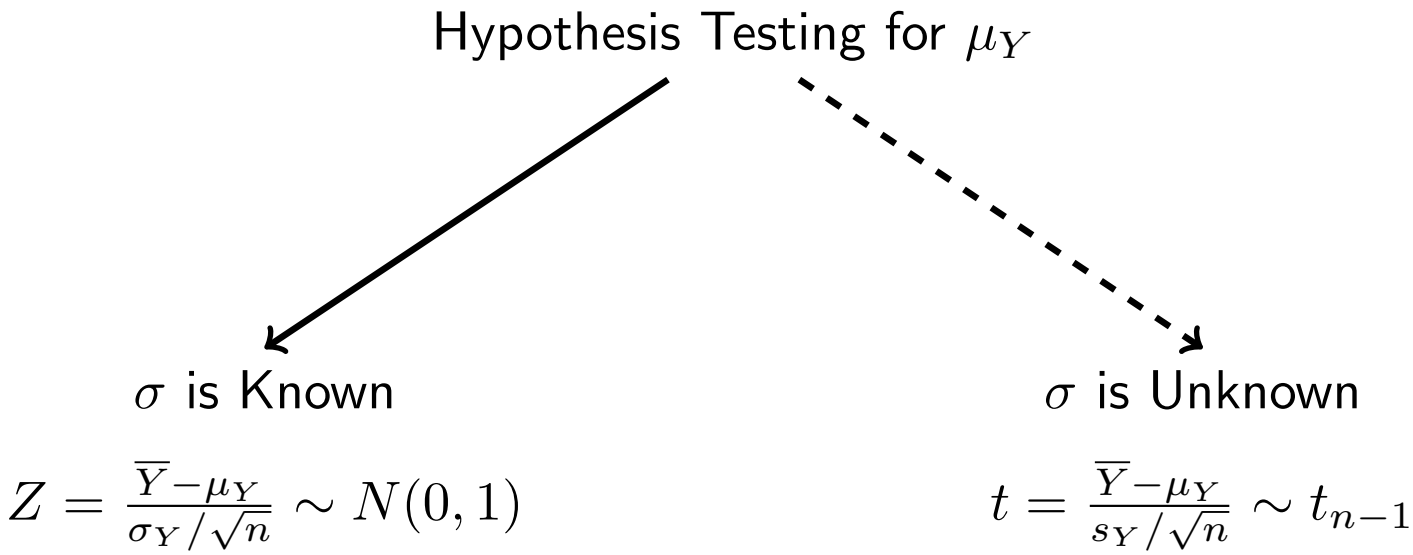
Hypothesis Testing Cycle



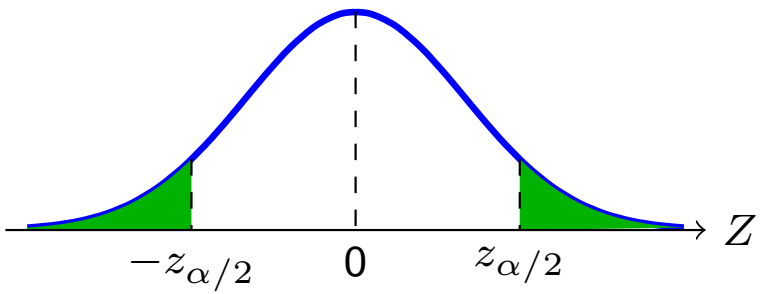
The Unlikelihoodness Criterion: Level of Significance, α

- **Level of Significance**, denoted by α, defines the unlikely values of the sample static if the null hypothesis is true
- In another words, defines the rejection region of the sampling distribution
- Typical values are 0.001, 0.05, or 0.10
- It is selected by the researcher at the beginning
- Provides the critical value(s) of the test
- **Type I Error**
  - is rejecting a true null hypothesis
  - considered a serious type of error
  - the probability of Type I Error is α, i.e. it is equal to the level of significance that was set by researcher in advance

Hypothesis Testing for Population Mean μ_Y



Tests and Decision Rules (σ_Y is Known)

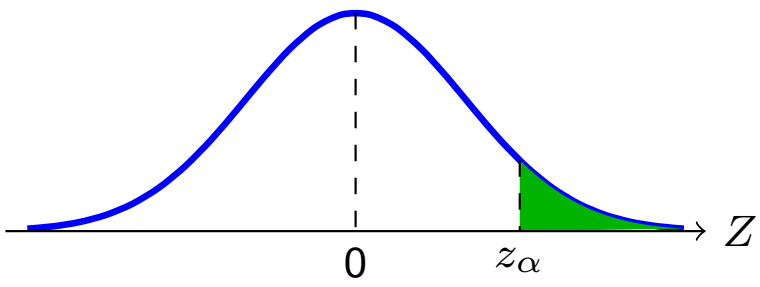


Two-sided Test

$H_0 : \mu_Y = \mu_0$
$H_1 : \mu_Y \neq \mu_0$

Reject H_0 if:

$$|Z^a = \frac{\overline{Y}^a - \mu_0}{\sigma_Y/\sqrt{n}}| > z_{\alpha/2}$$

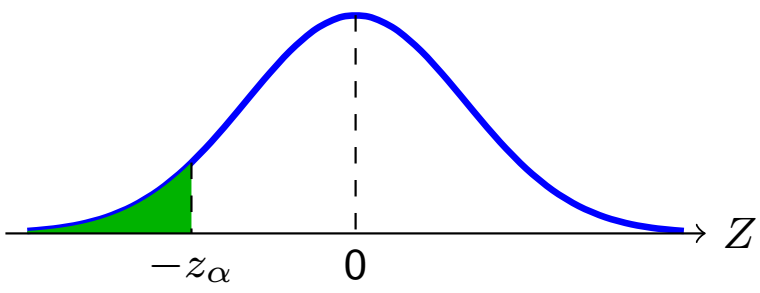


Upper-tail Test

$H_0 : \mu_Y \leq \mu_0$
$H_1 : \mu_Y > \mu_0$

Reject H_0 if:

$$Z^a = \frac{\overline{Y}^a - \mu_0}{\sigma_Y/\sqrt{n}} > z_{\alpha}$$



Lower-tail Test

$H_0 : \mu_Y \geq \mu_0$
$H_1 : \mu_Y < \mu_0$

Reject H_0 if:

$$Z^a = \frac{\overline{Y}^a - \mu_0}{\sigma_Y/\sqrt{n}} < -z_{\alpha}$$

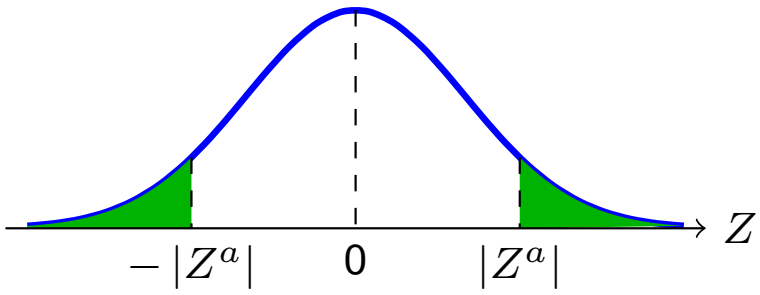
The p-Value Approach to Hypothesis Testing

- p-**value**: Smallest value of α for which H_0 can be rejected
- Alternatively, p-value is the probability of obtaining a test statistic more extreme than the observed sample value given H_0 is true
- p-value approach to testing (upper-tail test, σ_Y is known)
  - Convert sample result (e.g. $\overline{Y}^a$) to test statistic (e.g. Z statistic)
  - Obtain the p-value

$$p\text{-value} = P\left(Z > \frac{\overline{Y}^a - \mu_0}{\sigma_Y/\sqrt{n}}\right)$$
$$= 1 - \Phi\left(\frac{\overline{Y}^a - \mu_0}{\sigma_Y/\sqrt{n}}\right)$$

    where Φ is the cumulative standard normal distribution function.
  - Decision Rule: compare p-value to α
    - p-value $< \alpha \implies$ reject H_0
    - p-value $\geq \alpha \implies$ do not reject H_0

Summary: p-Value Approach (σ_Y is Known)

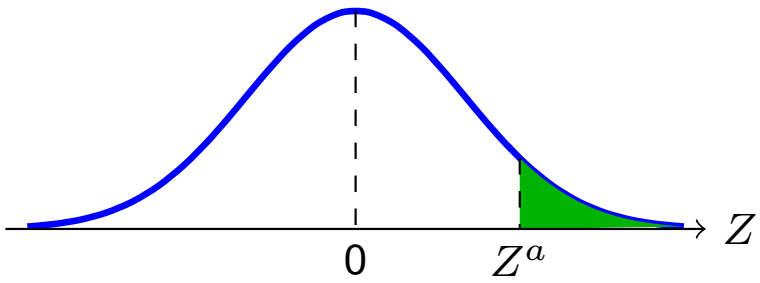


Two-sided Test

$H_0 : \mu_Y = \mu_0$
$H_1 : \mu_Y \neq \mu_0$

p-value $= 2\Phi\left(-|Z^a|\right)$, $\quad Z^a = \frac{\overline{Y}^a - \mu_0}{\sigma_Y/\sqrt{n}}$

Reject H_0 if p-value $< \alpha$

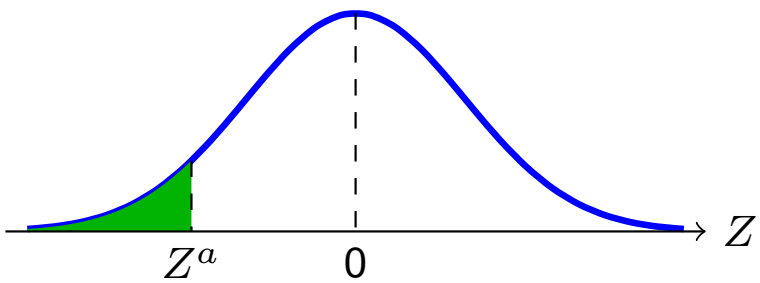


Upper-tail Test

$H_0 : \mu_Y \leq \mu_0$
$H_1 : \mu_Y > \mu_0$

p-value $= 1 - \Phi\left(Z^a\right)$, $\quad Z^a = \frac{\overline{Y}^a - \mu_0}{\sigma_Y/\sqrt{n}}$

Reject H_0 if p-value $< \alpha$



Lower-tail Test

$H_0 : \mu_Y \geq \mu_0$
$H_1 : \mu_Y < \mu_0$

p-value $= \Phi\left(Z^a\right)$, $\quad Z^a = \frac{\overline{Y}^a - \mu_0}{\sigma_Y/\sqrt{n}}$

Reject H_0 if p-value $< \alpha$

Example

- A phone industry manager thinks that customer monthly cell phone bill have increased, and now average over \$52 per month. The company wishes to test this claim. (Assume $\sigma_Y = 10$ is known).
- Suppose a sample is taken with following results:

$$\overline{Y} = 53.1 \quad \text{and} \quad n = 64$$

- Form the hypothesis test?

  $H_0 : \mu_Y \leq 52$ the average is **not** over \$52 per month

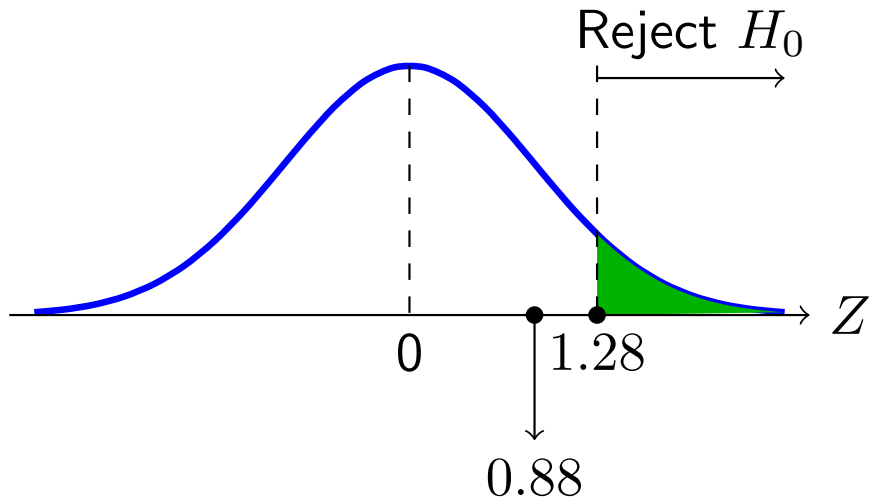
  $H_1 : \mu_Y > 52$ the average is greater than \$52 per month

- Find the rejection region for $\alpha = 0.10$?

  Upper-tail test and $\alpha = 0.10$ so find $z_{0.10} = 1.28$
  Rejection region: $Z^a > 1.28$

- Conclude the test?

$Z^a = \frac{\overline{Y}^a - \mu_0}{\sigma_Y/\sqrt{n}} = \frac{53.1-52}{10/\sqrt{64}} = 0.88 \leq 1.28 \implies$ Fail to reject H_0



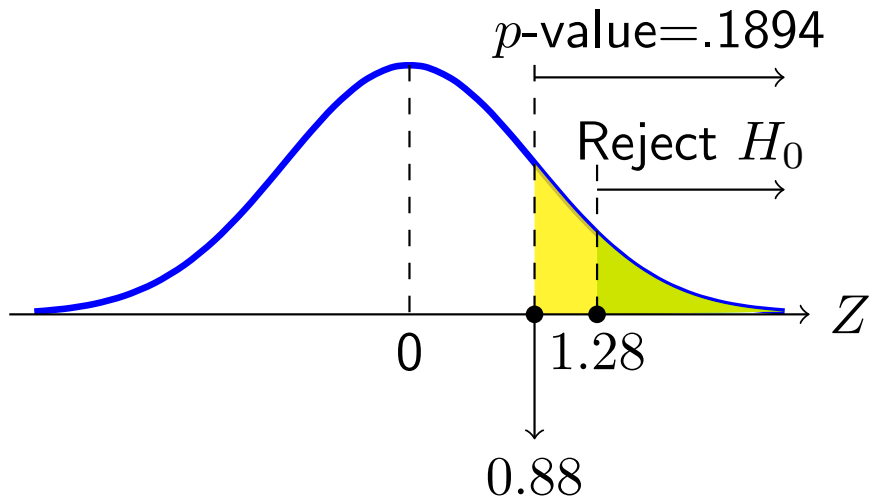
p-value Solution

- Calculate the p-value and compare to α?

  Upper-tail test and therefore p-value is

$$p\text{-value} = \Phi\left(-\left|Z^a\right|\right) = \Phi\left(-0.88\right) = 0.1894$$

  Since p-value ≥ 0.10, do not reject H_0.



Example

Test the claim that the true mean number of TV sets in US homes is equal to 3. Assume $\sigma_Y = 0.8$ is known and a sample of size 100 is selected and gave you $\overline{Y} = 2.84$.

- Form the hypothesis test?
- Find the rejection region for $\alpha = 0.05$?
- Conclude the test?
- Conclude the test using the p-value approach?

- Form the hypothesis test?

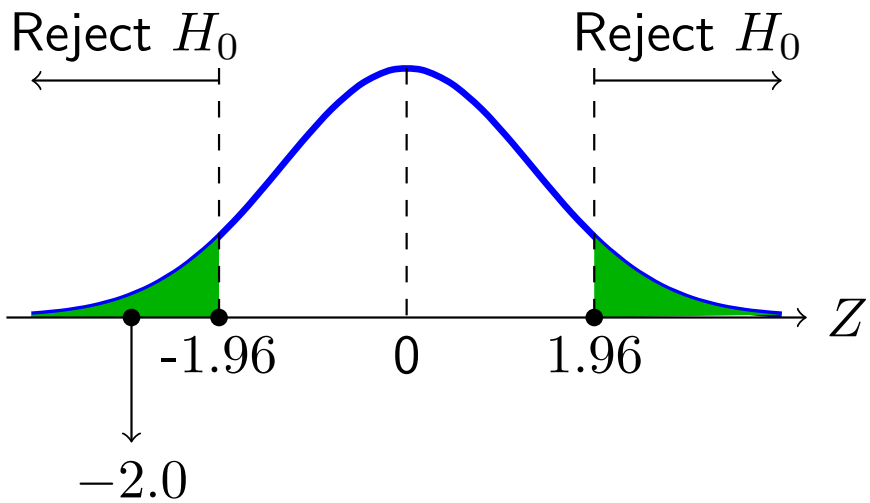
$$H_0 : \mu_Y = 3$$
$$H_1 : \mu_Y \neq 3$$

- Find the rejection region for $\alpha = 0.05$?

  Two-tail test and $\alpha = 0.05$ so find $z_{0.025} = 1.96$
  Rejection region: $Z^a > 1.96$

- Conclude the test?

$$|Z^a| = \left|\frac{\overline{Y}^a - \mu_0}{\sigma_Y/\sqrt{n}}\right| = \left|\frac{2.84-3}{0.8/\sqrt{100}}\right| = |-2.0| > 1.96 \implies \text{Reject } H_0$$



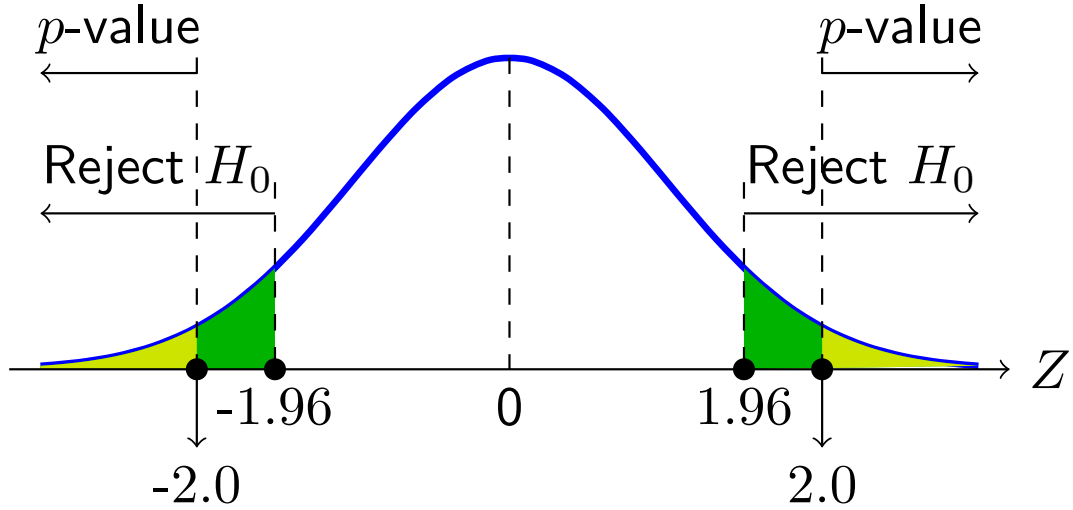
p-value Solution

- Calculate the p-value and compare to α to conclude the test.

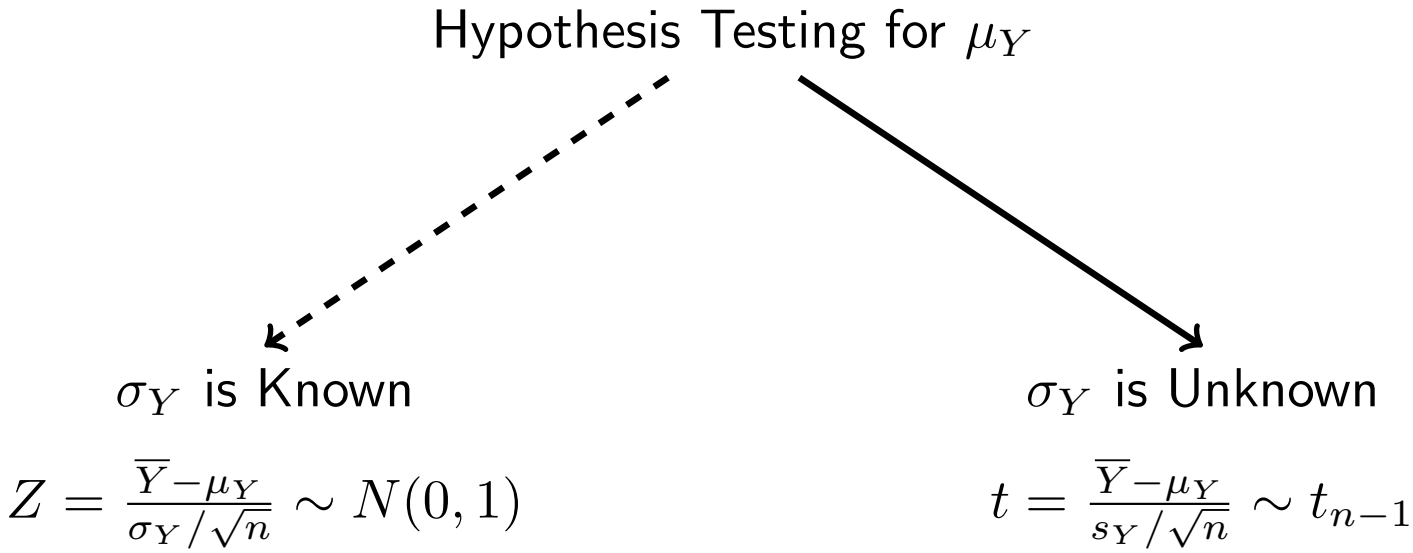
  Since it is a two-tail test,

$$p\text{-value} = 2\Phi\left(-\left|Z^a\right|\right) = 2\Phi\left(-2\right) = 2 \times 0.0228 = 0.0456$$

  Since p-value < 0.05, reject H_0.



Hypothesis Testing for μ_Y when σ_Y is Unknown



Hypothesis Testing for μ_Y when σ_Y is Unknown

- Contrary to what we have assumed so far, the population variance σ_Y is typically unknown (in $\sigma_{\overline{Y}} = \sigma_Y/\sqrt{n}$) and therefore it must be estimated in order to conduct hypothesis testing
- The sample variance

$$s_Y^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2$$

  is a consistent estimator of σ_Y^2, i.e. $s_Y^2 \longrightarrow \sigma_Y^2$.
- Therefore, a consistent estimator of $\sigma_{\overline{Y}} = \sigma_Y/\sqrt{n}$ can be obtained by replacing the population standard deviation σ_Y with the sample standard deviation s_Y:

$$\hat{\sigma}_{\overline{Y}} = s_Y/\sqrt{n}$$

  This is called the **standard error of** $\overline{Y}$ and is denoted also by $SE(\overline{Y})$.

- The distribution of the standardized sample average $(\overline{Y} - \mu_Y)/SE(\overline{Y})$ is known as the **student's t distribution** with $n-1$ degrees of freedom:

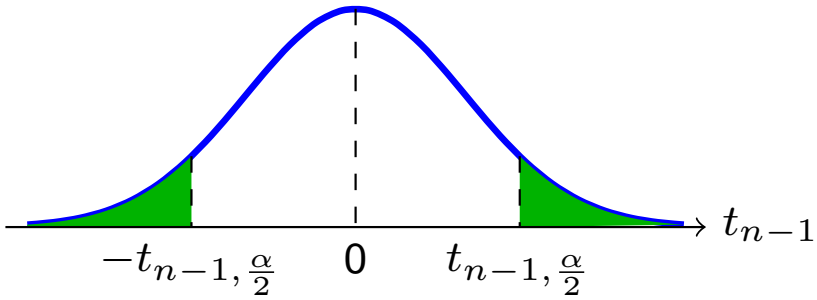
$$t = \frac{\overline{Y} - \mu_Y}{SE(\overline{Y})} \sim t_{n-1}$$

- Hypothesis testing for population mean μ_Y when σ_Y is unknown is similar to the case when σ_Y is known except:
  - Calculate the test statistics using the above formula instead of Z^a

$$t^a = \frac{\overline{Y}^a - \mu_0}{SE(\overline{Y})}$$

  - Then compare t^a to the critical values using the student's t distribution with $n-1$ degrees of freedom
- Decision rules for each type of test are summarized next:

Tests and Decision Rules (σ_Y is Unknown)

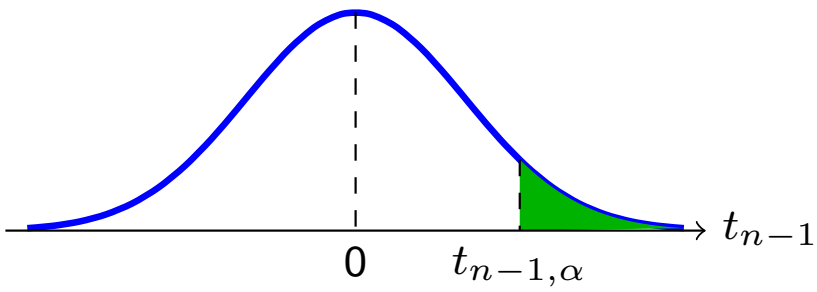


Two-sided Test

$H_0 : \mu_Y = \mu_0$
$H_1 : \mu_Y \neq \mu_0$

Reject H_0 if:

$$|t^a = \frac{\overline{Y}^a - \mu_0}{s_Y/\sqrt{n}}| > t_{n-1,\frac{\alpha}{2}}$$

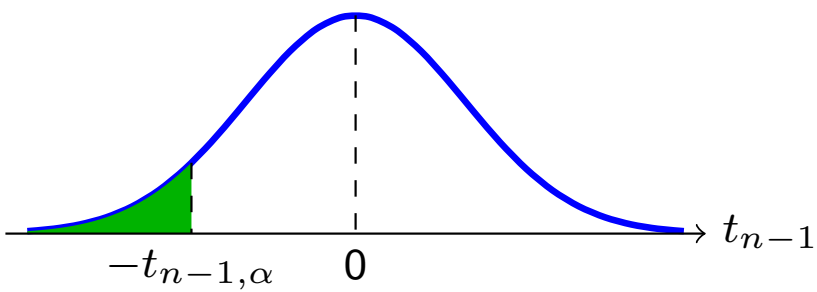


Upper-tail Test

$H_0 : \mu_Y \leq \mu_0$
$H_1 : \mu_Y > \mu_0$

Reject H_0 if:

$$t^a = \frac{\overline{Y}^a - \mu_0}{s_Y/\sqrt{n}} > t_{n-1,\alpha}$$



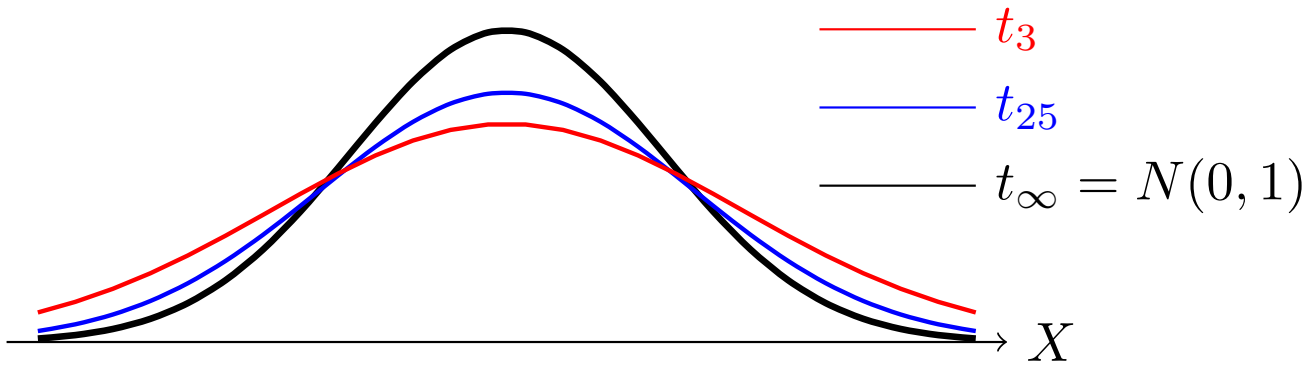
Lower-tail Test

$H_0 : \mu_Y \geq \mu_0$
$H_1 : \mu_Y < \mu_0$

Reject H_0 if:

$$t^a = \frac{\overline{Y}^a - \mu_0}{s_Y/\sqrt{n}} < -t_{n-1,\alpha}$$

- t distribution is approximately distributed $N(0,1)$ for large n:



- Therefore, to compute the p-value, when σ_Y is unknown, just replace Z^a by t^a.
- Accordingly, when n is large,
  - For a two-sided test: p-value $= 2\Phi\left(-\left|t^a\right|\right)$
  - For a one-sided test: p-value $= 1 - \Phi\left(\left|t^a\right|\right)$

Example

The average cost of a hotel room in Chicago is said to be \$168 per night. A random sample of 25 hotels resulted in

$$\overline{Y} = \$172.5, \quad s_Y = \$15.40$$

Test the claim at the $\alpha = 0.05$.

- Form the hypothesis test?
- Find the rejection region for $\alpha = 0.05$?
- Conclude the test?

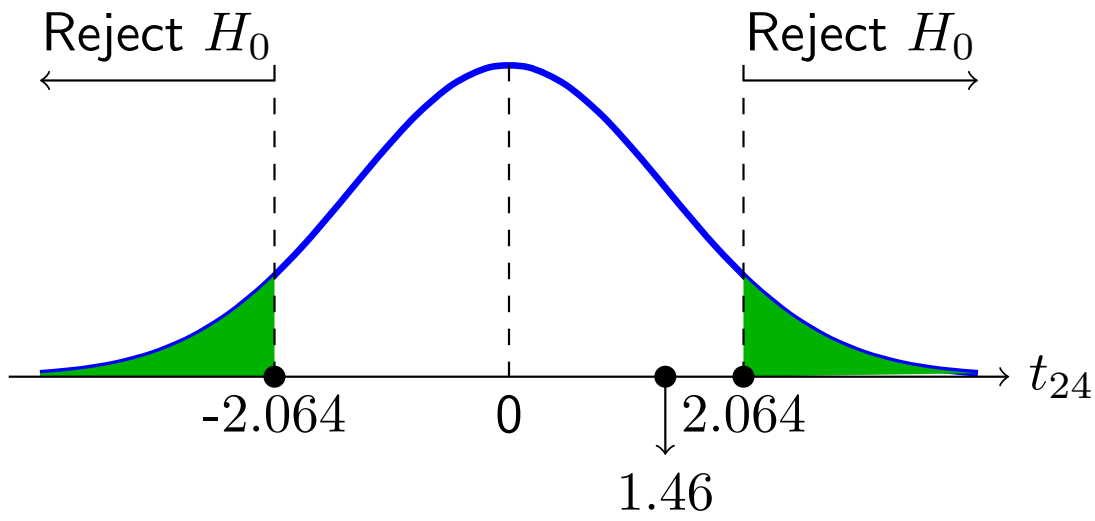
- Form the hypothesis test?

$$H_0 : \mu_Y = 168$$
$$H_1 : \mu_Y \neq 168$$

- Find the rejection region for $\alpha = 0.05$?

  Two-tail test and $\alpha = 0.05$ so find $t_{24,0.025} = 2.064$
  Rejection region: $|t^a| > 2.064$
- Conclude the test?

  $|t^a| = \left|\frac{\overline{Y}^a - \mu_0}{s_Y/\sqrt{n}}\right| = \left|\frac{172.5-168}{15.40/\sqrt{25}}\right| = |1.46| < 2.064$
  $\implies$ Fail to Reject H_0



Review Example

A simple random sample of 25 filtered cigarettes is obtained, and the tar content of each cigarette is measured. The sample has a mean of 13.2 mg and a standard deviation of 3.7 mg. We are interested in testing the claim that the mean tar content of filtered cigarettes is less than 21.1 mg, which is the mean for unfiltered cigarettes.

1. Formulate the appropriate null and alternative hypotheses.
2. Calculate the appropriate test statistic and conclude the test at $\alpha = 0.05$?
3. Find the p-value?

1. Formulate the appropriate null and alternative hypotheses.
   $H_0 : \mu_Y \geq 21.1$
   $H_1 : \mu_Y < 21.1$
2. Calculate the appropriate test statistic and conclude the test at $\alpha = 0.05$?

   Test statistic: $t^a = \frac{13.2-21.1}{3.7/5} = -10.68$

   Critical table value: $t_{24,0.05} = 1.711$

   Decision: $-10.68 < -1.711 \implies$ reject H_0.

   $\implies$ Filters are effective in reducing the amount of tar.
3. Find the p-value?
   - We can not find an exact p-value from the t table and since the sample size is not large ($n = 25$) we can not use the standard normal distribution to approximate either
   - But note that $t_{24,0.001} = 3.467$ so even with $\alpha = 0.001$ we would reject H_0
   - Therefore, we can say that p-value < 0.001, which means we can reject H_0 with high confidence

Confidence Intervals for the population Mean

- 90% Confidence Interval for μ_Y:

$$\overline{Y} - 1.64 \times SE(\overline{Y}) \leq \mu_Y \leq \overline{Y} + 1.64 \times SE(\overline{Y}), \qquad \left(SE(\overline{Y}) = s_Y/\sqrt{n}\right)$$

- 95% Confidence Interval for μ_Y:

$$\overline{Y} - 1.96 \times SE(\overline{Y}) \leq \mu_Y \leq \overline{Y} + 1.96 \times SE(\overline{Y})$$

- 99% Confidence Interval for μ_Y:

$$\overline{Y} - 2.58 SE(\overline{Y}) \leq \mu_Y \leq \overline{Y} + 2.58 \times SE(\overline{Y})$$

Interpretation:(for the first one) 90% of the time the true population mean μ_Y will be contained in the set $\{\overline{Y} \pm 1.64 \times SE(\overline{Y})\}$