

STATISTICAL ANALYSIS OF DATA FOR ENGINEERS



Joshua Steimel
Cal Poly Humbolt

Statistical Analysis of Data for Engineers

This text is disseminated via the Open Education Resource (OER) LibreTexts Project (<https://LibreTexts.org>) and like the thousands of other texts available within this powerful platform, it is freely available for reading, printing, and "consuming."

The LibreTexts mission is to bring together students, faculty, and scholars in a collaborative effort to provide an accessible, and comprehensive platform that empowers our community to develop, curate, adapt, and adopt openly licensed resources and technologies; through these efforts we can reduce the financial burden born from traditional educational resource costs, ensuring education is more accessible for students and communities worldwide.

Most, but not all, pages in the library have licenses that may allow individuals to make changes, save, and print this book. Carefully consult the applicable license(s) before pursuing such effects. Instructors can adopt existing LibreTexts texts or Remix them to quickly build course-specific resources to meet the needs of their students. Unlike traditional textbooks, LibreTexts' web based origins allow powerful integration of advanced features and new technologies to support learning.



LibreTexts is the adaptable, user-friendly non-profit open education resource platform that educators trust for creating, customizing, and sharing accessible, interactive textbooks, adaptive homework, and ancillary materials. We collaborate with individuals and organizations to champion open education initiatives, support institutional publishing programs, drive curriculum development projects, and more.

The LibreTexts libraries are Powered by [NICE CXone Expert](#) and was supported by the Department of Education Open Textbook Pilot Project, the California Education Learning Lab, the UC Davis Office of the Provost, the UC Davis Library, the California State University Affordable Learning Solutions Program, and Merlot. This material is based upon work supported by the National Science Foundation under Grant No. 1246120, 1525057, and 1413739.

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation nor the US Department of Education.

Have questions or comments? For information about adoptions or adaptations contact info@LibreTexts.org or visit our main website at <https://LibreTexts.org>.

This text was compiled on 05/15/2026

TABLE OF CONTENTS

[Licensing](#)

[Chapter 1: Importance of Statistics and Experimental Measurements](#)

[Chapter 2: Histograms, Statistical Measures, and Probability](#)

[Chapter 3: Probability Distributions and the Gaussian](#)

[Chapter 4: Central Limit Theorem and Confidence Intervals for Large and Small Sample Sizes](#)

[Chapter 5: Hypothesis Testing and P-Values](#)

[Chapter 6: Goodness of Fit Hypothesis Testing](#)

[Chapter 7: Total Uncertainty, Bias Error, and Propagation of Error](#)

[Chapter 8: F-Tests and Variable Independence](#)

[Chapter 9: Analysis of Variance \(ANOVA\)](#)

[Chapter 10: Fitting, Regressions, and Confidence Bands](#)

[Index](#)

[Glossary](#)

[Detailed Licensing](#)

[Detailed Licensing](#)

Licensing

A detailed breakdown of this resource's licensing can be found in [Back Matter/Detailed Licensing](#).

Chapter 1: Importance of Statistics and Experimental Measurements

Here you can find pre-recorded lecture videos that cover this topic here: https://youtube.com/playlist?list=PL...EANC_k_3z5Y5n8

1.1 Performing Error and Data Analysis in the Lab:

Here are some common comments that I will typically overhear when students are in lab and talking about data analysis

1. **How good is your data?**
2. **How well does your experimental data fit the theory?**
3. **What is the R^2 value?**
4. **What is the percent error?**

While well intentioned, these are **not the questions an experimentalist should ask**. And before any data set can be utilized for an application the **quality** of the data must first be established.

For example in the first question is the student assessing the **accuracy? Precision? How well it fits with theory? We must be specific and careful with our language**. In terms of the second question, data is **not necessarily good** just because it agrees with theory. The quality of the data must be assessed before any conclusions can be drawn. **What is the reason the theory and experiments disagree? Is that theory applicable in this situation?** We must ask these questions!

You all probably have experience with R^2 as a **measure of the goodness of a fit** but in this course we will do a **much more thorough error analysis**, we will go much deeper than R^2 so we will not stop there.

Finally, **never ever ever ever ever include percent error in any lab analysis** that we perform in this class. It is a meaningless error analysis. We never know the actual value or true value of a measurement so will not perform this error analysis.

When designing an experiment it is crucial that do the following

- Clearly defined goals and ideally graphs/figures for study
- Experimental plan, apparatus, and DAQ for collecting data
- Perform appropriate statistical analysis to effectively extract critical information from raw data
- Interpret results and draw conclusions that are statistically significant

Additionally, when we speak of quality what we mean is to find the **actual or true value** of the physical quantity being measured. And often this is different than the **measured value** of the physical quantity. The difference between the **measured value** and the **actual or true value** of a measurement is the **error**.

Here we encounter a problem as we cannot calculate the error exactly unless we know the true value of the quantity being measured. Instead we **estimate the bounds or the likelihood that the error exceeds a specific value**. For example say that 95% (19 out of 20) of the yield strength measurements will have an error less than 1 MPa. **Thus the reading has an uncertainty of 1 MPa at confidence level of 95%**. We will calculate confidence levels or confidence intervals when quantifying error.

As experimentalists we often work with large data sets with a large number of measurements. The **error or uncertainty can be estimated with statistical tools when a large number of measurements are taken**. Later on in the class we will learn how to use some software tools like **Matlab, R, Python, and/or Mathematica** to perform error analysis. Much more on this later, but let's get to some basics of experimentation and measurements.

1.2 Measurements:

A **measurement** is a **quantitative** comparison between a predefined **standard** and a **measurand**.

What are some examples of different types of measurements we can make?

- Temperature
- Length

- Time
- Stress
- Strain
- Viscosity
- Force
- Torque
-

Measurand: the particular physical parameter being observed and quantified, i.e., the **input quantity** of the **measuring process**

Standard: has a defined relationship to a unit of measurement for a given measurand. Typically a standard will be recognized by the NIST (National Institute of Standards and Technology)

What are some examples of standards?

- Tape Measure
- Balance
- Speed of Light
- Stopwatch
- Extensometer
- ...

2.1.1 What is the significance of measurements, why do we experimentally measure parameters?

- Obtain quantitative information on the actual state of physical variables and processes (i.e. T, P, V, etc.), this is critical especially for thermodynamics and finding thermodynamic equilibrium
- Vital for control processes which require measured discrepancy between actual and desired output (i.e. temperature control on sheet annealing line)
- Daily operations (i.e. temperature controls, heating and cooling, tire gauge, oil gauge, etc.)
- Establishing costs, Ti \$140/lb precision is important when dealing with tons of Ti.
- Temperature in bioreactor can influence cell growth
- Thermal conductivity, acceleration, material properties

2.2 Uncertainty in Measurements: Measurements Are Useless without Uncertainty Bounds or Error Bars

Measurements are only useful to us as engineers if we can quantify the **uncertainty** of a measurement. **No measurement is meaningful** or useful if they do not have an **associated error bar**. In this class and moving forward in your career as an engineer be sure to place error bars on all your reported values. Error bars are critical because no measurement is perfect. If you measure the length of a tensile specimen with a micrometer your measurement is likely to be different than that of your colleague. Thus, there will be some **uncertainty** in any measurement that one makes. In this section of this course we will develop a comprehensive framework for evaluating uncertainty and placing **tolerances and confidence intervals** on measured values. Much more on this later....

2.3 Using Correct Terminology When Assessing Measurement Quality/Apparatus Performance

- **Accuracy:** Difference between measured and true values. Manufacturer will specify maximum error but not confidence intervals typically.
- **Precision:** Difference between reported values during repeated measurements of same quantity.
- **Resolution:** Smallest increment of change in the measured value that can be determined by the readout scale. Often the same as precision.
- **Sensitivity:** The change of an instrument's output per unit change in the measured quantity. Higher sensitivity often indicates finer resolution, better precision, and higher accuracy.

2.4 Error and Uncertainty:

Error is the difference between the *measured* and *true* value of the quantity being measured. Unfortunately, the true value can never really be known so instead we quantify the **uncertainty** of a measurement at specific levels of confidence. Error can be categorized into two basic categories either **bias** or **precision** error/uncertainty. We will cover uncertainty and error analysis much more in-depth in the next several lectures...but never use percent error in this course.

Error

We can begin with our discussion of error by defining the following equation

$$Error = x_m - x_{true} \quad (\text{Chapter 1.1})$$

where x_m is the measured value and x_{true} is the true value.

When designing an experimental system or executing an experiment the experimentalist's primary objective should be to minimize the error. After the experiment is completed we need to estimate a bound on the error with a degree of certainty or a confidence interval (i.e. 90%, 95%, 99%), hopefully you have encountered confidence intervals before but if not no worries we will discuss this in depth.

There are many different sources of inaccuracy or error however most errors can be classified as:

- (1) **Precision**
- (2) **Bias**
- (3) **Illegitimate**
- (4) **Errors that are sometimes bias or sometimes precision errors**

In this course we will spend a considerable amount of time discussing **1. Precision** and **2. Bias** error in great detail. But we will discuss all these sources of error and how to account for them when designing experiments or presenting your results. Let's start with **Precision Error**.

2.4.1 Precision or Random Errors

I really do not like the term random error but you will occasionally see that term in literature but in this course we will use the term precision primarily.

Precision errors are **different for each successive measurement**. Thus if you create a histogram, which shows the probability of encountering or measuring a specific measurand, the width of your distribution (hopefully it is Gaussian) will be due to precision error if that is the only source of error in your experiment. From this distribution, **statistical analysis can be utilized to estimate the size of the precision error**. Typically bias and precision errors occur simultaneously and contributes to the **total error**, more on this fun calculation later, hope you remember Taylor expansions....

Why Do We Have Precision Error?

- (1) Disturbances to the equipment
- (2) Fluctuating experimental conditions

- (3) Insufficient measuring-system sensitivity
- (4) Stochastic (Random) nature of measurements even in static conditions

Examples of Precision Error:

- Measuring temperature with thermocouple as temperature fluctuates
- Measure the length of a beam with a micrometer you will get slightly different measurements than classmates
- Changes in temperature conditions can affect swimming of bacteria

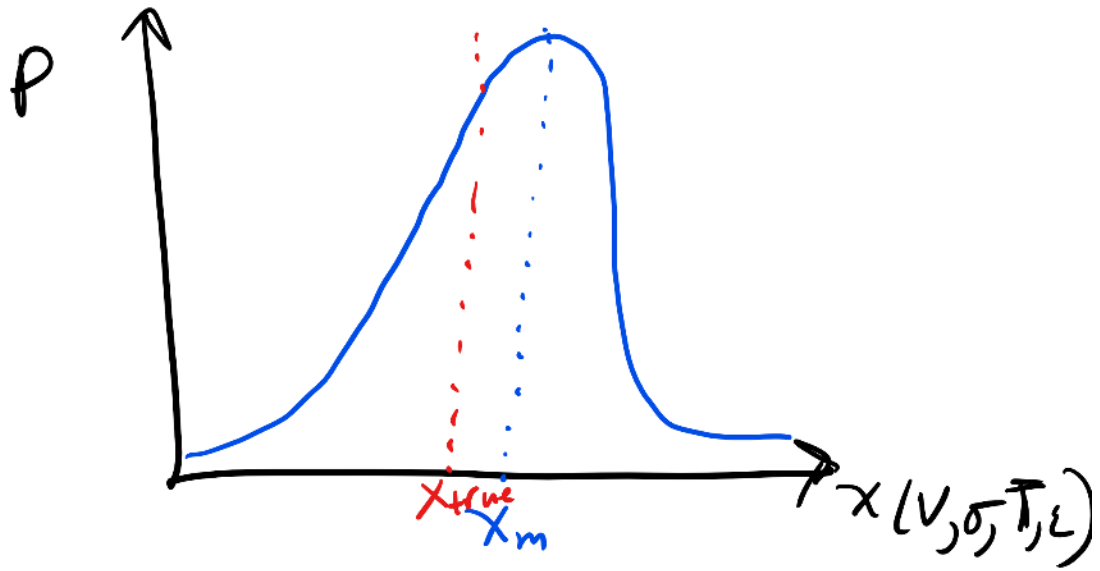


Figure Chapter1.1: Precision Error Distribution.

2.4.2 Bias or Systematic Errors

I like the term bias and systematic much better than random. However, in this course we will primarily use the term Bias error.

Bias errors will occur the same way each time a measurement is made. For example a balance that consistently reads 0.5 mg heavier. In some literature you might see this termed as a fixed offset error. The key thing to note about bias errors is they are fixed and thus **they do not show a distribution**. This is in stark contrast to precision errors which show a distribution, typically Gaussian. Therefore if there is no distribution we cannot apply any **statistical analysis**.

There are multiple types of bias errors:

(1) Calibration errors

- This is the most common error
- Typically calibration errors are **zero-offset errors** which cause all read- ing to be offset by constant amount or *scale errors*
- I.e. scale constantly light by 1lb or clock runs fast by 1 minute
- Can be reduced via calibration procedures and proving the system via comparison with *known* standard

(2) Consistently recurring human errors

- Ex. experimenter who consistently jumps the gun when recording synchronized readings either time, voltage, etc.

(3) Errors caused by defective equipment

(4) Loading errors

- i.e. not putting a DSC pan at the center of the DSC will skew temperature measurements

(5) System resolution limits

- i.e. Bias error for a ruler with cm tick marks will be cm, compared to a micrometer with $1\mu\text{m}$ resolution, that is 4 order of magnitude difference.

3. Illegitimate Error

Illegitimate error can and most likely will occur to you at some point as you perform experiments as an engineer. Illegitimate errors include

- Blunders and mistakes
- Computational or calculation errors
- Analysis errors

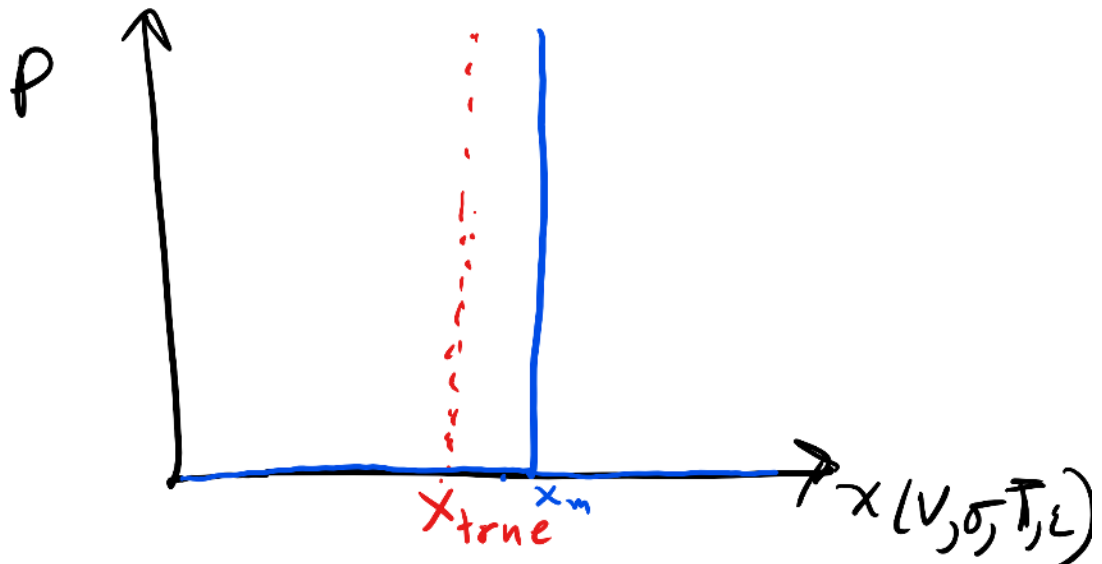


Figure Chapter1.2: Bias Error Distribution

While these things can occur in lab **you must always perform the experiment again**. You cannot simply present your results in either a presentation or a form of technical writing and attribute error to illegitimate error. You must perform the experiment again.

4. Errors That Are Sometimes Bias and Sometimes Precision Error:

These are the most difficult sources of error that we must analyze as it is difficult to determine what type of error analysis must be done and which error has a distribution and what does not.

- **Instrument Hysteresis**
 - Hysteresis error can be seen in stress strain measurements and is path dependent
- **Calibration drift and variation in test or environmental conditions**
 - Drift occurs when the response varies over time often due to sensitivity to temperature or humidity
- **Variations in procedure or definition among experimenters**

Quantifying Total Uncertainty, Comprehensive Error Analysis

Again typically we are concerned with **bias** and **precision** error when evaluating experimental data. And this data comes from two classes of experiments: i) *single-sample* and ii) *repeated-sample* experiments.

A **sample** refers to an individual measurement of a specific quantity. Ex. Measuring strain in a material several times under identical loading is a repeated-sample experiment. A single measurement of strain is a single-sample experiment and we cannot extract the distribution of precision error in this case. In this course and in your career as an engineer you will most likely be taking multiple measurements,

i.e. repeated sample experiments. Single sample experiments are more uncommon and typically occur when the measurement you are making is extremely difficult and multiple experiments are non realistic to conduct, typically due to economic or practicality reasons.

The total uncertainty U_x in a measurement of x is

$$U_x = \sqrt{B_x^2 + P_x^2} \quad (\text{Chapter 1.2})$$

where B_x is the bias uncertainty in a measurement of x and P_x is the precision uncertainty. Note that x is just a variable and we can talk about uncertainty in measuring volume (V), pressure (P), stress (σ), etc.

This is a largely empirically derived formula however it does depend on the **assumption that the bias and precision uncertainty are independent sources of error**. Also the bias and precision uncertainty should have the **same confidence level** when calculating the **total uncertainty**. We will come back to this a bit later in excruciating detail for this calculation so please be patient.

Minimizing Error in Designing Experiments

Best time to **minimize experimental error** is in the **design stage**. You should **perform a single-sample uncertainty analysis** of the proposed experimental arrangement **prior to building the apparatus**. This allows one to make a decision on whether the expected uncertainty is acceptable and to identify the sources of uncertainty. So when designing an experiment:

- **Avoid approaches** that require two large numbers to be measured to determine the small difference between them
- **Design experiments or sensors that amplify the signal strength** to improve sensitivity
- Build *null designs* in which the output is measured as a change from zero rather than as a change in nonzero value
- Avoid experiments with large *correction factors* must be applied as part of the data-reduction procedure
- Attempt to **minimize the influence of the measuring system on the measured variable**
- **Calibrate** entire system to minimize calibration related bias errors.

Once we have established the error and uncertainty with a measurement we also have to **report these results** and **explain our analysis**. **Technical writing** is a very distinct class of writing with many different formats and audiences. It can be difficult to strike the right **tone** and **level of technical information** that must be provided at times. In this class we will focus on **Laboratory Reports** but as you can see below there are many different form of technical writing.

Forms of Technical Writing:

- **Executive summary**

- Audience is a supervisor. Summary should be concise, usually only a few paragraphs, and highlight key aspects of work in first paragraph.
- Consequences can be drastic as in the case of the Challenger Shuttle tragedy.

- **Laboratory note or technical memo**

- Limited audience, immediate supervisor
- Sufficient information should be included so the experimentalist can re-construct the experiment

- **Progress report**

- Audience is supervisor of sponsoring agency
- Can vary in terms of length, structure, timelines, etc.

- **Full technical report**

- Wider audience typically not people closely associated with the work
- Relates all facts pertinent to project
- Contains enough detail so another professional can repeat experiments

- **Full technical report**

- Wider audience typically not people closely associated with the work
- Relates all facts pertinent to project
- Contains enough detail so another professional can repeat experiments

Line Fitting: Identify Outliers and Physical Interpretation

When you are fitting a curve or a line to a data set **one must be careful to select a fit that is appropriate for the experiment being conducted**. Often researchers are too consumed with obtaining a **high R^2 value**. While the goodness of a fit is important you must make sure that the equation that you are **fitting matches with the physical experiment that you are conducting**. For example if you are measuring how an electrochemical cell voltage changes as a function of temperature you must use the Nernst Equation. Looking at the data perhaps a 12th degree polynomial fit will give you a better R^2 value but that equation has no physical meaning in terms of your analysis. So always keep that in mind.

When identifying outliers and a particular point is in doubt, temporarily exclude the data point and fit a new line through the remaining data. **If the fit is better the point can possibly be dropped however it could indicate that the relationship is not linear it may indicate another scaling regime or a transition in behavior.**

We can also apply this same error analysis procedure and to optimize your experimental apparatus and minimize error/total uncertainty. One can identify what component or part of the measuring process is generating the most error and then one can re-design the apparatus to minimize that error.

Chapter 1: Importance of Statistics and Experimental Measurements is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

Chapter 2: Histograms, Statistical Measures, and Probability

Here you can find pre-recorded lecture videos that cover this topic here: <https://youtube.com/playlist?list=PL...gEHsGBZJpQhGzn>

Sample, Population, and Distribution Error

Before we start to analyze precision error we must understand two key concepts: i) **distribution of error** and ii) **population from which a sample is taken**.

- **Distribution of error:** characterizes the **probability** that an error of a given size will occur.
- **Population from which a sample is drawn:** experimentally we have a **limited set of observations, our sample**, from which we will infer the **characteristics of the larger population**.

Understanding Sample Vs. Population: Bag of Marbles

In any bag of marbles there will be a distribution of diameters. To estimate the mean diameter we can take a handful of marbles (**sample**) drawn from the bag (**population**).

No two handfuls will yield the exact same average value but **each should approximate the average of the population** to some level of uncertainty **if the sample is large enough to approximate the population**. The handful is a repeated sample whereas picking one marble would have been a single sample.

You can visually see this in the Mathematica Notebook simulation below where we create a population of marble diameters and take different sample sizes.

Visualizing Data: Histograms

One thing that you will immediately see here is that we are representing the data using a **histogram**. **Histograms** typically represent the frequency or number of times a measurand is measured. The y-axis typically represents frequency and the measurand value is on the x-axis. We will talk about distribution but the lower the frequency of a measurand the more rare it is to encounter such a measurand and the higher frequency the more likely it is to encounter a measurand.

Notice that the sample means vary from each sample grabbed from the bag and that as we grab larger and larger samples as you might anticipate we get a closer approximation to the true population mean.

```
1 popmarb<-c(rnorm(5000, mean = 10, sd = 1))
2 hist(popmarb)
3 mpopmarb<-mean(popmarb)
4 cat("Mean Popualtion",mpopmarb)
5 stdpopmarb<-sd(popmarb)
6 cat("SD Popualtion",stdpopmarb)
7 s1<-sample(popmarb,1)
8 cat("Mean n=1",s1)
9 s5<-c(sample(popmarb,5))
10 m5<-mean(s5)
11 cat("Mean n=5",m5)
12 s50<-c(sample(popmarb,50))
13 m50<-mean(s50)
14 cat("Mean n=50",m50)
15 s500<-c(sample(popmarb,500))
16 m500<-mean(s500)
17 cat("Mean n=500",m500)
```

run

run all

restart

restart & run all

Sampling:

- Sample of size n is drawn from a finite population of size p . We assume additional data cannot be added to the population and that $n \ll p$
- Finite number of items, n , is randomly drawn from a population of indefinite size and properties of population are inferred from the sample.

Sample Vs. Population Parameters

Population Mean and Standard Deviation

Typically, the population mean, μ , is an unknown since the population is infinite and as experimentalists we can never know the full population distribution. However it would be calculated as follows

$$\mu = \frac{x_1 + x_2 + \dots + x_n}{n} = \sum_{i=1}^n \frac{x_i}{n} \quad (\text{Chapter 2.1})$$

By averaging a **large sample** we are able to **estimate the true value of the population**. However, we will develop a much more systematic approach of estimating bounds on the true value of a population, coming soon....

Standard Deviation

The deviation, d is the amount by which a single measurement deviates from the mean value of the population

$$d = x - \mu \quad (\text{Chapter 2.2})$$

The mean squared deviation, σ^2 , is approximated by averaging the squared deviations of a very large sample:

$$\sigma \approx \sqrt{\frac{d_1^2 + d_2^2 + \dots + d_n^2}{n}} \quad (\text{Chapter 2.3})$$

σ is the **standard deviation** of the population and characterizes the deviation from the mean value and the **width of the Gaussian**, again more on this later.

Sample Mean and Standard Deviation

As we have previously discussed **it is often impractical or at times impossible to work with an entire population**, instead as experimentalists we work with samples from a population and **we use average values from the sample to estimate the mean or standard deviation of the population**.

The **sample mean** is defined as $\bar{x}$

$$\bar{x} = \sum_{i=1}^n \frac{x_i}{n} = \frac{x_1 + x_2 + \dots + x_n}{n} \quad (\text{Chapter 2.4})$$

and as we have previously discussed the **sample mean**, $\bar{x}$, can be used to approximate the **population mean**, μ , for large sample sizes, i.e. $n \geq 30$. Similarly we can calculate the **sample standard deviation**, S_x ,

$$S_x = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2}{n - 1}} \quad (\text{Chapter 2.5})$$

which can be used to approximate the **population standard deviation**, σ . Just a quick reminder that n is the number of data points/measurements in the sample.

But let's go back to this idea of a histogram and frequencies of encountering measurements because this reminds me of probabilities, let's refresh our understanding of this very fundamental and important concept.

Number of Possibilities

Before looking at probabilities it is often very important to determine the number of possibilities in a given scenario. If we have sets of $A_1, A_2, \dots, A_k$ which contain, respectively, $n_1, n_2, \dots, n_k$ elements then there are $n_1 \cdot n_2 \cdot \dots \cdot n_k$ ways of choosing A_1 , then A_2 and finally A_k .

Let's do an example to put this into practice.

Consider the following scenario where I am trying to maximize the yield of a field where I am planting lettuce there are two options of fertilizer, F1 and F2, 4 blocks of land to try to plant (B1-B4), and 3 types of seeding density (S1-S3). How possibilities of planting combination can we observe?

Well, from the given example there are $n = 2$ for our F scenario, 4 for B, and 3 for S so applying this rule we have

$$2 \cdot 4 \cdot 3 = 24 \quad (\text{Chapter 2.6})$$

How about another quick problem, I make a true-false exam (I will never do that by the way) with 15 questions, how many combinations of answers can we observe?

$$2^{15} = 32768 \quad (\text{Chapter 2.7})$$

Lot of grading options for me.

This concept is closely related to the number of ways to arrange objects distinctly, which you may remember was called.....

Permutations

If r objects are chosen from a set of n **distinct** objects any particular arrangement or order of these objects is a **permutation**. In such a scenario the number of permutations **P** is defined as

$${}_nP_r = \frac{n!}{(n-r)!} \quad (\text{Chapter 2.8})$$

Let's consider an example, where we have 15 candidates, how ways can we choose a president, vice president, chief of staff, and treasurer.

Well, what are the number of objects, r ? Here we see that $r = 4$ and the set of distinct objects here is $n = 15$.

$$\frac{15!}{11!} = 32760 \quad (\text{Chapter 2.9})$$

Combinations

If we care about selecting r objects that can be selected from n distinct objects the number of **combinations** of n objects taken r at a time is denoted by

$${}_nC_r = \frac{n!}{r!(n-r)!} \quad (\text{Chapter 2.10})$$

Let's say I have a class of 23 students how many ways can we choose teams of 3?

Here we have that $n = 23$ and $r = 3$

$$\frac{23!}{3!20!} = 1771 \quad (\text{Chapter 2.11})$$

How would you expect the number of combinations to change if we increase the team size to 8?

$$\frac{23!}{8!15!} = 490314 \quad (\text{Chapter 2.12})$$

Probability

Now that we have the number of possibilities and permutations defined we can now start to define **probability** (P) which is simply defined as

$$P = \frac{s}{m} \quad (\text{Chapter 2.13})$$

where there are m **equally likely possibilities** and s **are the successful or favorable outcomes** therefore the probability of a success is defined as above.

Let's consider a single deck of completely randomly shuffled deck of cards with 52 cards in total. What is the probability of pulling a red Queen? Well here $m = 52$ and there are only two red queens in the deck so $s = 2$

$$\frac{2}{52} = \frac{1}{26} \quad (\text{Chapter 2.14})$$

Probabilities: Mutually Exclusive Events

There are some scenarios where we may have some probabilities of multiple events occurring for example, if I have events $A_1, A_2, \dots, A_n$ which are all **mutually exclusive events** in a sample space S , then

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = P(A_1) + P(A_2) + \dots + P(A_n) \quad (\text{Chapter 2.15})$$

Let's see how this can be applied to an example, continuing with our card example what is the probability of pulling a red card, club, or 2 of spades?

Well these are all **mutually exclusive events**, let's make sure this is indeed the case which is valid here because pulling a red card, club, or 2 of spades are mutually exclusive and will not change the probabilities, so we need to look at each events probability which will be

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{52} = 0.7692308 \quad (\text{Chapter 2.16})$$

Probabilities: Not Mutually Exclusive

Now if A and B are any events in the sample space S , **not mutually exclusive** then

$$P(A \cup B) = P(A) + P(B) - P(A \cap B) \quad (\text{Chapter 2.17})$$

We would get an erroneous result if we applied the previous framework but we must exclude any overlapping probabilities.

Let's consider this example, what is the probability of pulling a card that is red or an ace?

So the probability of pulling a red card is $\frac{1}{2}$, the probability of pulling an ace is $\frac{1}{13}$, and the probability that we pull a red ace is $\frac{1}{26}$ so we find that

$$\frac{1}{2} + \frac{1}{13} - \frac{1}{26} = 0.5384615 \quad (\text{Chapter 2.18})$$

Probabilities: Conditional Probabilities

There are also conditional probabilities for example if A and B are any events in space S and $(P(B) \neq 0)$ then the conditional probability of A given B is

$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

Let's illustrate this with an example and keep going with our cards as long as we can, **so what is the conditional probability that we pull an ace given that it is also red?**

So the way that we can read this is given that

$$\frac{\frac{1}{26}}{\frac{1}{2}} = 0.07692308 \quad (\text{Chapter 2.19})$$

If we rearrange this conditional probability we can show that

$$P(A \cap B) = P(A) \cdot P(B|A) \quad (\text{Chapter 2.20})$$

$$P(A \cap B) = P(B) \cdot P(A|B) \quad (\text{Chapter 2.21})$$

this is assuming that $P(A) \neq 0$ and $P(B) \neq 0$. Then it follows that two events A and B will be independent events if and only if

$$P(A \cap B) = P(A) \cdot P(B) \quad (\text{Chapter 2.22})$$

Let's see if our card example was truly independent. We said that the probability for pulling a red ace was $\frac{1}{26}$ and the probability for a card being red was $P(A) = \frac{1}{2}$ and the probability of pulling an ace was $P(B) = \frac{1}{13}$ so we find that

$$\frac{1}{26} = \frac{1}{13} = \frac{1}{26} \quad (\text{Chapter 2.23})$$

So they are independent!! Now consider how this may change if the card is not replaced, a possible Pset question, hmmmmmm???

Bayes' Theorem

These multiplication rules for probabilities are extremely useful but we can also imagine several steps in calculating a probability so imagine $B_1, B_2, \dots, B_n$ are **mutually exclusive events** of which one must occur then the general expression which is sometimes referred to as the **rule of elimination** or the **rule of total probability**

$$P(A) = \sum_{i=1}^n P(B_i) \cdot P(A|B_i) \quad (\text{Chapter 2.24})$$

We can further generalize this problem by instead invoking the general **Bayes' theorem** where if $B_1, B_2, \dots, B_n$ are mutually exclusive events of which one must occur then

$$P(B_r|A) = \frac{P(B_r) \cdot P(A|B_r)}{\sum_{i=1}^n P(B_i) \cdot P(A|B_i)} \quad (\text{Chapter 2.25})$$

for $r = 1, 2, \dots, n$.

Let's see if we pull one more card out of our sleeve, excuse the horrible pun, to create one more example...

A card is lost from a pack of 52 cards, from the remaining cards two cards are randomly drawn and found to be hearts, what is the probability that the last card is also a heart.

We can define the following variables where the lost card can either be a heart, club, spade, or diamond so we can set up events B_1, B_2, B_3 , and B_4 of losing a card of hearts, clubs, spades, and diamonds respectively. Here we can also state that A is the event of drawing two cards that are clubs after a card is lost.

We can then determine that $P(B_1) = P(B_2) = P(B_3) = P(B_4) = \frac{1}{4}$.

We can also state that the probability of drawing two hearts given that the lost card is a heart $P(A|B_1) = \frac{12}{51} \cdot \frac{11}{50}$.

The probability of drawing two hearts given the lost card is a club is $P(A|B_2) = \frac{13}{51} \cdot \frac{12}{50}$. And we can also see that $P(A|B_2) = P(A|B_3) = P(A|B_4)$.

Now we can plug into our equation and we find that

$$P(B_1|A) = \frac{P(B_1) \cdot P(A|B_1)}{P(B_1) \cdot P(A|B_1) + P(B_2) \cdot P(A|B_2) + P(B_3) \cdot P(A|B_3) + P(B_4) \cdot P(A|B_4)} \quad (\text{Chapter 2.26})$$

and we find the probability is 0.22!!!

Look forward to some more of these problems in your Pset.

Chapter 2: Histograms, Statistical Measures, and Probability is shared under a [CC BY-NC-SA 4.0](#) license and was authored, remixed, and/or curated by LibreTexts.

Chapter 3: Probability Distributions and the Gaussian

We previously discussed about histograms and ways to represent data, these comes in various forms of distributions that are commonly encountered when performing experiments and gathering data.

Here you can find a series of pre-recorded lecture videos that cover this content: <https://youtube.com/playlist?list=PL...I2KSQZ3UtuLaf>

Random Distribution

If you are performing a random (stochastic) process or a Markov process that is truly random you can confirm this is true by observing whether the results are truly randomly distributed as seen below

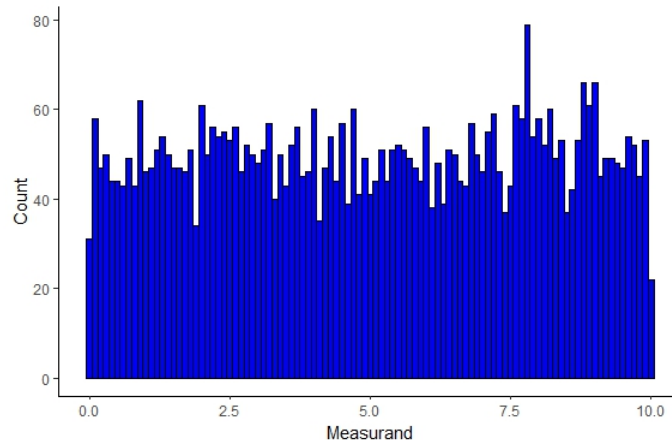


Figure Chapter3.1: Random/Uniform Sample Distribution.

As expected we see in this distribution each measurand value has an equal probability of being observed/measured. This type of distribution is somewhat uncommon experimentally unless you are looking at a component of a Markov process typically in a simulation system but this type of distribution can be observed.

Exponential or Negative Exponential Distribution

The **exponential** distribution is typically valid for processes in which events occur continuously and independently at a constant average rate. Such a distribution is rare to find in real-world experimental scenarios however it does capture well some physical systems such as the time until a radioactive particle decays, or the time until default on payment to debt hold in some financial models.

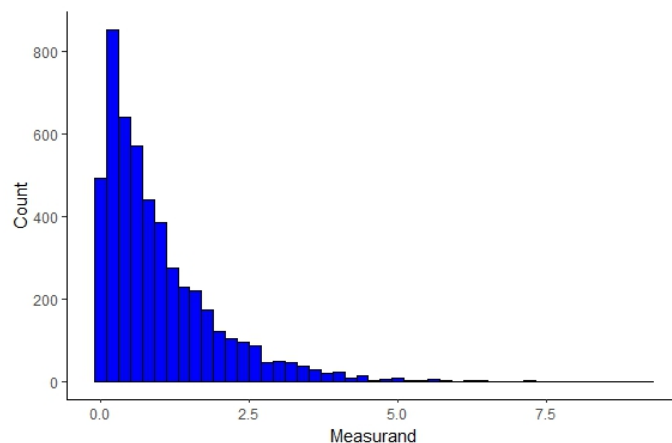


Figure Chapter3.2: Exponential Distribution.

The exponential distribution can be described as

$$f(x) = \lambda \exp^{-\lambda x} \quad (\text{Chapter 3.1})$$

where the population mean is represented as $\frac{1}{\lambda}$ as the standard deviation.

Poisson Distribution

The **Poisson** distribution is actually a discrete probability distribution, i.e. cannot be described by a continuous function. This distribution can be very useful to model the number and size of meteorites that hit the earth as well as modeling student exam score distributions. We can actually see the distribution below

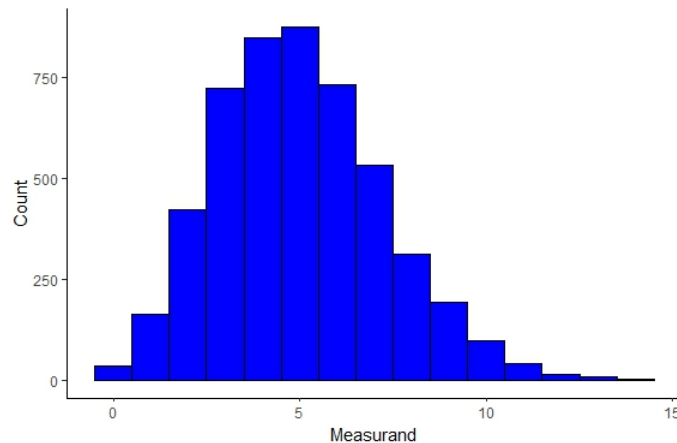


Figure Chapter3.3: Poisson Distribution.

Binomial Distribution

The binomial distribution is another distribution that is discrete and this distribution fits random number generators extremely well. This distribution can be seen below

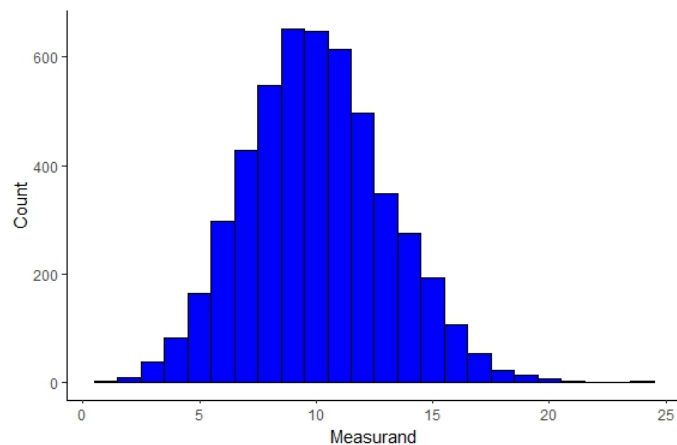


Figure Chapter3.4: Binomial Distribution.

Log-Normal Distribution

The **Log-normal** distribution is a continuous distribution and also can be referred to as the **Galton's** distribution. Unlike our previous distributions the Log-Normal distribution describes many natural phenomenon including the length of chess games, sizes of living tissues, many hydrology phenomenon, and many other phenomenon, which can be seen below

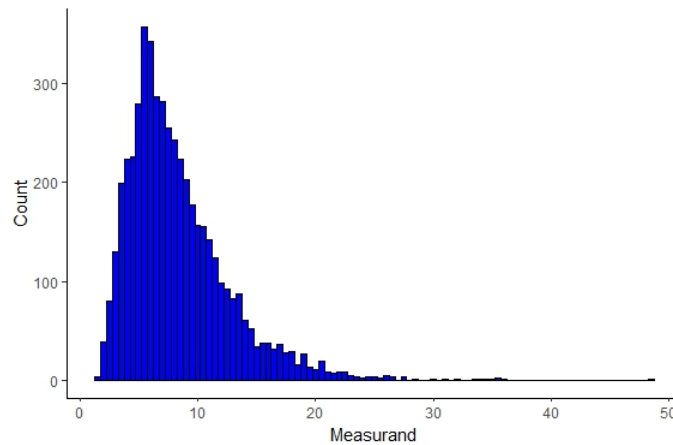


Figure Chapter3.5: Log-Normal Distribution

The distribution can be described as seen below:

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right) \quad (\text{Chapter 3.2})$$

Beta Distribution

The **beta** distribution is another continuous probability distribution and is defined by two parameters α and β which change the shape of the distribution as seen below

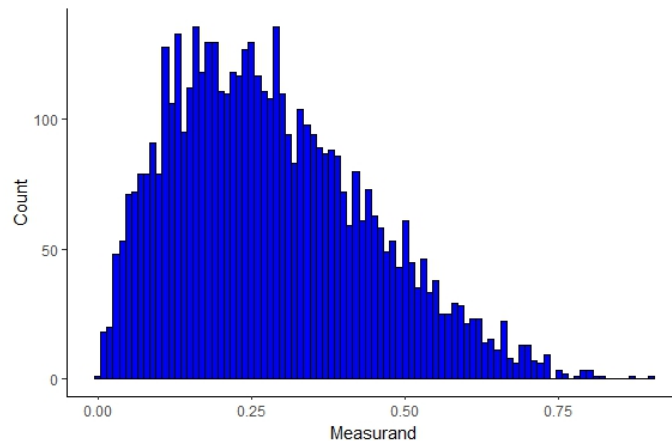


Figure Chapter3.6: Beta Distribution

The distribution can be described as

$$f(x) = \frac{1}{B} x^{\alpha-1} (1-x)^{\beta-1} \quad (\text{Chapter 3.3})$$

where B is a normalization factor. The beta distribution describes population genetics specifically allele frequencies, project management models, and other systems.

Gamma Distribution

The **gamma** distribution is another continuous distribution that can be described by two parameters and can be seen below

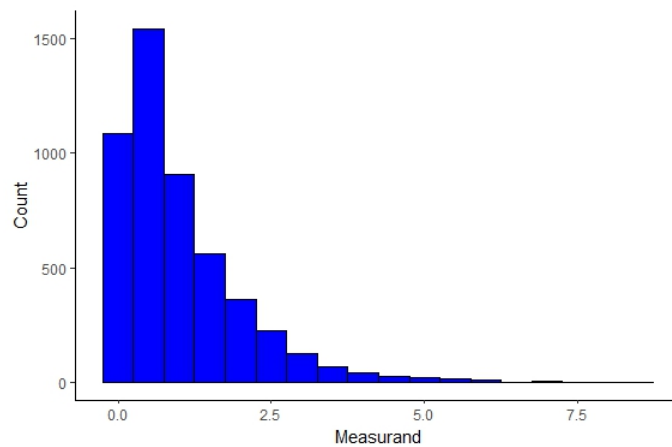


Figure Chapter3.7: Gamma Distribution

The gamma distribution model describes the size of insurance claims and rainfall quite well as well as the age distribution of cancer incidences. The gamma distribution can be described as

$$f(x) = \frac{1}{B\beta^\alpha} x^{\alpha-1} \exp^{-\frac{x}{\beta}} \quad (\text{Chapter 3.4})$$

where B is a constant normalization factor.

Weibull Distribution

The **Weibull** distribution is another continuous probability distribution governed by two parameters and this distribution famously describes fatigue failure of materials extremely well as well as many other physical systems and this distribution can be seen below

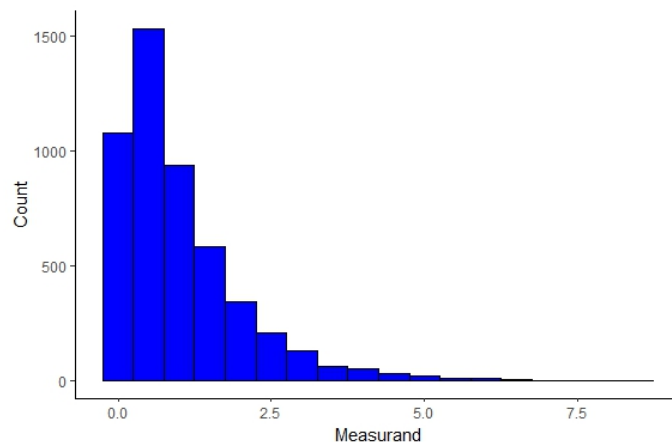


Figure Chapter3.8: Weibull Distribution

Gaussian and Normal Distribution

One of the most common distribution that you will encounter is the **Gaussian** distribution, often referred to as the normal distribution or bell-curve, which can be seen below

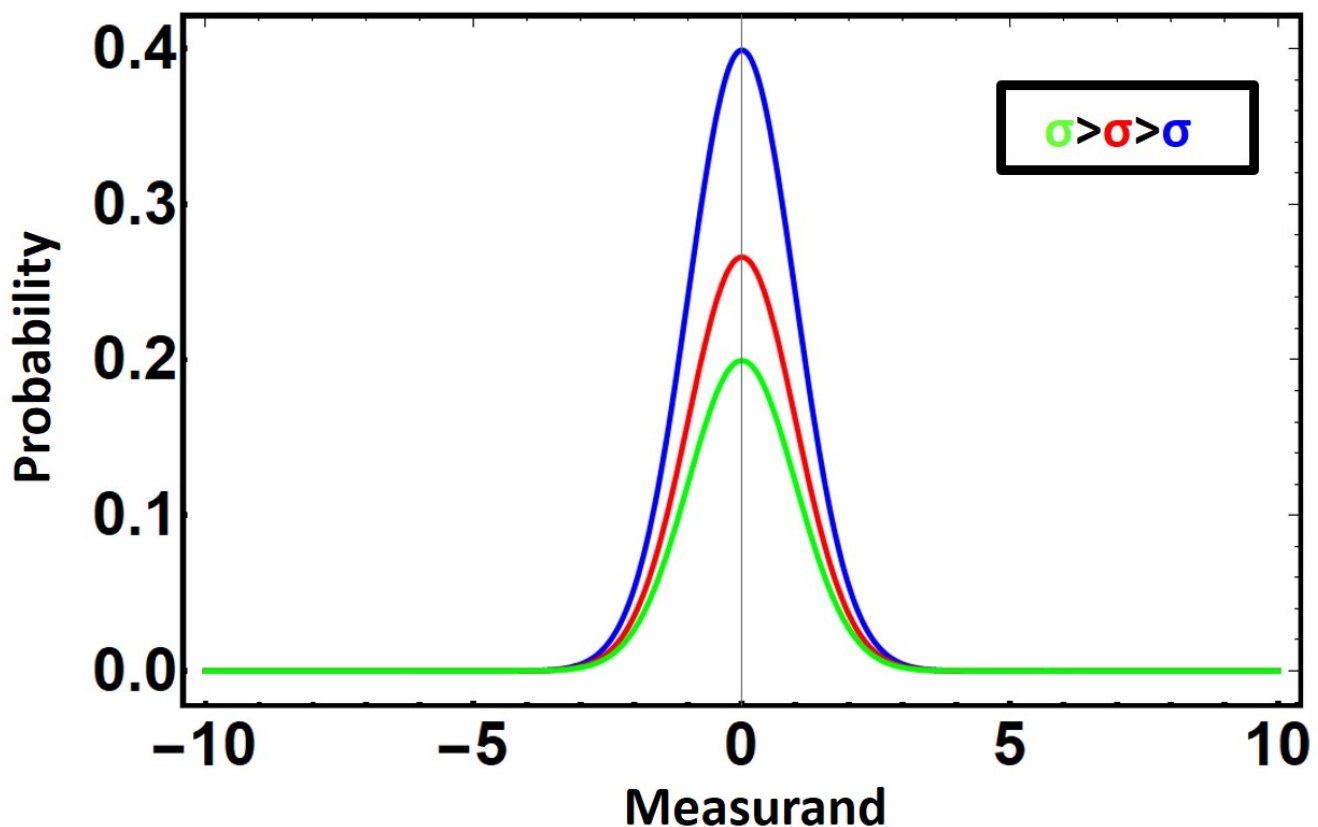


Figure Chapter3.9: Probability density function (PDF) or normal distribution curve. Notice the width of the Gaussian increases from blue to green.

The Gaussian will capture a myriad of physical systems and is by far the most likely distribution you will work with. Now remember, we are limited to working with samples of a population because it is either impractical or impossible to work with entire population **however let us consider an infinite population of data**, each datum representing a measurement of a single quantity, and assume that each datum, x , differs in magnitude only as a result of precision error. The probability of obtaining a specific value x depends on both the magnitude and distribution of x -values as described by the **probability density function or PDF**.

Since the population is infinite the PDF is a **continuous** curve. The PDF represents the probability of occurrence per unit change of x . The probability of measuring a given x is not $f(x)$ instead the probability of measuring an x in the interval $\delta x = x_2 - x_1$ is the area under the curve or:

$$P_{x_1 \rightarrow x_2} = \int_{x_1}^{x_2} f(x) dx \quad (\text{Chapter 3.5})$$

In any measurement some value of x must be observed so the total area under the curve is unity, 1, or alternatively the probability of measuring one value of x is 100% or equivalently

$$P = \int_{-\infty}^{\infty} f(x) dx = 1 \quad (\text{Chapter 3.6})$$

The form of the **Gaussian Probability Density Function** can be seen below

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left[-\frac{(x - \mu)^2}{2\sigma^2} \right] \quad (\text{Chapter 3.7})$$

where x is the magnitude of particular measurement, μ is the mean value of the **entire** population, and σ is the standard deviation of the **entire** population. μ is an unknown since the population is infinite.

Outliers

You have probably hear about outlier data points, i.e. data points that are much much higher or lower than the mean. There can be some controversy as an engineers in making the decision to neglect or disregard outlier data points. Typically in this class **you can disregard a data point as illegitimate if the data point exceeds 3σ** . However, it is good practice to always show this data point in your graphs and explain that when fitting a curve or determining a trend you disregarded this data point, do not just delete this point. The only time you can delete data you gather is if there was illegitimate error in which case all data must be disregarded and you must perform the experiment again.

We can simplify the form of the Gaussian PDF by defining an arbitrary variable z

$$z = \frac{x - \mu}{\sigma} \quad (\text{Chapter 3.8})$$

and the resulting Gaussian PDF becomes

$$f(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{z^2}{2}} \quad (\text{Chapter 3.9})$$

The above definition is the **standard curve** and using **tables** or **graphical integration** you can determine the area under the standard curve between 0 and various values of z .

Once we have this equation we can then estimate the size of errors with various levels of confidence. For example let us consider the most common confidence interval that you may encounter, the 95% confidence interval. One way we can quantify error is to make the statement that with 95% confidence if I make a measurement (stress for example) that measurement will fall between these two bounds. What we are saying here is that 95% of the measurements of a population should fall between these two bounds. So to calculate these bounds we can simply solve for the bounds where we will find 90, 95, or 99% of our data or whatever confidence interval that we are working with as seen in the Mathematica notebook below

```
Solve[(Integrate[pdf, {z, -x*σ, x*σ}] /. σ → 1 // N) * 100 == 90, x]
{{x → 1.64485}}
```

```
Solve[(Integrate[pdf, {z, -x*σ, x*σ}] /. σ → 1 // N) * 100 == 95, x]
{{x → 1.95996}}
```

```
Solve[(Integrate[pdf, {z, -x*σ, x*σ}] /. σ → 1 // N) * 100 == 99, x]
{{x → 2.57583}}
```

Figure Chapter3.10: Solved Confidence interval bounds in terms of $x \times \sigma$ for confidence intervals of 90,95, and 99%.

If you do not want to use mathematica you can instead use integration tables or for this function, **Z-Tables** as seen below. It should be noted that the table below lists values for only half the curve but since the standard curve is symmetric you can just multiply by 2 to get the total area. These **tabulated values** of areas is called the **Z-distribution**

Example Problems Confidence Intervals

Working with Z Tables

- What is the area under the standard normal distribution curve between $z = -1.43$ and 1.43 ?
- What percent of the population lies within this range?
- What range of z will contain 90% of the data?

Running and Tumbling E.coli

We measured the velocity of E.coli 135 times using Able Particle Tracker. The mean velocity was determined to be $3\mu m$ and the standard deviation was determined to be $0.1\mu m$. I'm particularly interested in velocity measurements between 2.857 and $3.143 \frac{\mu m}{s}$,

specifically the percentage of the population that will fall within this range because I plan to claim in my *Nature* paper that 95% of this data will fall within this range.

- What assumptions are you making about the data set?
 - Assume normal distribution
- Calculate z for this data set?
- What is the area under the Standard Normal Distribution for this data set range?
- What percent of my data will fall within this range?
- How many measurements fall within this range?
- What range of velocities will contain 95% of the data?

Marble Diameters

I have a bag of marbles and I grab a sample of 765 marbles. From this sample I measure the diameters and find that the population mean is 200mm, the population standard deviation is 20mm. I need the range of marble diameters where I can expect to find 60% of the marbles in the population.

Ultimate Tensile Strength Measurements

I'm working at a Materials company and I have measured the UTS for a new material to have a $\mu = 303\text{MPa}$ and the $\sigma = 33\text{MPa}$. What is the probability I find a sample that exceeds a UTS of 350MPa.

Voltage Measurements

I'm measuring the voltage in my Helmholtz Coil apparatus and I find, after 150 exhausting measurements, that the population mean is 10V with a population standard deviation of 3.4. How many voltage measurements will fall between 10-15 volts?

Under Pressure Measurements

I'm measuring the pressure in a cylindrical pressure vessel and find that $\mu = 39\text{Pa}$ and $\sigma = 4\text{Pa}$. I took 120 measurements. How many measurements will fall between 35-45Pa? How many less than 35? How many less than 45?

Chapter 3: Probability Distributions and the Gaussian is shared under a [CC BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/) license and was authored, remixed, and/or curated by LibreTexts.

Chapter 4: Central Limit Theorem and Confidence Intervals for Large and Small Sample Sizes

So far we have been calculating confidence intervals assuming that we know the population mean and standard deviation. But this is often never possible to do. Instead, recall that we will take measurements or samples and attempt to get an estimate or an approximate measure of population mean and standard deviation from our sample measurements. And we also want to **infer the probability distribution of the population from the sample**. Let's start our journey on how we can accomplish this.

Here you can find a series of pre-recorded lecture videos that cover this content: <https://youtube.com/playlist?list=PL...A9jQYH4GX6Ojeg>

Microwalker Rolling Velocity

As an undergraduate I performed experiments with super-paramagnetic particles and measured the velocity of the rolling particles my advisor was extremely and thorough to his undergraduate researchers so I had to measure the velocity of 138 particles. We predicted that the walkers would exhibit a mean velocity of $3 \frac{\mu m}{s}$ and that the walker velocity should not deviate by more than $0.5 \frac{\mu m}{s}$. The velocity measurements can be seen in the Mathematica Notebook below

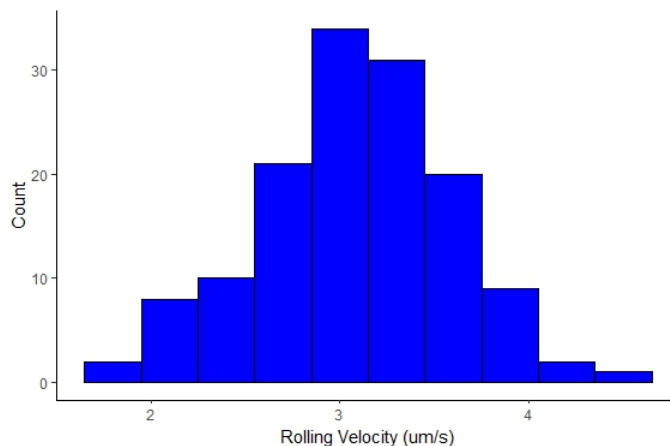


Figure Chapter4.1: Rolling Velocity of Microwalkers in $\frac{\mu m}{s}$.

So can we prove at a 95% confidence interval that the walker velocity will not deviate more than $0.5 \frac{\mu m}{s}$?

Well we have a lot of data so let's take a look at the data to make sure that our implicit assumption that we are working with a normally distributed population, i.e. Gaussian distribution. Let's plot the data in a histogram. **Does this distribution look Gaussian?** You may be wondering if there is a quantitative method to determine whether a distribution is Gaussian or not and the short answer yes and we will get to this in a few lectures.

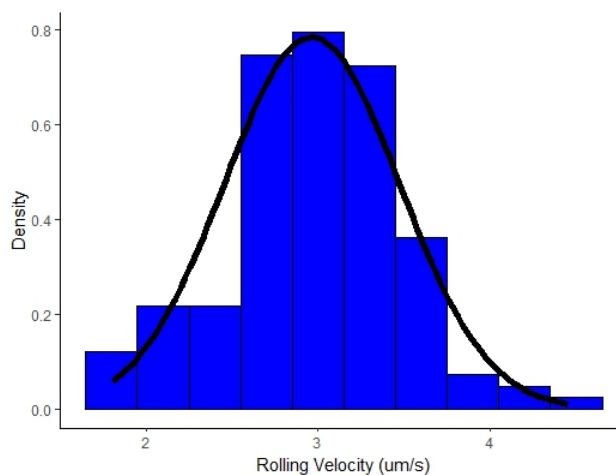


Figure Chapter4.2: Microwalker Rolling Velocity Histogram.

Let's calculate the sample mean and standard deviation and we find that

$$\bar{x} = 3.041031 \quad (\text{Chapter 4.1})$$

$$S_x = 0.5176254 \quad (\text{Chapter 4.2})$$

So you can see as anticipated the sample mean and standard deviation are slightly different from the assumed population mean and standard deviation.

So to see if we do not deviate more than $0.5 \frac{\mu\text{m}}{\text{s}}$ we can simply approximate the uncertainty in the population mean as follows:

$$\mu \pm 1.96\sigma \quad (\text{Chapter 4.3})$$

Since we are working with a large sample size, $n > 30$ we can use the following approximation that $x \approx \mu$ and that $S_x \approx \sigma$ so that we find that

$$\mu \pm 1.96\sigma \approx \bar{x} \pm 1.96S_x = 2.94 \pm 1.014546 \quad (\text{Chapter 4.4})$$

So we will in fact deviate more in this experiment.

Can we disregard $1.6 \frac{\mu\text{m}}{\text{s}}$ as an illegitimate data point or outlier? Justify.

Central Limit Theorem

We just examined the dispersion of sample values around the mean value of the sample, $\bar{x}$. But **we did not obtain an estimate for the uncertainty in $\bar{x}$ approximated to the true mean μ .** To do this we would have to **repeat the experiment test and compare the $\bar{x}$ of each sample test.** Thus we would also obtain a set of samples for the mean.

One of the most important and profound theories in statistics deals with this set of sample means and their distribution and is referred to as the **Central Limit Theorem**. So far we have been working on the assumption that the population that we are working with is normally distributed. While many times that may be true it is not always the case. Let think for instance if we were working with a uniform sample distribution like the one below

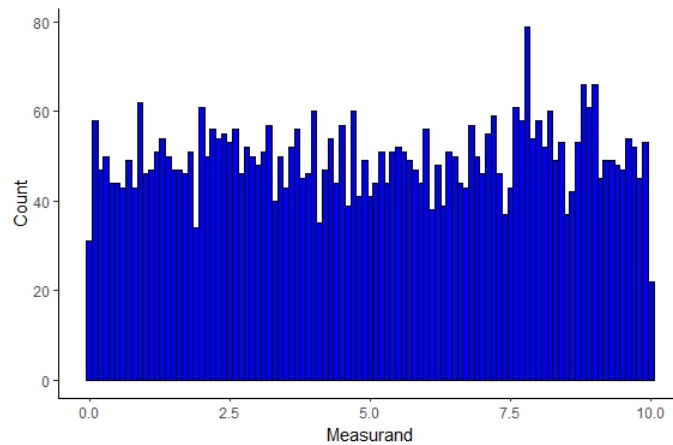


Figure Chapter4.3: Uniform Sample Distribution.

At this point one might think we must give up and we cannot perform any data analysis...noooo we can make it.

Instead the **Central Limit Theorem** postulates a very powerful idea that regardless of the shape of the population distribution the distribution of the mean values of a sample will be normally distributed as long as you obtain a large value of means, $n > 30$. We can see this visually with an example in the Mathematica Notebook for this lecture. If we randomly pick 6 samples from this distribution and average if we look at 2 or 10 averages as seen below the distribution do not look Gaussian but if we select 30 or more (sometimes 20 is enough but I would obtain 30 just to be safe) then we see the distribution of the mean measurand values becomes Gaussian!!!

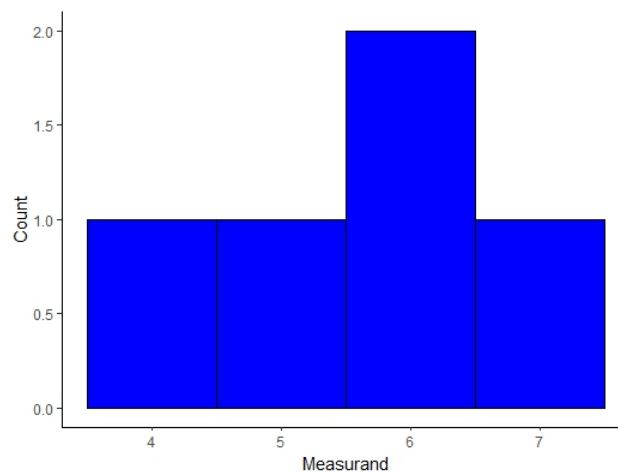


Figure Chapter4.4: Uniform Distribution Mean Sample Distribution with 5 Mean Samples.

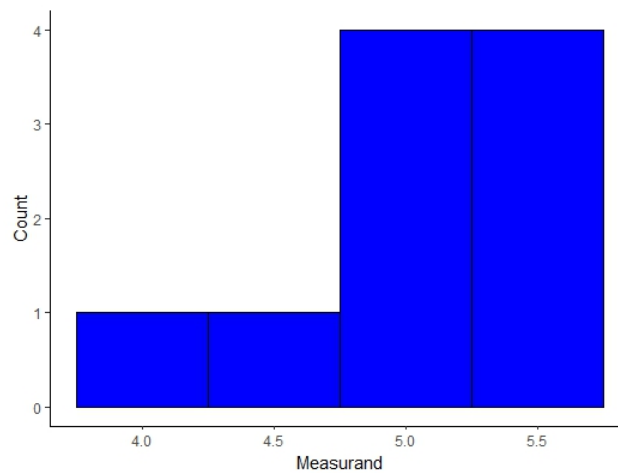


Figure Chapter4.5: Uniform Distribution Mean Sample Distribution with 10 Mean Samples.

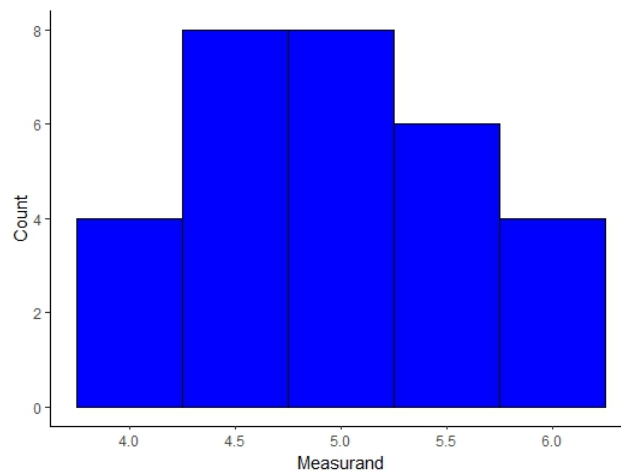


Figure Chapter4.6: Uniform Distribution Mean Sample Distribution with 30 Mean Samples.

This has implications for the analysis we can perform but let's prove this and show that it works for other types of distributions like an Exponential, χ^2 , and β . As you can see below the distribution of the means are clearly normal or Gaussian Distributions.

The practical application that is of interest to us as experimentalists is that we do not need to think about or worry about the shape of the population distribution. Since we know that the sample means are normally distributed we can simply use the normal distribution of the means to calculate confidence intervals and do more complex analysis and comparisons moving forward...you are so lucky t-test comparisons and hypothesis testing are coming soon!!!

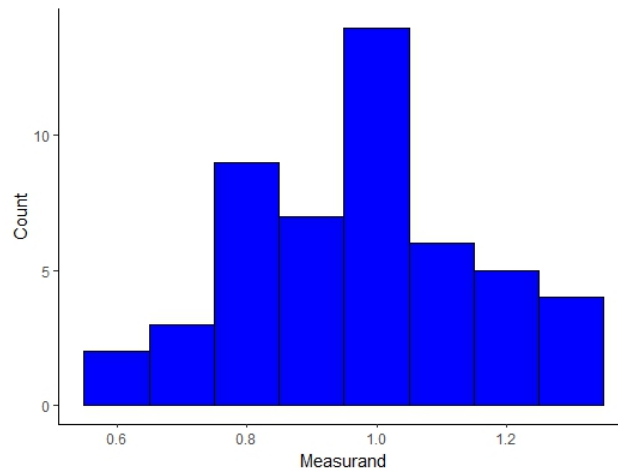


Figure Chapter4.7: Mean Distribution of 50 Samples for Exponential Distributions.

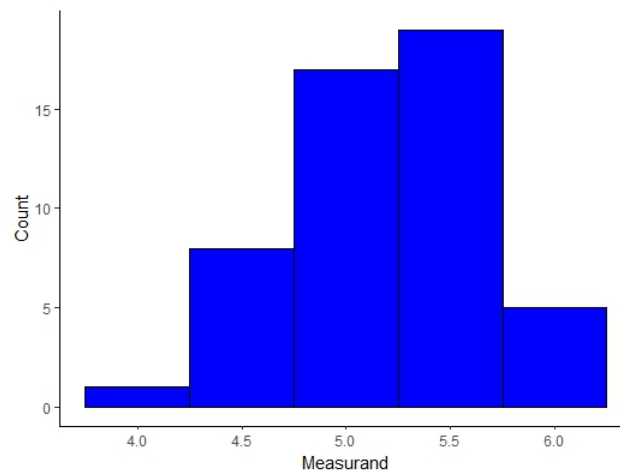


Figure Chapter4.8: Mean Distribution of 50 Samples for Poisson Distributions.

So when we invoke the central limit theorem we can make the following definitions where the standard deviation of the distribution of the means is

$$S_{\bar{x}} = \frac{S_x}{\sqrt{n}} \quad (\text{Chapter 4.5})$$

We can use the standard deviation of the means to now re-write our PDF or Gaussian function so z now becomes when describing the distribution of the means

$$z = \frac{\bar{x} - \mu}{S_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{S_x}{\sqrt{n}}} \quad (\text{Chapter 4.6})$$

This new definition of z allows us to re-write our confidence interval equation as follows but now we can explicitly determine the confidence level or uncertainty that in the population mean!!!

$$\mu = \bar{x} \pm z_{\frac{\alpha}{2}} \frac{S_x}{\sqrt{n}} \quad (\text{Chapter 4.7})$$

Whew!! That was a lot of work now let's go back to our rolling particles and calculate the 95% confidence interval for the population mean pressure.

Well that will simply be

$$\mu = \bar{x} \pm z_{\frac{\alpha}{2}} \frac{S_x}{\sqrt{n}} = 3.041031 \pm 0.08636387 \quad (\text{Chapter 4.8})$$

As you can see this is a much much much smaller range than our previous calculation due to the $\sqrt{\frac{1}{n}}$ factor. This is the key difference when trying to estimate the uncertainty in the population mean or using $\bar{x}$ as an estimate of the population mean. Previously we were just looking at the likelihood of observing a value that deviates from the population mean by a particular value.

In summary the c% interval for the mean value is narrower than the data by a factor of $\frac{1}{\sqrt{n}}$. This is because n observations have been used to average out the random deviations of individual measurements.

Confidence Intervals for Small Samples

We have been working with **large samples** thus far which is generally considered to be when $n \geq 30$ but there are many experiments when n will often be less than 30. For these situations we will utilize the **Student's t-distribution**, developed by an amateur statistician writing under the pseudonym Student:

$$t = \frac{\bar{x} - \mu}{\frac{S_x}{\sqrt{n}}} \quad (\text{Chapter 4.9})$$

Student was actually William Gosset a Guinness brewer who was applying statistical analysis to the brewing process. Guinness was against him publishing the data and thus giving away the Guinness secrets so he published under the pseudonym.

The assumption here is that the **underlying population satisfies the Gaussian distribution**. This distribution also depends on the **degrees of freedom**, $\nu = n - 1$.

t-distribution is similar to PDF, it is symmetric and the total area is unity. Moreover, the **t-distribution approaches the standard Gaussian PDF as ν or n becomes large** and for $n > 30$ the two distributions are identical as will be seen in the t-table soon.

The area beneath the t-distribution is tabulated at the end of these notes

The **area**, α , is between t and $t \rightarrow \infty$ and for a given sample sizes. This is very different from our Z-table where we were looking at the area between $z=0$ and z . α **corresponds to the probability** that for a given sample size, **t will have a value greater than that given in the table**. We can assert with $c\% = (1-\alpha)$ that the actual value of t does not fall in the shaded area.

A two sided confidence interval is then:

$$\bar{x} - t_{\frac{\alpha}{2}, \nu} \frac{S_x}{\sqrt{n}} < \mu < \bar{x} + t_{\frac{\alpha}{2}, \nu} \frac{S_x}{\sqrt{n}} \quad (\text{Chapter 4.10})$$

Sometimes α in literature will be referred to as the *level of significance*.

Designing Phase Change Materials

You are weighing out polymer samples to be integrated into a phase change material application to regulate temperature in textile applications. You are given 18 samples from your lab partner which should nominally be 2mg. You weight them and obtain the following results.

PCM
1.93614
1.969882
1.982175
2.051275
2.014028
2.027404
1.990602
2.070968
1.997488
2.036573
1.997471
2.047112
1.940512
2.056797
2.058416
2.033321
1.963191
1.998852

Figure *Chapter4.9*: PCM Measurements.

Based on these sample and the assumption that the parent population is normally distributed, what is the 95% confidence interval for the population mean.

$$\mu = 2.009567 \pm 0.020387mg \quad (\text{Chapter 4.11})$$

Calculating Confidence Intervals in R

You can actually calculate confidence intervals in R directly using the `t.test` functionality and we can get the confidence interval as seen below

One Sample t-test

```
data: PCMmine
t = 207.97, df = 17, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 1.989181 2.029954
sample estimates:
mean of x
 2.009567
```

Figure *Chapter4.10*: T-Test PCM Measurement.

and you can see that our results match our theoretical calculations which is fantastic!!!

But what is it do we see here about hypothesis tests??

Well we will discuss this next lecture, same bat time same bat place!

Chapter 4: Central Limit Theorem and Confidence Intervals for Large and Small Sample Sizes is shared under a [CC BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/) license and was authored, remixed, and/or curated by LibreTexts.

Chapter 5: Hypothesis Testing and P-Values

Here you can find a series of pre-recorded lecture videos that cover this content: <https://youtube.com/playlist?list=PL...I4vAR1y-IP1Ees>

Hypothesis Testing for a Single Mean for a Small Sample Size

The reason we are using these statistical tools is to make certain decisions regarding a measurand. One of the most common methods for doing this is **hypothesis testing**.

Typically we deal with two hypotheses:

- **Null Hypothesis**

- First step in hypothesis testing
- $H_0 : \mu = \mu_0$ where μ_0 is some constant specific value

- **Alternative Hypothesis**

- Second step
- Choice should reflect on what we are attempting to show
- * Two-tailed test: concerned with whether a population mean, μ is different from specific value μ_0 , i.e. $H_a : \mu \neq \mu_0$
- * Left-tailed test: concerned with whether a population mean is less than a specific value, $H_a : \mu < \mu_0$
- * Right-tailed test: concerned with whether a population mean is greater than a specific value, $H_a : \mu > \mu_0$

Procedure for Hypothesis Testing

- (1) Define null hypothesis, H_0
- (2) Define alternative hypothesis, H_a
- (3) Define c% interval
- (4) Calculate the value of t_{exp} from the data
- (5) Determine proper value of $t_{\alpha, \nu}$ or $t_{\frac{\alpha}{2}, \nu}$ using the degrees of freedom ν
- (6) If t_{exp} falls in the reject H_0 region, we reject H_0 and accept the alternative hypothesis H_a
- (7) If t_{exp} falls in the do not reject H_0 region, we conclude that we do not have sufficient data to reject H_0 at the level of confidence specified

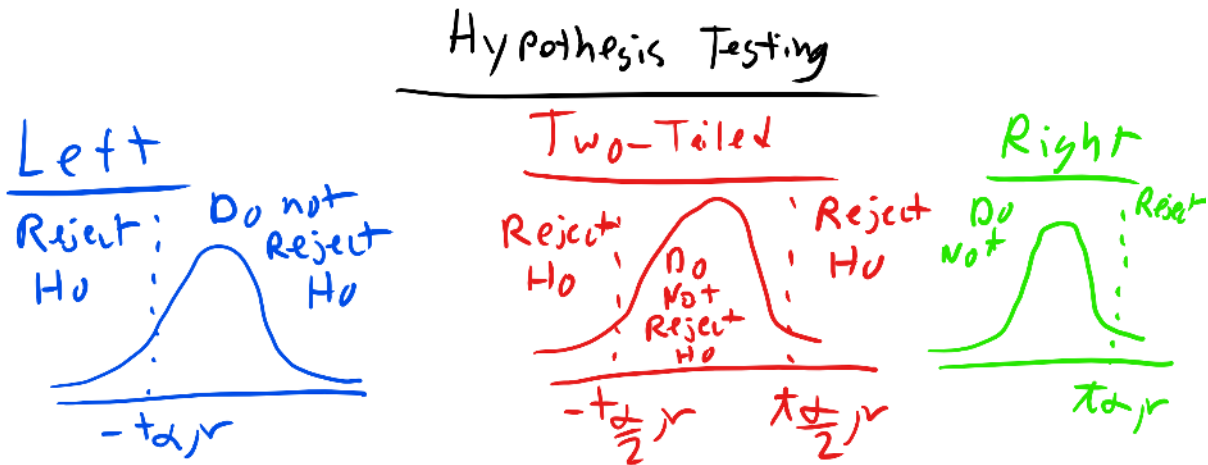


Figure Chapter 5.1: Rejecting null hypothesis for two, right, and left tailed test.

Let's start with a two-tailed test and ask ourselves the following question, does the PCM data from the example last lecture come from a population with a true mean of 2mg assuming a confidence level of 95%.

$$\begin{aligned}
 H_0 : \mu &= 2.00mg && \text{(Chapter 5.1)} \\
 H_a : \mu &\neq 2.00mg && \text{(Chapter 5.2)} \\
 t_{exp} &= \frac{\bar{x} - \mu_0}{\frac{s_x}{\sqrt{n}}} = 0.99011 && \text{(Chapter 5.3)} \\
 t_{\alpha, \nu} &= t_{0.025, 17} = 2.11 && \text{(Chapter 5.4)}
 \end{aligned}$$

We can look at our figure and we see that our t-value falls within the do-not reject H_0 region. Does this match with result from R???

I don't see any result only some mysterious thing called the P-value, but no worries we actually already intuitively know what this mysterious value is.

P-Values

Often if you have read scientific articles or if you have taken other statistics courses you may have heard of the term **P-Value**. This term is ubiquitous and is used much much more often than z or t-values. Simply put, the **P-value is the probability of getting a result that is more extreme than the value that is actually observed**. Let's see how it is used in the context of our previous hypothesis tests, starting with a two-tailed test.

Our P-value effectively gives us the probability of measuring a value greater than the observed, i.e. the tail ends. The way that we interpret the P-value is if we are running a two-tailed confidence interval at 5% confidence if the P-value is greater than 2.5% or 0.025 then we fall in the Do-Not Reject H_0 regime and if it less than that value then we fall in the Reject H_0 regime.

So we have previously found that we fall in the Do Not Reject region and looking at our calculated P-value of 0.336 that falls in the Do Not Reject region which matches our result!!!

Another PCM Application

Using the data from the previous example does the sample come from a population whose true mean weight is greater than 1.99mg, assuming a confidence level of 99%?

$$H_0 : \mu = 1.99mg \quad (\text{Chapter 5.5})$$

$$H_a : \mu > 1.99mg \quad (\text{Chapter 5.6})$$

$$t_{exp} = \frac{\bar{x} - \mu_0}{\frac{S_x}{\sqrt{n}}} = 2.025 \quad (\text{Chapter 5.7})$$

$$t_{\alpha, \nu} = t_{0.01, 17} = 2.567 \quad (\text{Chapter 5.8})$$

Clearly since it is a right tailed test it falls in the do not reject H_0 region. So we conclude with 99% confidence that the population mean was not significantly different than 1.99mg. Note the subtle distinction of the phrase was not significantly different than 1.99mg or we do not have sufficient data to reject H_0 with 99% confidence. We are **not** saying that the population mean is 1.99mg.

If we examine this from a P-value perspective we find that our P-value is 0.02943 and that also falls in the Do Not Reject Region as well!!

Yet Another PCM Application

Using the data from the previous example does the sample come from a population whose true mean weight is less than 2.01mg, assuming a confidence level of 90%?

$$H_0 : \mu = 2.01mg \quad (\text{Chapter 5.9})$$

$$H_a : \mu < 2.01mg \quad (\text{Chapter 5.10})$$

$$t_{exp} = \frac{\bar{x} - \mu_0}{\frac{S_x}{\sqrt{n}}} = -0.0448035 \quad (\text{Chapter 5.11})$$

$$t_{\alpha, \nu} = t_{0.01, 17} = -2.567 \quad (\text{Chapter 5.12})$$

Clearly since it is a left tailed test it does not fall in the do not reject H_0 region. So we conclude with 90% confidence that the population mean was not significantly different than 2.01mg.

If we examine this from a P-value perspective we find that our P-value is 0.4824 and that also falls in the Do Not Reject Region as well!!

Hypothesis Testing Single Mean for Large Sample Size

For larger sample sizes we follow the exact same procedure but we replace $t_{\alpha, \nu}$ by z_{α} and t_{exp} by:

$$z_{exp} = \frac{\bar{x} - \mu_0}{\frac{S_x}{\sqrt{n}}} \quad (\text{Chapter 5.13})$$

However, we can just use the t-table and use the value for degrees of freedom that corresponds to the $n > 30$ scenario i.e. $\nu \approx \infty$.

Looking Back at Rolling Velocity

Does the sample of our rolling magnetic beads come from a population with a velocity less than $3.1 \frac{\mu m}{s}$ at a confidence level of 99%.

$$H_0 : \mu = 3.1 \frac{\mu m}{s} \quad (\text{Chapter 5.14})$$

$$H_a : \mu < 3.1 \frac{\mu m}{s} \quad (\text{Chapter 5.15})$$

$$t_{exp} = \frac{\bar{x} - \mu_0}{\frac{S_x}{\sqrt{n}}} = -1.3383 \quad (\text{Chapter 5.16})$$

$$t_{0.01, Inf} = -2.362 \quad (\text{Chapter 5.17})$$

We can see we are in the Do Not Reject H_0 and we see that the P-Value is 0.09151 so we can confirm that we are in the Do Not Reject H_0 .

t-Test Comparison of Sample Means

To compare two samples solely based on their means:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{S_1^2}{n_1}\right) + \left(\frac{S_2^2}{n_2}\right)}} \quad (\text{Chapter 5.18})$$

where $\bar{x}_1, S_1$, and n_1 and $\bar{x}_2, S_2$, and n_2 are the mean, standard deviations, and sizes of the two samples. and the degrees of freedom can be approximated by:

$$\nu = \frac{\left[\left(\frac{S_1^2}{n_1}\right) + \left(\frac{S_2^2}{n_2}\right)\right]^2}{\frac{\left(\frac{S_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{S_2^2}{n_2}\right)^2}{n_2-1}} \quad (\text{Chapter 5.19})$$

and ν is rounded to nearest integer. If the value t falls into the interval $\pm t_{\frac{\alpha}{2}, \nu}$ then the two means are not significantly different at the chosen confidence level. One of the great things about this technique is that this equation is applicable for any combination comparing large and small samples sizes.

Are these Materials Significantly Stiffer

Consider Material A with a Young's Modulus average of 302.6 GPa, measured 12 times, and a sample standard deviation of 1.27 GPa and Material B with a Young's Modulus average of 302.3 GPa, measured 15 times, and a sample standard deviation of 1.7 GPa.

$$H_0 = \bar{x}_A = \bar{x}_B \quad (\text{Chapter 5.20})$$

$$H_a = \bar{x}_A \neq \bar{x}_B \quad (\text{Chapter 5.21})$$

$$\nu \approx 25 \quad (\text{Chapter 5.22})$$

$$t_{exp} = 0.547 \quad (\text{Chapter 5.23})$$

$$t_{0.025, 25} = 2.060 \quad (\text{Chapter 5.24})$$

The value falls in the do not reject region so there is not a significant difference in the stiffness of Material A and B.

We can also perform a P-value analysis to a t-Test comparison of different samples as well you will see the R code can be simply adapted to examine different samples. Let's simulate the example here and we find a P-value of 0.02194 so for this simulated example that would be a reject finding, again the difference here would be due to the simulated number values, as we can confirm the means and standard deviation values are slightly different. In fact if we were to re-run the analysis using our method by hand we can achieve the same result. Let's try that really quickly.

Chapter 5: Hypothesis Testing and P-Values is shared under a [CC BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/) license and was authored, remixed, and/or curated by LibreTexts.

Chapter 6: Goodness of Fit Hypothesis Testing

Here you can find a series of pre-recorded lecture videos that cover this content: <https://youtube.com/playlist?list=PLSKNWIZCsmjA4X-lkFzKySeumT1S5t3om&si=vYb0kaHmSIrHBBGn>

So far we have discussed Gaussian/normal distributions, t-distributions, and some other distributions briefly. Another useful and common distribution you may come across is the **chi-square (χ^2)** distribution. This distribution is distinct from the ones that we have dealt with previously as you can see below first we can see it has the shape of a **right-skewed curve**.

The **area under the curve is again 1.0** same as before and as the ν **becomes large** the distribution **approaches a symmetric distribution resembling the normal distribution**. χ^2 is often referred to as a goodness of fit test statistic is:

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \quad (\text{Chapter 6.1})$$

where O_i is the observed data, E_i is the expected data, k is the total number of variables being compared, and $\nu = k - 1 - r$, where r is the number of fitting parameters...more on this later. We will see in several examples that you can also think of this k as counting the number of possible outcomes. **k IS NOT THE NUMBER OF SAMPLES**. It is critical to remember that k is not n . The χ^2 is a right tailed test for hypothesis testing since χ^2 is a positive value. Also when the **observed data approaches the expected data the value of χ^2 approaches zero**.

Let's do a quick example right here to get a sense of when and how we can utilize this distribution with an example.

Flipping coins

Let's start off with flipping a coin. I asked one of my undergraduate research students to flip a coin 500 times as I was interested to see if my coin had any weight imbalance that would lead to a non-stochastic result in coin flipping, I know I am very mean. But my student flipped the coin 500 times and found 286 heads and 214 tails. **Is this coin fair or true using a confidence level of 95%?**

So to start with this problem I want to first point out this issue of fair or true. That just means that the probabilities or outcomes are what we expect. For example flipping a coin should be stochastic or random and thus the probability of getting heads or tails should be 50:50 if we see a significant deviation from our hypothesis test then we know that something has gone wrong. So let's get started first what is k here?

It is not 500! The number of possible outcomes is the coin will be either head or tails so $k = 2$. For the rest we just follow our hypothesis testing procedure

$$H_0 : \text{Frequencies are same} \quad (\text{Chapter 6.2})$$

$$H_a : \text{Frequencies are different} \quad (\text{Chapter 6.3})$$

$$\chi_{exp}^2 = 10.368 \quad (\text{Chapter 6.4})$$

$$\chi_{0.05,1}^2 = 3.841 \quad (\text{Chapter 6.5})$$

We are in the **reject region** so **these coins are not fair**, see I was right.

Note again that for this distribution we will always be dealing with a right tailed test due to the shape and nature of the distribution.

Fun with Random Numbers

As an expert *Mathematica* coder you generate 300 random numbers in the range from 0-10 and you get the following seen in the Mathematica Notebook

```
rand = {7, 6, 7, 6, 2, 5, 10, 5, 9, 7, 9, 1, 10, 3, 8, 0, 7, 9, 2, 7, 5, 2, 7, 3, 10, 5, 8, 10, 6, 4, 9, 7, 5, 7, 7, 4, 7, 9, 1, 1, 2, 0, 2, 6, 8, 2, 5, 1,
5, 4, 4, 9, 0, 3, 1, 8, 5, 7, 9, 10, 10, 8, 3, 5, 6, 6, 5, 10, 9, 4, 9, 9, 2, 4, 8, 6, 8, 8, 3, 6, 3, 7, 8, 8, 1, 3, 8, 7, 4, 8, 6, 8, 10, 8, 8, 1, 4,
6, 4, 8, 3, 10, 1, 4, 9, 2, 7, 6, 8, 8, 10, 10, 3, 1, 8, 1, 4, 6, 9, 10, 10, 5, 8, 6, 0, 5, 9, 5, 5, 0, 1, 5, 4, 5, 5, 6, 5, 5, 6, 1, 9, 2, 4, 4, 8, 8,
7, 9, 4, 6, 7, 5, 3, 9, 10, 10, 5, 1, 7, 1, 2, 6, 5, 9, 0, 2, 5, 1, 4, 2, 6, 7, 0, 6, 1, 5, 2, 8, 7, 7, 1, 0, 1, 9, 7, 5, 4, 4, 5, 6, 9, 0, 5, 1, 9,
7, 7, 10, 4, 5, 3, 2, 4, 4, 6, 1, 6, 5, 4, 4, 7, 7, 1, 1, 7, 0, 7, 3, 1, 5, 2, 2, 10, 7, 1, 1, 7, 1, 10, 8, 1, 3, 1, 3, 0, 4, 8, 3, 9, 3, 4, 6, 5, 1,
2, 1, 3, 1, 6, 1, 10, 10, 4, 7, 0, 7, 3, 8, 4, 10, 1, 5, 3, 0, 8, 5, 7, 7, 4, 8, 2, 9, 5, 10, 1, 2, 0, 4, 4, 10, 1, 2, 2, 0, 1, 10, 9, 1, 2, 8, 6, 3, 0, 5, 10, 4, 8, 1, 8};
```

Figure Chapter6.1: Random Number Generator Distribution 0-10.

Does the distribution differ significantly from the expected distribution at the 1% significance level?

$$\begin{aligned}
 H_0 : \text{Frequencies are same} & \quad (\text{Chapter 6.6}) \\
 H_a : \text{Frequencies are different} & \quad (\text{Chapter 6.7}) \\
 \chi^2_{exp} = 17.753 & \quad (\text{Chapter 6.8}) \\
 \chi^2_{0.01,10} = 23.209 & \quad (\text{Chapter 6.9})
 \end{aligned}$$

We **do not reject that hypothesis**. We can always trust in *Mathematica*! Note: Mention anecdote about *Mathematica* random number failing.

Why Am I So Old No Student Has Seen Casino

Ace Rothstein suspects a player at his casino of cheating so he asks you to test if his dice are **true** and he gives you the following data set for the outcome of rolling two 6 sided die. Ace wants to make sure there is no cheating so you can see he ran this experiment 800 times, he appreciates large data sets to work with as seen below

Face-Up Sum	Observed Frequency	Expected Frequency
2	8	10
3	25	20
4	29	30
5	42	40
6	53	50
7	55	60
8	46	50
9	39	40
10	29	30
11	22	20
12	12	10

Figure Chapter6.2: Dice Roll Distributions.

Determine if the dice are true at a 99% confidence level.

$$\begin{aligned}
 H_0 : \text{Frequencies are same} & \quad (\text{Chapter 6.10}) \\
 H_a : \text{Frequencies are different} & \quad (\text{Chapter 6.11}) \\
 \chi^2_{exp} = 297.826 & \quad (\text{Chapter 6.12}) \\
 \chi^2_{0.01,10} = 23.209 & \quad (\text{Chapter 6.13})
 \end{aligned}$$

So as you can see we fall in the reject region....uh oh!

Are My Rolling Velocities Normally Distributed?

You cannot escape my rollers!! Let's go back to our data set and see if our rolling velocities are normally distributed.

If you remember we found that sample mean rolling velocity was $2.938 \frac{\mu m}{s}$ with a standard deviation of 0.376 and we measured 138 samples. We find that there are 15 measurements with a velocity greater than 3.4 same units, 43 less than 2.8. We also find that there are 37 measurements between 2.2 and 2.8 and finally 136 measurements between 2 and 4.

Are the velocities normally distributed at the 5% significance level?

As usual we need to find what the expected values if they were normally distributed. To do this we need to go back to our old friend the Z-Table and calculate z values and look up areas. So let's break this problem up.

1. Greater than 3.4

$$\begin{aligned}
 z &= \frac{3.4 - \bar{x}}{S_x} = 1.23 & (\text{Chapter 6.14}) \\
 A_1 &= 0.5 - 0.3907 = 0.1093 & (\text{Chapter 6.15}) \\
 E_1 &= n \cdot A_1 \approx 15 & (\text{Chapter 6.16})
 \end{aligned}$$

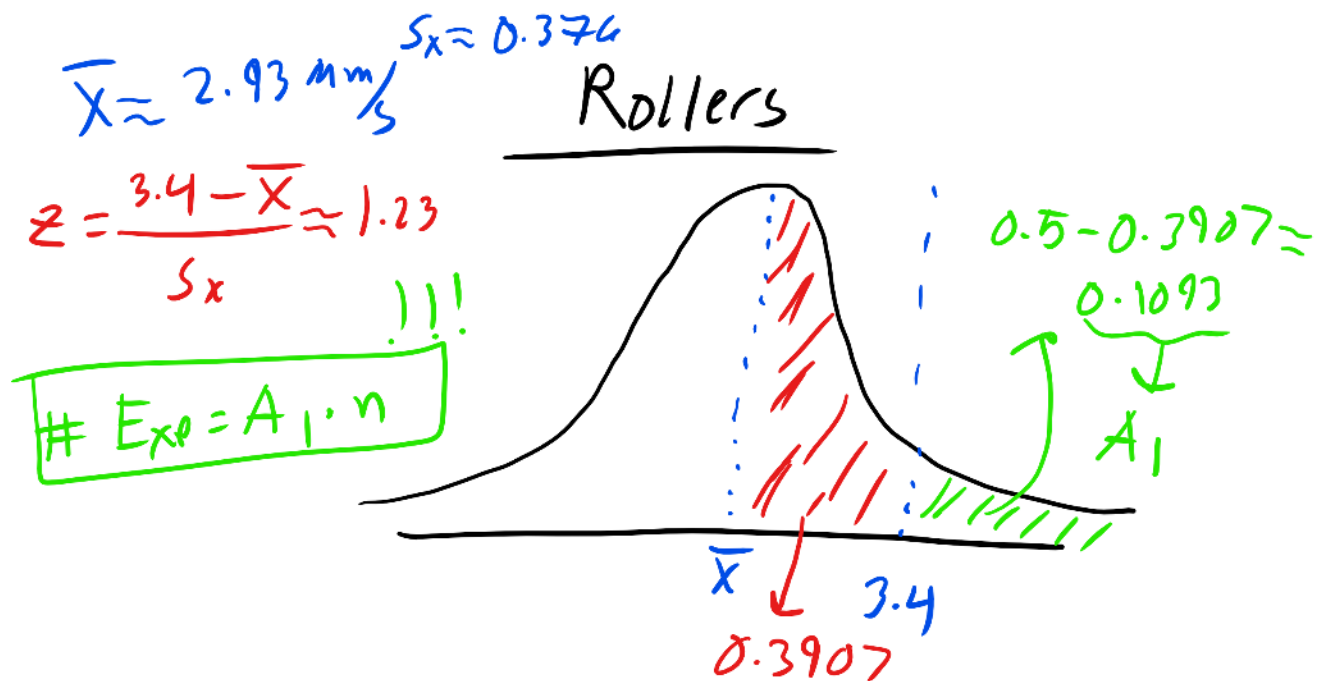


Figure Chapter6.3: Solving for Area 1.

2. Less than 2.8

$$z = \frac{2.8 - \bar{x}}{S_x} = -0.37$$

(Chapter 6.17)

$$A_2 = 0.5 - 0.1443 = 0.3557$$

(Chapter 6.18)

$$E_2 = n \cdot A_2 \approx 49.1$$

(Chapter 6.19)

3. Between 2.2 and 2.8

$$z = \frac{2.2 - \bar{x}}{S_x} = -1.96$$

(Chapter 6.20)

$$A_3 = 0.475 - 0.1443 = 0.3307$$

(Chapter 6.21)

$$E_3 = n \cdot A_3 \approx 45.6$$

(Chapter 6.22)

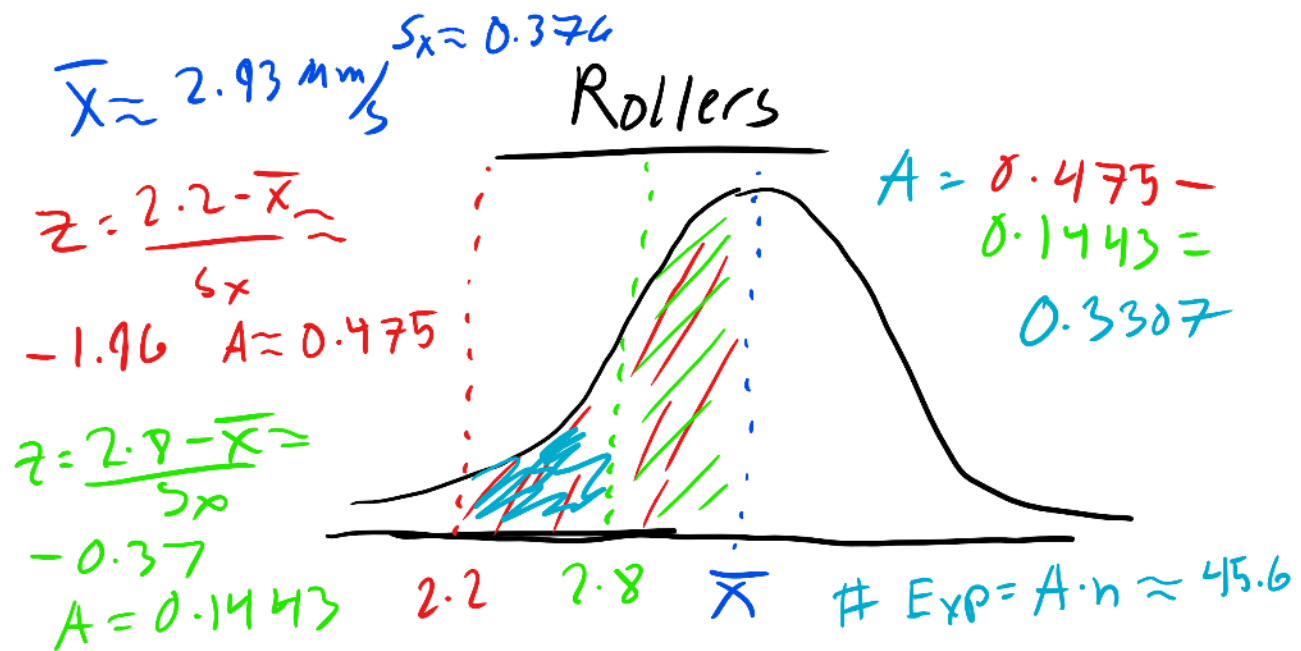


Figure Chapter6.4: Solving for Area 3.

4. Between 2 and 4

$$z = \frac{2 - \bar{x}}{S_x} = -2.49$$

(Chapter 6.23)

$$z = \frac{4 - \bar{x}}{S_x} = 2.82$$

(Chapter 6.24)

$$A_4 = 0.4936 + 0.4976 = 0.99$$

(Chapter 6.25)

$$E_4 = n \cdot A_4 \approx 136.7$$

(Chapter 6.26)

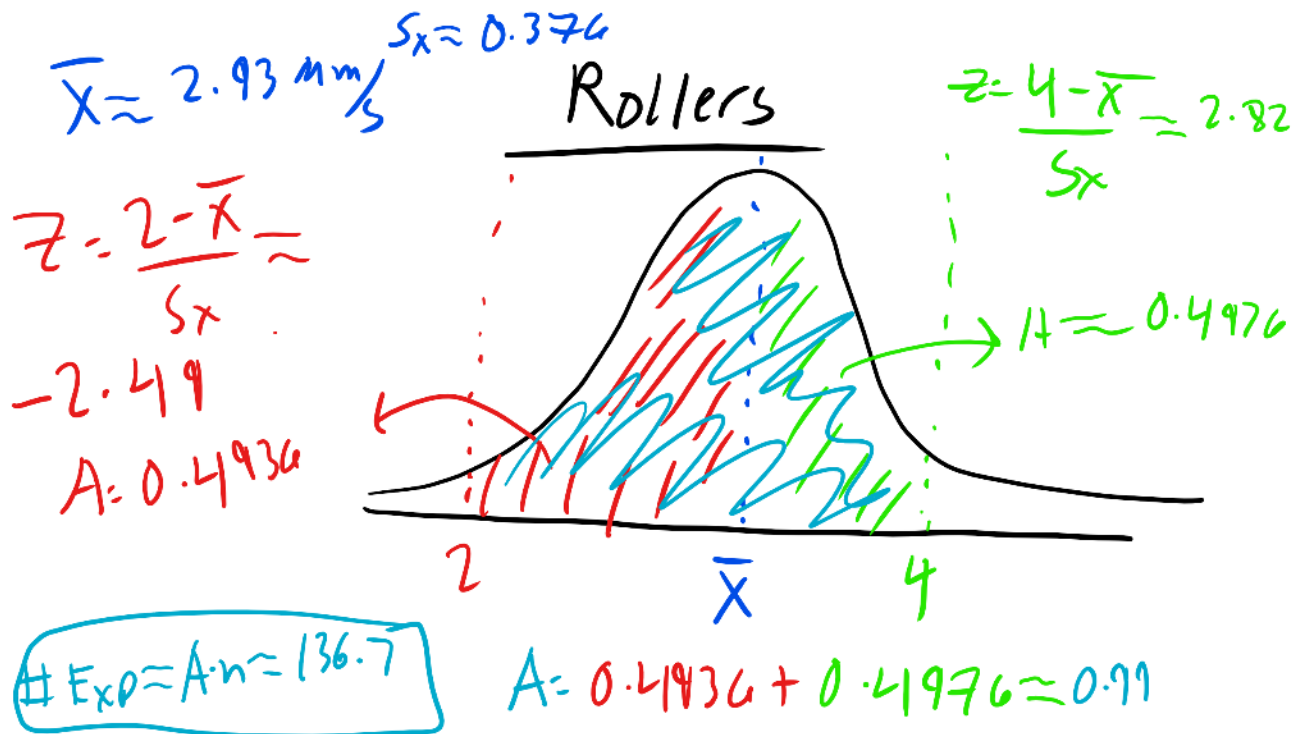


Figure Chapter6.5: Solving for Area 4.

Whew....that was quite a bit of work but now we can actually go ahead and perform our analysis via hypothesis testing but first let us figure out the degrees of freedom, v . We know the possible outcomes or variables k we are dealing with is 4 since we have 4 bins in our histogram. Now the question is what is r . Well when we are talking about a distribution or fitting r is the number of parameters used to fit the distribution, so when looking back at the Gaussian equation we fit based on values we calculate for μ and σ . If we were working with a 20th order polynomial fit r would be 20 but in this example for Gaussian distributions we have $r = 2$ thus we know that $v = 1$.

H_0 : Frequencies are same

(Chapter 6.27)

H_a : Frequencies are different

(Chapter 6.28)

$$\chi^2_{exp} = 2.4$$

(Chapter 6.29)

$$\chi^2_{0.05,1} = 3.841$$

(Chapter 6.30)

So **We do not reject**.

Goodness of Fit

Most data is Gaussian however there are certain data sets that are not approximated well by the Gaussian distribution. The most well known example begins perhaps **fatigue strength data which is approximated using the Weibull distribution**. So some estimate of the **goodness of fit** should be made before any critical decisions are based on statistical error calculations.

There is **no absolute or cure all check** in producing a figure of merit, at best a qualifying confidence level must be applied with final judgment left to the experimentalist. The quickest and easiest method is to *eyeball* the results but this is problematic and very subjective. Another method would be to graphically check using a *normal probability plot*.

Yet another method is to test for **goodness of fit based on the χ^2 distribution** although this requires considerable data manipulation and does not lend itself well to small samples of data.

Some limitations of this technique are that originally determined values must be numerical counts and integer frequencies, frequency values, should be equal to or greater than unity and there should be no unoccupied bins, and the use of χ^2 is questioned if 20% of the values in either the expected or original bins have counts less than 5.

Now you may have hear of another χ^2 test that gives you information about how variables can be correlated, don't worry we will get to that later.

Chapter 6: Goodness of Fit Hypothesis Testing is shared under a [CC BY-NC-SA 4.0](#) license and was authored, remixed, and/or curated by LibreTexts.

Chapter 7: Total Uncertainty, Bias Error, and Propagation of Error

Here you can find a series of pre-recorded lecture videos that cover this content: <https://youtube.com/playlist?list=PLSKNWlZCsmjAmuzoj5TCkHKIAJGx8gWZu&si=wV67ID1o8Gp2ihNH>

Bias and Single-Sample Uncertainty

Precision error can be calculated using **statistical methods** for repeat-sampled data however **bias error cannot** be uncovered using **statistical methods**. The only direct method is via comparison with measurements made using another more accurate apparatus. However, this is practically impossible. It is much too costly, time consuming, and I am going to show you another method to estimate the bias error right now. Instead we will utilize our **experience and knowledge of the device** and make an estimate on the **sizes of bias errors**. Then we will obtain an estimate of the precision and accuracy of our experimental apparatus and determine if we must re-design our apparatus to measure a particular quantity of interest.

Bias Error: Measuring Volume in a cylinder

What measurements do we need to measure the volume of a cylinder? Do you remember geometry?

The equation for the volume of a cylinder is

$$V = \pi r^2 l \quad (\text{Chapter 7.1})$$

where r is the radius of the cylinder and l is the length of the cylinder. When thinking about the uncertainty of these measurements there are many things that may come to mind but looking at the equation above we want to identify the number of independent variables as that is where uncertainty will originate from. So in the above equation we will have uncertainty in r and l . Now of course when we calculate volume we will have to measure r and l and there may be precision error in these measurements due to temperature fluctuation and thermal expansion and we know how to quantify precision error. However, there will also be **bias error** present as well. Bias error will manifest itself in the type of device utilized to measure these quantities. For example if you are using a yard stick with 1 yard intervals that will have a much larger uncertainty than a micrometer, or some type of quantum laser technique...we unfortunately do not have that available in lab.

Propagating Uncertainty

Usually we **measure several quantities** and use those measurements to **calculate a desired quantity** just like we saw in volume. Each measurement has some **associated uncertainty** which will create an **uncertainty in the calculated result**. **How do we quantify that uncertainty?**

We propagate the uncertainty!

We will utilize a statistical theorem to propagate uncertainty in independent variables. Let's consider a linear function y of several **several independent variables** x_i with standard deviation σ_i .

The standard deviation of this equation would be

$$\sigma_y = \sqrt{\left(\frac{\partial y}{\partial x_1} \sigma_1\right)^2 + \left(\frac{\partial y}{\partial x_2} \sigma_2\right)^2 + \dots + \left(\frac{\partial y}{\partial x_n} \sigma_n\right)^2} \quad (\text{Chapter 7.2})$$

Each variable x_i will have some associated uncertainty u_i which will then contribute to the total bias uncertainty in y which we will call u_y . In order to estimate u_y we will make a key assumption that each uncertainty is relatively small (reasonable) so that we can perform our favorite operation! A first order Taylor expansion of $y(x_i)$

$$\begin{aligned} & y(x_1 + u_1, x_2 + u_2, \dots, x_n + u_n) \\ & \approx y(x_1, x_2, \dots, x_n) + \frac{\partial y}{\partial x_1} u_1 + \frac{\partial y}{\partial x_2} u_2 + \dots + \frac{\partial y}{\partial x_n} u_n \end{aligned} \quad (\text{Chapter 7.3})$$

Thus y is a linear function of independent variables and the **uncertainty will behave like standard deviations**:

$$u_y = \sqrt{\left(\frac{\partial y}{\partial x_1} u_1\right)^2 + \left(\frac{\partial y}{\partial x_2} u_2\right)^2 + \dots + \left(\frac{\partial y}{\partial x_n} u_n\right)^2} \quad (\text{Chapter 7.4})$$

The key thing for this theorem to hold true and so that we can utilize this equation is that all uncertainties must be **independent of each other** and have the same odds.

Let's do some examples to make this clear

Volume of a Cylinder

A tube of circular cross section has a nominal length of $50 \text{ m} \pm 0.5 \text{ m}$ and a radius of $15 \text{ m} \pm 0.1 \text{ m}$. Determine the uncertainty in the volume.

So when we first approach this problem we must identify if there is precision error, bias error, or both. So when we read this problem if we want to identify precision error do we have any information about the number of samples? A list of measurements? Or anything else?

Well the answer in this problem is no so we have no precision error contribution, something to note when you are reporting your results in a lab report.

For identifying the bias error do we know that tool we were using to measure the independent variables, manufacturer's information, or if we are really luckily we are just given the uncertainties.

Well in this problem we are really lucky we are just given the uncertainties so we just have deal with bias error in calculating the total uncertainty!

The next step we should do is to identify what we need to measure, the equation required, and the number of independent variables.

In this problem

$$V = \pi r^2 l \quad (\text{Chapter 7.5})$$

So we have a function $V(r, l)$ that is a function of two independent variables r and l both of which will have some uncertainty associated with the measurement as seen below

So let's proceed with our solution to calculate the bias error in V , B_V we have to first take our differential

$$B_V = \sqrt{\left(\frac{\partial V}{\partial r} u_r\right)^2 + \left(\frac{\partial V}{\partial l} u_l\right)^2} \quad (\text{Chapter 7.6})$$

where we know from our problem statement that $u_r = 0.1 \text{ m}$ and that $u_l = 0.5 \text{ m}$.

And we also know that

$$\frac{\partial V}{\partial r} = 2\pi r l \quad (\text{Chapter 7.7})$$

$$\frac{\partial V}{\partial l} = \pi r^2 \quad (\text{Chapter 7.8})$$

We can then calculate B_V as seen below by plugging in the **nominal or mean** values for r and l as well as the uncertainties associated with them.

So we find

$$B_V = 589.05 \text{ m}^3 \quad (\text{Chapter 7.9})$$

and since in this problem we have no precision uncertainty we find that the total uncertainty in volume, U_V is just B_V as seen below

$$U_V = \sqrt{B_V^2 + P_r^2 + P_l^2} = B_V \quad (\text{Chapter 7.10})$$

where $P_r = P_l = 0$ since there is no precision error in this experiment.

It is critical when performing this analysis to keep your units consistent, this is where most people will make mistakes always always always convert to SI units.

Now often it is more useful to express your answers in terms of a fraction or a percent we can do this simply by dividing our differentials and B_V equation above and re-write as

$$\frac{B_V}{V} = \sqrt{\left(\frac{\frac{\partial V}{\partial r} u_r}{V}\right)^2 + \left(\frac{\frac{\partial V}{\partial l} u_l}{V}\right)^2} \quad (\text{Chapter 7.11})$$

We can clearly see here that

$$\frac{\frac{\partial V}{\partial r} u_r}{V} = \frac{2u_r}{r} \quad (\text{Chapter 7.12})$$

and that

$$\frac{\frac{\partial V}{\partial l} u_l}{V} = \frac{u_l}{l} \quad (\text{Chapter 7.13})$$

Plugging this back into our equation we will find that now

$$B_V = 0.01667 \quad (\text{Chapter 7.14})$$

This is a unitless fraction and you should check that this is the case for yourself if we want to convert this fraction to a percent we must simply multiply by 100 in this problem the nominal volume is 35343 m^3 so to double check our answer on the bias uncertainty we can find that if we multiply our fraction by the nominal volume we indeed get that the uncertainty is 589.05 m^3 . This is a nice way to double check your answer as the problems get more complex from here.

Let's do another example problem focusing on material mechanics

Uniaxial Tension Test

I am trying to measure the stress in a rectangular block of aluminum under uniaxial tension. I repeated the experiment 14 times and find the force to exhibit a mean of 48.9N and a standard deviation of 3.24. I also measured the length of the block and width 14 times and found a mean width of 0.1m and a standard deviation of 0.01m, the length had a mean of 0.5m and a standard deviation of 0.02m. I measured the length and width with a micrometer and the force applied is $\pm 1\text{N}$. What is the total uncertainty in stress?

So let's break down this problem will we have precision error, bias error, or both?

We will have both precision error and bias error in this problem. Precision error comes from the multiple measurements made in this problem and we are also given standard deviations as well. We will also have bias error because we are given an uncertainty in force plus the tool used to measure length and width.

What is the equation we are working with?

Well since we are dealing with uniaxial tension we can use Hooke's law, we will talk about tensor notation much much more in depth soon so

$$\sigma = \frac{F}{A} = \frac{F}{l \cdot w} \quad (\text{Chapter 7.15})$$

So we have a function $\sigma(F, l, w)$ that is a function of 3 independent variables so we will have to take 3 partials.

So at this point we have some options in terms of how to approach this problem. You can go directly to the partials but when you have precision error I would suggest calculating precision error first, expressing this uncertainty in percent, calculating bias error in percent, and then calculating total uncertainty in percent.

So to start with precision we have precision error in all 3 independent variables so we can calculate the precision error as we have done previously

$$P_F = t_{0.025,13} \frac{S_F}{\sqrt{n}} \quad (\text{Chapter 7.16})$$

$$P_l = t_{0.025,13} \frac{S_l}{\sqrt{n}} \quad (\text{Chapter 7.17})$$

$$P_w = t_{0.025,13} \frac{S_w}{\sqrt{n}} \quad (\text{Chapter 7.18})$$

But as I mentioned we want to work with percentages so we need to divide these values by the nominal value i.e. $\frac{P_F}{F}$, $\frac{P_l}{l}$, $\frac{P_w}{w}$.

So those values can be expressed as a dimensionless fraction as

$$\frac{P_F}{F} = 0.04 \quad (\text{Chapter 7.19})$$

$$\frac{P_l}{l} = 0.023 \quad (\text{Chapter 7.20})$$

$$\frac{P_w}{w} = 0.06 \quad (\text{Chapter 7.21})$$

Now we can move on to our partial equations so to calculate the bias error, B_σ we must use the following equation

$$B_\sigma = \sqrt{\left(\frac{\partial \sigma}{\partial F} u_F\right)^2 + \left(\frac{\partial \sigma}{\partial l} u_l\right)^2 + \left(\frac{\partial \sigma}{\partial w} u_w\right)^2} \quad (\text{Chapter 7.22})$$

Now here you can plug and check and get an uncertainty in units of stress, Pa , but I would highly suggest that you convert to dimensionless units and a fraction and you will find your answer looks much nicer as seen below

$$\frac{B_\sigma}{\sigma} = \sqrt{\left(\frac{u_F}{F}\right)^2 + \left(\frac{u_l}{l}\right)^2 + \left(\frac{u_w}{w}\right)^2} \quad (\text{Chapter 7.23})$$

Now you can plug in chug to solve this problem but wait I know that F, l, w are the nominal or mean values but what is the uncertainty? Well the uncertainty in F is given and we know the tool to measure length is a micrometer so the bias uncertainty for that tool is $1\mu m$ so we have our uncertainties.

Plugging in you will find that

$$\frac{B_\sigma}{\sigma} = \sqrt{\left(\frac{u_F}{F}\right)^2 + \left(\frac{u_l}{l}\right)^2 + \left(\frac{u_w}{w}\right)^2} = 0.02 \quad (\text{Chapter 7.24})$$

And now we can calculate the total uncertainty as a percent

$$U_\sigma = \sqrt{\left(\frac{B_\sigma}{\sigma}\right)^2 + \left(\frac{P_F}{F}\right)^2 + \left(\frac{P_l}{l}\right)^2 + \left(\frac{P_w}{w}\right)^2} \cdot 100 = 7.93\% \quad (\text{Chapter 7.25})$$

You will get much more practice with these problems during class, on YouTube, and also in Psets and exams but this is generally how you should approach these problems.

Suggested Procedure for Calculating Total Uncertainty

- (1) Determine if problem includes bias error, precision error, or both
- (2) Write out fundamental equation, i.e. what you are trying to calculate and identify independent variables
- (3) Write out known quantities, i.e. means, uncertainties, standard deviation, number of trials, etc, etc
- (4) Start with Precision Error first and convert to percentages or fractions
- (5) Take Partial derivatives and Calculate Bias Error and convert to percentages or fractions
- (6) Combine Precision and Bias error to calculate Total Uncertainty

Minimizing Error in Designing Experiments

Best time to **minimize experimental error** is in the **design stage**. You should **perform a single-sample uncertainty analysis** of the proposed experimental arrangement **prior to building the apparatus**. This allows one to make a decision on whether the expected uncertainty is acceptable and to identify the sources of uncertainty. So when designing an experiment:

- **Avoid approaches** that require two large numbers to be measured to determine the small difference between them
- **Design experiments or sensors that amplify the signal strength** to improve sensitivity
- Build *null designs* in which the output is measured as a change from zero rather than as a change in nonzero value
- Avoid experiments with large *correction factors* must be applied as part of the data-reduction procedure
- Attempt to **minimize the influence of the measuring system on the measured variable**
- **Calibrate** entire system to minimize calibration related bias errors.

When identifying outliers and a particular point is in doubt, temporarily exclude the data point and fit a new line through the remaining data. **If the fit is better the point can possibly be dropped however it could indicate that the relationship is not linear it may indicate another scaling regime or a transition in behavior.**

Chapter 7: Total Uncertainty, Bias Error, and Propagation of Error is shared under a [CC BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/) license and was authored, remixed, and/or curated by LibreTexts.

Chapter 8: F-Tests and Variable Independence

Here you can find a series of pre-recorded lecture videos that cover this content: <https://youtube.com/playlist?list=PLSKNWlZCsmjDNUfe0P58VkcTtwIfxpSw2&si=Xrau1wfJKX8-ZI-k>

Analyzing Samples Based on Variance

We have previously discussed the χ^2 goodness of fit test but we want to build upon our ANOVA analysis which looked at analyzing populations based on variances and build upon that and introduce a new statistical tool, specifically the χ^2 test. Here our key test metric for this type of test

$$\chi^2 = \frac{(n-1)S^2}{\sigma_o^2} \quad (\text{Chapter 8.1})$$

where σ_o is our null population standard deviation and S is our sample standard deviation previously defined. We can use this test just as we have for other statistical measures and for hypothesis testing as well, in fact let's go ahead and work with our old material A and test to see if this sample comes from a population with a standard deviation of 1.31 GPa at a 95% confidence interval.

$$H_0 : \sigma = 1.31 \text{ GPa} \quad (\text{Chapter 8.2})$$

$$H_a : \sigma \neq 1.31 \text{ GPa} \quad (\text{Chapter 8.3})$$

$$\chi^2 = \frac{(n-1)S^2}{\sigma_o^2} = 8.58 \quad (\text{Chapter 8.4})$$

$$\chi_{0.05,11}^2 = 19.675 \quad (\text{Chapter 8.5})$$

Here we can clearly see that we are in the do not reject regime. We can also see from our p-value is 0.68 which confirms we are in the reject regime.

We can also compare whether material B comes from a population with a standard deviation greater than 1.7 GPa at a 99% confidence interval?

$$H_0 : \sigma = 1.7 \text{ GPa} \quad (\text{Chapter 8.6})$$

$$H_a : \sigma > 1.7 \text{ GPa} \quad (\text{Chapter 8.7})$$

$$\chi^2 = \frac{(n-1)S^2}{\sigma_o^2} = 22.05 \quad (\text{Chapter 8.8})$$

$$\chi_{0.01,14}^2 = 29.141 \quad (\text{Chapter 8.9})$$

Here we can clearly see that we are in the do not reject regime. We can also see from our p-value is 0.078 which confirms we are in the reject regime.

However, as we have demonstrated several times in this course we often want to make statements to compare samples and to do this we have to invoke the **F-Test** to compare variances/standard deviations.

F-Test

One can also perform hypothesis testing on multiple samples by examining variances and this can be done using the **F-Test** metric as defined below

$$F = \frac{S_1^2}{S_2^2} \quad (\text{Chapter 8.10})$$

We can run a hypothesis test with this as well, let's go ahead and keep examining our material A and B at a 95% confidence interval.

$$H_0 : \sigma_A^2 = \sigma_B^2 \quad (\text{Chapter 8.11})$$

$$H_a : \sigma_A^2 \neq \sigma_B^2 \quad (\text{Chapter 8.12})$$

$$F = \frac{S_A^2}{S_B^2} = 0.29 \quad (\text{Chapter 8.13})$$

We can look up the theoretical F value but we can see from the p-value of 0.04798 and we are in the do not reject region because this is a two-sided test so the critical p-value is 0.025.

We can also run a test to see if the variance of A is smaller than B at a 90% confidence level

$$H_0 : \sigma_A^2 = \sigma_B^2 \quad (\text{Chapter 8.14})$$

$$H_a : \sigma_A^2 < \sigma_B^2 \quad (\text{Chapter 8.15})$$

$$F = \frac{S_A^2}{S_B^2} = 0.29 \quad (\text{Chapter 8.16})$$

We are in the reject region as the p-value is 0.02!

χ^2 Test of Independence

We can look for correlations between variables and specifically we can see whether variables are independent or dependent. The test works just the same as the goodness of fit test in that we will be comparing the observed frequencies to the expected frequencies.

Typically when performing these types of test the data can will be reported in a table not too dissimilar to the one we can see here below

	Defect 1	Defect 2	Defect 3
Annealed	120	90	40
Cold Rolled	110	95	45
Hot Rolled	230	185	85

Figure Chapter8.1: Material Processing Steps and Defects.

Here we are looking at 500 samples and classifying if the sample exhibits a particular type of defect, 1-3. We also correlate that sample according to the type of material processing step that was undergone for that particular sample. Now one question that we may ask ourselves is if there is a correlation between the material process and the defect that occurs in the sample.

We can then calculate the expected frequencies by multiplying the total number of observations for the row and the total number of observations for the column divided by the total number of observations.

Thus we will be looking effectively at a probability and one can find the relationship that

$$\sum_{i=1}^r p_{ij} = 1 \quad (\text{Chapter 8.17})$$

where r is the number of rows and p_{ij} is the probability of obtaining an observation belonging to the i th row and j th column.

Luckily, our χ^2 expression is similar with one caveat

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (\text{Chapter 8.18})$$

where here r is the number of rows and c is the number of columns in our table. Additionally, our degrees of freedom $\nu = (r - 1)(c - 1)$.

Therefore we can perform our hypothesis testing procedures at a 95% confidence interval

$$H_0 : p_{ij} = \sum_{i=1}^c \sum_{j=1}^c p_{ij} \quad (\text{Chapter 8.19})$$

$$H_a : p_{ij} \neq \sum_{i=1}^c \sum_{j=1}^c p_{ij} \quad (\text{Chapter 8.20})$$

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 0.86404 \quad (\text{Chapter 8.21})$$

$$\chi_{0.05,4}^2 = 9.488 \quad (\text{Chapter 8.22})$$

So we can see that we are in the Do Not Reject regime and our p-value confirms this as well!!

We can also confirm this by hand calculations as well.

[Chapter 8: F-Tests and Variable Independence](#) is shared under a [CC BY-NC-SA 4.0](#) license and was authored, remixed, and/or curated by LibreTexts.

Chapter 9: Analysis of Variance (ANOVA)

Here you can find a series of pre-recorded lecture videos that cover this content: https://youtube.com/playlist?list=PLSKNWIZCsmjAKitWVYeIIBT_Ktyi3DKhT&si=ovUKlw6lauQFoz6E

ANOVA

You may have heard of **ANOVA** analysis in some engineering or even more likely in the biological context. **ANOVA** is analysis of variance and it is an extremely useful statistical analytical tool.

ANOVA is particularly useful in experimental designs where there are multiple variables being varied simultaneously to see if they will have an effect on the measured outcome of the experiment. For example, perhaps I am creating a new high strength alloy and I want to design the new alloy material to maximize the ultimate tensile strength. I can vary a number of parameters when processing this material such as annealing time, solute concentration, amount of work hardening, and many more parameters. One question you may want to answer is do any of these variables have a statistically significant effect on the ultimate tensile strength? Which parameter has the largest effect the ultimate tensile strength? Are there factors that have a combined effect?

Some of these questions we can answer using the tools we know about but others are not as easy to answer. So, before we jump into the deep end so to speak let's start with developing a solid foundation and build a framework to analyze one a single variable and perform a **One-Way ANOVA**.

One-Way ANOVA

Let us consider that we run an experiment with k independent random samples from k different populations and there will be a number of observations of measurands y and we can classify them as y_{ij} where i will vary from 1– k and j will simply indicate the number of observations. Therefore we will have k number of means y_k .

Now in this analysis there are several critical values that we will need to calculate the first of which is the sum of all of some general observation for a some treatment/variable/parameter/etc, i.e. a

$$T = \sum_{i=1}^k \sum_{j=1}^{n_i} y_{ij} \quad (\text{Chapter 9.1})$$

where n_i is the number of observations/measurements. Also the total sample size N is

$$N = \sum_{i=1}^k n_i \quad (\text{Chapter 9.2})$$

The overall sample mean will then be simply

$$\bar{y} = \frac{T}{N} \quad (\text{Chapter 9.3})$$

The sum of squares (SS) for is defined as

$$SS_a = \sum_{i=1}^k n_i (\bar{y}_i - \bar{y})^2 \quad (\text{Chapter 9.4})$$

and a critically important measure, the **error sum of squares** is defined as

$$ESS_a = \sum_{i=1}^k \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_i)^2 \quad (\text{Chapter 9.5})$$

The degrees of freedom

$$df_a = \sum_{i=1}^k n_i - 1 = k - 1 + \sum_{i=1}^k n_i - k \quad (\text{Chapter 9.6})$$

The mean square can be defined as

$$\frac{SS_a}{df_a} \quad (\text{Chapter 9.7})$$

and the F ratio for treatments can be defined as

$$F_a = \frac{\frac{SS_a}{(k-1)}}{\frac{ESS_a}{N-k}} \quad (\text{Chapter 9.8})$$

The F ratio is a key parameter in performing ANOVA analysis as a larger F ratio value will indicate large differences between sample means and thus large differences between samples.

Finally, we will also obtain our P-value in this analysis as well.

We can summarize these key values that we must calculate when performing an ANOVA analysis in what we will typically read in our ANOVA table which we will see when we do some examples to come.

Let's go ahead and perform a One-Way ANOVA analysis for an example of crop yields as a function of type of fertilizer, crop density, and the location of where the crop has been planted. For a One-Way ANOVA analysis let us first just examine the effect of fertilizer and we can obtain our ANOVA table below

```

      Df Sum Sq Mean Sq F value Pr(>F)
fertilizer  2    6.07   3.0340   7.863 7e-04 ***
Residuals  93   35.89   0.3859
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figure Chapter9.1: R ANOVA Table Results.

Here we can see that the p-value is less than 0.05 and therefore we can conclude that fertilizer type has a statistically significant effect on the crop yield. W

Often times we want to go beyond the results from our ANOVA table and perform a post-hoc test. We can use this to go deeper into the data and specifically to see if we have differences between sub-types of fertilizer. One of the most common tests you may come across is a Tukey post-hoc test but this assumes that variances are equal so to be more general we will instead perform a Games Howell post-hoc test!!

Here we can see some very interesting details here. Looking at the p-value there is no statistically significant difference between fertilizer types 1 and 2 whereas there are statistically significant differences between 1 and 3 and 2 and 3!! This is a fantastic insight to observe as it can allow an experimentalist to design new experiments and explore new parameter spaces that yield new exciting results.

```

> games_howell_test(crop.data, yield ~ fertilizer)
# A tibble: 3 x 8
  .y. group1 group2 estimate conf.low conf.high p.adj p.adj.signif
*   <chr> <chr> <chr>      <dbl>    <dbl>    <dbl> <dbl> <chr>
1 yield 1     2      0.176  -0.203    0.556 0.509 ns
2 yield 1     3      0.599   0.213    0.986 0.001 ***
3 yield 2     3      0.423   0.070    0.775 0.015 *
>

```

Figure Chapter9.2: Games Howell Post-Hoc Test.

One of the other extremely valuable pieces of information that we can obtain from ANOVA analysis is not only if there are statistically significant differences due to a parameter but we can also determine if the parameter has a large effect on the outcome. We can quantify the effect by calculating our η^2 value which can be calculated as

$$\eta^2 = \frac{SS_a}{SS_a + SS_{error}} \quad (\text{Chapter 9.9})$$

where SS_a is the sum of squares of some parameter/effect/variable and if η^2 is 0.14 or greater we can say that is a large effective, between 0.06 and 0.14 is a medium effect, and 0.01 or lower is a small effect. So we can see in our analysis below that fertilizer has a large effect on the crop yield.

```
> eta_squared(one.way)
fertilizer
0.1446349
```

Figure Chapter9.3: Size effect of fertilizer.

Two-Way ANOVA

We can also run Two-Way ANOVA analysis to calculate the total sum of squares we can use the following identity

$$\sum_{i=1}^a \sum_{j=1}^b (y_{ij} - \bar{y}_{..})^2 \quad (\text{Chapter 9.10})$$

$$= \sum_{i=1}^a \sum_{j=1}^b (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2 + b \sum_{i=1}^a (\bar{y}_{i.} - \bar{y}_{..})^2 + a \sum_{j=1}^b (\bar{y}_{.j} - \bar{y}_{..})^2$$

We can also define our new sum of squares as

$$SS = b \sum_{i=1}^a (\bar{y}_{i.} - \bar{y}_{..})^2 \quad (\text{Chapter 9.11})$$

The block sum of squares is defined as

$$BSS = a \sum_{j=1}^b (\bar{y}_{.j} - \bar{y}_{..})^2 \quad (\text{Chapter 9.12})$$

The error sum of squares is

$$ESS = \sum_{i=1}^a \sum_{j=1}^b (y_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..})^2 \quad (\text{Chapter 9.13})$$

and the total sum of squares is

$$TSS = \sum_{i=1}^a \sum_{j=1}^b (y_{ij} - \bar{y}_{..})^2 \quad (\text{Chapter 9.14})$$

The degrees of freedom calculation for a Two-Way ANOVA will be

$$df = ab - a = (a - 1) + (b - 1) + (a - 1)(b - 1) \quad (\text{Chapter 9.15})$$

The F-ratio for the a variable will be

$$F_a = \frac{\frac{TSS}{(a-1)}}{\frac{ESS}{(a-1)(b-1)}} \quad (\text{Chapter 9.16})$$

and for the b variable

$$F_b = \frac{\frac{BSS}{(b-1)}}{\frac{ESS}{(a-1)(b-1)}} \quad (\text{Chapter 9.17})$$

We can now run our Two-Way ANOVA in R and Python and look at the ANOVA table result below.

```

> #Two-Way
> two.way <- aov(yield ~ dens + fertilizer + dens:fertilizer, data = crop.data)
> summary(two.way)

```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
dens	1	5.122	5.122	15.195	0.000186	***
fertilizer	2	6.068	3.034	9.001	0.000273	***
dens:fertilizer	2	0.428	0.214	0.635	0.532500	
Residuals	90	30.337	0.337			

```

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> eta_squared(two.way)

```

	dens	fertilizer	dens:fertilizer
	0.12207782	0.14463491	0.01019726

```

> games_howell_test(crop.data, yield ~ fertilizer)
# A tibble: 3 x 8

```

.y.	group1	group2	estimate	conf.low	conf.high	p.adj	p.adj.signif
yield 1	1	2	0.176	-0.203	0.556	0.509	ns
yield 1	1	3	0.599	0.213	0.986	0.001	***
yield 2	2	3	0.423	0.070	0.775	0.015	*

```

> games_howell_test(crop.data, yield ~ dens)
# A tibble: 1 x 8

```

.y.	group1	group2	estimate	conf.low	conf.high	p.adj	p.adj.signif
yield 1	1	2	0.462	0.208	0.716	0.000485	***

Figure Chapter9.4: R Two-Way ANOVA Result Table.

Here we can see some very interesting results. We again observe that fertilizer has a statistically significant effect on crop yield as does the density of crop planting. However, we see that the combined density and fertilizer variable does not have a statistically significant effect on the crop yield.

We can also confirm again that fertilizer has a large effect, density has a medium effect based on the η^2 value. And the combined variable of density and fertilizer is a small effect.

We can also see our post-hoc Games Howell test and confirm our previous results for fertilizer but also we see that for density there is a statistically significant difference between density type 1 and 2.

We can even extend this to three variables as well..

Three-Way ANOVA

One can perform another extension of the previous equations and we can run a **Three-Way ANOVA** and see the results below

Here we can perform a similar analysis and we can see that there is no statistically significant effect block location can have on the yield of the crop and neither does the two way variable of fertilizer and block, however density and block has a statistically significant effect on crop yield as does fertilizer, density, and block variables as well.

```
> # Three-Way
> three.way<-aov(yield~dens+fertilizer+block+dens:fertilizer+dens:block+fertilizer:block+fertilizer:block:dens,
data)
> summary(three.way)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
dens	1	5.122	5.122	14.854	0.000226 ***
fertilizer	2	6.068	3.034	8.799	0.000339 ***
block	2	0.486	0.243	0.705	0.497028
dens:fertilizer	2	0.428	0.214	0.620	0.540182
fertilizer:block	4	0.887	0.222	0.643	0.633043
Residuals	84	28.963	0.345		

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> games_howell_test(crop.data, yield ~ fertilizer)
# A tibble: 3 × 8
  .y. group1 group2 estimate conf.low conf.high p.adj p.adj.signif
* <chr> <chr> <chr>    <dbl>    <dbl>    <dbl> <dbl> <chr>
1 yield 1     2      0.176  -0.203    0.556 0.509 ns
2 yield 1     3      0.599   0.213    0.986 0.001 ***
3 yield 2     3      0.423   0.070    0.775 0.015 *
```

```
> games_howell_test(crop.data, yield ~ dens)
# A tibble: 1 × 8
  .y. group1 group2 estimate conf.low conf.high p.adj p.adj.signif
* <chr> <chr> <chr>    <dbl>    <dbl>    <dbl> <dbl> <chr>
1 yield 1     2      0.462   0.208    0.716 0.000485 ***
```

```
> games_howell_test(crop.data, yield ~ block)
# A tibble: 6 × 8
  .y. group1 group2 estimate conf.low conf.high p.adj p.adj.signif
* <chr> <chr> <chr>    <dbl>    <dbl>    <dbl> <dbl> <chr>
1 yield 1     2      0.460  -0.029    0.950 0.072 ns
2 yield 1     3     -0.144  -0.613    0.325 0.846 ns
3 yield 1     4      0.320  -0.172    0.811 0.318 ns
4 yield 2     3     -0.604  -1.08    -0.128 0.008 **
5 yield 2     4     -0.141  -0.639    0.357 0.874 ns
6 yield 3     4      0.463  -0.014    0.941 0.06 ns
```

```
> eta_squared(three.way)
```

	dens	fertilizer	block	dens:fertilizer	fertilizer:block
	0.12207782	0.14463491	0.01158736	0.01019726	0.02115164

Figure Chapter9.5: R Three-Way ANOVA Result Table.

We can also see that density has a medium effect, block has a small effect as does density and fertilizer and fertilizer and block variables.

Finally, we see that for the block variable there is only a statistically significant effect between block 2 and 3.

ANOVA is an incredibly useful tool!!

Chapter 9: Analysis of Variance (ANOVA) is shared under a CC BY-NC-SA 4.0 license and was authored, remixed, and/or curated by LibreTexts.

Chapter 10: Fitting, Regressions, and Confidence Bands

Here you can find a series of pre-recorded lecture videos that cover this content: https://youtube.com/playlist?list=PLSKNWlZCsmjBxk5lvZ_LaUX2hBGSEexpV&si=nOGeAaWzNwuNlz5X

As we touched on at the beginning of this course and have emphasized throughout is visualizing and plotting data and analyzing trend to find statistically significant results.

Often this may involve fitting data which we have only briefly touched upon up onto this point with the exception of our χ^2 goodness of fit test which can indeed be utilized to analyze distribution shapes and fits.

One of the most common methods to analyze fits and curves is with **regression analysis** and one of the most popular techniques is regression analysis via the **method of least squares**.

Method of Least Squares

In the method of least squares we will be working with two variables x , which will be considered the **independent** or **input** variable and y , the **dependent** or **response** variable. We will assume that the total uncertainty in y will be much greater than x which in most experimental conditions is a fairly reasonable assumption and we will start by analyzing the regression curve of y on x as a linear regression curve

$$y = a + bx \quad (\text{Chapter 10.1})$$

where a and b are fitting constants. Then the error e_i in predicting a y value will be given by

$$e_i = y_i - y(x_i) \quad (\text{Chapter 10.2})$$

We will thereby select a and b such that $\sum_{i=1}^n e_i$ is as close to 0 as possible to minimize the error.

So the principle of least squares will give us

$$\sum_{i=1}^n e_i = \sum_{i=1}^n [y_i - y(x_i)]^2 \quad (\text{Chapter 10.3})$$

Before we can proceed with this process it will be extremely convenient here to define some critical values specifically

$$S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2 \quad (\text{Chapter 10.4})$$

and similarly

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (\text{Chapter 10.5})$$

and

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \quad (\text{Chapter 10.6})$$

Then we can calculate the error sum of squares which we seek to minimize as

$$SSE = S_{yy} - \frac{S_{xy}^2}{S_{xx}} \quad (\text{Chapter 10.7})$$

This calculation will give us fitting parameters but we can also obtain estimates of the confidence intervals of those fitting parameters as well.

Let's look at some plots of raw data and then we can perform some linear fits

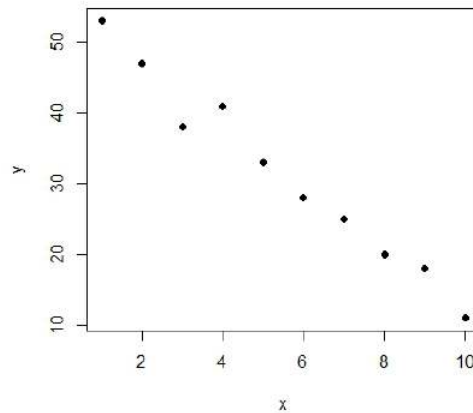


Figure Chapter10.1: R Raw Data Plots.

Here we can see some example of raw data and we can now proceed to attempt to fit the data based on functions that we believe will best fit the values. We set the function and then proceed to allow the program to run through the above calculations to determine the fitting parameters in the linear function

We can see the fits here and we can also plot the 95% confidence interval for these fits as well as seen here

[Non-Linear Fitting:](#)

We can extend these tools to also perform non-linear fits that can range from exponential expressions, logarithmic expressions, polynomials, and beyond.

Here we can see some nonlinear fits of the previous data set

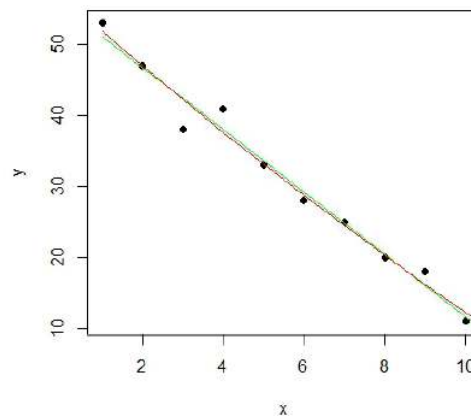


Figure Chapter10.2: R Linear Fit.

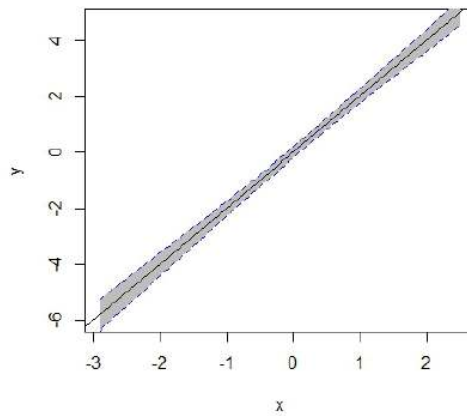


Figure Chapter10.3: R 95% Confidence Interval Fit.

Non-Linear Fitting:

We can extend these tools to also perform non-linear fits that can range from exponential expressions, logarithmic expressions, polynomials, and beyond.

Here we can see some nonlinear fits of the previous data set

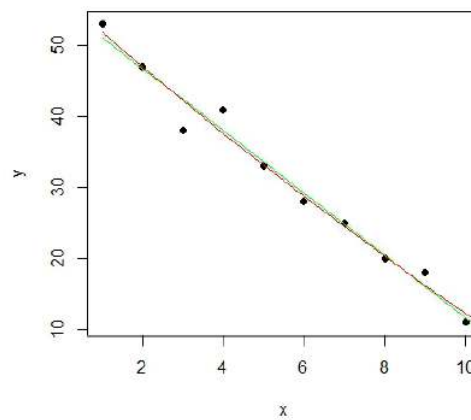


Figure Chapter10.2: R Linear Fit.

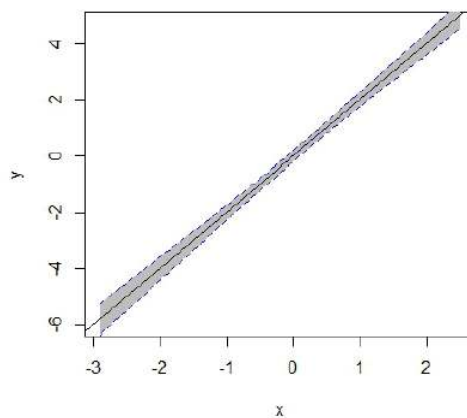


Figure Chapter10.3: R 95% Confidence Interval Fit.

Correlation

So far we have assumed that the total uncertainty in x was extremely small and close to 0, while this holds for my real life scenarios there are many many many scenarios where this is not the case in which not true.

In which case the purpose of plotting x vs y in the most general sense is to observe some type of **correlation** between the variables. One can quantitatively calculate this via the **sample correlation coefficient** r which is defined as

$$r = \frac{S_{xy}}{\sqrt{S_{xx} \cdot S_{yy}}} \quad (\text{Chapter 10.8})$$

here if $r = 1$ we have perfect positive correlation if $r = -1$ we have perfect negative correlation, and if $r = 0$ we have no correlation, i.e. there is no linear association. Additionally the limits of r is between -1 and 1 .

We can also calculate our r^2 value which varies from 0-1 and is defined as

$$r^2 = \frac{S_{xy}^2}{S_{xx}S_{yy}} \quad (\text{Chapter 10.9})$$

We can see these values tabulated here for our previous fits to show correlations, both r and r^2 values.

Looking at our previous fits we can run regression analysis to investigate the correlations as well.

Critical Consideration When Fitting

Regardless of the statistical tool and measure that you are using when fitting value be sure that you do not lose the physical intuition of the system that you are investigating. There are very very few physical systems that can be represented by 12th order polynomial function despite the fact that that particular function may yield the strongest correlation factor. Instead, a more physically intuitive function that yields a lower correlation coefficient may be more appropriate and one may spend more time analyzing why this experimental system deviated from this theoretical function/representation.

Chapter 10: Fitting, Regressions, and Confidence Bands is shared under a [CC BY-NC-SA 4.0](https://creativecommons.org/licenses/by-nc-sa/4.0/) license and was authored, remixed, and/or curated by LibreTexts.

Index

D

dire

Glossary

Sample Word 1 | Sample Definition 1

Detailed Licensing

Overview

Title: [Statistical Analysis of Data for Engineers](#)

Webpages: 21

Applicable Restrictions: Noncommercial

All licenses found:

- [CC BY-NC-SA 4.0](#): 90.5% (19 pages)
- [Undeclared](#): 9.5% (2 pages)

By Page

- [Statistical Analysis of Data for Engineers](#) - [CC BY-NC-SA 4.0](#)
 - [Front Matter](#) - [CC BY-NC-SA 4.0](#)
 - [TitlePage](#) - [CC BY-NC-SA 4.0](#)
 - [InfoPage](#) - [CC BY-NC-SA 4.0](#)
 - [Table of Contents](#) - [Undeclared](#)
 - [Licensing](#) - [CC BY-NC-SA 4.0](#)
 - [Chapter 1: Importance of Statistics and Experimental Measurements](#) - [CC BY-NC-SA 4.0](#)
 - [Chapter 2: Histograms, Statistical Measures, and Probability](#) - [CC BY-NC-SA 4.0](#)
 - [Chapter 3: Probability Distributions and the Gaussian](#) - [CC BY-NC-SA 4.0](#)
 - [Chapter 4: Central Limit Theorem and Confidence Intervals for Large and Small Sample Sizes](#) - [CC BY-NC-SA 4.0](#)
 - [Chapter 5: Hypothesis Testing and P-Values](#) - [CC BY-NC-SA 4.0](#)
 - [Chapter 6: Goodness of Fit Hypothesis Testing](#) - [CC BY-NC-SA 4.0](#)
 - [Chapter 7: Total Uncertainty, Bias Error, and Propagation of Error](#) - [CC BY-NC-SA 4.0](#)
 - [Chapter 8: F-Tests and Variable Independence](#) - [CC BY-NC-SA 4.0](#)
 - [Chapter 9: Analysis of Variance \(ANOVA\)](#) - [CC BY-NC-SA 4.0](#)
 - [Chapter 10: Fitting, Regressions, and Confidence Bands](#) - [CC BY-NC-SA 4.0](#)
 - [Back Matter](#) - [CC BY-NC-SA 4.0](#)
 - [Index](#) - [CC BY-NC-SA 4.0](#)
 - [Glossary](#) - [CC BY-NC-SA 4.0](#)
 - [Detailed Licensing](#) - [CC BY-NC-SA 4.0](#)
 - [Detailed Licensing](#) - [Undeclared](#)

Detailed Licensing

Overview

Title: [Statistical Analysis of Data for Engineers](#)

Webpages: 21

Applicable Restrictions: Noncommercial

All licenses found:

- [CC BY-NC-SA 4.0](#): 90.5% (19 pages)
- [Undeclared](#): 9.5% (2 pages)

By Page

- [Statistical Analysis of Data for Engineers](#) - [CC BY-NC-SA 4.0](#)
 - [Front Matter](#) - [CC BY-NC-SA 4.0](#)
 - [TitlePage](#) - [CC BY-NC-SA 4.0](#)
 - [InfoPage](#) - [CC BY-NC-SA 4.0](#)
 - [Table of Contents](#) - [Undeclared](#)
 - [Licensing](#) - [CC BY-NC-SA 4.0](#)
 - [Chapter 1: Importance of Statistics and Experimental Measurements](#) - [CC BY-NC-SA 4.0](#)
 - [Chapter 2: Histograms, Statistical Measures, and Probability](#) - [CC BY-NC-SA 4.0](#)
 - [Chapter 3: Probability Distributions and the Gaussian](#) - [CC BY-NC-SA 4.0](#)
 - [Chapter 4: Central Limit Theorem and Confidence Intervals for Large and Small Sample Sizes](#) - [CC BY-NC-SA 4.0](#)
 - [Chapter 5: Hypothesis Testing and P-Values](#) - [CC BY-NC-SA 4.0](#)
 - [Chapter 6: Goodness of Fit Hypothesis Testing](#) - [CC BY-NC-SA 4.0](#)
 - [Chapter 7: Total Uncertainty, Bias Error, and Propagation of Error](#) - [CC BY-NC-SA 4.0](#)
 - [Chapter 8: F-Tests and Variable Independence](#) - [CC BY-NC-SA 4.0](#)
 - [Chapter 9: Analysis of Variance \(ANOVA\)](#) - [CC BY-NC-SA 4.0](#)
 - [Chapter 10: Fitting, Regressions, and Confidence Bands](#) - [CC BY-NC-SA 4.0](#)
 - [Back Matter](#) - [CC BY-NC-SA 4.0](#)
 - [Index](#) - [CC BY-NC-SA 4.0](#)
 - [Glossary](#) - [CC BY-NC-SA 4.0](#)
 - [Detailed Licensing](#) - [CC BY-NC-SA 4.0](#)
 - [Detailed Licensing](#) - [Undeclared](#)