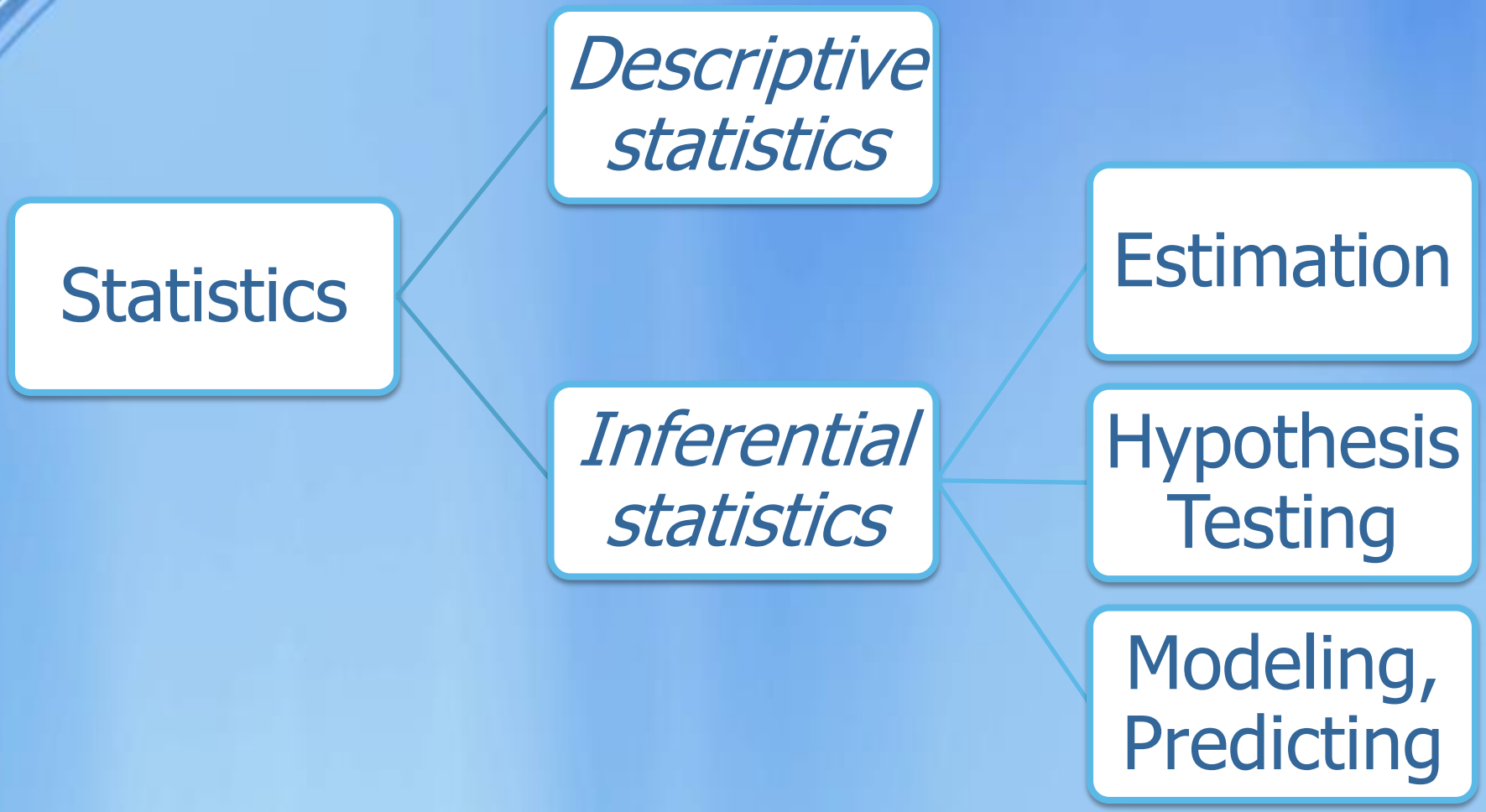
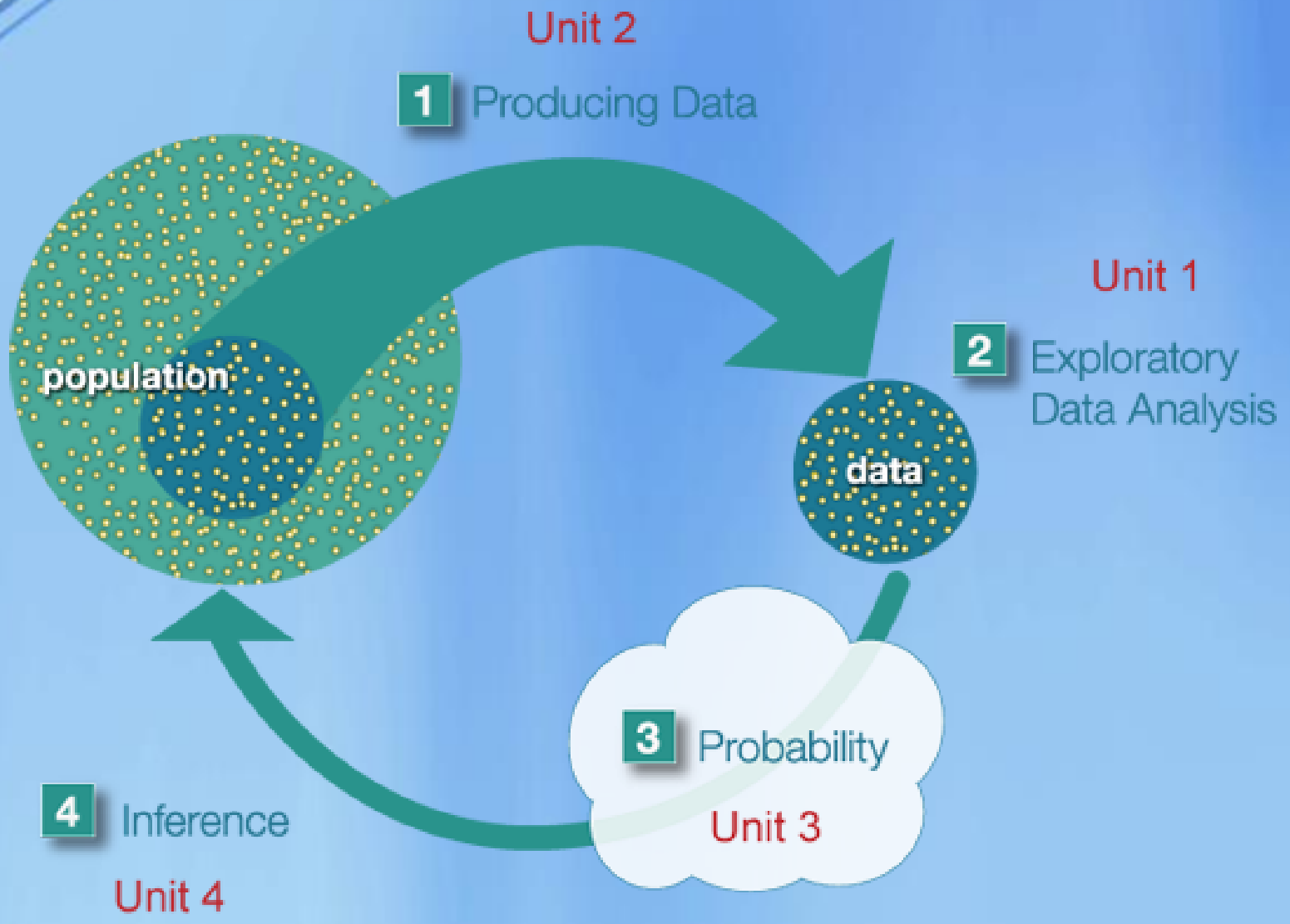
Inferential Statistics

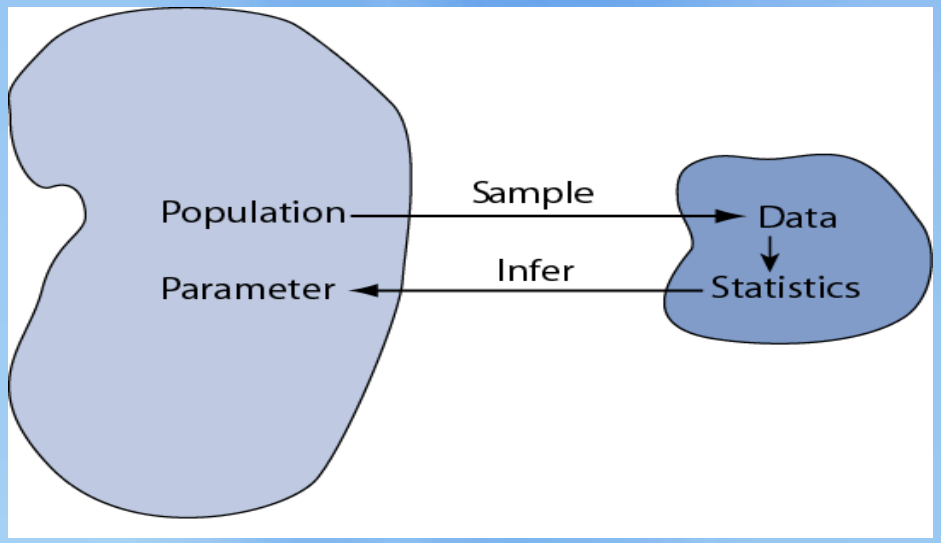

Statistics
Descriptive statistics
Inferential statistics
Estimation
Hypothesis Testing
Modeling, Predicting


Unit 2
1 Producing Data
population
Unit 1
2 Exploratory Data Analysis
data
3 Probability
Unit 3
4 Inference
Unit 4

Inferential Statistics

Statistical inference is the act of generalizing from a **sample** to a **population** with calculated degree of certainty.

We want to learn about population *parameters…*



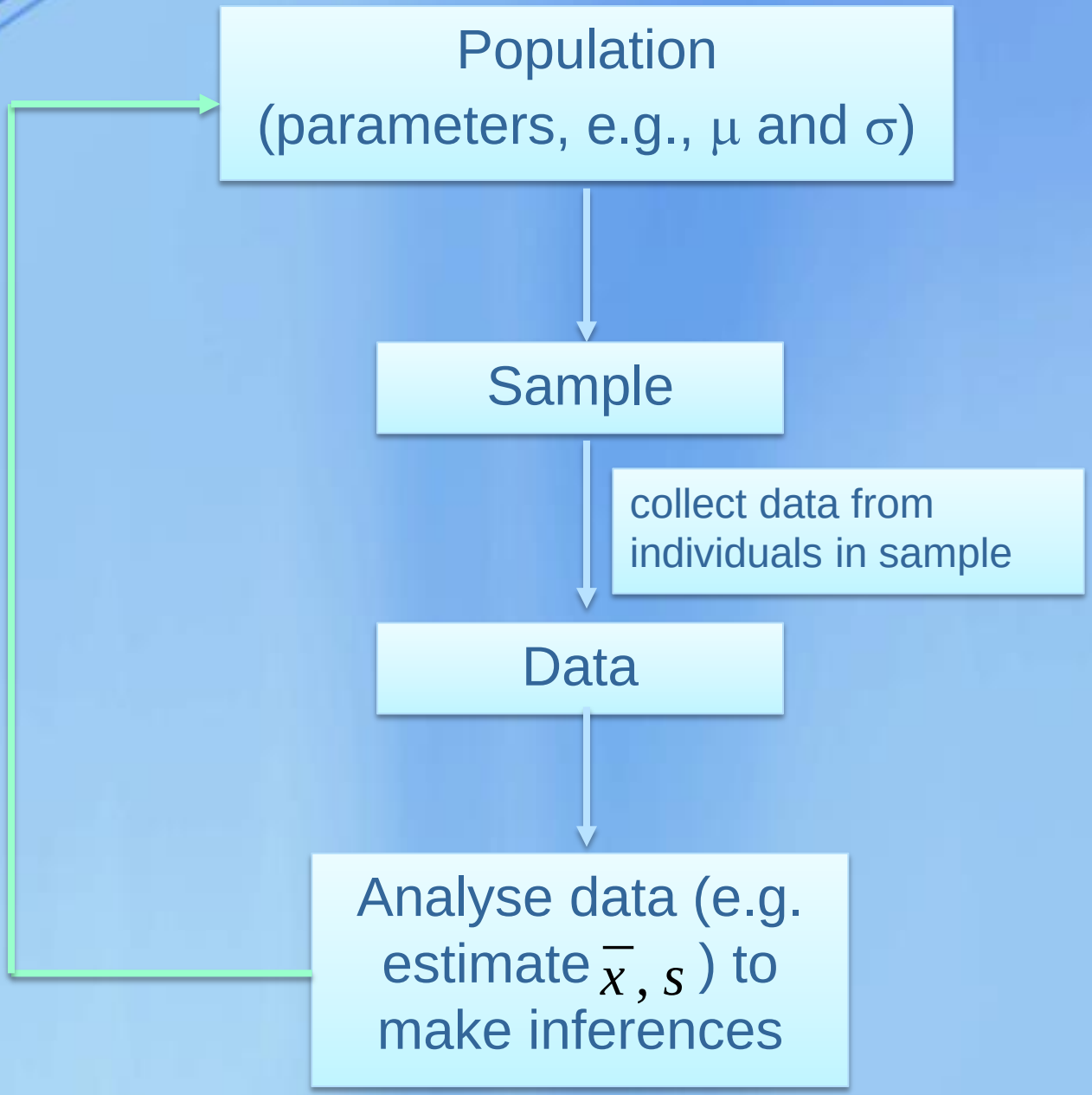
but we can only calculate *sample statistics*

Methods for drawing conclusions about a population from sample data are called

Parameters and Statistics

It is essential that we draw distinctions between parameters and statistics

	Parameters	**Statistics**
Source	Population	Sample
Calculated?	No	Yes
Constants?	Yes	No
Examples	μ, σ, p	$\bar{x}$, s, $\hat{p}$


Population
(parameters, e.g., μ and σ)
Sample
collect data from individuals in sample
Data
Analyse data (e.g. estimate $\bar{x}$, s) to make inferences

Estimation

That act of guessing the value of a population parameter

Point Estimation

Estimating a specific value

Interval Estimation

determining the range or interval in which the value of the parameter is thought to be

Point Estimates

- One of our fundamental questions is: "**How well does our sample statistic estimate the value of the population parameter**?"

- In other word: **How close is Sample Statistic to Population Parameter ?**

Choosing a good estimator

Good estimators have certain desirable properties:

- ***Unbiasedness****:* The sampling distribution for the estimator 'centers' around the parameter. (On average, the estimator gives the correct value for the parameter.)
- ***Efficiency:*** If at the same sample size one unbiased estimator has a smaller sampling error than another unbiased estimator, the first one is more efficient.
- ***Consistency:*** The value of the estimator gets closer to the parameter as sample size increases. Consistent estimators may be biased, but the bias must become smaller as the sample size increases if the consistency property holds true.

Estimating the population mean

- A natural estimator to use for estimating the population mean is the sample mean.
- We will use the mean of a sample $\bar{x}$ (statistic) as an estimate of the mean of the population μ (parameter).
- What properties of $\bar{x}$ make it a desirable estimator of μ? How close is $\bar{x}$ to μ?
  - Cannot answer question for a particular sample!
  - Can answer if we can find out about the distribution that describes the variability in the random variable $\bar{X}$.

Sampling Distribution

- Consider the set of all possible samples of size *n* that could have been selected from the population.
- The values of $\bar{x}$ in each of these samples will, in general, be different! These values will be denoted by $\bar{x}_1$, $\bar{x}_2$, and so forth.
- What is the sampling distribution of $\bar{x}$?
- The sampling distribution of $\bar{X}$ is the distribution of values of $\bar{x}$ over all possible samples of size *n* that could have been selected from the reference population.

Example

- The mean height of women age 20 to 30 is normally distributed (bell-shaped) with a mean of 65 inches and a standard deviation of 3 inches.
- A random sample of 200 women was taken and the sample mean $\bar{x}$ recorded.
- Now IMAGINE taking MANY samples of size 200 from the population of women.
- For each sample we record the $\bar{x}$. What is the sampling distribution of $\bar{x}$?

- We can show that the average of these sample means ($\bar{x}$'s), when taken over a large number of random samples of size *n*, approximates μ as the number of samples selected becomes large.
- In other words, the expected value of $\bar{X}$ over its sampling distribution is equal to μ:

  $$\mathrm{E}(\bar{X}) = \mu$$

- We refer to $\bar{X}$ as an unbiased estimator of μ.

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

- Standard Error of the Mean
  - We can estimate how much variability there is among potential sample means by calculating the standard error of the mean.
  - Thus, it is measure of how good our point estimate is likely to be!
  - The *standard error of the mean* is the standard deviation of the sampling distribution.

- It is given by:

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

- This equation implies that sampling error decreases as sample size increases.
- Thus the larger the sample size, the more precise an estimator $\bar{X}$ is.
- **In fact**: as the sample gets bigger, the sampling distribution…
  - stays centered at the population mean.
  - becomes less variable.
  - becomes more normal.

Central Limit Theorem

For any population with mean μ and standard deviation σ, the distribution of sample means for sample size n …

1. will have a mean of μ
2. will have a standard deviation (standard error) of $\frac{\sigma}{\sqrt{n}}$
3. will approach a normal distribution as n gets large (n ≥ 30).

Example

Given the population of women has normally distributed weights with a mean of 143 lbs and a standard deviation of 29 lbs,

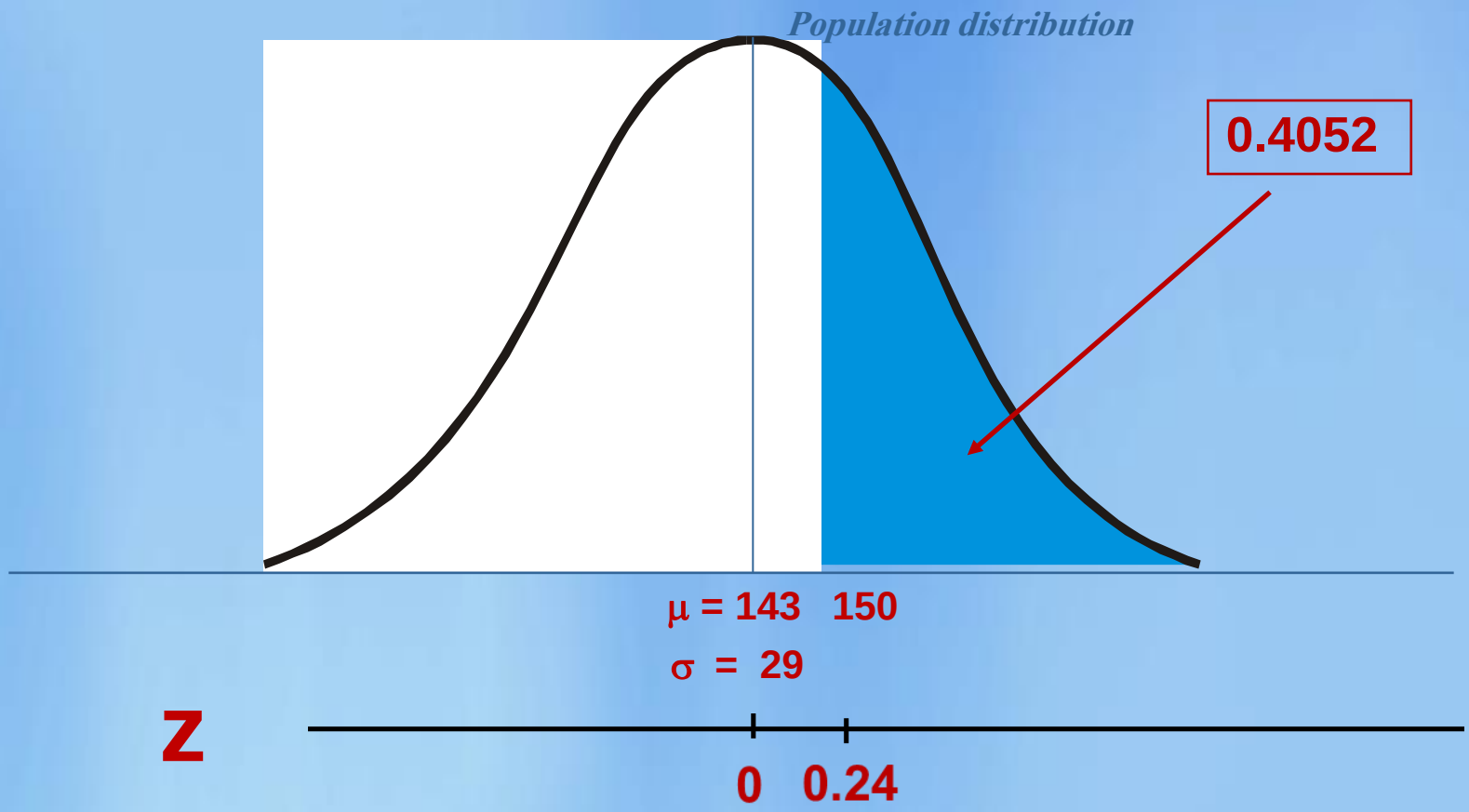
1. if one woman is randomly selected, find the probability that her weight is greater than 150 lbs.

2. if 36 different women are randomly selected, find the probability that their mean weight is greater than 150 lbs.

1. The probability that her weight is greater than 150 lbs.

$$P(X > 150) = .41$$

$$z = \frac{150-143}{29} = 0.24$$

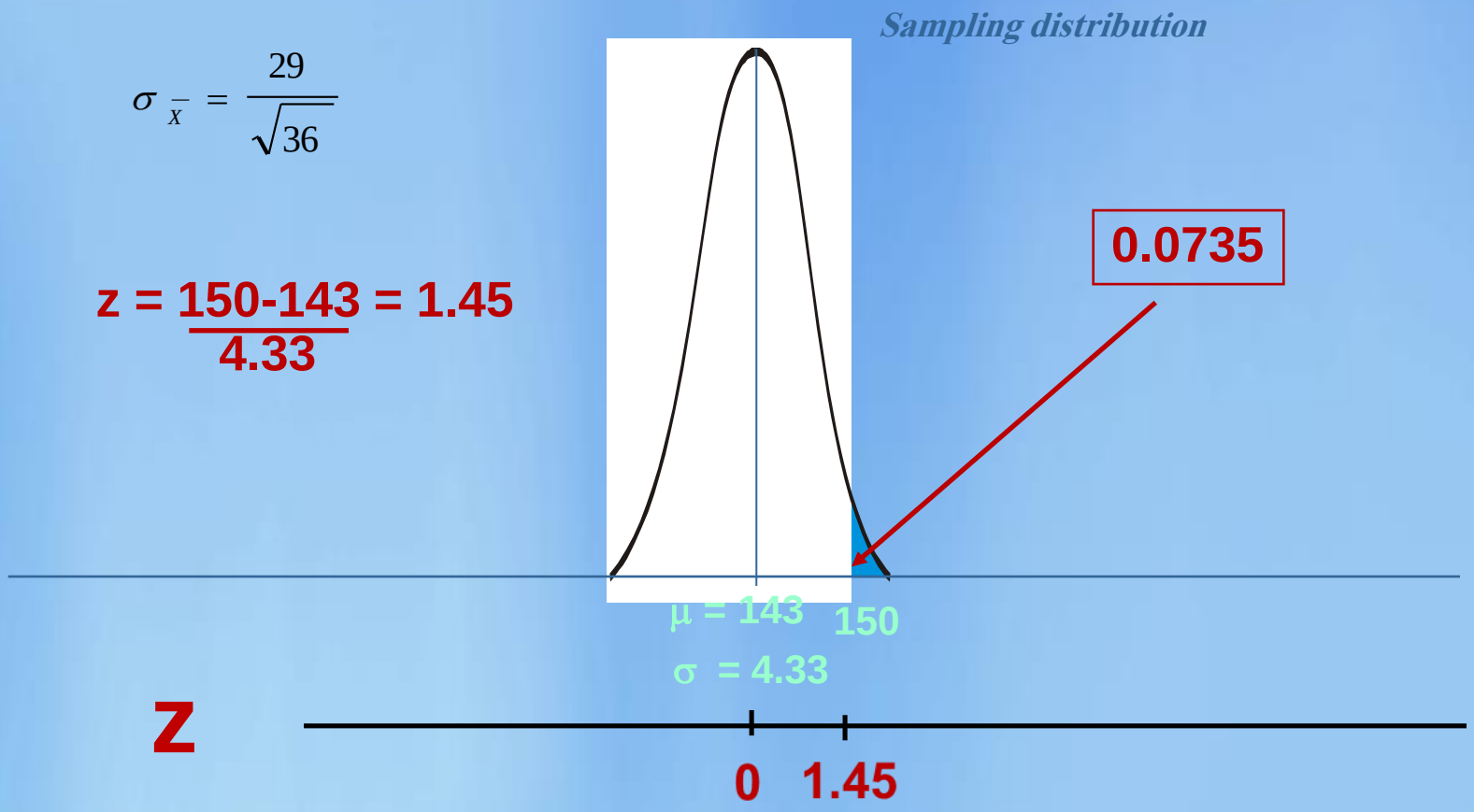


2. if 36 different women are randomly selected, find the probability that their mean weight is greater than 150 lbs.

$$P(\bar{X} > 150) = .07$$

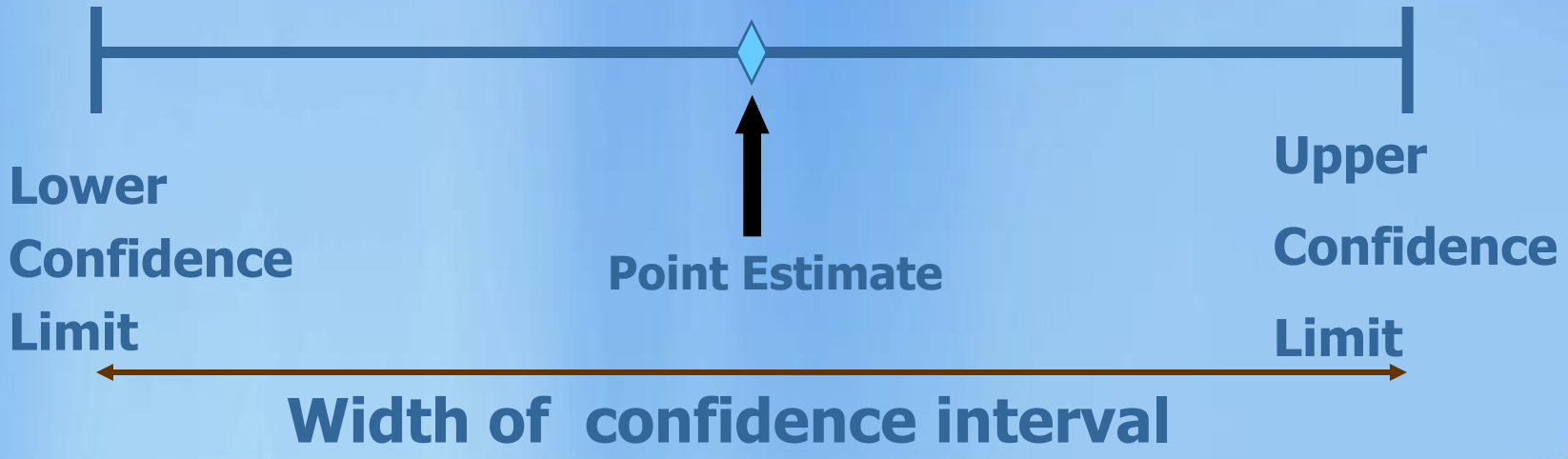
$$\sigma_{\bar{X}} = \frac{29}{\sqrt{36}}$$

z = 150-143 = 1.45
4.33



Interval Estimation

- A **point estimate** is a single number,
  - How much uncertainty is associated with a point estimate of a population parameter?
- An **interval estimate** provides more information about a population characteristic than does a point estimate. It provides a confidence level for the estimate. Such interval estimates are called **confidence intervals.**



Interval Estimation

- A **confidence interval** for a population characteristic is an interval of plausible values for the characteristic. It is constructed so that, with a chosen degree of confidence (the **confidence level**), the value of the characteristic will be captured inside the interval
- **Confidence** in statistics is a measure of our surety that a key value falls within a specified interval.
- A Confidence Interval is always accompanied by a probability that defines the risk of being wrong.
- This probability of error is usually called α (alpha).

- How confident do we want to be that the interval estimate contains the population parameter?
- The level of confidence $(1-\alpha)$ is the probability that the interval estimate contains the population parameter.
- The most common choices of level of confidence are 0.95, 0.99, and 0.90.
- For example: If the level of confidence is 90%, this means that we are 90% confident that the interval contains the population.

- The general formula for all confidence intervals is equal to:

Point Estimate ± Margin of Error

The general formula for all confidence intervals is:

The value of the statistic in my sample (eg., mean, odds ratio, etc.)

point estimate ± (measure of how confident we want to be) × (standard error)

From a Z table or a T table, depending on the sampling distribution of the statistic.

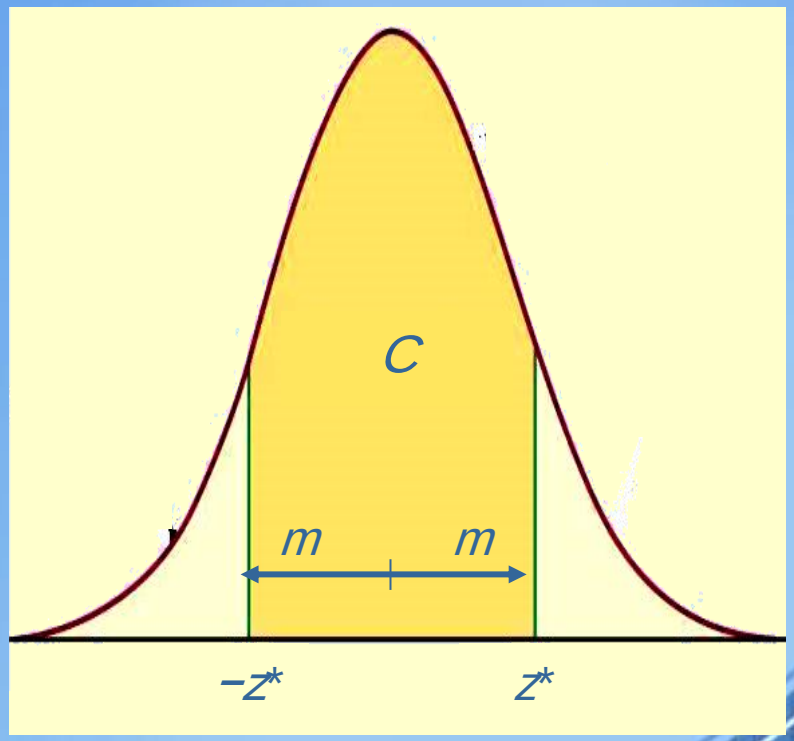
Standard error of the statistic.

Confidence Interval for the Mean μ

The (1- α) % confidence interval for μ is given by:

$$\bar{x} \pm \text{margin of error (m)}$$

- Higher confidence ***C*** implies a larger margin of error ***m*** (thus less precision in our estimates).
- A lower confidence level ***C*** produces a smaller margin of error ***m*** (thus better precision in our estimates).



There are 3 different cases:

1. Interval Estimation of μ when X is normally distributed.
2. Interval Estimation of μ when X is not normally distributed but the population Standard Deviation is known
3. Interval Estimation of μ when X is normally distributed and the population Standard Deviation unknown

Case 1: Interval Estimation of μ when X is normally distributed

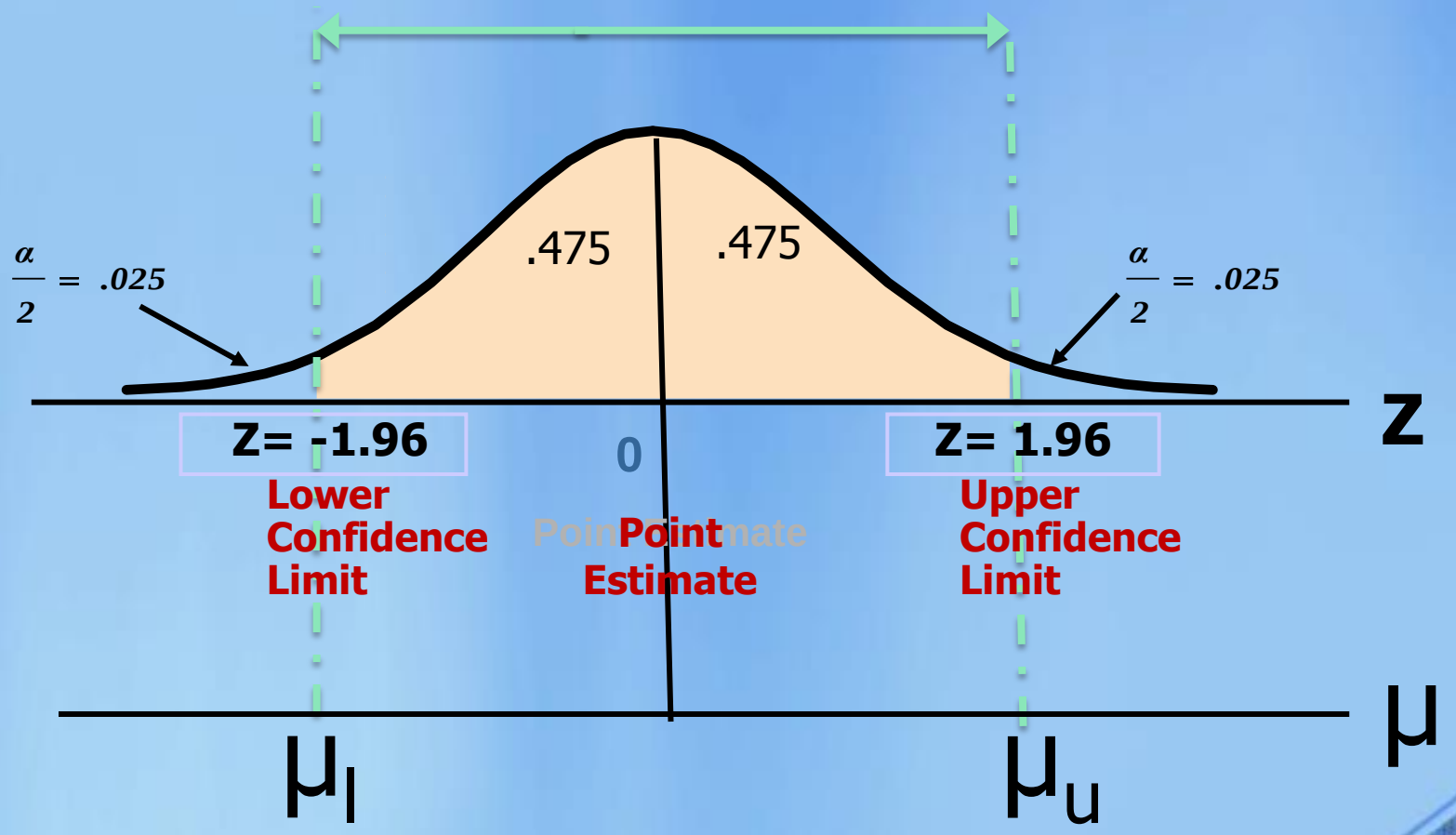
- This is the **standard** situation and you simply use this equation to estimate the population mean at the desired confidence interval:

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

- where Z is the standard normal distribution's critical value for a probability of α/2 in each tail.

- Consider a 95% confidence interval:

$1 - \alpha = .95$ $\alpha = .05$ $\alpha / 2 = .025$



Case 2: Interval Estimation of μ when X is not normally distributed but σ is known

- If the distribution of X is not normally distributed, then the **central limit theorem** can only be applied loosely and we can only say that the sampling distribution is approximately normal.
- When n is greater than 30, this approximation is generally good.
- Then we use the previous formula to calculate confidence interval.

Case 3: Interval Estimation of μ when X is normally distributed and σ is unknown

- Nearly always σ is unknown and is estimated using sample standard deviation ***s.***
- If population standard deviation is unknown then it can be shown that sample means from samples of size n are **t-distributed** with n-1 degrees of freedom.
- As an estimate for standard error we can use

$$\frac{s}{\sqrt{n}}$$

Student's *t* Distribution

- Closely related to the standard normal distribution *Z*
  - Symmetric and bell-shaped
  - Has mean = 0 but has a larger standard deviation
  - As n increases, the t-dist. approached the Z-dist.

- Exact shape depends on a parameter called **degrees of freedom** (df) which is related to sample size
  - In this context df = *n-1*

- A 100% × (1 − **α**) CI for the mean **μ** of a normal distribution with unknown variance is given by:

$$\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \le \mu \le \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

where $t_{\alpha/2,n-1}$

is the critical value of the t distribution with n-1 d.f. and an area of α/2 in each tail

Example: Left ventricular ejection fraction (LVEF)

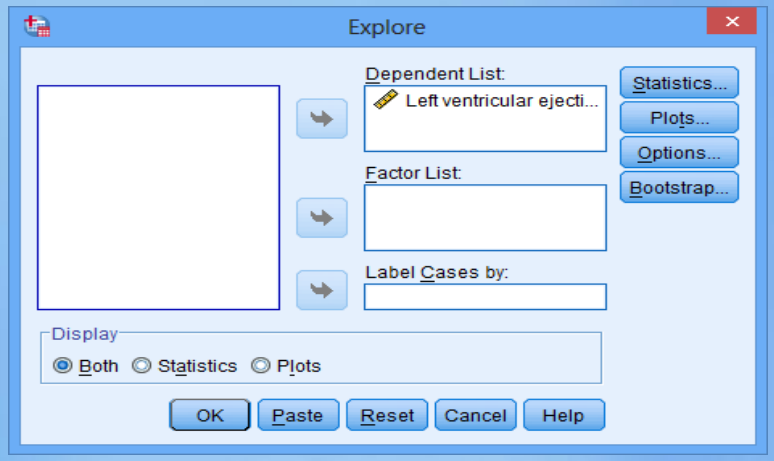
- We have 27 patients with acute dilated cardiomyopathy.

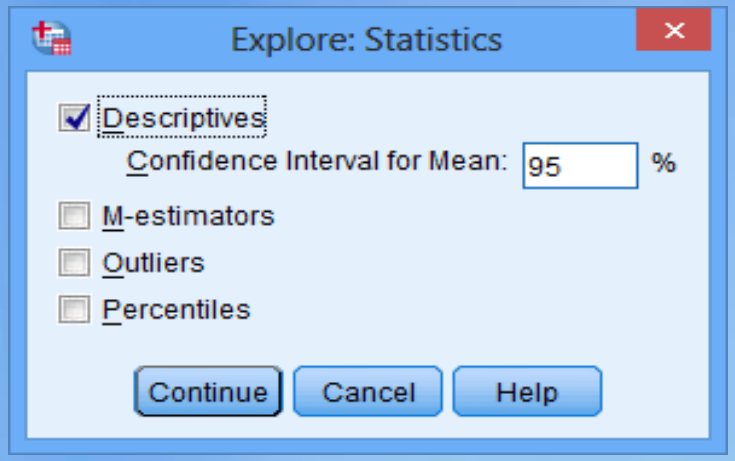
Compute a 95% CI for the true mean LVEF.

1	.19	15	.24
2	.24	16	.18
3	.17	17	.22
4	.40	18	.23
5	.40	19	.14
6	.23	20	.14
7	.20	21	.30
8	.20	22	.07
9	.30	23	.12
10	.19	24	.13
11	.24	25	.17
12	.32	26	.24
13	.32	27	.19
14	.28		

Use SPSS to do this!

Analyze ⇨ Descriptive ⇨ Statistics ⇨ Explore





Descriptive Statistics

	N	Minimum	Maximum	Sum	Mean		Std. Deviation
	Statistic	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic
Left ventricular ejection fraction	27	.07	.40	6.05	.2241	.01538	.07992
Valid N (listwise)	27						

S = 0.07992

$$\frac{s}{\sqrt{n}} = \frac{0.07992}{\sqrt{27}} = 0.01538$$

Descriptives

			Statistic	Std. Error
Left ventricular ejection fraction	Mean		.2241	.01538
	95% Confidence Interval for Mean	Lower Bound	.1925	
		Upper Bound	.2557	
	5% Trimmed Mean		.2221	
	Median		.2200	
	Variance		.006	
	Std. Deviation		.07992	
	Minimum		.07	
	Maximum		.40	
	Range		.33	
	Interquartile Range		.11	
	Skewness		.518	.448
	Kurtosis		.239	.872

Conclusion:

We are 95% confident that the true mean of LVEF will be contained in the interval (0.1925, 0.2557)

Estimation of the Variance of a Distribution

Point Estimation

- Let $X_1, \ldots, X_n$ be a random sample from some population with mean μ and variance σ^2.
- The sample variance S^2 is an **unbiased estimator** of σ^2 over all possible random samples of size n that could have been drawn from this population; that is, $E(S^2) = \sigma^2$.

Estimation of a population proportion

- We are frequently interested in estimating the proportion of a population with a characteristic of interest, such as:
  - the proportion of smokers
  - the proportion of cancer patients who survive at least 5 years
  - the proportion of new HIV patients who are female

- The *population of interest* has a certain characteristics that appears with a certain proportion **p**. This proportion is a quantity that we usually don't know, but we would like to estimate it. Since it is a characteristic of the population, we call it a ***parameter***.
- To estimate p we draw a representative sample of size n from the population and count the number of individuals (X) in the sample with the characteristic we are looking for. This gives us the estimate **X/n**.

Point estimation of P

- This estimate, $\hat{p}$ **= X/n**, is called the sampling proportion and it is denoted by

$$\hat{p} = \frac{X}{n} = \frac{\text{total number of successes}}{\text{total number of observations in the sample}}$$

Sampling Distribution of Proportion

- You learned that a binomial can be approximated by the normal distribution if

  $np \geq 5$ and $nq \geq 5$.

- Provided we have a large enough sample size, all the possible sample proportions will be approximately normally distributed with:

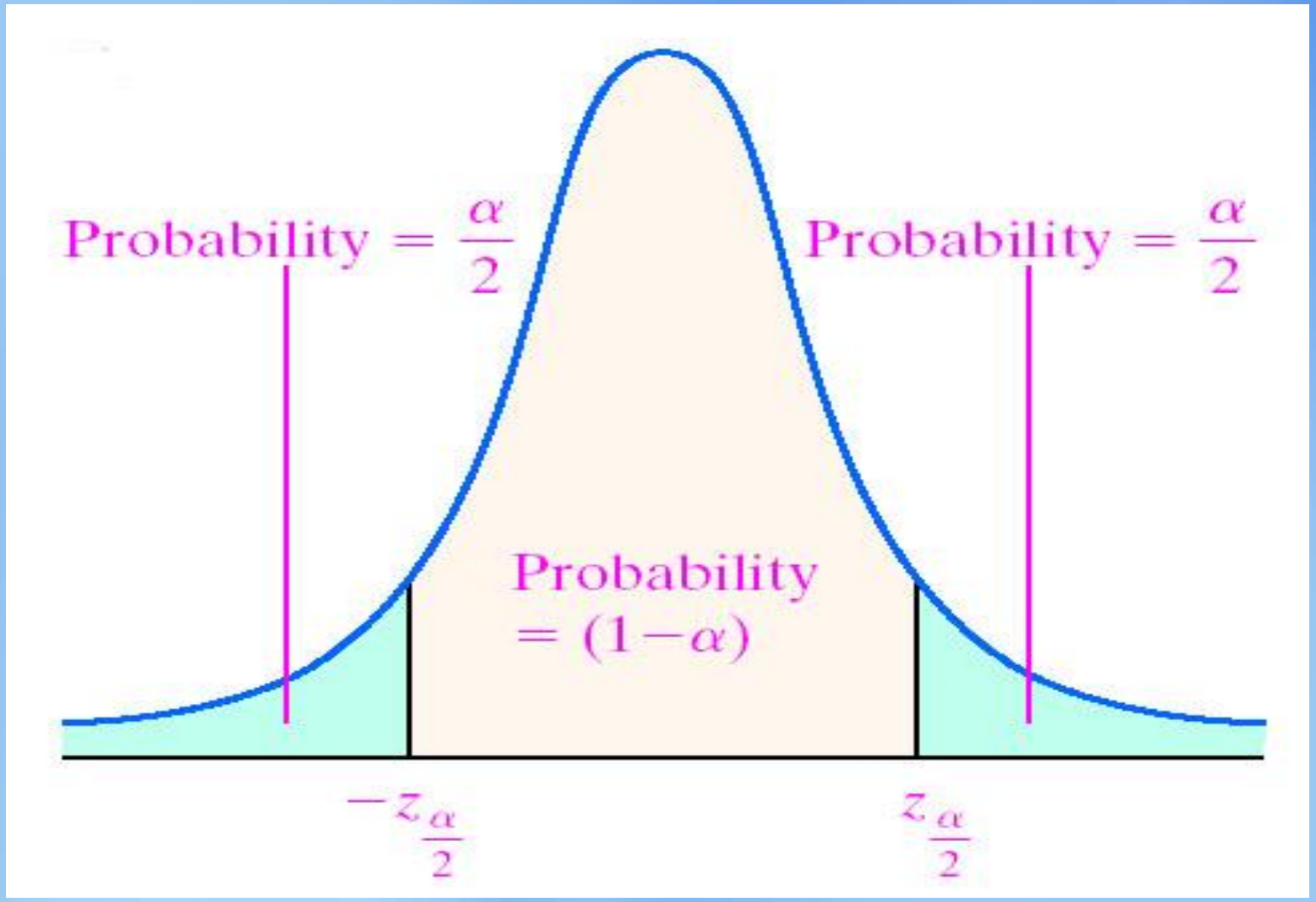
$$\mu_{\hat{p}} = p$$

$$\sigma_{\hat{p}} = \sqrt{\frac{pq}{n}}$$

Confidence Interval for the Population Proportion

An approximate 100% × (1 − α) CI for the binomial parameter p based on the normal approximation to the binomial distribution is given by:

$$\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$


Probability = $\frac{\alpha}{2}$
Probability = $\frac{\alpha}{2}$
Probability = $(1-\alpha)$
$-z_{\frac{\alpha}{2}}$
$z_{\frac{\alpha}{2}}$

Example

- Suppose we are interested in estimating the prevalence rate of breast cancer among 50- to 54-year-old women whose mothers have had breast cancer. Suppose that in a random sample of 10,000 such women, 400 are found to have had breast cancer at some point in their lives.
  1. Estimate the prevalence of breast cancer among 50- to 54-year-old women whose mothers have had breast cancer.
  2. Derive a 95% CI for the prevalence rate of breast cancer among 50- to 54-year-old women.

Solution:

1. The best point estimate of the prevalence rate p is given by the sample proportion

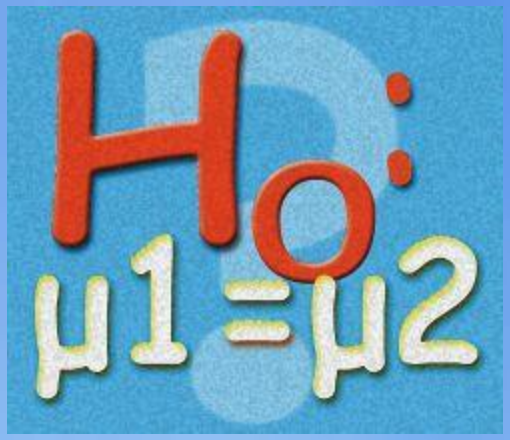
$$\hat{p} = \frac{400}{10000} = .040 .$$

2. $\hat{p} = .040$ $\quad a = .05$ $\quad Z_{1-a/2} = 1.96$

$n = 10{,}000$

Therefore, an approximate 95% CI is given by

$$0.04 \pm 1.96 \sqrt{\frac{(.04)(.96)}{10000}} = (0.036, 0.044)$$



Hypothesis Testing

Introduction

- Researchers often have preconceived ideas about what the values of these parameters might be and wish to test whether the data conform to these ideas.
- Used to investigate the validity of a claim about the value of a population characteristic.
- For example:
  1. The hospital administrator may want to test the hypothesis that the average length of stay of patients admitted to the hospital is 5 days!
  2. Suppose we want to test the hypothesis that mothers with low socioeconomic status (SES) deliver babies whose birth weights are lower than "normal."

- Hypothesis testing is widely used in medicine, dentistry, health care, biology and other fields as a means to draw conclusions about the nature of populations.
- **Why is hypothesis testing so important**? Hypothesis testing provides an objective framework for making decisions using probabilistic methods, rather than relying on subjective impressions.

Hypothesis testing is to provide information in helping to make decisions.

General Concepts

- A hypothesis is a statement about one or more populations. There are research hypotheses and statistical hypotheses.
- A research hypothesis is the supposition or conjecture that motivates the research.
- Statistical hypotheses are stated in such a way that they may be evaluated by appropriate statistical techniques

Elements of a hypothesis test

- **Null hypothesis (H_0):** Statement regarding the value(s) of unknown parameter(s). It is the hypothesis to be tested.
- **Alternative hypothesis (H_1):** Statement contradictory to the null hypothesis. It is a statement of what we believe is true if our sample data cause us to reject the null hypothesis.
- **Test statistic -** Quantity based on sample data and null hypothesis used to test between null and alternative hypotheses
- **Rejection region -** Values of the test statistic for which we reject the null in favor of the alternative hypothesis

Four possible outcomes in hypothesis testing

Truth

Decision

	H_0	H_1
Accept H_0	H_0 is true and H_0 is accepted Correct decision	H_1 is true and H_0 is accepted Type II error (β)
Reject H_0	H_0 is true and H_0 is rejected Type I error (α)	H_1 is true and H_0 is rejected Correct decision

- The probability of a type I error is the probability of rejecting the null hypothesis when *H_0 is true,* denoted by α and is commonly referred to as the significance level of a test.
- The probability of a type II error is the probability of accepting the null hypothesis when *H_1 is true, and* usually denoted by β.

- The general aim in hypothesis testing is to use statistical tests that make α and β as small as possible.
- This goal requires compromise because making α small will increase β.
- Our general strategy is to **fix** α at some specific level (for example, .10, .05, .01, . . .) and to use the test that minimizes β or, equivalently, maximizes the power.

One-Sample Test for the Mean of a Normal Distribution: One-Sided Alternatives

- Hypothesis:
  - Null Hypothesis $H_0 : \mu = \mu_0$
  - Alternative hypothesis $H_1: \mu < \mu_0$ *or* $H_1: \mu > \mu_0$
- Identify level of significance
  - α is a predetermined value (e.g., $\alpha = .05$)

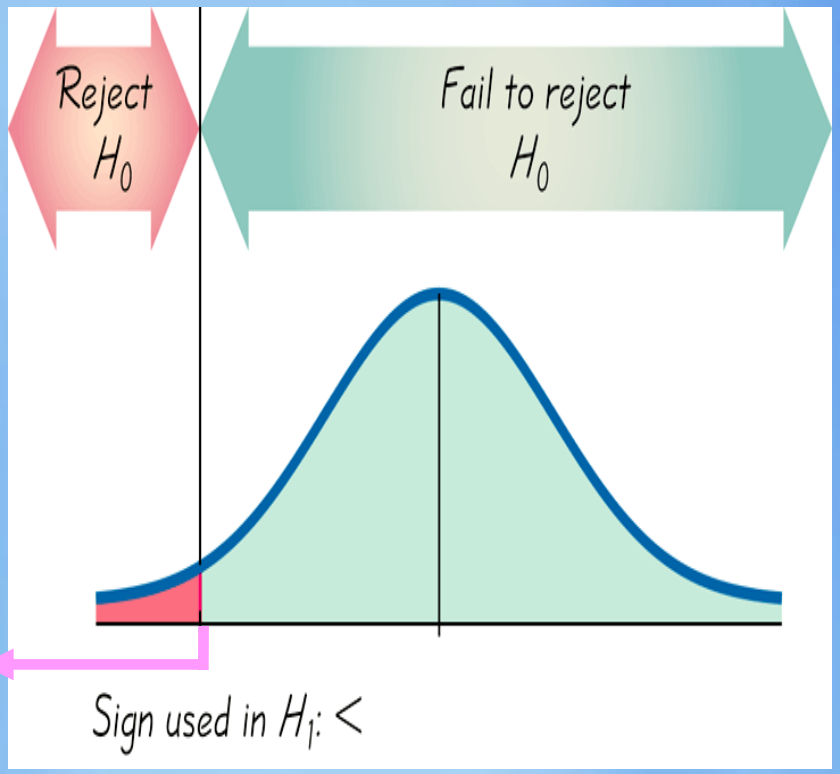
- Test statistic:
  - If σ is known, or if sample size is large

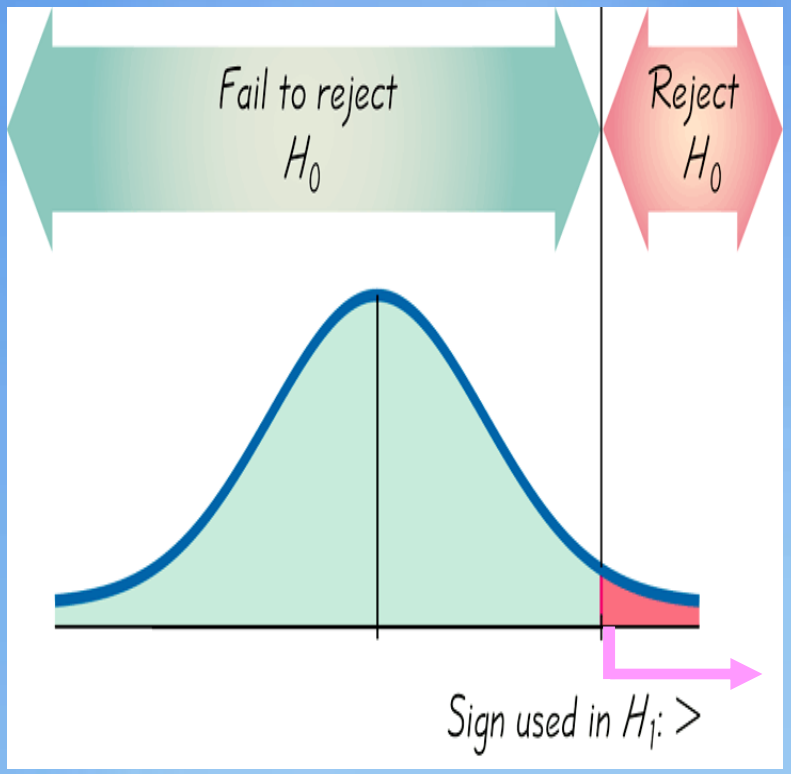
$$Z = \frac{\overline{X} - \mu_o}{\sigma / \sqrt{n}}$$

  - If σ is unknown, or if sample size is small

$$t = \frac{\overline{X} - \mu_o}{s / \sqrt{n}}$$

Critical Value: the critical region





Conclusion:

- Left-tailed Test:
  - Reject H_0 if $Z < Z_{1-\alpha}$ (when use Z - test)
  - Reject H_0 if $t < t_{n-1,\,\alpha}$ (when use T - test)
- Right-tailed Test
  - Reject H_0 if $Z > Z_{1-\alpha}$ (when use Z - test)
  - Reject H_0 if $t > t_{1-\alpha,n-1}$ (when use T - test)

- An Alternative Decision Rule using the p - value Definition

P-Value

- The P-value (or p-value or probability value) is the probability of getting a value of the test statistic that is at least as extreme as the one representing the sample data, assuming that the null hypothesis is true.
- The p-value is defined as the smallest value of α for which the null hypothesis can be rejected

- If the p-value is less than or equal to α ,we reject the null hypothesis ($p \leq \alpha$)
- If the p-value is greater than α ,we do not reject the null hypothesis ($p > \alpha$)

One-Sample Test for the Mean of a Normal Distribution: Two-Sided Alternatives

- A two-tailed test is a test in which the values of the parameter being studied (in this case μ) under the alternative hypothesis are allowed to be either *greater than or less than the values of the parameter under the null hypothesis (μ_0).*
- To test the hypothesis

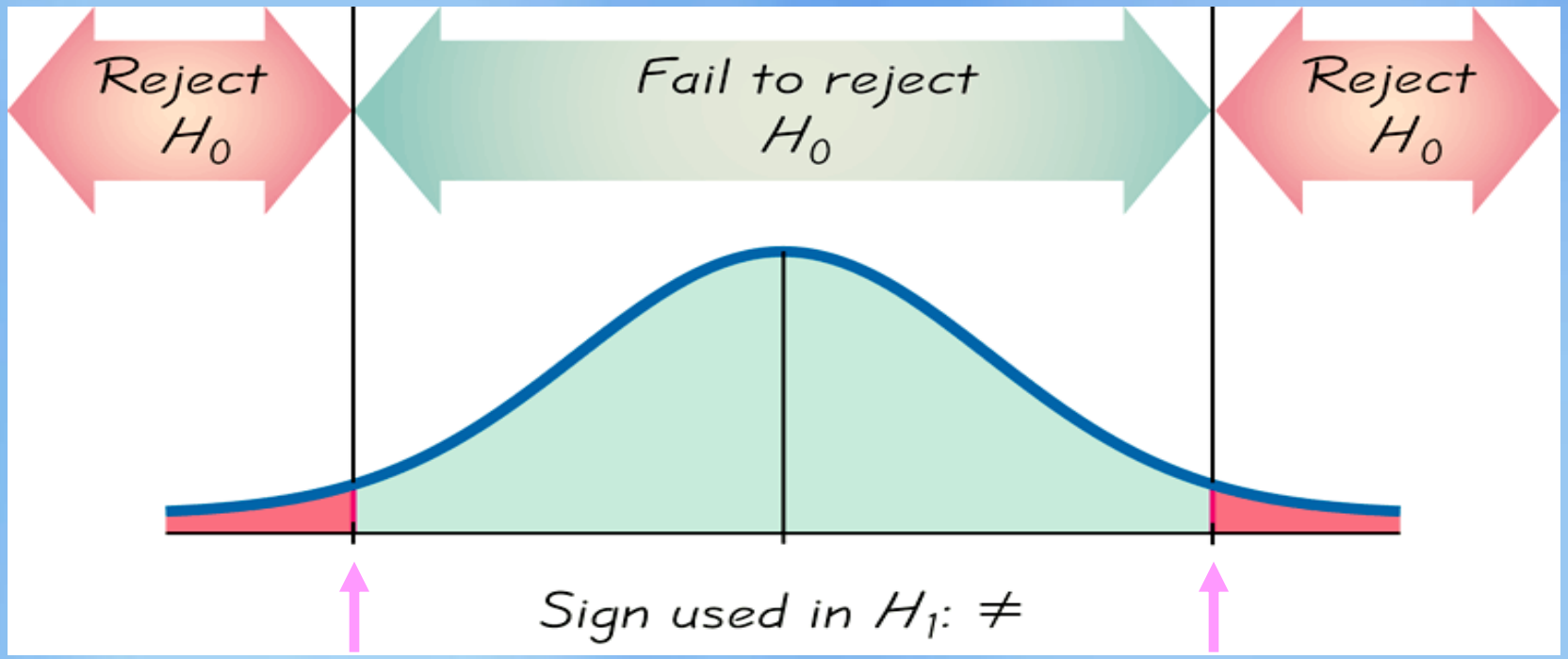
$$H_0: \mu = \mu_0 \text{ vs. } H_1: \mu \neq \mu_0$$

- Using the Test statistic:
  - If σ is known, or if sample size is large
  - If σ is unknown, or if sample size is small

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

$$t = \frac{\overline{X} - \mu_o}{s / \sqrt{n}}$$

- **Decision Rule:**
  - Reject H_0 if $Z > Z_{1-\alpha/2}$ or $Z < -Z_{1-\alpha/2}$ (when use Z - test)
  - Reject H_0 if $T > t_{1-\alpha/2,n-1}$ or $t < -t_{1-\alpha/2,n-1}$ (when use T- test)



Values that differ significantly from H_0

Example 1

- Among 157 Saudi men ,the mean systolic blood pressure was 146 mm Hg with a standard deviation of 27. We wish to know if on the basis of these data, we may conclude that the mean systolic blood pressure for a population of Saudi is greater than 140. Use $\alpha=0.01$.

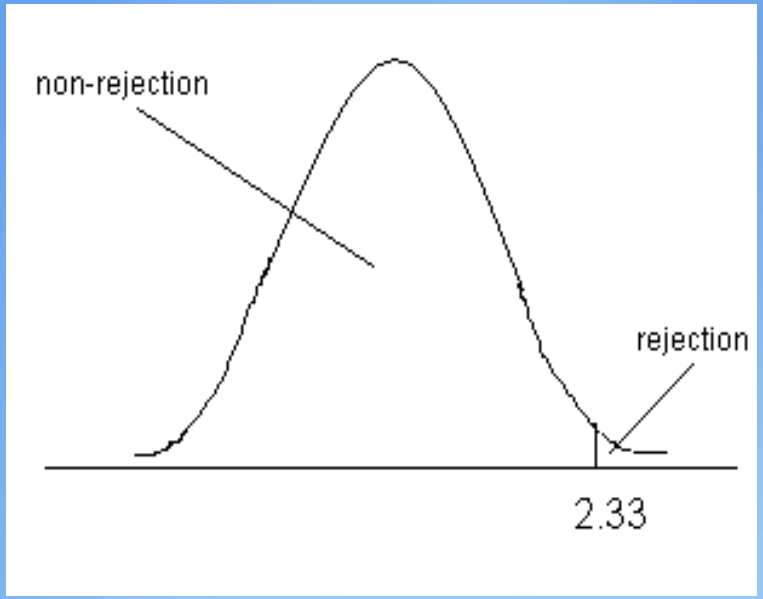
- Data: Variable is systolic blood pressure,
  n=157 , $\bar{X}$ =146, s=27, α=0.01.
- Assumption: population is not normal, σ^2 is unknown
- Hypotheses: H_0 :μ = 140
  H_1: μ > 140
- Test Statistic:

$$Z = \frac{\bar{X} - \mu_o}{\frac{s}{\sqrt{n}}} = \frac{146 - 140}{\frac{27}{\sqrt{157}}} = \frac{6}{2.1548} = \mathbf{2.78}$$

- Decision Rule:
  - we reject H_0 if $Z > Z_{1-\alpha}$
  - $Z_{0.99}$= 2.33 from table Z

- Decision:
  - Since 2.78 > 2.33.

then we reject H_0



Example 2

- The following are the systolic blood pressures of 12 patients undergoing drug therapy for hypertension:

  183, 152, 178, 157, 194, 163, 144, 114, 178, 152, 118, 158

Can we conclude on the basis of these data that the population mean of SBP is less than 165? Use α=0.05.

Use SPSS to do this!

Here σ is unknown and sample size is small!

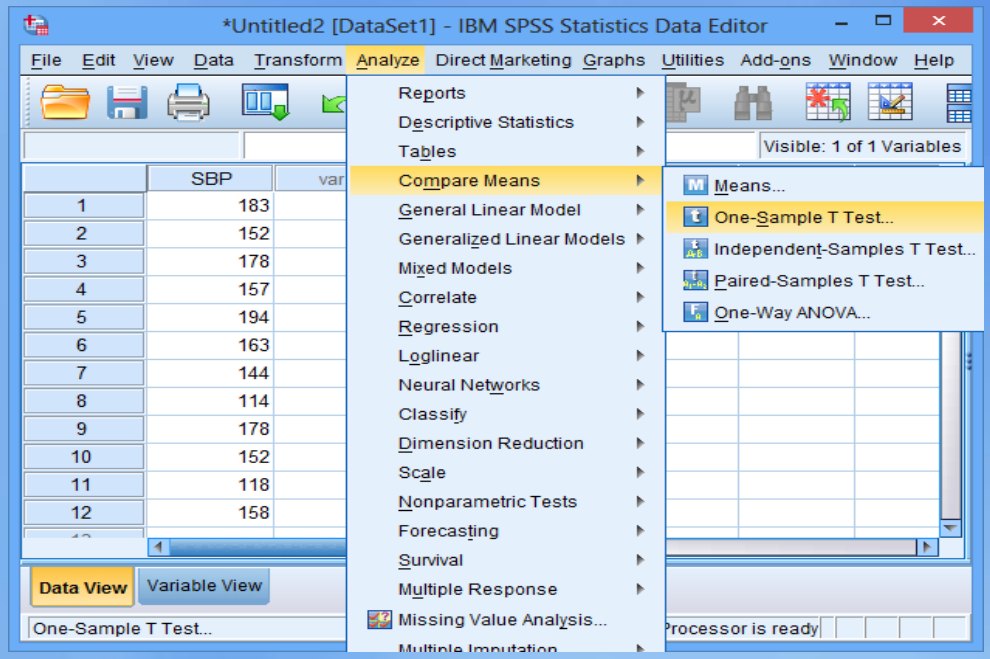
Solution:

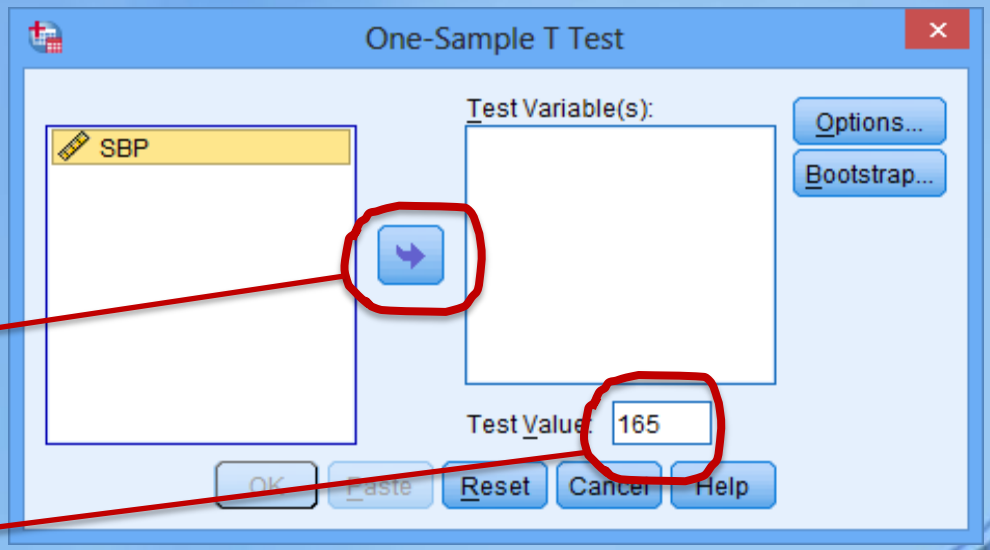
State your hypotheses:

$H_0 : \mu = 165$

$H_1 : \mu < 165$

Analyze → Compare Means → One-Sample T Test





Select the variable

Set the test value μ_0

The results of this test

Sample mean

Sample standard dev.

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
SBP	12	157.58	24.400	7.044

One-Sample Test

	Test Value = 165					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
SBP	-1.053	11	.315	-7.417	-22.92	8.09

T test statistic

P-value (two-tailed)

Mean from null hyp.

Here, the two-tailed P-value is 0.315. Since we are conducting a one-tailed test, the P-value is 0.315/2 = 0.1575.

Since p-value > 0.05. Then we fail to reject H_0.

One-Sample Inference for the Binomial Distribution

- **Recall that**: The sampling distribution of a sample proportion $\hat{p}$ is approximately normal (normal approximation of a binomial distribution) when the sample size is large enough.
- Thus, we can easily test the null hypothesis:

  H_0: $p = p_0$ (a given value we are testing).
- **Test Statistic**

$$z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1 - p_0)}{n}}}$$

- **Alternative Hypotheses:**

$$H_A : p \neq p_o \qquad H_1 : p < p_o \qquad H_1 : p > p_o$$

- **Decision Rule:**

If $H_1: P \neq P_0$
Reject H_0 if $Z > Z_{1-\alpha/2}$ or $Z < -Z_{1-\alpha/2}$

If $H_1: P > P_0$
Reject H_0 if $Z > Z_{1-\alpha}$

If $H_1: P < P_0$
Reject H_0 H_0 if $Z < -Z_{1-\alpha}$

Example

- Historically, one in five kidney cancer patients survive 5 years past diagnosis, i.e. 20%.
- An oncologist using an experimental therapy treats $n = 40$ kidney cancer patients and 16 of them survive at least 5 years.
- Is there evidence that patients receiving the experimental therapy have a higher 5-year survival rate?

p = the proportion of kidney cancer patients receiving the experimental therapy that survive at least 5 years.

H_o : $p \leq .20$ (5 - yr. survival rate is not better)

H_A : $p > .20$ (5 - yr. survival rate is better)

Determine test criteria

Choose $\alpha = .05$ (may want to consider smaller?)

Use large sample test since:

$np = (40)(.20) = 8 > 5$ and $n(1-p) = (40)(.80) = 32 > 5$

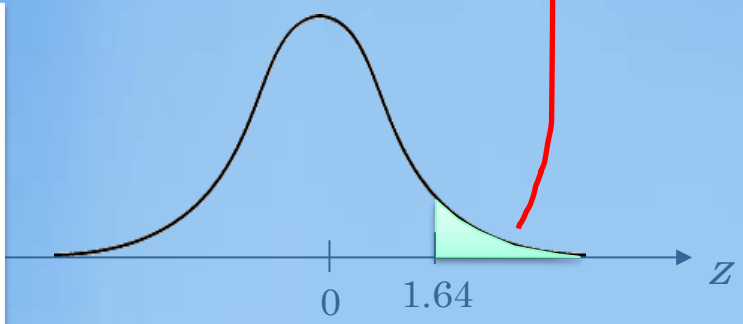
$$\hat{p} = \frac{16}{40} = .40 \quad \text{or a } 40\% \text{ 5 - yr. survival rate}$$

$$z = \frac{\hat{p} - p_o}{\sqrt{\frac{p_o(1 - p_o)}{n}}} = \frac{.40 - .20}{\sqrt{\frac{.20\,(1 - .20)}{40}}} = 3.16$$

Conclusion:

$Z_{1-\alpha} = Z_{0.95} = 1.64$

Since $Z > Z_{1-\alpha}$ ➡ Reject H_0



Hypothesis Testing for Variance and Standard Deviation

- **Why test variance ?**
  - In real life, things vary. Even under strictest conditions, results will be different between one another. This causes variances.
  - When we get a sample of data, apart from testing on the mean to find if it is reasonable, we also test the variance to see whether certain factor has caused greater or smaller variation among the data.
  - In nature, variation exists. Even identical twins have some differences. This makes life interesting

- Sometimes we are interested in making inference about the variability of processes.
- Examples:
  - The consistency of a production process for quality control purposes.
  - Investors use variance as a measure of risk.
- To draw inference about variability, the parameter of interest is σ^2.

- We know that the sample variance s^2 is an unbiased, consistent and efficient point estimator for σ^2.
- The statistic χ^2 has a distribution called Chi-squared, if the population is normally distributed.
- The test statistic for testing hypothesis about a population variance is:

$$\chi^2 = \frac{(n-1)s^2}{{\sigma_0}^2}$$

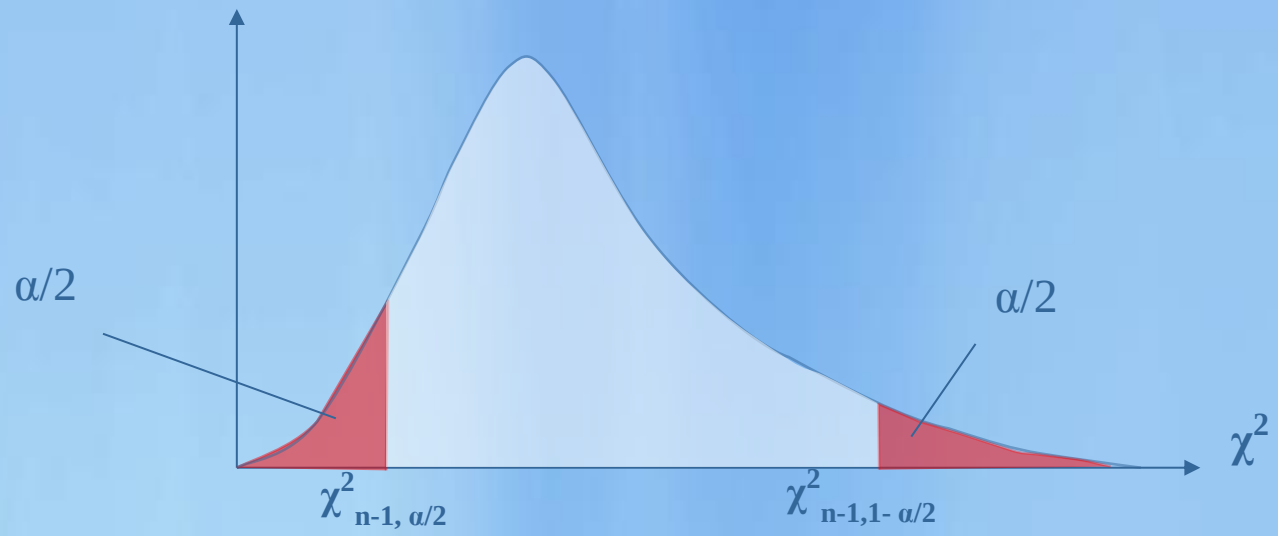
- follows a chi-square distribution with degrees of freedom d.f. = $n - 1$.

- If we want to test the following hypothesis:

$$H_o : \sigma = \sigma_o$$

$$H_1 : \sigma \neq \sigma_o$$

- reject H_0 if: $\chi^2 < \chi^2_{n-1,\, \alpha/2}$ or $\chi^2 > \chi^2_{n-1,1-\alpha/2}$



Example

- A local balloon company claims that the variance for the time one of its helium balloons will stay afloat is 5 seconds.
- A disgruntled customer wants to test this claim.
- She randomly selects 23 customers and finds that the variance of the sample is 4.5 seconds.
- At $\alpha = 0.05$, does she have enough evidence to reject the company's claim?
- Hypothesis to be tested:

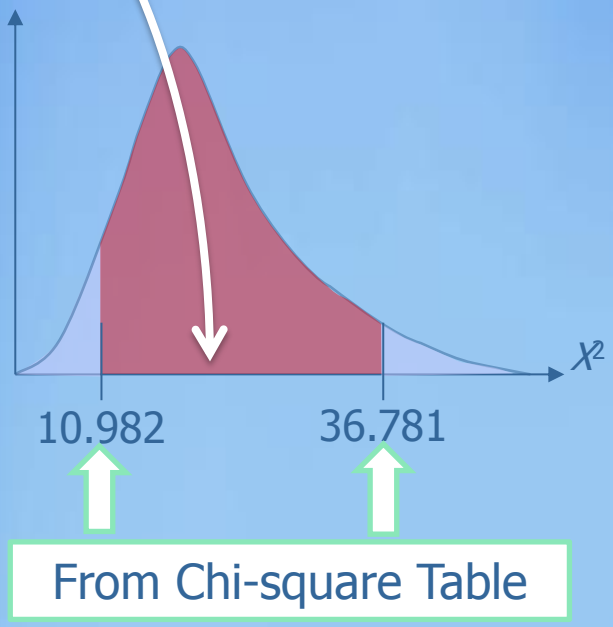
  $H_0: \sigma^2 = 5$ (Claim)

  $H_1: \sigma^2 \neq 5$

- Test statistic:

$$X^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{(23-1)(4.5)}{5} = 19.8$$

Fail to reject H_0.



At $\alpha = 0.05$, there is not enough evidence to reject the claim that the variance of the float time is 5 seconds.

Hypothesis Testing: Two-Sample Inference

- A more frequently encountered situation is the two-sample hypothesis-testing problem.
- In a two-sample hypothesis-testing problem, the underlying parameters of two different populations, neither of whose values is assumed known, are compared.
- Distinguish between **Independent** and **Dependent** Sampling.

- Two samples are said to be **independent** when the data points in one sample are unrelated to the data points in the second sample.
- Two samples are said to be **Dependent** (**paired)** when each data point in the first sample is matched and is related to a unique data point in the second sample.
- Dependent samples are often referred to as **matched-pairs** samples.

The Paired *t* Test

- Statistical inference methods on matched-pairs data use the same methods as inference on a single population mean with σ unknown, except that the **differences** are analyzed.

$$d_i = x_{i2} - x_{i1}.$$

- The hypothesis testing problem can thus be considered a *one-sample t test based on the differences* (d_i).

- To test hypotheses regarding the mean difference of matched-pairs data, the following must be satisfied:

1. the sample is obtained using simple random sampling
2. the sample data are matched pairs,
3. the differences are normally distributed with no outliers or the sample size, n, is large ($n \geq 30$).

Step 1:

- Determine the null and alternative hypotheses. The hypotheses can be structured in one of three ways, where μ_d is the population mean difference of the matched-pairs data.

Two-Tailed	Left-Tailed	Right-Tailed
$H_0: \mu_d = 0$	$H_0: \mu_d = 0$	$H_0: \mu_d = 0$
$H_1: \mu_d \neq 0$	$H_1: \mu_d < 0$	$H_1: \mu_d > 0$

Step 2:

- Select a level of significance, α, based on the seriousness of making a Type I error.

Step 3:

- Compute the test statistic

$$t_0 = \frac{\bar{d}}{\frac{s_d}{\sqrt{n}}}$$

which approximately follows Student' t-distribution with n-1 degrees of freedom

- The values of $\bar{d}$ and s_d are the mean and standard deviation of the differenced data.

Step 4:

- Use Table t to determine the critical value using n-1 degrees of freedom.

Step 5:

- Compare the critical value with the test statistic:
- If P-value < α, reject the null hypothesis.

Two-Tailed	Left-Tailed	Right-Tailed
If $t_0 < -t_{\frac{\alpha}{2}}$ or $t_0 > t_{\frac{\alpha}{2}}$, reject the null hypothesis.	If $t_0 < -t_\alpha$, reject the null hypothesis.	If $t_0 > t_\alpha$, reject the null hypothesis.

Example: Hypertension

- Let's say we are interested in the relationship between oral contraceptive (OC) use and blood pressure in women.
- Identify a group of non-pregnant, premenopausal women of childbearing age (16–49 years) who are not currently OC users, and measure their blood pressure, which will be called the ***baseline blood pressure***.
- Rescreen these women 1 year later to ascertain a subgroup who have remained non-pregnant throughout the year and have become OC users. This subgroup is the **study population**.
- Measure the blood pressure of the study population at the follow-up visit.

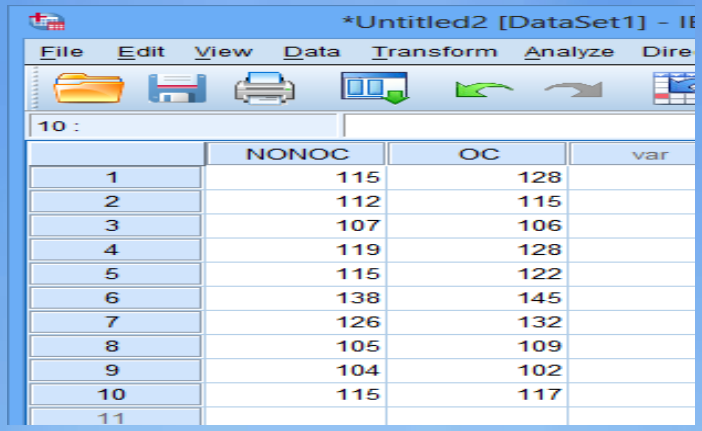
- Compare the baseline and follow-up blood pressure of the women in the study population to determine the difference between blood pressure levels of women when they *were* using the pill at follow-up and when they *were not* using the pill at baseline.
- Assume that the SBP of the i_{th} woman is normally distributed at baseline with mean μ_i and variance σ^2 and at follow-up with mean $\mu_i + \Delta$ and variance σ^2.
- If $\Delta = 0$, then there is no difference between mean baseline and follow-up SBP. If $\Delta > 0$, then using OC pills is associate with a raised mean SBP. If $\Delta < 0$, then using OC pills is associated with a lowered mean SBP.
- We want to test the hypothesis:

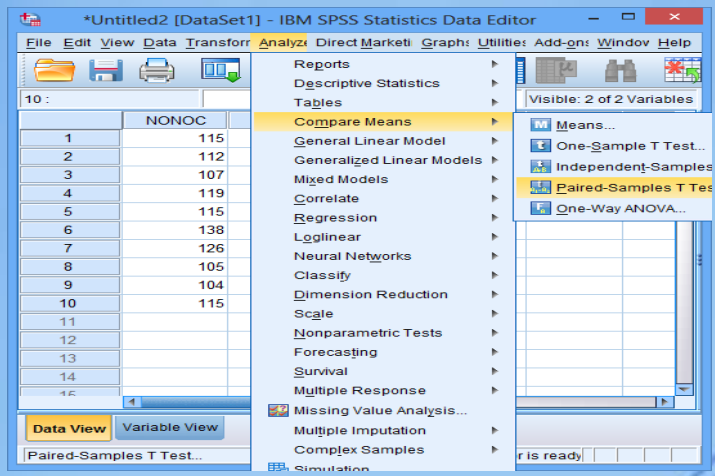
$$H_0: \Delta = 0 \quad \text{vs.} \quad H_1: \Delta \neq 0.$$

- **How should we do this?**

Solution:

- Using SPSS
  - After Importing your dataset, and providing names to variables, click on:
  - ANALYZE → COMPARE MEANS → PAIRED SAMPLES T-TEST
  - For PAIRED VARIABLES, Select the two dependent (response) variables (the analysis will be based on first variable minus second variable)



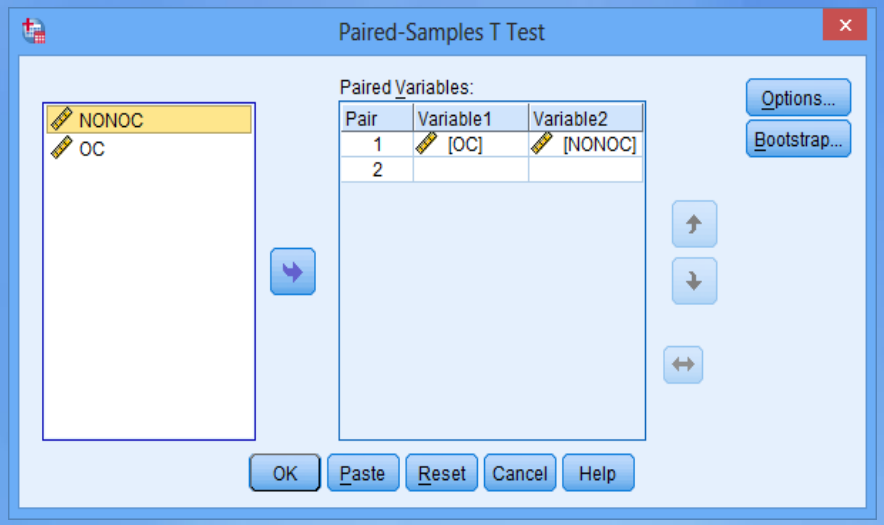


Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	OC	120.40	10	13.226	4.183
	NONOC	115.60	10	10.309	3.260

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	OC & NONOC	10	.955	.000



$\bar{d}$ S_d

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	OC - NONOC	4.800	4.566	1.444	1.534	8.066	3.325	9	.009

p-value

Since the exact two-sided *p*-value = .009 < 0.05, we reject the null hypothesis.Therefore, we can conclude that starting OC use is associated with a significant change in blood pressure.

- A two-sided 100% × (1 − α) CI for the true mean difference (Δ) between two paired samples is given by:

$$\bar{d} - t_{n-1,1-\frac{\alpha}{2}} \cdot \frac{s_d}{\sqrt{n}}, \quad \bar{d} + t_{n-1,1-\frac{\alpha}{2}} \cdot \frac{s_d}{\sqrt{n}}$$

Compute a 95% CI for the true increase in mean SBP after starting OCs.

(1.53, 8.07)

Note that the conclusion from the hypothesis test and the confidence interval are the same!

Two-Sample t Test for Independent Samples with Equal Variances

- The independent samples t-test is probably the single most widely used test in statistics.
- It is used to compare differences between separate groups.
- We want to test the hypothesis:

$$H_0: \mu_1 = \mu_2 \text{ vs. } H_1: \mu_1 \neq \mu_2.$$

- Sometimes we will be willing to assume that the variance in the two groups is equal:

$$\sigma_1^2 = \sigma_2^2 = \sigma^2$$

- If we know this variance, we can use the z-statistic

$$z = \frac{\left(\overline{x_1} - \overline{x_2}\right) - \left(\mu_1 - \mu_2\right)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- Often we have to estimate σ^2 with the sample variance from each of the samples, s_1^2, s_2^2
- Since we have two estimates of one quantity, we pool the two estimates

- The average of s_1^2 and s_2^2 could simply be used as the estimate of σ^2 :

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- The t-statistic based on the pooled variance is very similar to the z-statistic as always:

$$t = \frac{(\overline{x_1} - \overline{x_2}) - (\mu_1 - \mu_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- The t-statistic has a t-distribution with $n_1 + n_2 - 2$ degrees of freedom

Constructing a (1-α)100 % Confidence Interval for the Difference of Two Means: equal variances $\sigma_1^2 = \sigma_2^2$

- If the two populations are normally distributed or the sample sizes are sufficiently large ($n_1 \geq 30$ and $n_2 \geq 30$), and a (1-α)100 % confidence interval about μ_1 - μ_2 is given by:

$$\left[(\bar{x}_1 - \bar{x}_2) - t_{n_1+n_2-2,1-\frac{\alpha}{2}} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \;,\; (\bar{x}_1 - \bar{x}_2) + t_{n_1+n_2-2,1-\frac{\alpha}{2}} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \right]$$

Unequal variance

- Often, we are unwilling to assume that the variances are equal!
- The test statistic as:

$$t = \frac{\left(\overline{x_1} - \overline{x_2}\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- The distribution of this statistic is difficult to derive and we approximate the distribution using a t-distribution with ν degrees of freedom

$$\nu = \frac{\left[\left(s_1^2/n_1\right) + \left(s_2^2/n_2\right)\right]^2}{\left[\frac{\left(s_1^2/n_1\right)^2}{(n_1 - 1)} + \frac{\left(s_2^2/n_2\right)^2}{(n_2 - 1)}\right]}$$

- This is called the Satterthwaite or Welch approximation

Constructing a (1-α)100 % Confidence Interval for the Difference of Two Means

- If the two populations are normally distributed or the sample sizes are sufficiently large ($n_1 \geq 30$ and $n_2 \geq 30$), a (1-α)100 % confidence interval about μ_1 - μ_2 is given by:

$$\left[(\bar{x}_1 - \bar{x}_2) - t_{\nu, 1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad , \quad (\bar{x}_1 - \bar{x}_2) + t_{\nu, 1-\frac{\alpha}{2}} \cdot \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \right]$$

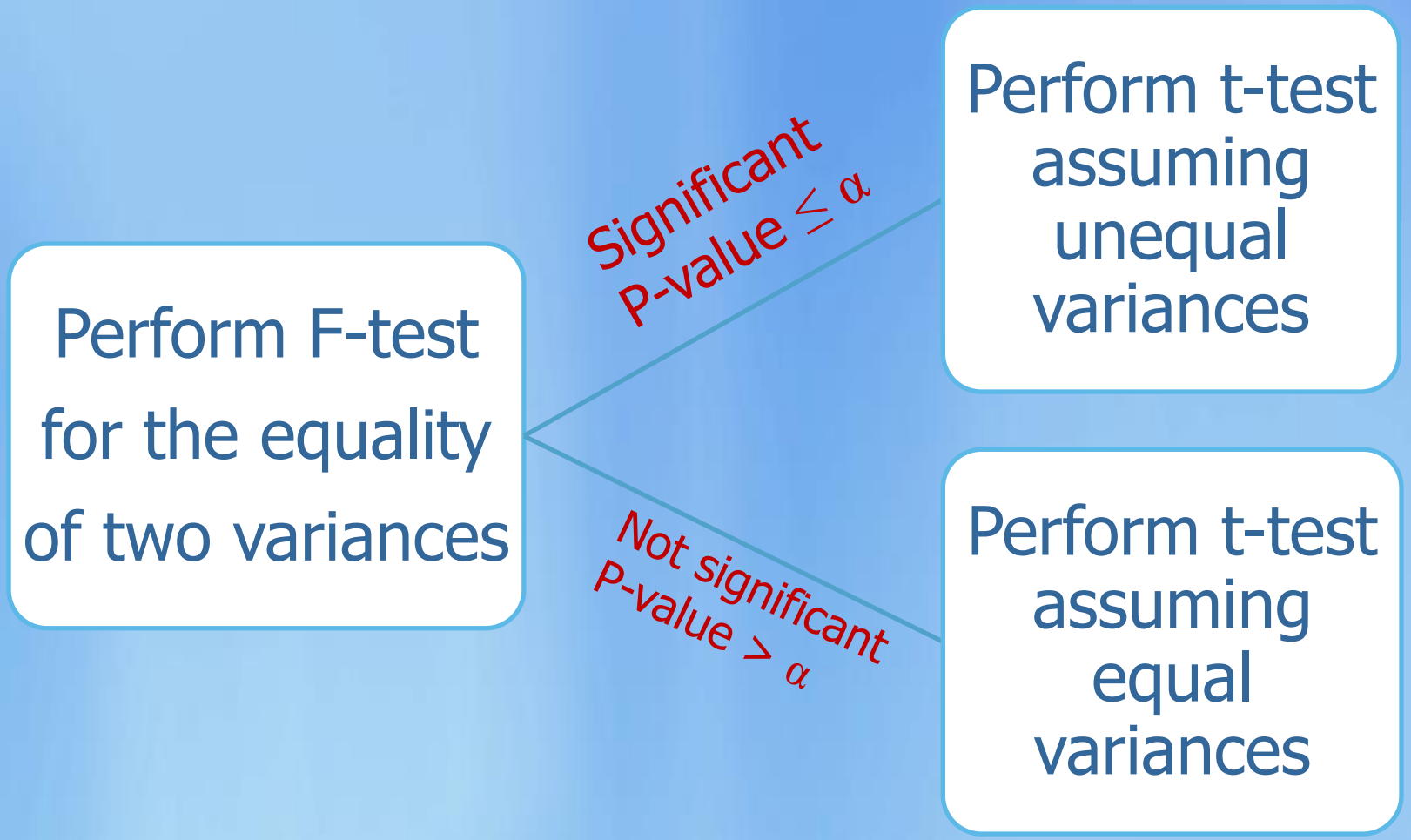
Equality of variance

- Since we can never be sure if the variances are equal, could we test if they are equal?
- Of course we can!!!
  - But, remember there is error in every statistical test
  - Sometimes it is just preferred to use the unequal variance unless there is a good reason.
- Test: H_0: $\sigma_1^2 = \sigma_2^2$ vs. H_1: $\sigma_1^2 \neq \sigma_2^2$ with significance level α.
- To test this hypothesis, we use the sample variances: s_1^2, s_2^2

Testing for the Equality of Two Variances

- One way to test if the two variances are equal is to check if the **ratio** is equal to 1 (H_0: ratio=1)
- Under the null, the ratio simplifies to $\frac{s_1^2}{s_2^2}$
- The ratio of 2 chi-square random variables has an **F-distribution.**
- The F-distribution is defined by the numerator and denominator degrees of freedom.
- Here we have an F-distribution with n_1-1 and n_2-1 degrees of freedom

Strategy for testing for the equality of means in two independent, normally distributed samples



Example

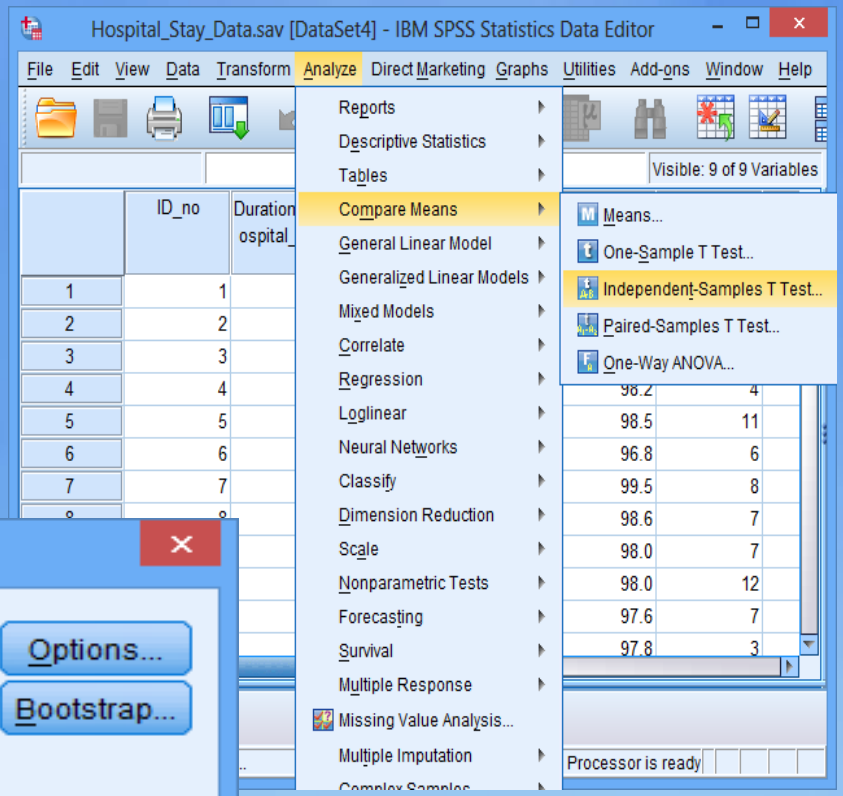
- Use **hypothesis testing** for **Hospital-stay** data in Table 2.11 (page 33 in the book) to answer (2.3).
- Use SPSS to do the analysis!

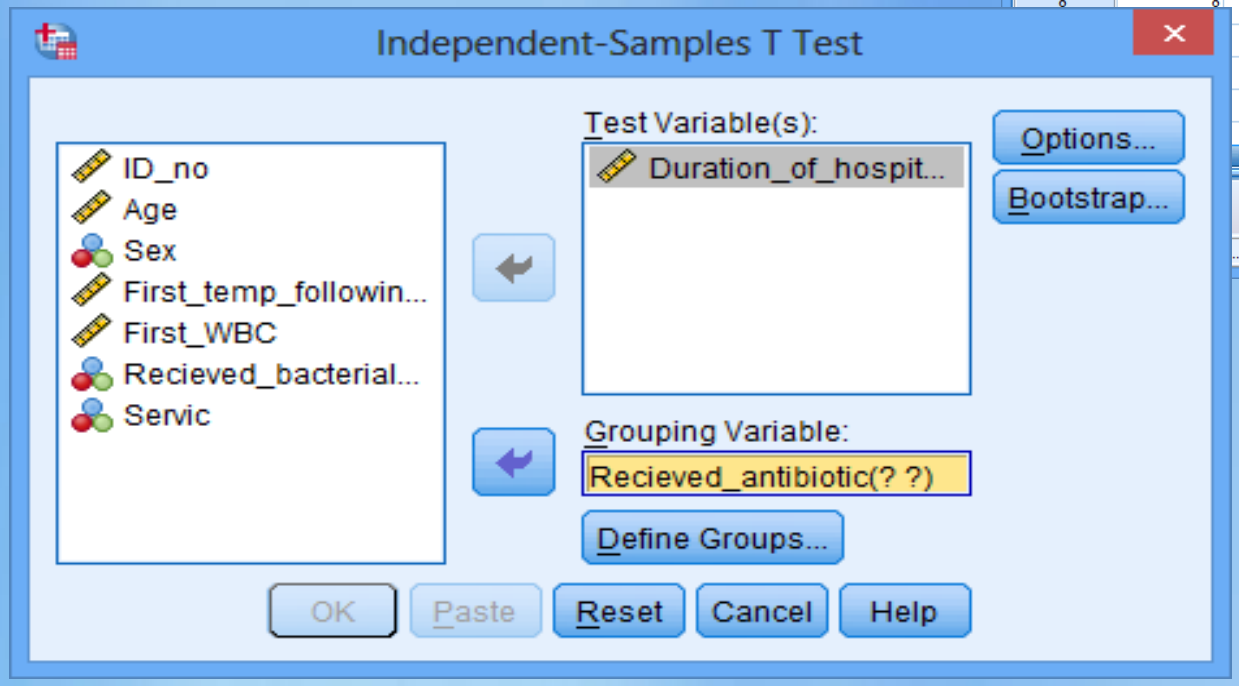
2.3: It is of clinical interest to know if the duration of hospitalization is affected by whether a patient has received antibiotics.

That is to compare the mean duration of hospitalization between antibiotic users and non-antibiotic users.

Solution:

Analyze → Compare Means → Independent Samples T Test





SPSS output:

Group Statistics

	Recieved_antibiotic	N	Mean	Std. Deviation	Std. Error Mean
Duration_of_hospital_stay	yes	7	11.57	8.810	3.330
	no	18	7.44	3.698	.872

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Duration_of_hospital_stay	Equal variances assumed	3.478	.075	1.682	23	.106	4.127	2.454	-.950	9.204
	Equal variances not assumed			1.199	6.839	.270	4.127	3.442	-4.051	12.305

$H_0: \sigma_1^2 = \sigma_2^2$ vs. $H_1: \sigma_1^2 \neq \sigma_2^2$

$H_0: \Delta = 0$ vs. $H_1: \Delta \neq 0.$

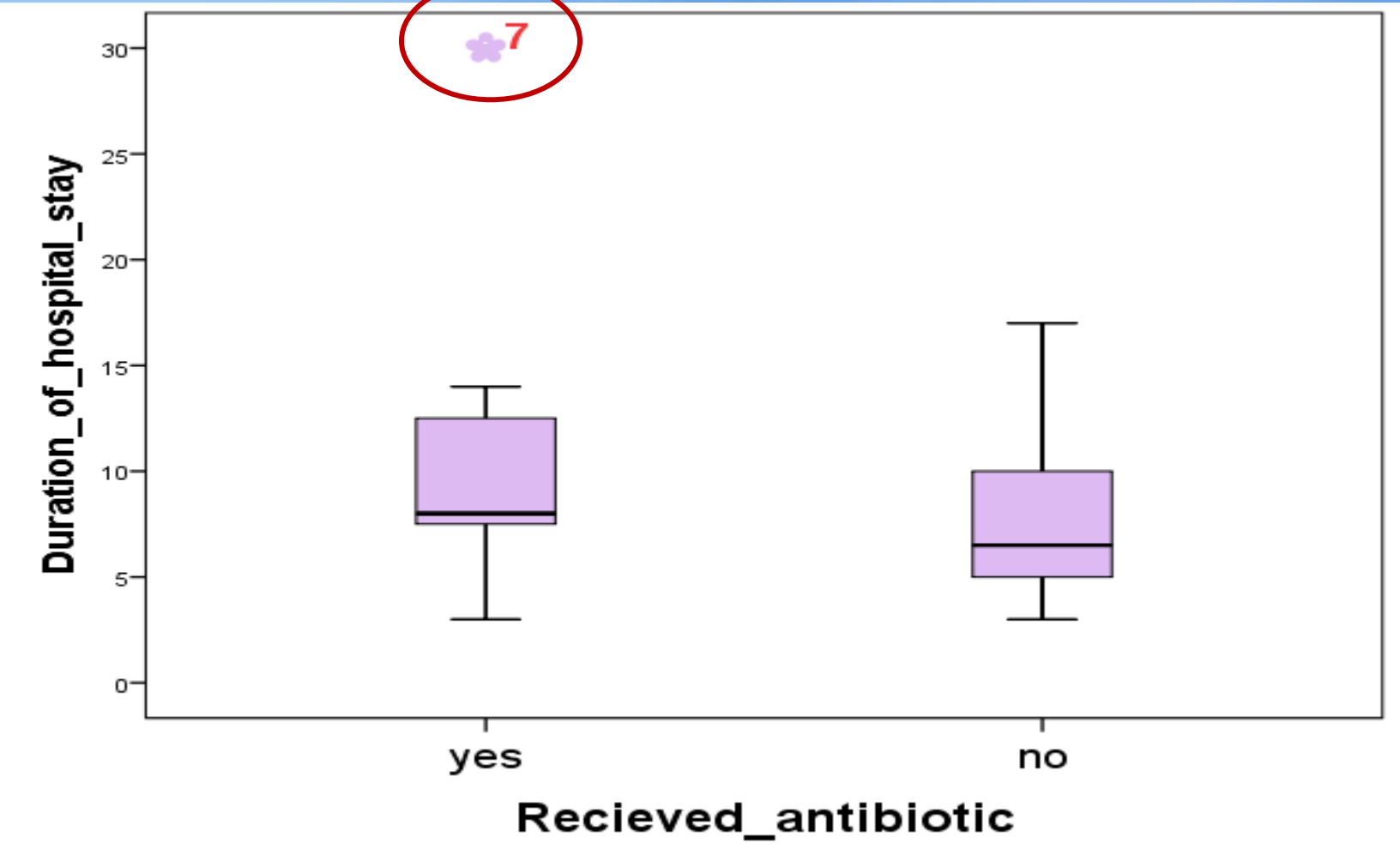
Conclusion

- First step in comparing the two means is to perform the F-test for the equality of two variances in order to decide whether to use the t-test with equal or unequal variances.
- The *p*-value = 0.075 > 0.05. Thus the variances DO NOT differ significantly, and a two-sample t-test with **equal variances** should be used.
- Therefore, refer to the Equal Variance row, where the t-statistic is 1.68 with degrees of freedom d'(df). = 23.
- The corresponding two-tailed *p*-value = 0.106 > 0.05.
- Thus no significant difference exists between the mean duration of hospitalization in these two groups.

The Treatment of Outliers

- Outliers can have an important impact on the conclusions of a study.
- It is important to definitely identify outliers and either:
  - Exclude them! (NOT recommended) or
  - Perform alternative analyses with and without the outliers present.
  - Use a method of analysis that minimizes their effect on the overall results.

Box-Plot


7
30
25
20
15
10
5
0
Duration_of_hospital_stay
yes
no
Recieved_antibiotic

Inference about Two Population Proportions

- Testing hypothesis about two population proportion (P_1, P_2) is carried out in much the same way as for difference between two means when condition is necessary for using normal curve are met.
- We have the following steps:

  Data: sample size (n_1 ,n_2), sample proportions ($\hat{p}_1$, $\hat{p}_2$), $\overline{p} = \frac{x_1 + x_2}{n_1 + n_2}$

  Characteristic in two samples (x_1 , x_2),

Assumption: Two populations are independent.

Hypotheses: Determine the null and alternative hypotheses. The hypotheses can be structured in one of three ways

Two-Tailed	Left-Tailed	Right-Tailed
$H_0: p_1 = p_2$	$H_0: p_1 = p_2$	$H_0: p_1 = p_2$
$H_1: p_1 \neq p_2$	$H_1: p_1 < p_2$	$H_1: p_1 > p_2$

Note: p_1 is the population proportion for population 1, and p_2 is the population proportion for population 2.

Select a level of significance α, based on the seriousness of making a Type I error.

- Compute the test statistic:

$$Z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n_1} + \frac{\overline{p}(1-\overline{p})}{n_2}}}$$

- Decision Rule:
  - Compare the critical value with the test statistic:

Two-Tailed	Left-Tailed	Right-Tailed
If $z_0 < -z_{\frac{\alpha}{2}}$ or $z_0 > z_{\frac{\alpha}{2}}$, reject the null hypothesis.	If $z_0 < -z_\alpha$, reject the null hypothesis.	If $z_0 > z_\alpha$, reject the null hypothesis.

Example

- An economist believes that the percentage of urban households with Internet access is greater than the percentage of rural households with Internet access.
- He obtains a random sample of 800 urban households and finds that 338 of them have Internet access. He obtains another random sample of 750 rural households and finds that 292 of them have Internet access.
- Test the economist's claim at the α =0.05 level of significance.

Solution:

- The samples are simple random samples that were obtained independently.

  x1=338, n1=800, x2=292 and n2=750, so

$$\hat{p}_1 = \frac{338}{800} = 0.4225 \quad \text{and} \quad \hat{p}_2 = \frac{292}{750} = 0.3893 \quad . \text{ Thus,}$$

- We want to determine whether the percentage of urban households with Internet access is greater than the percentage of rural households with Internet access.

- So, we test the hypothesis:

$$H_0: \ p_1 = p_2 \quad \text{versus} \quad H_1: p_1 > p_2$$

or, equivalently,

$$H_0: \ p_1 - p_2 = 0 \quad \text{versus} \quad H_1: p_1 - p_2 > 0$$

The level of significance is $\alpha = 0.05$

The pooled estimate:

$$\overline{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{338 + 292}{800 + 750} = 0.4065 \ .$$

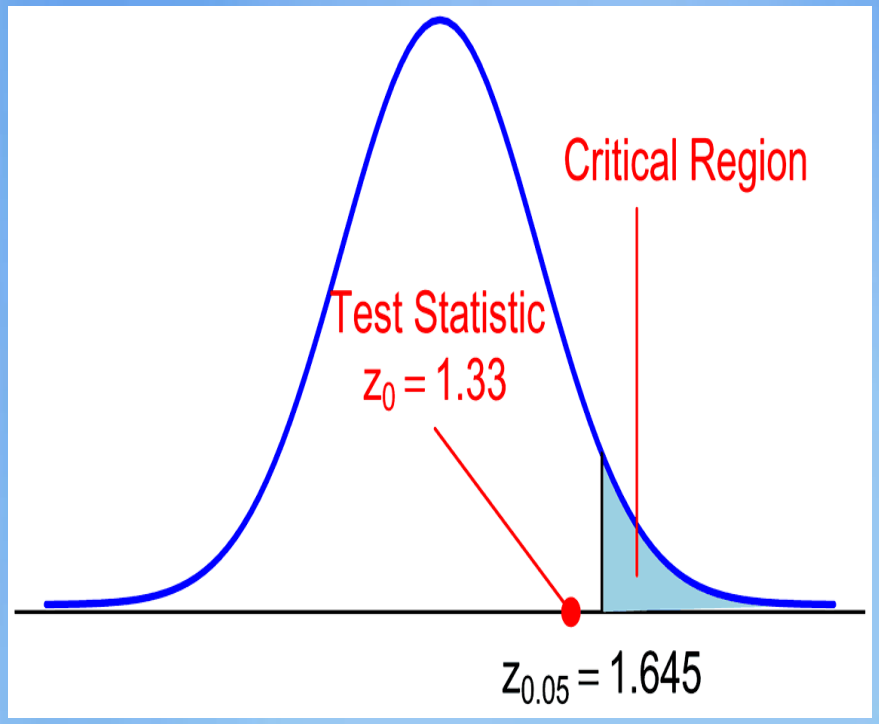
The test statistic is:

$$Z = \frac{0.4225 - 0.3893}{\sqrt{0.4065\ (1 - 0.4065\ )}\sqrt{\frac{1}{800} + \frac{1}{750}}} = 1.33 \ .$$

- This is a right-tailed test with α=0.05.
  The critical value is $Z_{0.05}$=1.645.

Since the test statistic, z_0=1.33 is less than the critical value $z_{.05}$=1.645,

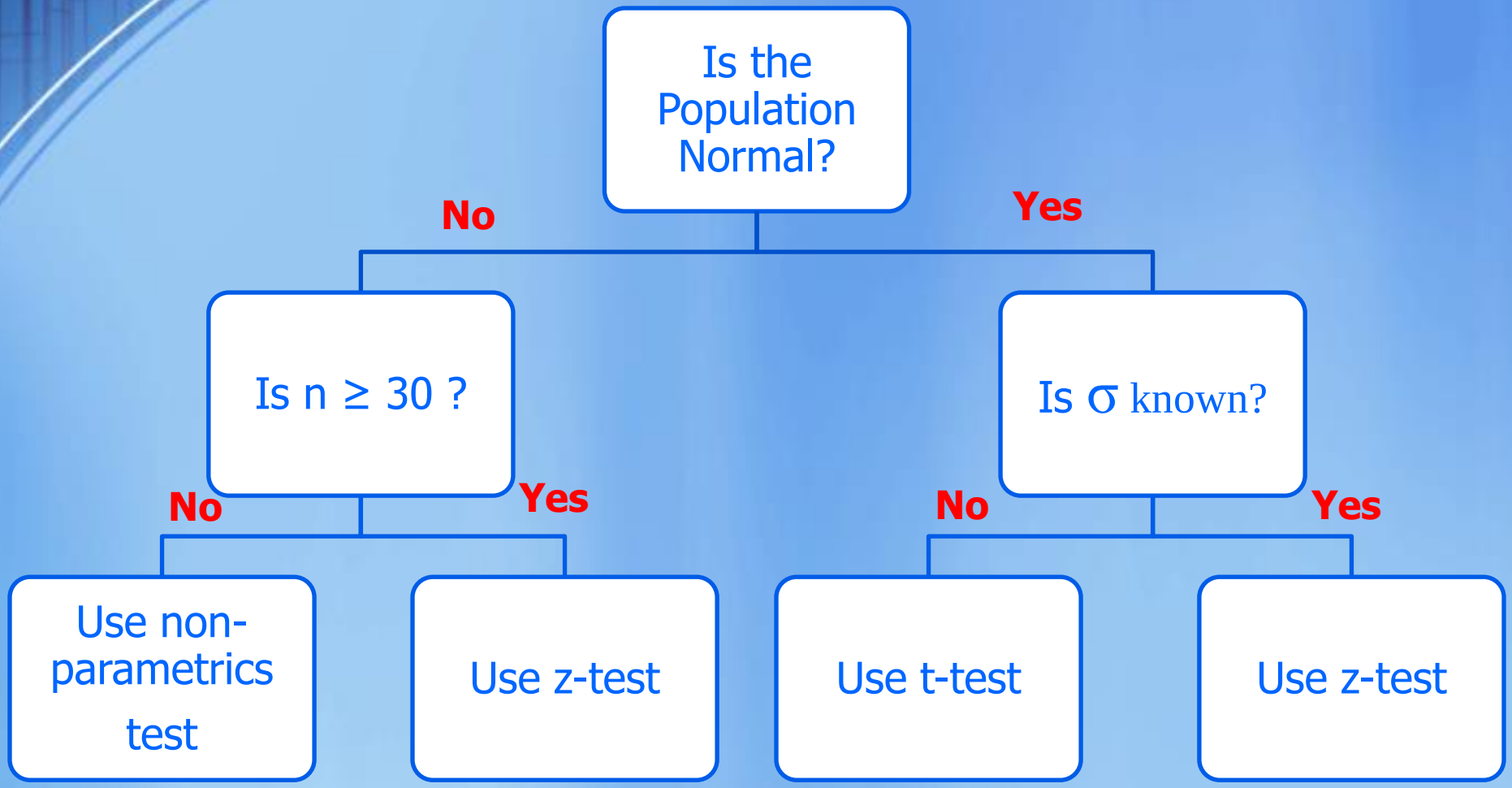
We fail to reject the null hypothesis.



Constructing a (1-α)100% Confidence Interval for the Difference between Two Population Proportions

- To construct a (1-α)·100% confidence interval for the difference between two population proportions:

$$\left(\hat{p}_1 - \hat{p}_2\right) - z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}_1\left(1-\hat{p}_1\right)}{n_1} + \frac{\hat{p}_2\left(1-\hat{p}_2\right)}{n_2}}, \quad \left(\hat{p}_1 - \hat{p}_2\right) + z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}_1\left(1-\hat{p}_1\right)}{n_1} + \frac{\hat{p}_2\left(1-\hat{p}_2\right)}{n_2}}$$


Is the Population Normal?
No
Yes
Is n ≥ 30 ?
Is σ known?
No
Yes
No
Yes
Use non-parametrics test
Use z-test
Use t-test
Use z-test