

Statistics Reference Document Series Part 3: Inferential Statistics

June 2023  
  
Corey Thrush, Contractor

Dr. Gina Sigler, Contractor

DISTRIBUTION STATEMENT A. Approved for public release; distribution is unlimited.   
CLEARED on 28 Jul 2023. Case Number: 88ABW-2023-0709

The STAT COE provides independent STAT consultation to designated acquisition programs and special projects to improve Test & Evaluation (T&E) rigor, effectiveness, & efficiency.

About this Publication:

This work was conducted by the Scientific Test & Analysis Techniques Center of Excellence under contract FA8075-18-D-0002, Task FA8075-21-F-0074.

For more information:­­  
Visit, [www.AFIT.edu/STAT](http://www.AFIT.edu/STAT)  
Email, [CoE@afit.edu](mailto:CoE@afit.edu)  
Call, 937-255-3636 x4736

Copyright Notice: No Rights Reserved

Scientific Test & Analysis Techniques Center of Excellence  
2950 Hobson Way   
Wright-Patterson Air Force Base, Ohio

The views expressed are those of the author(s) and do not necessarily reflect the official policy or position of the Department of the Air Force, the Department of Defense, or the U.S. government.

Version: 1, FY23

Executive Summary

The field of statistics revolves around the study and analysis of data. Data analysis can be further broken down into descriptive statistics and inferential statistics. While descriptive statistics attempt to summarize data, inferential statistics instead make deductions about a population based on a sample. The statistical method used is highly dependent on the type of conclusion that needs to be drawn. Three of the most popular statistical inference methods are introduced: statistical intervals, hypothesis testing, and linear regression. These methods are described at a high level and an appendix is provided with detailed examples for an assortment of the hypothesis tests and a linear regression model.

***Keywords: Estimation, Statistical Intervals, Hypothesis Testing, Linear Regression, Inferential Statistics***

**Table of Contents**

Executive Summary i

Introduction 1

Statistical Intervals 1

Confidence 1

Prediction 1

Tolerance 2

One-Sided 2

Hypothesis Testing 2

Hypotheses 2

Errors 3

Significance Level ( 3

Power (1- ) 3

Test-Statistic 3

P-Value 4

Rejecting or Failing to Reject the Null 5

Types of Tests 6

Non-Parametric Tests 7

Regression 8

Model 8

Residuals 8

R-Squared 8

Conclusion 9

Appendix 12

Z-Test 12

Example of Z-Test 12

P-Value Computation for Z-Test 12

t-Tests 14

Single 14

Example of One Sample t-Test 15

Paired 15

Example of Paired t-Test 15

Two-Sample 16

Example of Two-Sample t-Test 16

Chi-Square 17

F-Test 17

Analysis of Variance (ANOVA) 18

Example of ANOVA 21

Regression 21

R-Squared 22

Example of Regression 22

Introduction

In the first two parts of the Statistics Reference Series descriptive statistics, and probability theory were covered. We now move on to the field of inferential statistics. Inferential Statistics is the process of collecting a sample to infer something from a population. It can be confusing, as there are many ways of analyzing the same thing. This best practice will walk through options of calculating statistical intervals, drawing conclusions with hypothesis tests, and predicting values with linear regression. Mathematical formulas and worked examples of each hypothesis test and linear regression are included in the appendix.

Statistical Intervals

In statistics, we frequently use a single value to estimate a parameter of a distribution called a point estimate. For example, the sample mean estimates the population mean, which is the center of a symmetric distribution. While point estimates have their uses, more often practitioners need to understand the uncertainty of the estimate. Statistical intervals combine uncertainty with the point estimate to give a range of plausible values for the parameter. Three distinct intervals are discussed: confidence, prediction, and tolerance intervals. Even though they are all intervals, the intent of each is different.

Confidence

A popular type of statistical interval is a confidence interval. This type of interval attempts to quantify the precision of a population parameter (such as a mean, standard deviation, proportion, etc.) based on a random sample (Meeker et al., 2017). The most common confidence intervals involve the population mean. For information on confidence intervals specifically for percentiles please see *Confidence Intervals for the Median and Other Percentiles* (Burke, 2016).

Confidence intervals are notoriously difficult to interpret because they use the vague term “confidence”. Confidence is defined in terms of the process used to construct the interval. “For example, a 95% confidence interval means that in the long run 95% of confidence intervals constructed in this manner will contain the true parameter” (Kensler & Cortes, 2014). Intervals can be calculated at any confidence level, but typically in analysis the confidence level will be 95%. Consider a free dataset proved by R, a statistical software, called Mtcars. The dataset includes information about fuel consumption and 10 aspects of automobile design and performance for 32 automobiles from the 1974 *Motor Trend* US magazine. This dataset can be used to answer the question: What was the average miles per gallon (mpg) of cars in 1974? A 95% confidence interval for the mean miles per gallon (mpg) can be constructed with a lower bound of 17.9 and upper bound for 22.3 mpg. An interval is typically written: (17.9, 22.3) mpg. The interpretation for this confidence interval would be, “We are 95% confident that the population mean falls between 17.9 and 22.3 miles per gallon.” It may be tempting to talk about the probability that the population mean falls in the interval, but this would indicate that the population mean wasn’t a fixed number. Instead, there is a 95% chance that a given confidence interval includes the population mean. This is a common challenge of confidence intervals, since the interval is typically constructed based only on a single sample.

Prediction

Another type of statistical interval is a prediction interval. This type of interval is commonly used to estimate the lower and upper bounds of a potential new observation under some prediction equation. Additionally, it can be extended to more than one observation or even a point estimate from a future sample (Meeker, 2017).

Prediction intervals can be computed with the predict() function with in R and “hot spots” (red triangles with contextual drop down) in JMP, both common statistical software tools. These intervals are wider than confidence intervals because they include estimates of uncertainty in both the mean and standard error. Consider the miles per gallon example mentioned previously. A car manufacturer could ask, “If a new car had 80 horsepower, what could the miles per gallon be?” A prediction interval would be able to answer this question. The software gives a 95% prediction interval of (16.5, 31.8) mpg. The prediction interval estimates for a new car with 80 horsepower, the miles per gallon will be between 16.5 and 31.8. Again, 95% in this case refers to confidence, not a probability that a new observation will land in the interval.

Tolerance

The purpose of a tolerance interval is to contain a specified proportion of a population. These are often the widest intervals since they account for uncertainty in the parameters (typically mean and variance) as well as the observations. Further discussion of tolerance intervals can be found in *Tolerance Intervals Demystified* (Splinter et al., 2020).

Typically, requirements specify a percentage of the time a system must meet a metric. A tolerance interval covers a specified proportion of the population for a given confidence level. Consider a requirement that states 90% of the time, filled potato chip bags must weigh between 7.5 and 8.5 ounces, with 95% confidence. Using the specified proportion as 90% and the confidence as 95%, we could collect a sample and calculate a tolerance interval (7.82, 8.18) ounces. We expect 90% of the population weight of filled potato chip bag to be within 7.82 and 8.18 ounces with 95% confidence. Since the computed tolerance interval is inside the requirement bounds, we would have evidence to conclude the requirement is met (at least 90% of bags weigh between 7.5 and 8.5 ounces with 95% confidence).

One-Sided

In the examples above, we covered two-sided intervals, but any interval can be one-sided. One-sided intervals provide an estimate for an upper or lower bound for a parameter or data observation. Consider the requirement: a software system must respond to a cyber-threat within 5 seconds 99.5% of the time. A one-sided tolerance interval should be used as the requirement suggests only a maximum amount of time is considered. If the calculated upper bound is smaller than 5 seconds, then the software satisfies the requirement. The two-sided intervals that we saw earlier will not be the same one-sided upper and lower bounds because the equations differ for each interval. For more details of one-sided tolerance intervals look at *Tolerance Intervals Demystified* (Splinter et al., 2020). For calculations on any type of intervals please see *Statistical Intervals: A Guide for Practitioners and Researchers* (Meeker, 2017).

Hypothesis Testing

Unlike summary statistics and intervals, hypothesis testing allows a formal comparison to a requirement(s) to be conducted. Data sets can be compared to certain standards, each other, or distributions. All of these involve different types of tests, which will be talked about in the next section. However, each hypothesis test has many of the same general parts. A deep dive into each of these topics can be found in *Statistical Hypothesis Testing* (Kensler, 2018).

Hypotheses

To perform any type of hypothesis test, a test goal must first be established. In statistics, there are two hypotheses, the null hypothesis ( and the alternative hypothesis (. The null hypothesis is set to be the status quo or the value that is already believed to be true. The alternative hypothesis is what we wish to prove in the hypothesis test.

For example, if a friend claimed that the average mpg of cars in the 1980’s was less than 20 mpg, a hypothesis test could confirm this theory. Using this claim, the null hypothesis is the mean mpg is equal to 20 . While more complex hypotheses exist, this best practice focuses on simple hypothesis tests where the null will always set a parameter equal to a value. The alternative is the claim that the mean miles per gallon is less than 20 . If the claim had been that the miles per gallon were greater than 20, the alternative hypotheses would have been . Or, if the claim was that the mean miles per gallon was not equal to 20, the alternative hypothesis would have been .

Errors

When making conclusions in hypothesis tests, there are two ways to make an error. A Type I error is made when the null hypothesis is incorrectly rejected when it was true. A Type II error is made when the null hypothesis is not rejected when it was false. The probabilities of making one of these two types of errors are related to one another.

Significance Level (

The Type I error level (significance level, alpha, ) is directly chosen by the tester before the test. The significance level in its simplest form can be thought of as a threshold where if the system under test performs well it is no longer due to a random chance. The probability of committing a Type I error is called the significance level of the test. A smaller is typically better, but an level too small inflates the probability of making a Type II error (beta, ) given the sample size remains the same. is directly related to the confidence level mentioned in the interval section: . This is because confidence and Type I error are both probabilities.

Power (1- )

Power is something that we indirectly choose by modifying sample size, signal to noise ratio, and Type I error. How to calculate power is beyond the scope of this best practice. For more details on the signal to noise ratio refer to *Understanding the Signal to Noise Ratio in Design of Experiments* (Ramert, 2019). Statistical power is the probability of rejecting the null hypothesis when the alternative hypothesis is correct. Power is the complement of :. A minimum power of 80% is typically used in the Department of Defense (DOD), but higher power is almost always better. However, as mentioned before, there is a balance between having high power and low Type I error rate (). When designing a test, you may want to conduct power analysis to investigate minimum sample sizes with an acceptable power. The relationships between the different error probabilities ( and , p-value, and power are shown in Figure 1 below. The null and alternative hypotheses have distributions associated with them. You can see that changing alpha (the dotted vertical line in Figure 1) impacts many other values in the Figure. If alpha is changed from 0.05 to 0.01 (shifting it to the right) you are deeper in the tail of the null distribution allowing for more evidence to reject the null hypothesis. By lowering alpha you are decreasing Type I error (when the null hypothesis is incorrectly rejected when it was true), but increasing Type II error when the null hypothesis is not rejected but it should have been. You can visualize the darker gray error labeled False Negative becoming larger and the lighter gray error labeled False Positive becoming smaller, if the dotted vertical moved to the right, and vice versa when it is moved to the left.

Diagram

Description automatically generated

Figure 1

Summary of the different errors in Hypothesis Testing (Cowan, 2020)

Test-Statistic

A test statistic is a way of describing the extremeness of an observed value compared to the hypothesized value or another sample. Consider a simple form of hypothesis testing: comparing a sample mean to a hypothesized value. In this scenario, the test statistic can be thought of as the number of standard deviations from the center of the null distribution. Figure 2 shows a hypothetical test statistic and its distance from center of the hypothetical distribution. The more extreme a test statistic is, the more evidence there is against the null hypothesis (or status quo). The formula for the test statistic is based on the type of test chosen, but will include an assumption on underlying distribution. Almost any test statistic can be calculated using statistical software such as JMP, R, or even Excel in the Analysis ToolPak.

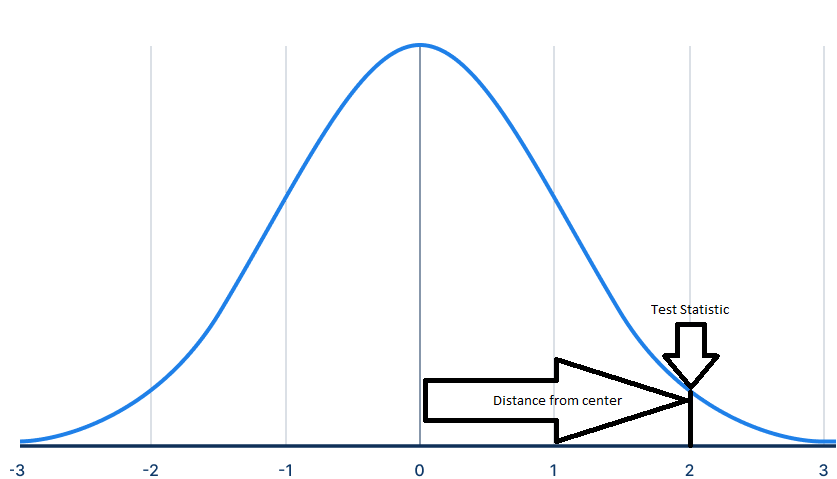


Figure 2

Graphical display of a Test Statistic modified from (Bhandari, 2020)

P-Value

To make conclusions in a hypothesis test we need the help of something called a p-value. From *Design and Analysis of Experiments* (Montgomery, 2017, p. 35)*,* “the p-value is the probability that the test statistic will take on a value that is at least as extreme as the observed value of the statistic when the null hypothesis is true.” If the p-value is smaller (closer to 0), the test statistic is further in the tail of the null distribution and the chance of a null hypothesis being true is low. With a larger p-value (closer to 1) the test statistic is closer to the center of the null distribution and the chance of the null hypothesis being true is increased. If your pre-specified alpha was 0.05, then a p-value of 0.048 and 0.015 will give you the same conclusion, which is to reject the null hypothesis. This resembles a guilty or not guilty ruling where the amount of evidence may make a defendant guilty, but there is no such thing as being “more guilty.” Figure 3 shows an example of performing a hypothesis test where the alternative hypothesis is the mean is greater than X () with a calculated test statistic of 2. The definition of a p-value states “the test statistic will take on a value at least as extreme” which is the red shaded area. The test statistic could have been any value from the line to the right. If we were testing the mean is less than X () the red area would be in the left tail. For when testing that the mean is different than X () the red area would be divided equally in both tails.

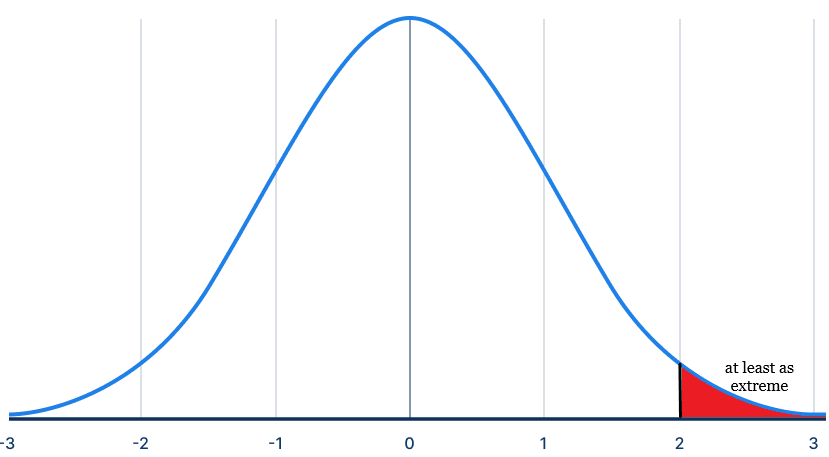


Figure 3

Graphical display of p-value modified from (Bhandari, 2020)

Rejecting or Failing to Reject the Null

If the p-value summarizes the evidence against a null hypothesis, then when do you reject or fail to reject? Decision rules for hypothesis tests are:

* if the p-value is smaller than your specified significance level (alpha), then you reject the null hypothesis
* if the p-value is larger than your specified significance level (alpha), then you fail to reject the null hypothesis

Note that we do not “accept” the null hypothesis. It is always possible to go back and gather more evidence against the null.

Rejecting and failing to reject the null hypothesis can be related back to the confidence interval. Earlier we calculated the 95% confidence interval for the mean miles per gallon in the mtcars data set (Motor Trend, 1974) to be (17.92, 22.26). If there was a claim that the mean is 20 mpg, we would believe the claim because 20 mpg is within our interval. Along with this claim, the p-value would be larger than our alpha of 0.05 and the test statistic would be small (in magnitude), so we would fail to reject the null hypothesis. This would be associated with a small (in magnitude) test statistic. If someone claimed that the mean mpg was 17, we do not believe the claim since 17 is outside of the interval. Additionally, the p-value is smaller than our alpha of 0.05 and the test statistic would be large (in magnitude), so we would reject the null hypothesis.

Types of Tests

Within hypothesis testing, there are a myriad of different tests to choose depending on types of data or test purposes. The ones described in this document are commonly seen. It might be helpful to start with this list of questions to decide which test-statistic to use:

* Do you want to compare center or spread parameters?
* Is there a claim that you want to test against?
* Do you wish to see an effect of treatment?
* Is your data normally distributed?
* Do you know the population variance?

Often, testing will involve a comparison of a distribution center. If you remember from the *Statistics Reference Series Part 1: Descriptive Statistics* (Sigler, 2018)*,* examples of distribution centers include means and medians. Distribution centers are used to summarize a distribution and its characteristics to compare to a hypothesized value or to another distribution. Several helpful hypothesis test methods are a z-test, one sample t-test, paired t-test, and two sample t-test. In more extreme cases, a distribution center can be compared to many others, as is done with an ANOVA test. Other popular tests include those that compare measures of spread to a hypothesized value or each other such as the chi-square test or F test. All these tests have associated assumptions that lead testers to choose one over another. A summary of the tests and their assumptions has been laid out in Table 1. The details as well as examples for all these tests can be found in the Appendix.

Table 1

Table of Types of Hypothesis Tests

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Test** | **Mean/ Variance** | **Claim** | **Distributed** | **Population Variance** |
| Z-test | Mean | Compare to single value | Normal | Known |
| One sample t-test | Mean | Compare to single value | Normal | Unknown |
| Paired t-test | Mean | Compare two samples on a single subject separated by time | Normal | Unknown |
| Two sample t-test | Mean | Compare two samples | Normal | Unknown |
| Chi-square test | Variance | Compare to single value | Chi-squared | NA |
| F test | Variance | Compare two samples | Chi-squared | NA |
| ANOVA test | Mean | Compare many samples | Normal | Unknown |

Notice that in all the tests in Table 1, the distribution of the underlying data is known. In practice, this is not always the case. Details on all these tests and more can be found in *An Introduction to Statistical Methods & Data Analysis* (Ott & Longnecker, 2016)or many other introductory statistics textbooks.

Non-Parametric Tests

Although the test statistics described above are suitable most of the time, there can be constraints or limitations on the data collected. The test statistics we used before all assumed that the data were randomly sampled from a known distribution. If the data that we collect does not have a known distribution or sample size is not large, the median may be a better measure of center for our distribution than the mean. These tests are also appropriate if we do not want to assume normality as was the case in the tests from Table 1. Detailed discussions of non-parametric statistics are beyond the scope of this best practice, but Table 2 shows the non-parametric alternatives for the mentioned parametric tests. Additionally, the *Interactive Inferential Statistics Flowchart* (Burke, 2019) walks through all these inferential statistics options and more. Non-parametric tests can be performed in R or JMP and with limited capability within Excel.

Table 2

Comparison of Parametric and Non-parametric Hypothesis Tests

|  |  |
| --- | --- |
| **Parametric tests of means** | **Non-parametric tests of medians** |
| 1-sample t-test | 1-sample Sign, 1-sample Wilcoxon |
| 2-sample t-test | Mann-Whitney test |
| matched pairs t-test | Signed Rank test |
| One-Way ANOVA | Kruskal Wallis |

Regression

Regression is used to model the relationship between two or more variables. Linear regression creates the line of best fit between a continuous response and a predictor(s). For example, what if you wanted to know the relationship between temperature and ice cream sales? We think that ice cream sales will increase as the temperature increases, but by how much? Linear regression can help answer these questions. Linear regression also requires three assumptions: 1) observations are independent from one another, 2) constant variance (homoscedasticity), and 3) the errors are normally distributed with a mean of zero. For more details on what those assumptions are and how to check them, refer to *Model Building Process Part 1: Checking Model Assumptions* (Burke, 2017). When you use only one predictor this is called simple linear regression (SLR). But linear regression is not restricted to only one predictor. Multiple linear regression (MLR) is when more than one predictor is used. MLR is used for when we wish to know the impact of more than one variable on the response.

Model

The goal is to model a relationship between the response and the predictor(s) that minimizes the distance between the line of best fit to the data observations. Consider the ice cream sales example again. What additional variables might impact ice cream sales other than temperature? We could build a model that used temperature, amount of rain, and number of houses with children in the area. In this instance, we expect each of the three predictors to have a different impact on the ice cream sales. Models may get more complicated by adding functions of the predictor variables or interactions between predictor variables.

Residuals

The linear regression model may not perfectly fit your data as some points are higher or lower than the fitted line. The best fit line is one that minimizes the vertical distance between the line and data points. Figure 4 shows the green line and red line to be examples of residual distance. These distances are balanced between positive and negative values by the model to create the line of best fit; this balance between values is called least squares. In regression, this line is often used to predict the value of the response variable based on the inputs of the predictors.

Chart, line chart, scatter chart

Description automatically generated

Figure 4

Residuals shown graphically (Khan Academy, 2016)

R-Squared

There are a multitude of metrics to see if your model fits the data set well, but one of the most prevalent is R2. Also called the coefficient of determination, R2 is the percentage of the total variation in your data that your model explains. All R2 values are between 0 and 1. A higher R2 means a better fit of the model, but in some situations the threshold for a good R2 will be different. Over-fitting is an issue because the linear model will predict poorly on a new set of data. Details on this and other metrics for model goodness can be found in *Model Building Process Part 3: Model Goodness Metrics* (Burke, 2020).

Conclusion

Statistical inference can be challenging as there are many potential tools to be used. Details of intervals, hypothesis testing, and linear regression were explained along with practical objectives. Intervals provide a range of plausible estimates for parameters, next data observation, and where proportion of population falls. Hypothesis tests allow formal comparisons to requirements, and linear regression provides context about an expected increase or decrease for every unit of a predictor. After reading this best practice readers should understand when these inference methods could be used. To find an appropriate method, we encourage readers to turn to the *Interactive Inferential Statistics Flowchart* (Burke, 2019). Further questions on choosing an analysis method can be directly towards your local STAT Expert.

References

Bhandari, P. (2020, November 5). *The Standard Normal Distribution | Calculator, Examples & Uses*. Scribbr. https://www.scribbr.com/statistics/standard-normal-distribution/

Boston University School of Public Health. (2016, January 6) *SAS – One Sample t-test*.Paired T-Test. Retrieved November 21, 2022, from https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/sas/sas4-onesamplettest/SAS4-OneSampleTtest7.html

Burke, S. (2016). *Confidence Intervals for the Median and Other Percentiles.* STAT COE,

Department of Defense. Air Force Institute of Technology.

Burke, S. (2017). *Model Building Process Part 1: Checking Model Assumptions*. STAT COE,

Department of Defense. Air Force Institute of Technology.

Burke, S. (2019) *Interactive Inferential Statistics Flowchart.* STAT COE, Department of Defense.

Air Force Institute of Technology.

Burke, S. (2020). *Model Building Process Part 3: Model Goodness Metrics*. STAT COE,

Department of Defense. Air Force Institute of Technology.

Cowan, D. (2020). *Statistical Power of a Test*. The Science of Machine Learning. https://www.ml-science.com/statistical-power-of-a-test

Donnelly, R., & Kelley, W. M. (2009). *The Humongous Book of Statistics Problems*. Alpha.

JMP Statistical Discovery. (n.d.). *The Two-Sample t-Test.* Statistics Knowledge Portal.

Retrieved November 21, 2022, from https://www.jmp.com/en\_us/statistics-knowledge-

portal/t-test/two-sample-t-test.html

Khan Academy. (2016). *Introduction to Residuals.* Khan Academy.

https://www.khanacademy.org/math/statistics-probability/describing-relationships-quantitative-data/regression-library/a/introduction-to-residuals

Kensler, J., & Cortez, L. A. (2014). *Interpreting Confidence Intervals*. STAT COE, Department of

Defense. Air Force Institute of Technology.

Kensler, J. (2018). *Statistical Hypothesis Testing.* STAT COE, Department of Defense. Air

Force Institute of Technology.

Meeker, W. Q., Hahn, G. J., & Escobar, L. A. (2017). *Statistical intervals: A Guide for*

*Practitioners and Researchers (*2nd ed.). John Wiley & Sons.

Montgomery, D. C. (2017). *Design and Analysis of Experiments* (10th ed.). John Wiley & Sons.

Motor Trend (1974). *Motor Trend Car Road Tests*[Data set]. Motor Trend

Magazine.  1974 *Motor Trend* US magazine

Natoli, C. (2017). *Understanding Analysis of Variance*. STAT COE, Department of Defense. Air

Force Institute of Technology.

Ott, R. L. & Longnecker, M. (2016). *An Introduction to Statistical Methods & Data Analysis*

*(7th edition).* Cengage Learning.

Sigler, G. (2018). *Statistics Reference Series Part 1: Descriptive Statistics*. STAT

COE, Department of Defense. Air Force Institute of Technology.

Splinter, K., Sigler, G., Harman, M., & Kolsti, K. (2020). *Tolerance Intervals Demystified*. STAT

COE, Department of Defense. Air Force Institute of Technology.

Ramert, A. (2019). *Understanding the Signal to Noise Ratio in Design of Experiments*. STAT

COE, Department of Defense. Air Force Institute of Technology.

Appendix

Details of Hypothesis Tests

Z-Test

To perform a Z-Test, we assume the data is normally distributed. Normality is checked by viewing a histogram of the data or a Q-Q plot. Details on checking normality of your data is contained in *Model Building Process Part 1: Checking Model Assumptions* (Burke, 2017). This test is used to compare the sample mean in your data against a hypothesized mean or claim. Additionally, the variance of the population should be known.

The Z-Test statistic is computed:

where is the sample mean, is the claim you want to test against, is the known population variance, and is the sample size.

Example of Z-Test

You wish to test the claim that the average height of males in your department is greater than the United States average of 5 feet 9 inches with a known population standard deviation of 3 inches. You calculate that the sample mean of heights at your workplace is 71.8 inches.

Begin by setting up hypotheses that support the notion that males in your work department are taller on average than the American male population. The null would be that the mean height of males in your workplace is equal to the mean height of males in the United States. The alternative is your claim of males are taller on average than American males.

We decide to set a significance level of = 0.05. Next, the test statistic is computed from Equation 1 to see how extreme our sample mean is compared to the 69 inches (the mean American male height).

Since the p-value of 0.0183 is less than = 0.05, we have significant evidence to reject (the null). Therefore, we can conclude that mean height of males in your work department is greater than the American mean height.

P-Value Computation for Z-Test

In the previous example the z-score was 2.09, but the p-value was already given. Usually, p-values need to be calculated using a computer or a z-table. Z-tables provide the probability of observing a score as or more extreme than the calculated z-score (the cumulative probability- from the left end to the associated z-score). In Figure 5, the z-scores are around the outside of the table and probabilities are in the middle. To find the associated cumulative probability for a z-score of 2.09 from the normal table you first look for 2.0, which is circled in red on the left side of Figure 5. Next, 0.09 is needed and can be found in the top right of Figure 5 and circled in red. Where these two intersect is at 0.9817, which is also circled. This is the probability of having a z-score less than 2.09. Since the alternative hypothesis from the example above is mean height of males in your workplace is greater than the mean height of males in the United States (, the p-value is calculated by subtracting from one; thus, . If instead the alternative hypothesis was the mean height of males in your workplace was less than the mean height of males in the United States (), then the value of 0.9817 is the correct p-value. This is because we were interested in the left side. These are one-sided tests because they are concerned with less than or greater than. A two-sided test would be where the alternative hypothesis is the mean height of males in your workplace is different than the mean height of males in the United States (). To calculate this p-value is to find the smaller p-value between both one-sided tests and multiply by two. In this case the p-value, would be . The concept of calculating a p-value is the same for all other tests in this best practice, but they are beyond the scope of this best practice. Tables or tools will differ depending on the test used. However, statistical software will calculate the p-value no matter the complexity of the test.

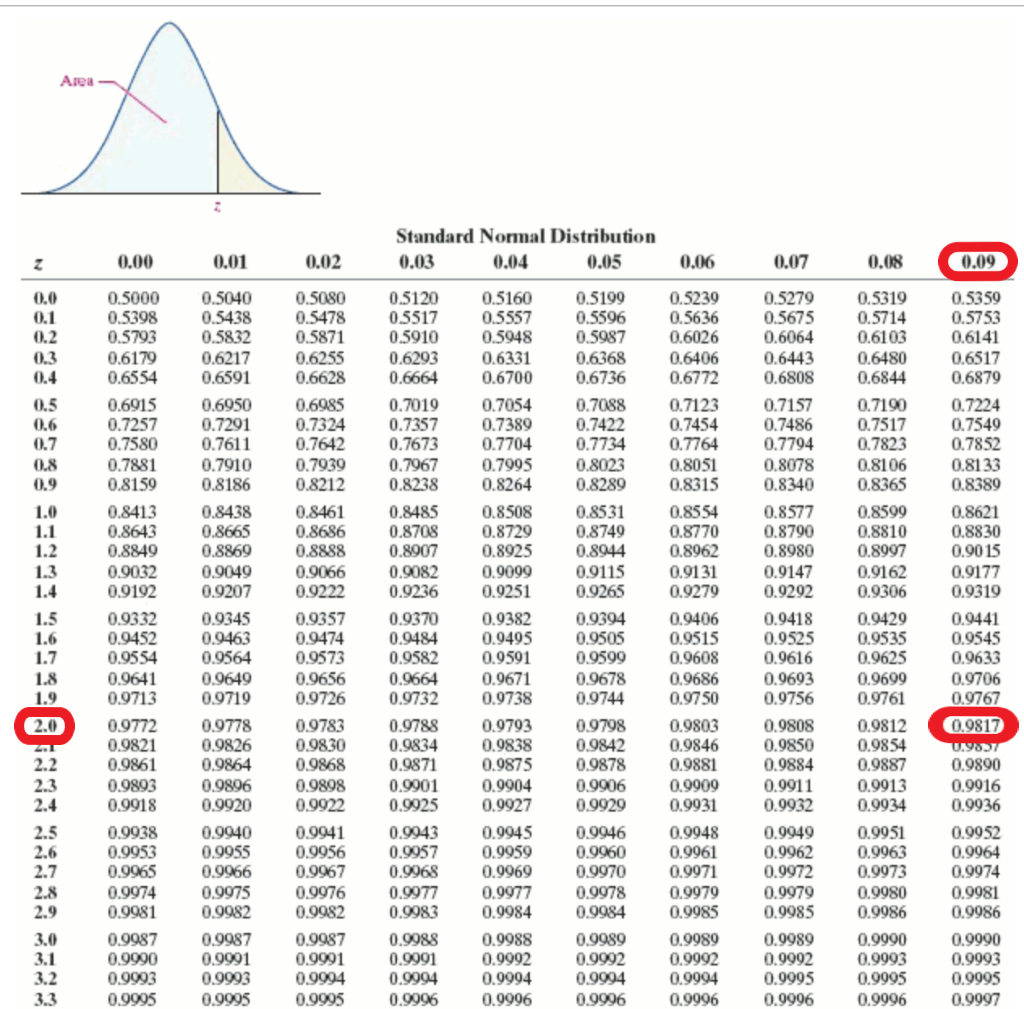


Figure 5

Example of Normal Table for z-score of 2.09

t-Tests

Single

The one sample t-test is not very different from the Z-test. The data is assumed to be normally distributed, but the population variance is unknown, so we estimate the variance from the data. The one sample t-test still conducts a comparison of a sample mean against a claim or hypothesized mean.

The t-statistic is computed:

where is the sample mean, is the claim you want to test against, is the sample standard deviation, and is the sample size.

Example of One Sample t-Test

We can use the same example on the average height of American males. You calculate that the sample mean of heights at your workplace is 71.8 inches. This time we do not know the population standard deviation of heights. Instead, we must use the sample standard deviation, which is 3.2711 inches.

The hypotheses do not change. The null would be that the mean height of males in your workplace is equal to the mean height of males in the United States. The alternative is your claim that males are taller on average than American males.

We decide to set a significance level of = 0.05. Next, the test statistic is computed from Equation 2 to see how extreme our sample mean is compared to 69 inches (the mean American male height).

Since the p-value of 0.064 is greater than = 0.05, we do not have significant evidence to reject the . Therefore, we cannot conclude that mean height of males in your work department is greater than the American male mean height.

Paired

A paired t-test would be used for when we are interested in the difference between two variables for the same subject. Often data are collected pre and post treatment on a particular subject. The variables of interest are typically separated by time. An example of that case would be to compute the pre-pill cholesterol level and post-pill cholesterol level for each person. The normality assumption in this situation is applied to the differences before and after a treatment. To compute the paired t-statistic is the following:

where is the average difference before and after treatment in the data, is the hypothesized difference (often 0 because you are testing against no difference), is the standard deviation of the difference between before and after treatment in the data, and is the sample size.

Example of Paired t-Test

Consider a scenario where you wanted to test whether a pill lowered cholesterol levels over a group of subjects. For the example of a paired t-test, let’s investigate data that was collected on cholesterol levels of men in a study over a duration of 10 years (Boston University School of Public Health, 2016). We are interested in the difference of pre-pill and post-pill cholesterol levels for each subject. We found that after ten years of taking the pill the sample mean of differences is -69.8 and standard deviation is 46.6.

The hypotheses for this case are that we want to test if there is a difference in cholesterol levels after taking a drug. Therefore, our null hypothesis is that there is no mean difference in cholesterol when using the drug. The alternative is the mean difference of cholesterol level is not zero.

We decide to set a significance level of = 0.05. Next, the test statistic is computed from Equation 3 to see how extreme the sample difference is compared to no difference.

Since the p-value < 0.0001 is less than = 0.05, we have significant evidence to reject the . Therefore, we can conclude that the difference in average cholesterol is not 0.

Two-Sample

Consider instead we wish to know if one treatment was different than another. A two-sample t-test should be utilized to compare the means of two different samples. We assume that each sample is normally distributed. However, we need to consider another assumption on whether the variance is the same in both samples. If this assumption is not satisfied, a slightly different test statistic is computed. In the case of assuming equal variance the test statistic is:

where is the sample mean of the first sample, is the sample mean the second sample, and is the pooled variance, which is the weighted average of standard deviation for two or more groups (because equal variance is assumed, standard deviation is pooled). For exact formulas, please see *An Introduction to Statistical Methods & Data Analysis* (Ott, 2016).

Example of Two-Sample t-Test

For the example of a two-sample t-test, let’s investigate body fat percentage by gender. The data used was from JMP Statistics Knowledge Portal (JMP Statistics Knowledge).

We found that the sample mean for men was 14.95 and 22.29 for women with a pooled standard deviation of 6.24.

The hypotheses for this problem are that we want to test to investigate whether women have a higher amount of body fat percentage than men. The null would be that mean body fat percentage for men and women are equal, and the alternative is that the mean body fat percentage is lower for men than women.

We decide to set a significance level of = 0.05. Next, the test statistic from Equation 4 is computed to see how extreme the body fat sample means are between men and women.

Since the p-value = 0.0054 is less than = 0.05, we have significant evidence to reject the . Therefore, we can conclude that men have a lower body fat percentage than women on average.

Chi-Square

The tests discussed before were comparisons of distribution centers to a hypothesized mean or even another distribution. The chi-square test is used to compare a sample variance against a hypothesized variance (or standard deviation) for normally distributed data. Consider an example of where the variability of the strength of a manufactured material is tested against a maximum variability standard. The chi-square test statistic is computed as follows:

where is the sample size, is the sample variance, and is the claim that you are testing against.

**Example of Chi-Square Test**

For this example, from *The Humongous Book of Statistics Problems* (Donnelly, 2009), let’s test whether the variance of customer service wait time is less than a certain amount. A company claims that the standard deviation of their wait times is less than five minutes. Therefore, the hypotheses should be the following:

We found that the sample variance for the 7 randomly sampled wait times is 14.81. A significance level of = 0.05 is chosen. The test statistic is computed from Equation 5 to see how extreme the sample standard deviation was compared to 5.

Since our p-value of 0.2627 is larger than = 0.05, we fail to reject the null . Therefore, we cannot conclude that the standard deviation of customer service wait times are less than 5 minutes.

F-Test

An F-test is like a two-sample t-test where we want to compare estimates of two population parameters of different samples. Instead of means, an F-Test compares variances from two distributions. An example when an F-Test is implemented to compare variances of between two similar but separate normally distributed populations. An F-Test would answer the question, “Do men’s heights have less variability than women’s?” The F-test statistic is computed as a ratio of the sample standard deviations:

where is the sample standard deviation of the first sample and is the sample standard deviation of the second sample. Another piece of information the F-test statistic requires is degrees of freedom (df) for each sample. In an F-test the degrees of freedom is the sample size of each group subtracted by one. In the case of two samples, the degrees of freedom are and . Further explanations of degrees of freedom are beyond the scope of this best practice and statistical software calculates it by default.

**Example of F-Test**

A tire company developed a new tire that should have a more consistent tread life than their existing leading product. The data set in this example comes from *The Humongous Book of Statistics Problems* (Donnelly, 2009). Table 3 shows the summary statistics for tire life.

**Table 3**

*Dataset for Tire Tread Life*

|  |  |  |
| --- | --- | --- |
|  | **Existing** | **New** |
| Sample Standard Deviation | 5325 miles | 3560 miles |
| Sample Size | 15 | 18 |

The company specifies to be the significance level. Since the company claims that the tread life should be more than their existing product, the hypotheses are the following:

We found that the sample standard deviation for the existing tire product is 5325 miles and 3560 miles for the new tire. The company chooses a significance level of = 0.10. The test statistic is computed from Equation 6 to see how extreme of a difference between the two tires are.

Since our p-value is 0.058401 is less than , we can reject the null hypothesis. Therefore, there is sufficient evidence to conclude that existing tire has more variance than the newly developed tire.

Analysis of Variance (ANOVA)

Earlier in this best practice, comparisons between as many as two samples were discussed, but Analysis of Variance (ANOVA) can be used to compare more than two sample means. For example, we want to know if there was a difference in mean flavor scores for three different cookies. Performing an ANOVA test requires some prior work to get to a test statistic by decomposing variation in your data into a table. First, we need to understand what information the ANOVA test uses to make conclusions.

ANOVA requires knowing the difference between the overall mean of the data and each individual data point, this is referred to as the Sum Squares Total (SST). Visually in Figure 6 below, it is the distance from the points to the black line, and the length of all fifteen-line segments are squared and summed up. Mathematically expressed as , where indicates summing over the 15 data points, is a respective data point, and is the overall mean flavor score.

Chart, scatter chart

Description automatically generated

Figure 6

Visual Representation of Overall Mean Cookie Flavor Scores

ANOVA calculations also require the difference between the mean of each group (or cookie type in our case) and the overall mean, known as Sum Squares Between (SSB). Graphically, it is the length of the line segment from each group’s mean to the overall mean, with each distance multiplied by the number of observations in each group, then squared and summed. This is expressed in Figure 7. The equation for this is , where is the sum, is the respective group mean, and is the overall mean.

Chart, line chart, scatter chart

Description automatically generated

Figure 7

Visual Representation of Mean Flavor Scores for Each Cookie

Finally, ANOVA requires an error estimate (Sum of Squares Error, SSE), which is the difference between the mean of each group compared to individual data points. Graphically, shown below in Figure 8 as each type of cookie’s distance to the respective group mean, each distance is squared and summed. Mathematically, , where is the double sum over the respective group and data points within each group, is a respective data point, and is the respective group mean.

Chart, scatter chart

Description automatically generated

Figure 8

Visual Representation of Errors for Cookie Flavor Scores

These distances are reduced to a test statistic and p-value. For more details on the ANOVA process including the table and all required calculations, please see *Understanding Analysis of Variance* (Natoli, 2017).

Example of ANOVA

Consider the cookie example where judges scored three types of cookies on a scale of 1-10: Chocolate Chip (CC), Peanut Butter (PB), and Gingersnaps (GS).

Our corresponding hypotheses are:

The summary statistics that are used to decompose the variance are , . We begin by calculating the Sum Square Total (SST):

Then, we compute the Sum of Squares Regression (SSR), essentially the variance accounted for by each group’s respective means:

Finally, we subtract SSR from SST to calculate the Sum of Squares Error (SSE), which is the leftover variance that is not accounted for by each group mean.

Using the respective Sum of Squares for each source we receive an F-Statistic of 39.529 and a p-value of approximately . Since our p-value is less than , we can reject the . Therefore, we have enough evidence to conclude that at least one cookie’s flavor score mean is different. Note that the test does not determine which one is different, just that at least one is different.

Regression

The general form of SLR is the following:

where is your predicted response variable, the y-intercept (the value on the y-axis where the line crosses), is the slope of the line, and is the value of your explanatory variable to predict .

The equation for multiple linear regression (MLR) is the following:

where is your response variable, is your y-intercept, and are the expected increase or decrease of the respective x variable, and is the value of your explanatory variable to predict y.

R-Squared

Recall R2 is the percentage of the total variation in your data that your model explains. The sample mean of the response variable is our initial guess for predicting its value based on any combination of predictors. We want to improve that by using linear regression. The distance of the observed y-values to the mean of the response is summed and squared. That is called the total sum of squares (TSS), which is exactly how SST is computed for ANOVA.

Where is the Total Sum of Squares, is the sum over all n data points, is the observed response value, and is the mean response value.

Next, we sum and square the residuals from our model, called the residual sum of squares or RSS.

Where is the Residual Sum of Squares, is the observed response value, and is the predicted response value from your regression model

The ratio of is the amount of variation that is not explained by your model, so subtracting the ratio from 1 will give you

Where is the Residual Sum of Squares, is the Total Sum of Squares also denoted , and is the Model Sum of Squares.

Example of Regression

Earlier in this best practice we alluded to linear regression in the prediction intervals section where a range of *miles per gallon* could be predicted from a given amount of horsepower. Let’s model that relationship. The scatter of the relationship between Horsepower and MPG is plotted below in Figure 9.

Chart, scatter chart

Description automatically generated

Figure 9

Scatterplot of Horsepower and MPG with Line of Best Fit

The line that best fits is one that minimizes residuals, or vertical distance to a proposed line. The resulting model equation with that line:

To interpret this model: for every unit increase in horsepower the expected miles per gallon decreases by about 0.068. How well did this model fit the data? The R2 is 0.5892, so only about 59% of the variation in miles per gallon is explained solely by horsepower. We can add more predictors to increase this amount.

Another variable that we can add to the model is weight of the vehicle. The equation for that line is:

To interpret this model: for every unit increase in horsepower the expected miles per gallon decreases by about 0.03177 holding all other variables constant. Additionally, we have another variable to interpret, for every pound increase in weight the expected miles per gallon decreases by about 3.87783 holding all other variables constant. By adding another predictor to the model, how much did R2 increase? It increased from 0.5892 to 0.8148, so about 81% of the variation in miles per gallon is explained by horsepower and weight.