Hypothesis test

In statistics, a **hypothesis** is a claim or statement about a property of a population. A **hypothesis test** (or **test of significance**) is a standard procedure for testing a claim about a property of a population.
Null Hypothesis $H_0$

- The null hypothesis (denoted by $H_0$) is a statement that the value of a population parameter (such as proportion, mean, or standard deviation) is equal to some claimed value.
- We test the null hypothesis directly.
- Either reject $H_0$ or fail to reject $H_0$. 
Alternative Hypothesis $H_1$

- The **alternative hypothesis** (denoted by $H_1$ or $H_a$ or $H_A$) is the statement that the parameter has a value that somehow differs from the null hypothesis.

- The symbolic form of the alternative hypothesis must use one of these symbols: ≠, <, or >.
Example:

Consider the claim that the mean weight of airline passengers (including carry-on baggage) is at most 195 lb (the current value used by the Federal Aviation Administration). Follow the three-step procedure outlined to identify the null hypothesis and the alternative hypothesis.
Example:

Step 1: The claim that the mean is at most 195 lb is expressed in symbolic form as $\mu \leq 195$.

Step 2: If $\mu \leq 195$ is false, then $\mu > 195$ must be true.

Step 3: Of the two symbolic expressions, we see that $\mu > 195$ does not contain equality, so we let the alternative hypothesis $H_1$ be $\mu > 195$. Also, the null hypothesis must be a statement that the mean equals 195 lb, so we let $H_0$ be $\mu = 195$. 
Forming Your Own Hypotheses

If you are conducting a study and want to use a hypothesis test to support your claim, the claim must be worded so that it becomes the alternative hypothesis.
Mechanism of hypothesis test

If, under the null hypothesis, the probability of a particular observed event is exceptionally small, we conclude that the null hypothesis is probably not correct, and it should be rejected in favor of the alternative hypothesis.
Conclusions in Hypothesis Testing

We always test the null hypothesis. The initial conclusion will always be one of the following:

1. Reject the null hypothesis.
2. Fail to reject the null hypothesis.
Example:
Using the weights of the simple random sample of men, we obtain these sample statistics: $n = 40$ and mean $= 172.55$ lb, and $s = 26.33$ lb. Use these results to test the claim that men have a mean weight greater than 166.3 lb, which was the weight in the National Transportation and Safety Board’s recommendation. Use a 0.05 significance level, and the traditional method for the test.
Example:
Requirements are satisfied: simple random sample, $\sigma$ is not known, sample size is 40 ($n > 30$)

Step 1: Express claim as $\mu > 166.3$

Step 2: The opposite to claim is $\mu \leq 166.3$

Step 3: $\mu > 166.3$ does not contain equality, it is the alternative hypothesis:

$H_0: \mu = 166.3$ is null hypothesis

$H_1: \mu > 166.3$ is alternative hypothesis and original claim
Example:

Step 4: significance level is $\alpha = 0.05$

Step 5: claim is about the population mean, so the relevant statistic is the sample mean, 172.55 lb

Step 6: calculate $t$

$$ t = \frac{\bar{x} - \mu}{s} = \frac{172.55 - 166.3}{\frac{26.33}{\sqrt{40}}} = 1.501 $$

$df = n - 1 = 39$, area of 0.05, one-tail yields $t = 1.685$;
Example:

Step 7: \( t = 1.501 \) does not fall in the critical region bounded by \( t = 1.685 \), we fail to reject the null hypothesis.

\[ \mu = 166.3 \]

or

\[ z = 0 \]

\[ \bar{x} = 172.55 \]

or

\[ t = 1.5 \]

Critical value \( t = 1.685 \)
Conclusion:

Because we fail to reject the null hypothesis, we conclude that there is not sufficient evidence to support a conclusion that the population mean is greater than 166.3 lb, as in the National Transportation and Safety Board’s recommendation.
Critical Region

The critical region (or rejection region) is the set of all values of the test statistic that cause us to reject the null hypothesis. For example, see the red-shaded region in the previous figure.
Significance Level

The \textit{significance level} (denoted by $\alpha$) is the probability that the test statistic will fall in the critical region when the null hypothesis is actually true. This is the same $\alpha$ introduced in confidence intervals. Common choices for $\alpha$ are 0.05, 0.01, and 0.10.
Critical Value

A critical value is any value that separates the critical region (where we reject the null hypothesis) from the values of the test statistic that do not lead to rejection of the null hypothesis. The critical values depend on the nature of the null hypothesis, the sampling distribution that applies, and the significance level \( \alpha \). See the previous figure where the critical value of \( z = 1.645 \) corresponds to a significance level of \( \alpha = 0.05 \).
**P-Value**

The null hypothesis is rejected if the *P*-value is very small, such as 0.05 or less.

Here is a memory tool useful for interpreting the *P*-value:

- If the *P* is low, the null must go.
- If the *P* is high, the null will be plausible.
Two-tailed Test

\[ H_0: = \]
\[ H_1: \neq \]

is divided equally between the two tails of the critical region

Means less than or greater than

Reject \( H_0 \)  
Fail to reject \( H_0 \)  
Reject \( H_0 \)

Sign used in \( H_1: \neq \)
Left-tailed Test

\[ H_0: = \]
\[ H_1: < \]

Points Left

Sign used in \( H_1: < \)

the left tail
Right-tailed Test

\[ H_0: = \]
\[ H_1: > \]

Fail to reject \( H_0 \)

Reject \( H_0 \)

Sign used in \( H_1: > \)
Types of Hypothesis Tests: Two-tailed, Left-tailed, Right-tailed

The tails in a distribution are the extreme regions bounded by critical values.

Determinations of $P$-values and critical values are affected by whether a critical region is in two tails, the left tail, or the right tail. It therefore becomes important to correctly characterize a hypothesis test as two-tailed, left-tailed, or right-tailed.
Procedure for Finding \( P \)-Values

1. Start
2. What type of test?
   - Left-tailed
   - Right-tailed
   - Two-tailed
3. Is the test statistic to the right or left of the center?
   - Left
   - Right
4. \( P \)-value = area to the left of the test statistic
5. \( P \)-value = twice the area to the left of the test statistic
6. \( P \)-value = twice the area to the right of the test statistic
7. \( P \)-value = area to the right of the test statistic

Test statistic

\( P \)-value

\( P \)-value is twice this area.
**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.

**Critical region in the left tail:**

\[ P\text{-value} = \text{area to the left of the test statistic} \]

**Critical region in the right tail:**

\[ P\text{-value} = \text{area to the right of the test statistic} \]

**Critical region in two tails:**

\[ P\text{-value} = \text{twice the area in the tail beyond the test statistic} \]
Decision Criterion

$P$-value method:
Using the significance level $\alpha$:

If $P$-value $\leq \alpha$, reject $H_0$.

If $P$-value $> \alpha$, fail to reject $H_0$. 
Accept Versus Fail to Reject

• Some texts use “accept the null hypothesis.”

• We are not proving the null hypothesis.

• Fail to reject says more correctly

• The available evidence is not strong enough to warrant rejection of the null hypothesis (such as not enough evidence to convict a suspect).
Wording of Final Conclusion

Start

Does the original claim contain the condition of equality?

Yes
(Original claim contains equality)

No
(Original claim does not contain equality and becomes $H_1$)

Do you reject $H_0$?

Yes
(Reject $H_0$)

No
(Fail to reject $H_0$)

Do you reject $H_0$?

Yes
(Reject $H_0$)

No
(Fail to reject $H_0$)

Wording of final conclusion

“There is sufficient evidence to warrant rejection of the claim that . . . (original claim):”

(This is the only case in which the original claim is rejected.)

“There is not sufficient evidence to warrant rejection of the claim that . . . (original claim):”

(This is the only case in which the original claim is supported.)

“The sample data support the claim that . . . (original claim):”

“There is not sufficient sample evidence to support the claim that . . . (original claim):”
Caution

Never conclude a hypothesis test with a statement of “reject the null hypothesis” or “fail to reject the null hypothesis.” Always make sense of the conclusion with a statement that uses simple nontechnical wording that addresses the original claim.
Type I Error

- A Type I error is the mistake of rejecting the null hypothesis when it is actually true.
- The symbol $\alpha$ (alpha) is used to represent the probability of a type I error.
Type II Error

- A **Type II error** is the mistake of failing to reject the null hypothesis when it is actually false.
- The symbol $\beta$ (beta) is used to represent the probability of a type II error.
## Type I and Type II Errors

<table>
<thead>
<tr>
<th>Decision</th>
<th>True State of Nature</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>The null hypothesis is true</td>
</tr>
<tr>
<td></td>
<td>The null hypothesis is false</td>
</tr>
<tr>
<td>We decide to reject the null hypothesis</td>
<td><strong>Type I error</strong> (rejecting a true null hypothesis) $P(\text{type I error}) = \alpha$</td>
</tr>
<tr>
<td>We fail to reject the null hypothesis</td>
<td><strong>Correct decision</strong></td>
</tr>
</tbody>
</table>
Controlling Type I and Type II Errors

• For any fixed $\alpha$, an increase in the sample size $n$ will cause a decrease in $\beta$.

• For any fixed sample size $n$, a decrease in $\alpha$ will cause an increase in $\beta$. Conversely, an increase in $\alpha$ will cause a decrease in $\beta$.

• To decrease both $\alpha$ and $\beta$, increase the sample size.
Power of a test

The power of a hypothesis test is the probability \((1 - \beta)\) of rejecting a false null hypothesis. The value of the power is computed by using a particular significance level \(\alpha\) and a particular value of the population parameter that is an alternative to the value assumed true in the null hypothesis.

That is, the power of the hypothesis test is the probability of supporting an alternative hypothesis that is true.
Power and the Design of Experiments

Just as 0.05 is a common choice for a significance level, a power of at least 0.80 is a common requirement for determining that a hypothesis test is effective. (Some statisticians argue that the power should be higher, such as 0.85 or 0.90.) When designing an experiment, we might consider how much of a difference between the claimed value of a parameter and its true value is an important amount of difference. When designing an experiment, a goal of having a power value of at least 0.80 can often be used to determine the minimum required sample size.
Confidence Interval with Hypothesis Test

A confidence interval estimate of a population parameter contains the likely values of that parameter. We should therefore reject a claim that the population parameter has a value that is not included in the confidence interval.

<table>
<thead>
<tr>
<th>Table 8-2</th>
<th>Confidence Level for Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Two-Tailed Test</td>
</tr>
<tr>
<td>Significance</td>
<td>0.01</td>
</tr>
<tr>
<td>Level for Hypothesis Test</td>
<td>0.05</td>
</tr>
<tr>
<td>Test</td>
<td>0.10</td>
</tr>
</tbody>
</table>
Summary: Decision Criterion

Traditional method:

If the test statistic falls within the critical region, reject $H_0$ in favor of $H_1$.

If the test statistic does not fall within the critical region, fail to reject $H_0$. 
Another option:

Instead of using a significance level such as 0.05, simply identify the $P$-value and leave the decision to the reader.
Summary: Decision Criterion

Confidence Intervals:

A confidence interval estimate of a population parameter contains the likely values of that parameter. If a confidence interval does not include a claimed value of a population parameter, reject that claim.
ANOVA

- Analysis of variance (ANOVA) is a method for testing the hypothesis that three or more population means are equal.

- For example:

  $H_0: \mu_1 = \mu_2 = \mu_3 = \ldots \mu_k$

  $H_1$: At least one mean is different
Example:

Use the chest deceleration measurements listed in Table 12-1 and a significance level of $\alpha = 0.05$ to test the claim that the three samples come from populations with means that are all equal.

<table>
<thead>
<tr>
<th>Table 12-1</th>
<th>Chest Deceleration Measurements (in g) from Car Crash Tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small Cars</td>
<td>44 43 44 54 38 43 42 45 44 50 $\Rightarrow \bar{x} = 44.7$ g</td>
</tr>
<tr>
<td>Medium Cars</td>
<td>41 49 43 41 47 42 37 43 44 34 $\Rightarrow \bar{x} = 42.1$ g</td>
</tr>
<tr>
<td>Large Cars</td>
<td>32 37 38 45 37 33 38 45 43 42 $\Rightarrow \bar{x} = 39.0$ g</td>
</tr>
</tbody>
</table>
ANOVA
Fundamental Concepts

1. The variance **between** samples (also called variation due to treatment) is an estimate of the common population variance $\sigma^2$ that is based on the variability among the sample means.

2. The variance **within** samples (also called variation due to error) is an estimate of the common population variance $\sigma^2$ based on the sample variances.
Key Components of Analysis of Variance

$SS(\text{total})$, or total sum of squares, is a measure of the total variation (around $\bar{x}$) in all the sample data combined.

$$SS(\text{total}) = \sum (x - \bar{x})^2$$
Key Components of Analysis of Variance

SS(treatment/group), also referred to as Sum of Squares between groups, is a measure of the variation between the sample means.

\[
SS(\text{treatment}) = n_1 (\bar{x}_1 - \bar{x})^2 + n_2 (\bar{x}_2 - \bar{x})^2 + \cdots + n_k (\bar{x}_k - \bar{x})^2 \\
= \sum n_i (x_i - \bar{x})^2
\]
Key Components of Analysis of Variance

SS(error), also referred to as Sum of Squares within groups, is a sum of squares representing the variability that is assumed to be common to all the populations being considered.

\[
SS(\text{error}) = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \cdots + (n_k - 1)s_k^2
\]

\[
= \sum (n_i - 1)s_i^2
\]
Key Components of Analysis of Variance

Given the previous expressions for \( SS(\text{total}) \), \( SS(\text{treatment}) \), and \( SS(\text{error}) \), the following relationship will always hold.

\[
SS(\text{total}) = SS(\text{treatment}) + SS(\text{error})
\]
Test Statistic for One-Way ANOVA

\[ F = \frac{\text{variance between samples}}{\text{variance within samples}} \]

Degree of freedom with overall sample size \( n \) among \( k \) groups:

- numerator \( df = k - 1 \)
- denominator \( df = n - k \)
**F-Distribution**

1. The *F*-distribution is not symmetric; it is skewed to the right.
2. The values of *F* can be 0 or positive; they cannot be negative.
3. There is a different *F*-distribution for each pair of degrees of freedom for the numerator and denominator.
**F-Distribution**

Not symmetric (skewed to the right)

Nonnegative values only
Hypothesis test

An Approach to Understanding ANOVA

1. Understand that a small \( P \)-value (such as 0.05 or less) leads to rejection of the null hypothesis of equal means.
   With a large \( P \)-value (such as greater than 0.05), fail to reject the null hypothesis of equal means.

2. Develop an understanding of the underlying rationale by studying the examples in this section.
Relationship Between $F$ Test Statistic / $P$-Value

Sample means are all close.

Small $F$ test statistic, large $P$-value

Large $F$ test statistic, small $P$-value

Compare the sample means.

At least one sample mean is very different.

Fail to reject equality of population means.

Reject equality of population means.
Procedure for testing

\[ H_0: \mu_1 = \mu_2 = \mu_3 = \ldots \]

If the \( P \)-value \( \leq \alpha \), reject the null hypothesis of equal means and conclude that at least one of the population means is different from the others.

If the \( P \)-value \( > \alpha \), fail to reject the null hypothesis of equal means.
Example (continued):

Requirements are satisfied: distributions are approximately normal (normal quantile plots); population variances appear to be about the same; simple random samples; independent samples, not matched; categorized according to a single factor of size

\[ H_0: \mu_1 = \mu_2 = \mu_3 \]

\[ H_1: \text{At least one of the means is different from the others} \]

significance level is \( \alpha = 0.05 \)
Example:

Step 1: Display $P$-value = 0.028 when rounded

Step 2: Because the $P$-value of 0.028 is less than the significance level of $\alpha = 0.05$, we reject the null hypothesis of equal means.

There is sufficient evidence to warrant rejection of the claim that the three samples come from populations with means that are all equal.
Conclusion:

Based on the samples of measurements listed in Table 12-1, we conclude that those values come from populations having means that are not all the same. On the basis of this ANOVA test, we cannot conclude that any particular mean is different from the others, but we can informally note that the sample mean is smallest for the large cars. Because small measurements correspond to less trauma experienced by the crash test dummies, it appears that the large cars are safest, but this conclusion is not formally justified by this ANOVA test.
Caution

When we conclude that there is sufficient evidence to reject the claim of equal population means, we cannot conclude from ANOVA that any particular mean is different from the others.

There are several other tests that can be used to identify the specific means that are different, and some of them are discussed later.
Identifying Means That Are Different

After conducting an analysis of variance test, we might conclude that there is sufficient evidence to reject a claim of equal population means, but we cannot conclude from ANOVA that any particular mean is different from the others.
Identifying Means That Are Different

Informal methods for comparing means

1. Use the same scale for constructing boxplots of the data sets to see if one or more of the data sets are very different from the others.

2. Construct confidence interval estimates of the means from the data sets, then compare those confidence intervals to see if one or more of them do not overlap with the others.
Example (continued):

Using the data in Table 12-1, we concluded that there is sufficient evidence to warrant rejection of the claim of equal means. Use the Bonferroni C.I. for the difference with a 0.05 confidence level to identify which mean is different from the others.

The Bonferroni CI is equivalent to give a separate test for each different possible pair of samples. Here are the null hypotheses to be tested:

\[ H_0 : \mu_1 = \mu_2 \quad H_0 : \mu_1 = \mu_3 \quad H_0 : \mu_2 = \mu_3 \]
Conclusion:

The result shows that the mean for Sample 1 (small cars) is significantly different from the mean for Sample 3 (large cars). Based on the Bonferroni CI (-1.9, 7.1) for the difference $\mu_1 - \mu_2$, (1.2, 10.2) for $\mu_1 - \mu_3$, (-1.4, 7.6) for $\mu_2 - \mu_3$, it appears that the measurements from small cars have a mean that is significantly different from the mean for large cars.