Tests of significance involve testing a claim about a population parameter, such as the mean or proportion. You can use your knowledge of the sampling distribution of the point estimate to test a claim that was made about the population.

Remember, you cannot gather information from *every* member of the **population**, so you collect information about a portion of the population (a **sample**). If you collected your data appropriately (**SRS**) and the appropriate conditions are satisfied, then the sample should adequately represent the population.

**EXAMPLE SCENARIOS**

The operating guidelines for a cooling system specify that the water temperature be maintained at approximately 70°F with a standard deviation of approximately 0.3°F. Records of prior temperature readings have indicated that the water temperature has an approximate normal distribution. However, a technician suspects that the water temperature is significantly higher than 70°F and decides to take 9 temperature readings over a 3-hour period. The sample mean temperature is 70.4°F. Do the data support the technician’s suspicion that the water temperature is significantly higher than 70°F … OR, can the difference between 70.4°F and 70°F simply be explained by random variation?

The birth weight of babies born full term has an (unknown) distribution with a mean of 120 ounces. A public health official makes the claim that the birth weight of babies born in a certain community is significantly less than 120 ounces. To gain support for the claim, a random sample of 49 babies from the community is selected. The sample mean birth weight was 118 ounces with a standard deviation of 21 ounces. Does the statistic support the claim that the true average birth weight of babies born in the community is significantly less than 120 ounces … OR, can the difference between 118 ounces and 120 ounces simply be explained by random variation?

In a 2012 study, Ogden, et. al., found that 17% of United States youths are obese. Another researcher claims that the percent of obese youths in Pennsylvania is significantly lower than 17% and selects a random sample of 1411 youths from across the state to gain statistical support for his claim. The researcher found 219 obese youths in the sample. Do the data support the researcher’s claim that the percent of obese youths in Pennsylvania is significantly lower than 17% … OR, can the difference between 17% and the sample proportion of  simply be explained by random variation?

What is a “normal” body temperature? For decades it was thought that the normal body temperature was 98.6°F. This number was converted to Fahrenheit from a study in Germany, which reported normal at 37°C. However, this number was an average rounded to the nearest degree. The fact that 98.6°F is measured to the tenth degree is simply because of the conversion from Fahrenheit to Celsius. In other words, temperature readings were really only accurate to two significant digits, not the three digits we have with 98.6°F. Data was gathered to determine whether the true average body temperature is significantly different from 98.6°F.

Each scenario is an example of a **test of significance**. The purpose of a test of significance is to evaluate a claim about the value of a population parameter: a population mean μ or a population proportion *p*.

**THE BIG IDEA OF TESTING A HYPOTHESIS**

A hypothesis test is a standard procedure for testing some claim about a property in a population.

**Hypothesis**

A statement or a claim about one or more populations that challenges a given assumption

**Rare Event Rule for Inferential Statistics**

If under a given assumption, the probability of a particular observed event is extremely small, we conclude that this assumption is probably not correct, i.e., we reject explanations when they are based on extremely small probabilities.

**Flying Monkey Example: The Hypothesis**

***Claim:*** The Labor Day 5K Classic and the Flying Money 5K are both 5K courses, however, some members of the Washington County Road Runners Club (WCRRC) claim that the Flying Monkey 5K (FM5K) is a faster course.

***Hypothesis:*** The Flying Monkey 5K (FM) is a faster race than the Labor Day 5K Classic (LDC).

The given assumption is that the mean difference between race times is zero, i.e., the Flying Monkey is not a faster race. Under this assumption, if we sample data and determine that the probability of collecting that sample is very small, we reject the null hypothesis and conclude that the mean difference in race times is different. If that difference is positive, the Flying Monkey can be considered a faster race.

**COMPONENTS OF A FORMAL HYPOTHESIS TEST**

1. **Null and Alternative Hypotheses**
2. **Test Statistic from the Data**
3. **Critical Region**
4. **Decision and In-context Conclusion**

**NULL AND ALTERNATIVE HYPOTHESES**

**Null Hypothesis**, H0

The **null hypothesis** is the hypothesis that is tested directly. The null hypothesis is assumed to be true unless the data supports the rejection of the null hypothesis.

* By definition, the null hypothesis contains a statement of equality, i.e., =, ≤, or ≥.
* We make a decision to either “reject H0” or “fail to reject H0.”

**Alternative Hypothesis**, HA

The **alternative hypothesis** is a statement of what we believe to be true if our sample data cause us to reject H0.

* The alternative hypothesis is the complement of H0, i.e., it’s opposite

By definition, the alternative hypothesis contains a statement of inequality, i.e., ≠, <, or >.

**Flying Monkey Example: The Null and Alternative Hypotheses**

Let μD represent the mean difference in race times: Labor Day 5K Classic minus Flying Monkey times.

***Alternative Hypothesis:***Ha: μD > 0

If the difference is greater than zero, the average time to complete the Flying Monkey is less than the time to complete the Labor Day 5K Classic, and the Flying Monkey can be considered a faster race.

***Null Hypothesis:*** Ho: μD ≤ 0

If the difference is not greater than zero, i.e., if it is less than or equal to zero, then the average time to complete each race is either the same or the average time to complete the Flying Monkey is greater than that of the Labor Day 5K Classic. Either way, the Flying Monkey is considered not to be a faster race.

**Analogy**

Consider the way our judicial system works. A defendant is brought before a jury of his peers on some criminal charge. It is up to the jury to decide whether the defendant is innocent or guilty. The jury is instructed to **assume the defendant is innocent** (null hypothesis) until the **evidence** (data from the sample) suggests otherwise. The jury is then presented with the evidence, some of which supports the defendant’s guilt. The jury weighs this evidence and makes a conclusion on the basis of the evidence: guilty or not guilty.

|  |  |  |
| --- | --- | --- |
| True State    Jury’s Decision | The Defendant is **innocent** | The Defendant is **guilty** |
| The Defendant is **guilty**. |  |  |
| The Defendant is **not guilty**. |  |  |

If the defendant is truly innocent and the jury finds the defendant not guilty, then the jury has made a correct decision. Similarly, the jury is also correct in handing down a guilty verdict to a defendant who is in fact guilty. However, there are two ways that the jury can be in error. The first type of error is the situation in which the jury convicts an innocent defendant (in other words, the jury concludes that the defendant is guilty when in fact the defendant is really innocent). The second type of error occurs when a jury acquits a guilty defendant (in other words, the jury concludes that the defendant is not guilty when the defendant is really guilty). In our society, an error of the first type is considered much worse than an error of the second type.

The theory of hypothesis testing is very much like our judicial system. We (the jury) must **make a decision as to the validity of a claim** **about an unknown population parameter** (like accusations against an alleged criminal). This claim is called the research or alternative hypothesis. At this time, we also introduce the opposite claim called the null hypothesis. **Like the jury, we proceed under the assumption that the null hypothesis is true until such time as our statistical evidence suggests otherwise.** The evidence we are presented with are statistics; the point estimate of the population parameter and the accompanying standard error. From these values, we compute a test statistic and reach a final conclusion as to whether or not the test supports our claim of the alternative hypothesis. We will conclude either to “reject the null hypothesis” (analogous to “guilty”) or to “fail to reject the null hypothesis” (analogous to “not guilty”).

|  |  |  |
| --- | --- | --- |
| True State    Statistician’s Decision | **H0 is true** | **HA is true** |
| **Reject the Null Hypothesis**  **(The data support the alternative hypothesis)** |  |  |
| **Fail to Reject the Null Hypothesis**  **(The data do not support the alternative hypothesis)** |  |  |

**CONSTRUCT NULL AND ALTERATIVE HYPOTHESES**

1. All hypotheses are written symbolically in terms of population parameters, i.e., we use μ or p NOT .
2. Write the original claim symbolically and then write it’s opposite symbolically.
3. Label the claims as null or alternative. The null hypothesis is the one with the equality, =, ≤, or ≥.

**Example 1**

The operating guidelines for a cooling system specify that the water temperature be maintained at approximately 70°F with a standard deviation of 0.3°F. A technician suspects that the water temperature is significantly higher than 70°F. Set-up the appropriate test of hypothesis using the appropriate notation.

***HO:***  = 70°F

***HA:***  > 70°F

**Example 2**

A public health official makes the claim that the birth weight of babies born in a certain community is significantly less than the reported mean of 120 ounces. To gain support for the claim, a random sample of 49 babies from the community is selected. The sample mean birth weight was 118 ounces with a standard deviation of 21 ounces. Set-up the appropriate test of hypothesis using the appropriate notation.

***HO:*** μ < 120 oz.

***HA:***  ≥ 120 oz.

**Example 3**

In a 2012 study, Ogden, et. al., found that 17% of United States youths are obese. Another researcher claims that the percent of obese youths in Pennsylvania is significantly lower than 17%. Set-up the appropriate test of hypothesis using the appropriate notation.

***HO:*** *p* = 0.17

***HA:*** *p* < 0.17

**Example 4**

What is a “normal” body temperature? Data was gathered to determine whether the true average body temperature is significantly different from 98.6°F. Set-up the appropriate test of hypothesis using the appropriate notation.

***HO:***  = 98.6 °F

***HA:***  ≠ 98.6 °F

**TEST STATISTIC FROM THE DATA**

To determine if we have enough evidence to reject the null hypothesis, we compute a **test statistic** that tells us where our sample would fall in the sampling distribution of the statistic if the null hypothesis were true. When the sampling distributions of population means and proportions are normally distributed, we can determine the probability that we would pull a sample with a statistic as extreme or more extreme than the one we pulled. If the probability of pulling a sample is small, we reject the null hypothesis. We decide how small the probability has to be before we will reject the null hypothesis, which is a process similar to choosing the level of confidence for confidence intervals and explained in the Critical Region section next.

**Test Statistic**

A value computed from your sample data that is used in making a decision to reject the null hypothesis.

|  |  |  |
| --- | --- | --- |
| **Mean with σ known** | **Mean with σ unknown** | **Proportion** |
|  |  |  |

**Flying Monkey Example: Test Statistic from the Data**

The mean difference in 50 race times, between the Flying Monkey 5K and the Labor Day 5K Classic is 61.336 seconds with a standard deviation, s, of 18.4295 seconds. The null hypothesis states that there is no difference in the mean race times so the population mean difference is zero,

= = 23.5

**CRITICAL REGION**

Also referred to as the rejection region, the **critical region** is the set of all test statistic values that would cause us to reject the null hypothesis. The threshold or boundary for the critical region is a **critical value**, z\* or t\*, that is based on a chosen **level of significance**.

**Level of Significance**, α

The level of significance, α (pronounced “alpha”), is the probability of rejecting a true null hypothesis (Type I error). The levels of significance correspond to the levels of confidence for confidence intervals—90%, 95%, 99%—but α is the compliment of these so α=0.01 for 99%, 0.05 for 95%, and 0.1 for 90%. Essentially, α defines how rare a result, measured by the sample **test statistic**, must be before you reject the null hypothesis.

**Statistically significant at α**

If the test statistic falls within the critical region, we say that the results of a study are **statistically significant** at α. In that case, we **reject the null hypothesis** H0 and conclude that there is convincing evidence in favor of the alternative hypothesis HA. If the test statistic falls outside the critical region, then we cannot conclude that there is convincing evidence in favor of the alternative hypothesis HA, and we **fail to reject the null hypothesis**.

**p-value**

A **p-value** (or probability-value) of the test is the probability, assuming H0 is true, that the statistic (such as or) takes a value as extreme as or more extreme than the one actually observed, in the direction specified by HA. Small P-values provide support for HA because they suggest that the observed statistic is unlikely to occur when H0 is true. Large P-values fail to provide support for HA because they suggest that the observed result is likely to occur by chance alone when H0 is true.

**Decision Rules**

|  |  |  |
| --- | --- | --- |
|  | **Reject the null hypothesis** | **Fail to reject the null hypothesis** |
| Test statistic, t, and critical value, t\* | |z| ≥ |z\*| OR |t| ≥ |t\*| | |z| < |z\*| OR |t| ≥ |t\*| |
| p-value and level of significance, α | p-value ≤ α | p-value > α |

**Two-tailed Test**

Use when HA specifies not equal, ≠ (Ex. ***HA:***  ≠ 98.6 °F)

**Right-tailed Test**

Use when HA specifies strictly greater than, > (Ex. ***HA:***  > 70°F)

**Left-tailed Test**

Use when HA specifies strictly less than, < (Ex. ***HA:*** *p* < 0.17)

**Flying Monkey Example: Critical Region, Critical Value, and p-value**

Based on a 0.05 level of significance and a right-tailed test, the **critical region** is the upper 0.05 portion of the t-distribution. The **critical value** that is the threshold for this critical region is t\* = 2.01. The **test statistic** is t = 23.5, which is significantly more extreme than 1.68. The **p-value** for a test statistic of 23.5 is less than 0.0001.

**DECISION AND IN-CONTEXT CONCLUSION**

If the test statistic falls in the critical region, we have enough evidence to **reject the null hypothesis**. If the test statistic does not fall in the critical region, we **fail to reject the null hypothesis**.

**Flying Monkey Example: Decision and In-context Conclusion**

The probability of observing an average difference of 61.336 seconds or more assuming there is no difference in the average running time between the two races is less than 0.0001. Therefore, we reject the null hypothesis. There is enough evidence from the data to conclude that the Flying Monkey 5K is, on average, a faster race than the Labor Day 5K Classic.

**Example 5**

Researchers are interested in the mean age of a certain normally distributed population with an unknown standard deviation. Suppose the researcher selects 10 individuals from this population and finds the mean age of this sample to be 25 with a standard deviation of 4.47 years. Based on this data, can we conclude that the true mean of this population is different from 30 years?

1. Construct the null and alternative hypotheses.
2. Compute the test statistic.
3. Choose the level of significance and define the critical region.
4. Make a decision and write an in-context conclusion.