Hypothesis Testing

Author: John M. Cimbala, Penn State University
Latest revision: 04 May 2022

Introduction

- An important part of statistics is ***hypothesis testing*** – ***making a decision about some hypothesis (reject or accept), based on statistical methods***.
- The four basic steps in any kind of hypothesis testing are:
  - Determine the *null hypothesis* and the *alternative hypothesis*.
  - Collect data and determine a *test statistic*.
  - Determine how *unlikely* the test statistic would be *if the null hypothesis were true*.
  - Make a *decision* based on the test statistic.
- These steps will become clearer when we do some example problems.

Hypotheses [Watch a 3-Minute video by Professor Cimbala about this: **https://youtu.be/ZlBNhpXTdQE**]

- First, we need to clarify the difference between a *null hypothesis* and an *alternative hypothesis*:
  - A ***null hypothesis*** is defined as ***a theory that is being considered or tested***. The null hypothesis has not been *proven* by a test and may or may not be true. In many cases, the null hypothesis represents "nothing is happening," hence the adjective "null." [This is how a ***statistician*** thinks.]
  - An ***alternative hypothesis***, also called a ***research hypothesis***, is defined as ***the complement of the null hypothesis***. The alternative hypothesis is typically the *opposite* of the null hypothesis, and is the theory that the researchers are trying to prove (or disprove). [This is how an ***engineer*** thinks.]
- A common application is in testing the **relationship between two variables**, as in clinical studies of drugs or lifestyles. Here are some examples:
  - Is there a statistically significant relationship between taking an aspirin a day and the risk of heart attack?
    - *Null hypothesis*: There is *no* relationship between taking an aspirin a day and the risk of heart attack.
    - *Alternative hypothesis*: There *is* a relationship between taking an aspirin a day and the risk of heart attack.
  - Is there a statistically significant relationship between drinking one can of beer before driving, and having a car accident?
    - *Null hypothesis*: There is *no* relationship between drinking one can of beer before driving, and having a car accident.
    - *Alternative hypothesis*: There *is* a relationship between drinking one can of beer before driving, and having a car accident.
- Note: A statistically significant correlation between two items does not necessarily imply causation. In other words, two items may be correlated, but by coincidence, not because one item is the cause of the other.

- The χ^2 distribution is typically used for relationship studies; the test statistic is called a ***chi-square statistic***.
- Another common application is to test hypotheses about the **population mean of a variable**.
- In these types of problems, there are *two parts* to the null hypothesis:
  - The **critical value** of a variable under consideration (we set the null hypothesis to this value).
  - The **type** and **side of the tail(s)** of concern in the hypothesis test:
    - **Two-tailed test**: We are concerned about values *less than or greater than* the critical value.
    - **Left-tailed test**: We are concerned about values *less than* the critical value.
    - **Right-tailed test**: We are concerned about values *greater than* the critical value.
  - For any of the above three types, we choose the ***least likely scenario*** as this part of the null hypothesis.

- Here are some examples:
  - Is the year-round average temperature in a room equal to 25.6°C?
    - *Null hypothesis*: First we set the critical value: $\mu_0 = 25.6$°C. Then we consider the type of test. Here we set $\mu = \mu_0$. In other words, the actual population mean is *equal to* the claimed value. The least likely scenario is that the population mean is *exactly* equal to the claimed value.
    - *Alternative hypothesis*: $\mu \neq 25.6$°C, which is the opposite of the null hypothesis. This is the most likely scenario. (The null hypothesis fails (is rejected) if the population mean temperature is either *less than* or *greater than* 25.6°C to some specified confidence level.)
    - Some authors call this a ***two-sided alternative hypothesis***, since *we are interested in values on both sides of the claimed value* (to the *left* of (less than) or *right* of (greater than) the claimed value). The test we perform in this case is called a ***two-sided test*** or a ***two-tailed test***.
  - Can a strut hold a load of *at least* 800.0 kg without breaking?
    - *Null hypothesis*: First we set the critical value: $\mu_0 = 800.0$ kg. Then we consider the type of test. There are two possibilities, depending on the experimental results:
      - If the sample mean $\bar{x}$ is *less than* the critical value ($\bar{x} < \mu_0$), we consider $\mu > \mu_0$ as the null hypothesis. In other words, the null hypothesis is that the actual population mean is *greater than* the critical value. This is the *least likely scenario*, since our measurements show that $\bar{x} < \mu_0$. This is a ***left-tailed test***, based on this choice of the null hypothesis.
      - If the sample mean $\bar{x}$ is *greater than* the critical value ($\bar{x} > \mu_0$), we consider $\mu < \mu_0$ as the null hypothesis. In other words, the null hypothesis is that the actual population mean is *less than* the critical value. This is the *least likely scenario*, since our measurements show that $\bar{x} > \mu_0$. This is a ***right-tailed test***, based on this choice of the null hypothesis.
    - *Alternative hypothesis*: We choose the *opposite* of the null hypothesis, which is the *most likely scenario* based on our experimental observations. In other words,
      - If the null hypothesis is $\mu > \mu_0$, the alternative hypothesis is $\mu < \mu_0$.
      - If the null hypothesis is $\mu < \mu_0$, the alternative hypothesis is $\mu > \mu_0$.
    - In either case, the null hypothesis fails (is rejected) if the probability of it happening is very slim.
    - Some authors call this a ***one-sided alternative hypothesis***, since *we are interested only in values on one side of the claimed value* (either to the *left* or to the *right* of the critical value). The test we perform in this case is called a ***one-sided test***, or a ***one-tailed test***.
    - The side of the alternative hypothesis is *opposite* that of the null hypothesis:
      - For a ***left-tailed null hypothesis test***, we have a ***right-tailed alternative hypothesis test***.
      - For a ***right-tailed null hypothesis test***, we have a ***left-tailed alternative hypothesis test***.
- The student's *t* distribution is typically used for such studies of the population mean, and the test statistic is called the ***experimental t statistic***, or simply the ***t statistic***.

The *p*-value [Watch a 4-Minute video by Professor Cimbala about this: **https://youtu.be/wn8WqvY5zjM**]

- The ***p-value*** (short for probability value) is formally defined as ***the probability of observing a test statistic as extreme (or more extreme) as the one observed if the critical value of the null hypothesis were true***.
- We can think of the *p*-value as the probability of *wrongly* rejecting the null hypothesis, if it is in fact true.
- Another way to think of it: *p*-value is the probability that the null hypothesis *could occur by pure chance*.
- A ***small*** *p*-value means that the null hypothesis is ***unlikely to be true*** – we reject the null hypothesis if the *p*-value is less than some level, which we choose.
- A ***large*** *p*-value means that the null hypothesis is ***more likely to be true*** – we do not reject the null hypothesis if the *p*-value is larger than some level, which we choose.
- In most cases, we choose the *p*-value level as the ***level of significance*** α. [Recall, $\alpha = 1 - c$, where c is the confidence level.] The standard engineering level of significance is $\alpha = 0.05$ (95% confidence level):
  - If $p < 0.05$, we ***reject the null hypothesis*** because it has less than a 5% chance of being true.
  - If $0.05 < p$, we ***cannot reject or accept the null hypothesis*** because it has more than a 5% chance of being true. The ***results are inconclusive – we should conduct more tests***.
- A level of significance of $\alpha = 0.01$ (99% confidence level) is also popular in research studies. In that case, substitute 0.01 for 0.05 in the two criteria above.
- As will be seen when we do some example problems, we need to be very careful about factors of 2 that pop up, depending on whether the test is one-tailed or two-tailed; these factors of 2 can lead to trouble.

Types of errors in hypothesis testing

- There are two types of errors that can occur in hypothesis testing:
  - A ***type I error*** (also called an ***error of the first kind***) occurs when ***the null hypothesis is wrongly rejected***. In other words, the null hypothesis is in fact *true*, but it is rejected erroneously.
  - A ***type II error*** (also called an ***error of the second kind***) occurs when ***the null hypothesis is wrongly accepted***. In other words, the null hypothesis is in fact *false*, but it is *accepted* (not rejected) erroneously.
- In any given hypothesis test, type I and type II errors are inversely related. In other words, the smaller the risk (probability) of a type I error, the greater the risk of a type II error, and vice-versa.
- The probability of making a type I error in a two-tail test is equal to α, the ***level of significance*** chosen for the test. Conversely, the confidence level that a type I error will *not* occur is $c = 1 - \alpha$.
- The probability of making a type II error is given the symbol β. Conversely, the probability that a type II error will *not* occur is $\eta = 1 - \beta$, where η is called the ***power***.
- The type of error of particular concern in engineering is usually determined by safety. **Example**: A strut is claimed to withstand a load of *at least* 800.0 kg before failing. We set the critical value to $\mu_0 = 800.0$ kg and our tests yield $\bar{x} > \mu_0$. We perform a one-tailed hypothesis test, in which we choose the side of the null hypothesis as $\mu < \mu_0$. The null hypothesis is that the strut *cannot* withstand the desired load without failing.
  - Type I error: The strut *cannot* withstand the required load (the null hypothesis *is* correct), but we *reject it erroneously*. This could lead to a catastrophic failure, which is undesirable in terms of safety. *Sound engineering decisions are based on the assurance that this type of error is highly unlikely*.
  - Type II error: The strut *can* withstand the required load (the null hypothesis *is not* correct), but we *accept it erroneously*. This forces the manufacturer to beef up the strut, which may cost more, but is generally *desirable in terms of safety*. [There are other scenarios in which Type II errors are more critical.]

- **Example (Two-Sided Hypothesis Test)** [As engineers, we typically do *not* do this kind of test]:
  Given: A manufacturer claims that the population mean of their resistors is 1.00 kΩ. We test this claim by measuring the resistance (which we call x for convenience) with a multimeter. We do 20 tests ($n = 20$) and calculate the sample mean resistance $\bar{x} = 1.20$ kΩ, and the sample standard deviation $S = 0.30$ kΩ.
  To do: Should we accept or reject the manufacturer's claim, and to what confidence level?
  Solution: This is a ***two-sided test***. For these types of problems in which the hypothesis involves population means and sample means, we use the student's t distribution. There are two ways to solve this problem, and we show both for completeness. The first one is based on methods we learned previously, and the second is based on methods that we learn here for the first time.
  - Method 1: [This is a review of what we previously learned – we do this for comparison only.] We estimate the population mean and its confidence interval using the student's t distribution.
    - Since only *one* measurement is necessary to estimate the mean, df = $n - 1 = 20 - 1 = 19$.
    - For 95% confidence level, $\alpha = 1 - 0.95 = 0.05$.
      - To calculate $t_{\alpha/2}$, we need the value of t such that the area under the PDF between t and ∞ is equal to $\alpha/2 = 0.025$. We use Excel's TINV(α,df) function: $t_{\alpha/2}$ = TINV(0.05,19) = 2.0930.
      - Our estimate for the population mean then becomes
        $$\mu = \bar{x} \pm t_{\alpha/2}\frac{S}{\sqrt{n}} = 1.20 \pm (2.0930)\frac{0.3}{\sqrt{20}} = 1.20 \pm 0.1404 \text{ kΩ}.$$
      - Our prediction to 95% confidence is **population mean resistance = 1.20 ± 0.14 kΩ**. In other words, we are 95% confident that the actual mean resistance of the whole population of resistors lies between 1.06 and 1.34 mm/s. That is, **$1.06 < \mu < 1.34$ kΩ**.
      - Since the manufacturer's claim does *not* lie within our calculated range, we *reject* the manufacturer's claim (with 95% confidence level).
    - We repeat the analysis for other confidence levels for comparison:
      - For 98% confidence level, $\alpha = 1 - 0.98 = 0.02$, and **$1.03 < \mu < 1.37$ kΩ**. Notice that the range of μ for 98% confidence is *wider* than that for 95% confidence. Nevertheless, the manufacturer's claim still does *not* lie within our calculated range, and we must *reject* the claim.
      - Similarly, for 99% confidence level, $\alpha = 1 - 0.99 = 0.01$, and **$1.01 < \mu < 1.39$ kΩ**. The range is wider still. We *reject* the claim.
      - Similarly, for 99.9% confidence level, $\alpha = 1 - 0.999 = 0.001$, and **$0.94 < \mu < 1.46$ kΩ**. This time, the manufacturer's claim *does* lie within our predicted range of μ. We must *accept* the manufacturer's claim because we cannot reject it to this great of a confidence level.

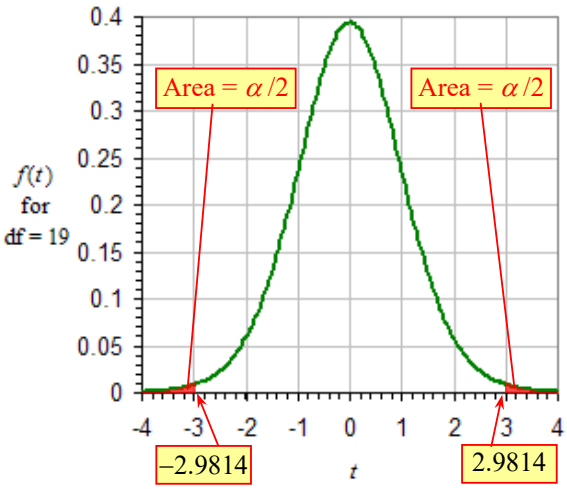
- In conclusion, we are somewhere between 99% and 99.9% confident that the manufacturer's claim is false. If we continue by trial and error, we determine that the critical confidence level is **99.2%**.
- The second method determines this level much more quickly, and without any trial and error.

○ Method 2: [This is new material.] Here is the procedure for determining the *p*-value and the confidence level for rejecting the manufacturer's claim.

- The null hypothesis is stated as follows: The critical value is $\mu_0 = 1.00$ kΩ, and the *least likely scenario* is that the population mean resistance is exactly 1.00 kΩ. **We write the null hypothesis as $\mu = \mu_0$.** The alternative hypothesis is that $\mu \neq \mu_0$.
- We calculate the *t*-statistic at the critical value, based on the definition of *t* in the student's *t* PDF:

$$t = \frac{\bar{x} - \mu_0}{S/\sqrt{n}} = \frac{1.20 - 1.00}{0.30/\sqrt{20}} = 2.9814.$$

- This *t*-statistic is indicated on the plot of the student's *t* PDF for df = 20 – 1 = 19, shown below.
- From the *t*-statistic, we calculate the *p*-value, based on whether this is a one-sided or two-sided test.
  - For a one-sided test, the *p*-value equals *one* of the shaded areas or ***tails*** on the plot.
  - For a two-sided test, the *p*-value is the sum of *both* shaded areas or tails on the plot.
- Since this is a two-sided test, we define the *p*-value as the probability that $|t| > 2.9814$. This represents the area under the student's *t* PDF between $-\infty < t < -t$-statistic *plus* the area under the student's *t* PDF between *t*-statistic $< t < \infty$ (both shaded areas or tails in the plot).



- Fortunately, Excel has a function called **TDIST(ABS(*t*), df, tails)** that calculates this probability, where *t* is the *t*-statistic, df is degrees of freedom, and tails is either 1 for a one-sided test or 2 for a two-sided test. We type "=TDIST(2.9814,19,2)" in a cell, and Excel returns the *p*-value: ***p*-value = 0.00767**.
- If Excel is not available, the *p*-value can be calculated on some hand calculators, and tables of *p*-value are available in statistics books. Several of these tables are also posted on the Exams tab of the website for this course.
- This *p*-value represents a level of significance $\alpha = p$-value = 0.00767, corresponding to a confidence level of $c = 1 - 0.00767 = 0.9923$ or about 99.2%.
- Final result: **We are 99.2% confident that the null hypothesis should be rejected.**

○ Note that the result of Method 2 agrees with that of Method 1, but no iteration is required.

○ The *p*-value is often listed in a report about hypothesis testing. For example, some authors would write, "We reject the manufacturer's claim because *p*-value = 0.0077." Alternately, some authors simply compare their *p*-value to standard levels of significance, e.g., "We reject the manufacturer's claim because *p*-value < 0.05." (95% confidence) or "We reject the manufacturer's claim because *p*-value < 0.01." (99% confidence), without actually reporting the *p*-value.

Discussion: The 99.2% confidence level is a ***critical value*** – we *reject* the null hypothesis for any confidence level *below* this value (e.g., we are more than 95% confident that the null hypothesis is false). We *could not reject* the null hypothesis for any confidence level *above* this value (e.g., we cannot be 99.9% confident that the null hypothesis is false). Exactly the same procedure is used for one-sided tests, except that the variable "tails" in Excel's TDIST function is set to 1 instead of 2. It turns out that the one-sided TDIST value is exactly half of the two-sided TDIST value, since the student's *t* PDF is symmetric.

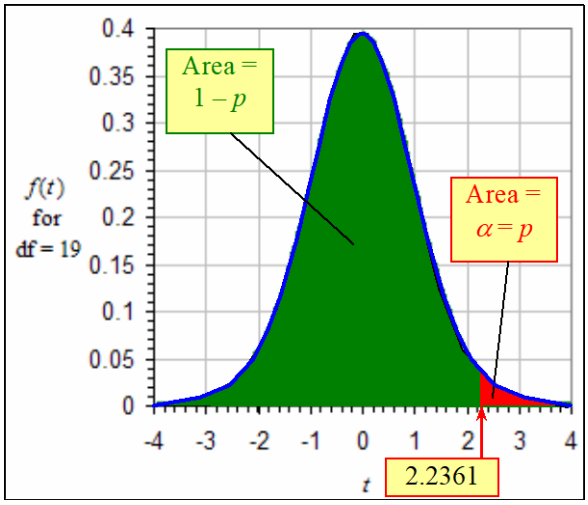
- **Example (One-Sided Hypothesis Test)** [As engineers, this is the kind of test we typically do]:

Given: The same resistors and experiments as those of the previous problem. This time, the manufacturer claims that the population of resistors has an average resistance of *at least* 1.05 kΩ.

To do: To what confidence level can we accept the manufacturer's claim?

Solution: This is a ***one-sided test*** since we are concerned only that the actual population mean is greater than 1.05 kΩ.

- The null hypothesis is stated: The critical value is set to $\mu_0 = 1.05$ kΩ. Since $\bar{x} = 1.20$ kΩ is *greater than* the critical value ($\bar{x} > \mu_0$), this a right-tailed hypothesis test. We set the *side* of the null hypothesis as $\boldsymbol{\mu < \mu_0}$, which is the ***least likely scenario***.
- The alternative hypothesis is: $\mu > \mu_0$.
- The *t*-statistic is $t = \dfrac{\bar{x} - \mu_0}{S/\sqrt{n}} = \dfrac{1.20 - 1.05}{0.30/\sqrt{20}} = 2.2361$. Notice that $\bar{x} > \mu_0$ here, so $t > 0$.
- A sketch of the student's *t* PDF is shown here.



- We type "=TDIST(2.2361,19,1)" in a cell in Excel, and Excel returns the *p*-value: **p-value = 0.01877**.
- The critical level of significance is $\alpha = p$-value = 0.01877, corresponding to a critical confidence level of $c = 1 - 0.01877 = 0.98123$ or about 98.1%.
- Statistically, there is about a 1.9% chance that the null hypothesis ($\mu < \mu_0$) is true, even though our experiments indicate otherwise ($\bar{x} > \mu_0$). Since the null hypothesis is the *opposite* of the manufacturer's claim in this case, **We are 1.9% confident that the manufacturer's claim is false.**
- Alternatively, we use the simpler "foolproof" method introduced in the previous example: Since our experiments show that *the sample mean* $\bar{x}$ *is greater than* μ_0, we know that it is highly unlikely that the *population mean* would turn out to be *less than* μ_0. Therefore, the manufacturer's claim must have a probability *greater than* 50%, so we pick the larger of the two probabilities (1.9% and 98.1%) as the confidence level that the manufacturer's claim is true. So, we state our final result: **We are 98.1% confident that the manufacturer's claim is true.** To standard engineering confidence level (95%), we *accept* the manufacturer's claim, since 98.1% > 95%.
- In summary, we accept the manufacturer's claim to standard engineering confidence level only if we are confident to at least 95%, which *is* the case here. Our conclusion is thus: ***We accept the manufacturer's claim since* 98.1% > 95%**.
- The bottom line from this hypothesis test: **We accept the manufacturer's claim.**
- [If our requirement were more strict (e.g., 99% confidence), we could *not* accept the claim, since the *p*-value is not less than 0.01.]