Module 1C - Statistics

Omar Betancourt, Payton Goodrich, Emre Mengi

July 24, 2021

Contents

Objectives: To learn the fundamentals of statistics and understand the strengths and limitations of sampling data from a larger population.

Prerequisite Knowledge: General arithmetic

Prerequisite Modules: N/A

Difficulty: Easy

Summary: In this module, you will learn when and when not statistical analysis can be applied to data, and how to use statistics as a tool for data analytics.

1 Theory

1.1 Random and Systematic Error

One of the best ways to assess the reliability of a measurement is to perform it several times and consider the different values obtained. Experience has shown that no measurement - no matter how carefully it is made - will obtain values that are exactly the same. Error analysis is the study and evaluation of uncertainty in a measurement. Note that when we use the term error in statistics, it is not used to mean that a mistake or blunder was made. Rather, it is used to describe the inevitable uncertainty of scientific measurements, and hereafter uncertainty and error will be used interchangeably.
Uncertainties can be classified into two groups: *random* uncertainties and *systematic* uncertainties. Systematic uncertainties always push the measured results in a single direction, while random uncertainties are equally likely to push the results in any direction. Consider trying to time an event with a stopwatch: one source of error will be the reaction time of the user starting and stopping the watch. The user may delay more in starting the stopwatch, thereby underestimating the duration of the event, but they are equally likely to delay more in stopping the stopwatch, resulting in an overestimate of the event. This is an example of a random uncertainty. Now consider if the stopwatch consistently runs slow - in this case all events will be underestimates. This is an example of systematic uncertainty. Systematic uncertainties are hard to evaluate and sometimes even difficult to detect. However, the use of statistics described hereafter give a reliable estimate of random uncertainties.

1.2 Common Statistical Terms

1.2.1 Mean

Suppose we want to measure some quantity, x. Assume we have identified and reduced all systematic uncertainties to a negligible level. Because all remaining sources of uncertainty are random, we should be able to identify them by repeating the measurement several times. After we have taken the measurements, we will see that the values differ. In statistics, the best representation of the measured value then would be the average or the *mean* of the measured values:

$$\bar{x} = \frac{x_1 + x_2 + ... + x_N}{N} = \frac{\sum_{n=1}^{N} x_i}{N} \tag{1.1}$$

where x_i is an individual measurement and N is the total number of measurements that were made.

1.2.2 Standard Deviation

The standard deviation is an estimate of the average uncertainty of the measurements. Consider any single measurement, x_i, to the mean, $\bar{x}$. The measurement deviates from the best estimate by $x_i - \bar{x}$. If our measurements are precise, the value of any deviation is likely small. If some of the deviations are large, than the measurements are not so precise. To estimate the average reliability, we square the value of each deviation and take the square root of the result to evaluate the magnitude of the deviation, and then calculate the mean of these values. The result is the standard deviation, σ_x:

$$\sigma_x = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(x_i - \bar{x})^2} \tag{1.2}$$

1.2.3 Standard Error

We have seen how the best estimate for a quantity x is $\bar{x}$ and that the average uncertainty of the seperate measurements is σ_x. However, $\bar{x}$ represents a combination of all N measurements, and we have reason to believe it is better than any single measurement taken alone. In fact, the uncertainty of $\bar{x}$ is given by the standard error, $\sigma_{\bar{x}}$:

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}} \tag{1.3}$$

An important feature of the standard error is that as we increase the number of measurements we make, the value of the standard error will decrease. This makes sense: as we increase the number of measurements we make, we are less and less uncertain of the average result of the measurements.

1.3 Distributions

A distribution is a function that describes the frequency that a repeated measurement yields each of its various possible answers. As the number of measurements increase, the distribution will begin to take on a definite shape and become more and more accurate.

1.3.1 The Normal Distribution

If a measurement is subjected to many small sources of random error and negligible sources of systematic error, then the measured values will be distributed in a bell-curve shape centered on $\bar{x}$. When this distribution is normalized such that the total area under the bell-curve equals 1 and is centered at $x = \bar{x}$, this distribution can be plotted as a normal (also known as the Gauss or Gaussian) distribution:

$$G_{\bar{x},\sigma_x}(x) = \frac{1}{\sigma_x\sqrt{2\pi}} e^{\frac{(x-\bar{x})^2}{2\sigma_x^2}} \tag{1.4}$$

Here, the subscripts of G denote the center and the width of the distribution. There are many uses for the normal distribution, particularly in probability theory. It's also useful in statistical analysis to graphically communicate findings, find confidence intervals, and various other tests that will not be covered in this module.

1.3.2 The Binomial Distribution

The binomial distribution describes the probability of achieving k in n trials, where the probabilities of each individual event is known and independent. Classic example of the binomial distribution involve flipping coins, rolling dice, and flipping coins. However, binomial distributions also describe and visualize the likelihood of multiple failures occurring in food manufacturing, distribution, or misgendering of plants. Alternatively, binomial distributions could be used to analyze the efficacy of different plant treatments by 'racing' plant growth against untreated plants and seeing if the treated plants won more times than would be expected.
The binomial distribution is given by:

$$B_{n,p}(k) = \binom{n}{k} p^k q^{n-k} \tag{1.5}$$

where:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{1.6}$$

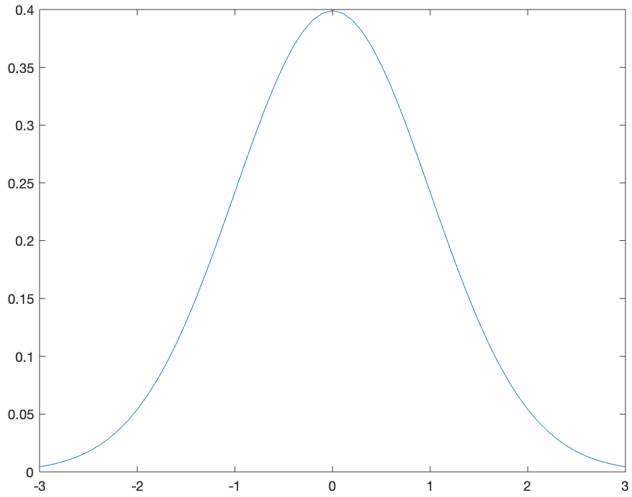

Figure 1.1: The normal distribution centered on $x = 0$ and $\sigma = 1$.

Here, the subscripts n, p refer to the number of trials and the probability of 'success' in one trial, and $q = 1 - p$. It is called the binomial distribution because the $\binom{n}{k}$ term is the binomial coefficient.

1.3.3 The Poisson Distribution

The Poisson distribution describes the results of experiments where randomly-occuring events with an average rate are counted. For example, the number of babies born in a single hospital over the course of a week would fit a Poisson distribution.
The Poisson distribution is given by:

$$P_\mu = e^{-\mu} \frac{\mu^v}{v!} \tag{1.7}$$

where $\mu > 0$ is the number of expected counts in the time interval.

1.4 Central-Limit Theorem

How is it that researchers can make overarching claims about large populations of people using data collected from a study performed on far fewer people? Many people overlook this phenomena and accept it as fact without knowledge of it's innerworkings. In reality, the ability to make conclusions about a large population from a small sample comes from statistics.

In probability theory, the central limit theorem establishes that the normalized sum of independent random variables tend toward a normal distribution. This key concept of a normal distribution makes statistical methods that even if the original variables themselves are not normally distributed.

Mean values obtained from a sample group of adequate size is statistically representative of the mean value of the larger population. In this module's assignment, you will investigate this first-hand.

2 Example

2.1 Binomial Distribution

Suppose we toss four coins and count the number of heads obtained. In this example, the population is four (n=4) and the probability of a success is one-in-two ($p = \frac{1}{2}$). What then, is the probability of the coin landing on heads 0, 1, 2, 3, or 4 times? The probability is simply the binomial distribution:

$$\binom{4}{k}\binom{1}{2}^4 = \frac{n!}{k!(n-k)!} \tag{2.1}$$

Which can be calculated and plotted as shown in Figure 2.1

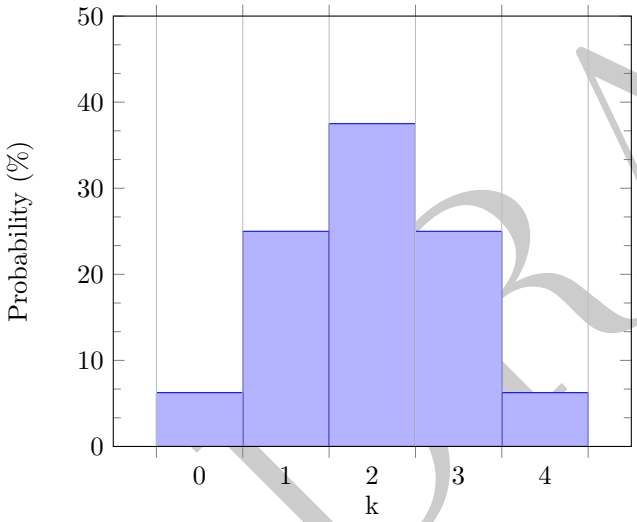

Figure 2.1: Binomial distribution with $n = 4$ and $p = \frac{1}{2}$. This gives the result of flipping heads k times in four coin tosses.

2.2 Poisson Distribution

It has been found that Minnesota experiences 112 rainy days each year. Using this information alone, what is the probability of the next $v = 0, 1, 2$, and $v \geq 3$ days will be rainy?
The expected average count is just $\mu = \frac{112}{365} \approx 0.31$. The probability or experiencing v rainy days is given by the poisson distribution:

$$P_{0.31} = e^{-0.31}\frac{0.31^v}{v!} \tag{2.2}$$

The values for $v = 0, 1, 2$ can be calculated from the formula, and the value for $v \geq 3$ can be calculating by subtracting the sum of the three previous values from one. The poisson distribution for this example is shown in Figure 2.2.

3 Assignment

In this assignment, you will demonstrate how sub-sampling a population can provide statistically meaningful data that is representative of a larger population using the central-limit theorem.

1. Use some established algorithm to randomly generate a population of 10,000 numbers that fit a normal distribution, having a mean value of 500 and a standard deviation of 100. Verify your results by

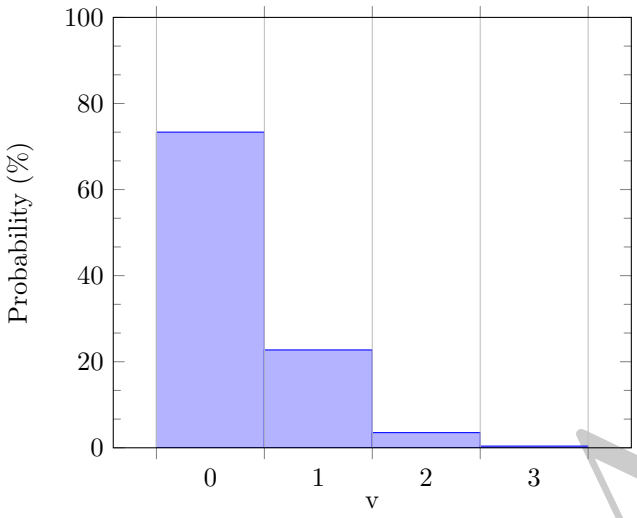

Figure 2.2: Poisson distribution with $\mu = 0.31$. This shows the odds of there being v consecutive days of rain in Minnesota (using only statistics and no meteorological or historical data).

plotting a histogram of your numbers to see its shape and to check the mean and SD against the specified values.

2. Starting with n=100, randomly sample n numbers from the full set of 10,000 and calculate the mean and standard deviation of that sample. Repeat this 1,000 times and plot a histogram of: 1) the resulting distribution of means; and 2) the resulting distribution of standard deviations. From that, find the mean and standard deviation values of the distribution of means, and find the mean and standard deviation values of the distribution of standard deviations. How do these values compare against the population mean and SD?

3. Repeat Step 2 using the following sample sizes: n=50, 30, 20, 10, 6, and 4.

4. Interpret these results and consider how this information can be used for comparing a mean value of a sample against some known mean for a population. Also explain what is the mean by a distribution of means and a distribution of standard deviations. Most critically, explain how one makes estimates of the population mean and standard deviations from measurements of a sample, and the role of the sample size on those estimates. Finally, what does all this mean for testing hypotheses, comparing two measurement techniques, and validating model predictions?

5. Write all this up in one document: Introduction (0.5 page), Methods (1 page), Results (1 page), Discussion (1 page), and show any graphs or tables of your data. Assume the audience is a first-year engineering student.

4 Solution

The assignment solution is encoded in Matlab below.

```
N = normrnd(500,100,[10000,1]); %Generate a population that fits a normal
                                %distribution with mean 500 and SD 100

figure(1)
H = histogram(N);
mn = mean(N);
sd = std(N);

```

```
mnlabel=sprintf('Mean — %3.2d', mn);
stdlabel=sprintf('Std Deviation — %3.2d', sd);
h=annotation('textbox',[0.58 0.75 0.1 0.1]);
set(h,'String',{mnlabel, stdlabel});

for i = 1:1000

    %Sample n data points from the population N
    n_100(:,i) = datasample(N,100);
    n_50(:,i) = datasample(N,50);
    n_30(:,i) = datasample(N,30);
    n_20(:,i) = datasample(N,20);
    n_10(:,i) = datasample(N,10);
    n_6(:,i) = datasample(N,6);
    n_4(:,i) = datasample(N,4);

    %Take the mean of n-samples
    m_100(i) = mean(n_100(:,i));
    m_50(i) = mean(n_50(:,i));
    m_30(i) = mean(n_30(:,i));
    m_20(i) = mean(n_20(:,i));
    m_10(i) = mean(n_10(:,i));
    m_6(i) = mean(n_6(:,i));
    m_4(i) = mean(n_4(:,i));

    %Find the standard deviation of n-samples
    sd_100(i) = std(n_100(:,i));
    sd_50(i) = std(n_50(:,i));
    sd_30(i) = std(n_30(:,i));
    sd_20(i) = std(n_20(:,i));
    sd_10(i) = std(n_10(:,i));
    sd_6(i) = std(n_6(:,i));
    sd_4(i) = std(n_4(:,i));


end

%Plot a histogram of the resulting distribution of means of several
%sampling sizes

figure(2)
subplot(7,1,1)
H_m100 = histogram(m_100);
mn = mean(m_100);
sd = std(m_100);

mnlabel=sprintf('Mean — %3.2d', mn);
stdlabel=sprintf('Std Deviation — %3.2d', sd);
h=annotation('textbox',[0.7 0.81 0.1 0.1]);
set(h,'String',{mnlabel, stdlabel});
title('Distribution of means n = 100')

subplot(7,1,2)
H_m50 = histogram(m_50);
```

```
mn = mean(m_50);
sd = std(m_50);

mnlabel=sprintf('Mean — %3.2d', mn);
stdlabel=sprintf('Std Deviation — %3.2d', sd);
h=annotation('textbox',[0.68 0.69 0.1 0.1]);
set(h,'String',{mnlabel, stdlabel});
title('Distribution of means n = 50')

subplot(7,1,3)
H_m30 = histogram(m_30);
mn = mean(m_30);
sd = std(m_30);

mnlabel=sprintf('Mean — %3.2d', mn);
stdlabel=sprintf('Std Deviation — %3.2d', sd);
h=annotation('textbox',[0.68 0.57 0.1 0.1]);
set(h,'String',{mnlabel, stdlabel});
title('Distribution of means n = 30')

subplot(7,1,4)
H_m20 = histogram(m_20);
mn = mean(m_20);
sd = std(m_20);

mnlabel=sprintf('Mean — %3.2d', mn);
stdlabel=sprintf('Std Deviation — %3.2d', sd);
h=annotation('textbox',[0.68 0.45 0.1 0.1]);
set(h,'String',{mnlabel, stdlabel});
title('Distribution of means n = 20')

subplot(7,1,5)
H_m10 = histogram(m_10);
mn = mean(m_10);
sd = std(m_10);

mnlabel=sprintf('Mean — %3.2d', mn);
stdlabel=sprintf('Std Deviation — %3.2d', sd);
h=annotation('textbox',[0.68 0.33 0.1 0.1]);
set(h,'String',{mnlabel, stdlabel});
title('Distribution of means n = 10')

subplot(7,1,6)
H_m6 = histogram(m_6);
mn = mean(m_6);
sd = std(m_6);

mnlabel=sprintf('Mean — %3.2d', mn);
stdlabel=sprintf('Std Deviation — %3.2d', sd);
h=annotation('textbox',[0.68 0.21 0.1 0.1]);
set(h,'String',{mnlabel, stdlabel});
title('Distribution of means n = 6')

subplot(7,1,7)
```

```
H_m4 = histogram(m_4);
mn = mean(m_4);
sd = std(m_4);

mnlabel=sprintf('Mean — %3.2d', mn);
stdlabel=sprintf('Std Deviation — %3.2d', sd);
h=annotation('textbox',[0.68 0.09 0.1 0.1]);
set(h,'String',{mnlabel, stdlabel});
title('Distribution of means n = 4')

%Plot a histogram of the resulting distribution of standard deviations

figure(3)
subplot(7,1,1)
H_sd100 = histogram(sd_100);
mn = mean(sd_100);
sd = std(sd_100);

mnlabel=sprintf('Mean — %3.2d', mn);
stdlabel=sprintf('Std Deviation — %3.2d', sd);
h=annotation('textbox',[0.7 0.81 0.1 0.1]);
set(h,'String',{mnlabel, stdlabel});
title('Distribution of standard deviations n = 100')
xlim([0 200])

subplot(7,1,2)
H_sd50 = histogram(sd_50);
mn = mean(sd_50);
sd = std(sd_50);

mnlabel=sprintf('Mean — %3.2d', mn);
stdlabel=sprintf('Std Deviation — %3.2d', sd);
h=annotation('textbox',[0.7 0.69 0.1 0.1]);
set(h,'String',{mnlabel, stdlabel});
title('Distribution of standard deviations n = 50')
xlim([0 200])

subplot(7,1,3)
H_sd30 = histogram(sd_30);
mn = mean(sd_30);
sd = std(sd_30);

mnlabel=sprintf('Mean — %3.2d', mn);
stdlabel=sprintf('Std Deviation — %3.2d', sd);
h=annotation('textbox',[0.7 0.57 0.1 0.1]);
set(h,'String',{mnlabel, stdlabel});
title('Distribution of standard deviations n = 30')
xlim([0 200])

subplot(7,1,4)
H_sd20 = histogram(sd_20);
mn = mean(sd_20);
sd = std(sd_20);

```

```
mnlabel=sprintf('Mean — %3.2d', mn);
stdlabel=sprintf('Std Deviation — %3.2d', sd);
h=annotation('textbox',[0.7 0.45 0.1 0.1]);
set(h,'String',{mnlabel, stdlabel});
title('Distribution of standard deviations n = 20')
xlim([0 200])

subplot(7,1,5)
H_sd10 = histogram(sd_10);
mn = mean(sd_10);
sd = std(sd_10);

mnlabel=sprintf('Mean — %3.2d', mn);
stdlabel=sprintf('Std Deviation — %3.2d', sd);
h=annotation('textbox',[0.7 0.33 0.1 0.1]);
set(h,'String',{mnlabel, stdlabel});
title('Distribution of standard deviations n = 10')
xlim([0 200])

subplot(7,1,6)
H_sd6 = histogram(sd_6);
mn = mean(sd_6);
sd = std(sd_6);

mnlabel=sprintf('Mean — %3.2d', mn);
stdlabel=sprintf('Std Deviation — %3.2d', sd);
h=annotation('textbox',[0.7 0.21 0.1 0.1]);
set(h,'String',{mnlabel, stdlabel});
title('Distribution of standard deviations n = 6')
xlim([0 200])

subplot(7,1,7)
H_sd4 = histogram(sd_4);
mn = mean(sd_4);
sd = std(sd_4);

mnlabel=sprintf('Mean — %3.2d', mn);
stdlabel=sprintf('Std Deviation — %3.2d', sd);
h=annotation('textbox',[0.7 0.09 0.1 0.1]);
set(h,'String',{mnlabel, stdlabel});
title('Distribution of standard deviations n = 4')
xlim([0 200])
```

5 References

1. Bishop, Christopher M. Pattern recognition and machine learning. springer, 2006.
2. Taylor, John. Introduction to error analysis, the study of uncertainties in physical measurements. 1997.
3. Zohdi, T. I. (2003). Large-scale statistical inverse computation of inelastic accretion in transient granular flows. The International Journal of Nonlinear Mechanics. Vol. 8, Issue 38, 1205-1219
4. Zohdi, T. I. (2005). Statistical ensemble error bounds for homogenized microheterogeneous solids.

Journal of Applied Mathematics and Physics. (Zeitschrift fÂ¨ur Angewandte Mathematik und Physik). Volume 56, Number 3. 497-515

5. Zohdi, T. I. (2015). Rapid computation of statistically-stable particle/feature ratios for consistent substrate stresses in printed flexible electronics. Journal of Manufacturing Science and Engineering, ASME. MANU-14-1476 http://dx.doi.org/10.1115/1.4029327

6. Zohdi, T. I. (2003). Genetic design of solids possessing a random-particulate microstructure. Philosophical Transactions of the Royal Society: Mathematical, Physical and Engineering Sciences. Vol: 361, No: 1806, 1021-1043.

7. Zohdi, T. I. (2003). Constrained inverse formulations in random material design. Computer Methods in Applied Mechanics and Engineering.1-20. 192, 28-30, 18, 3179-3194.

8. Zohdi, T. I. (2004). Staggering error control for a class of inelastic processes in random microheterogeneous solids. The International Journal of Nonlinear Mechanics. 39, 281-297.

9. Keaveny, T. M., Pinilla, T. P., Crawford, R. P., Kopperdahl, D. L., Lou, A. (1997). Systematic and random errors in compression testing of trabecular bone. Journal of orthopaedic research, 15(1), 101-110.