Surveying FE 208
Lecture 3

Statistics of Errors in Measurements

Learning Objectives for this Lecture

1. **Define the terms *population, sample, observation, parameter,* and *statistic***
2. **Understand statistical notation**
3. **Define the terms *mean, standard deviation,* and *standard error***
4. **Know the Central Limit Theorem**
5. **For a data set, Be able to interpret the standard deviation**
6. **For a data set, Be able to interpret the standard error**
7. **Be able to state why replication is necessary for statistical inference**

Statistics of Errors

Repeating the most important concept from last lecture:

All Measurements Have Error

Whenever we take a measurement then, the question in mind should be, "How close is this measurement to the true value, what I am going to call ***"The Truth"***

There are a few fundamental questions that we should be concerned with in regard to the measurement taken and "*The Truth*":

1. How accurate is our measurement? (mean)
2. How repeatable is our measurement? (standard deviation)
3. How confident are we in our measurement being close to the truth? (standard error)

We can answer these questions by calculating statistics on our measurements but first we need to understand the things that may be involved in what our statistics may mean.

The accuracy of our measurement can depend on a number of things. In particular, what instrument are we using to make the measurement? For example, are we measuring distance with a highly calibrated laser range finder; or are we measuring distance by pacing (counting footsteps). We would hardly expect to be as accurate with pacing.

In addition, is our practice or protocol for measuring one that places a value on accuracy or is it one that places a value on speed. For example, if using a steel tape, are we insuring that proper tension is being used and double checking for placement of the pins?

If turning angles, are we doubling or tripling angles and checking by back sighting, or just taking the first shot and moving ahead.

Another factor that plays into this is the question, "does the project have a need for high accuracy?". For example, the survey of anchor pins on a construction foundation may have highly accurate expectations with thousandths of a foot, whereas the location points for a low order forest survey may have accuracies of plus or minus several feet.

It would not be practical to follow a procedure of repeating measurements and calculating statistics each time we make a single survey measurement. However, understanding how error calculations work is important in our understanding of the importance placed on each measurement.

It is important to understand that ***all measurements are estimates***. There are no exact measurements. Because every measurement is an estimate there will be some *variance* and *distribution* around the measurement. The measures of variance will allow us to calculate some estimate of *confidence* in how good our measurement is.

Definitions

Population - A population is best defined as the entire collection of measurements that one wishes to draw some conclusion about.

Observation – The basic element of a sample. A number of observations make up the sample = n.

Sample – A group of observations made that we hope is representative of a population we have previously *well-defined*.

Parameter – Some characteristic of a *population* that we wish to know.

Statistic – A characteristic of a *sample* that we use as an estimate of some population parameter.

Let's look at how these definitions can be used. Suppose we are interested in the average length of all of the bridges in the United States.

Populations are often extremely large and trying to measure all of the elements in a population. In this case, the population would be all of the bridges in the United States. Finding and then measuring each and every bridge would be a nearly impossible task.
So we would create a list of sample bridges to measure. Our hope is that we will have chosen a sample of bridges that are representative of all the bridges in the country.

The average length of the population of bridges is what we wish to know and this is our parameter. However, without measuring each and every bridge we will never know what this value is. However, we can measure our sample bridges and calculate an average statistic that will be an estimate of the parameter. We start with our first bridge and this measurement would be an observation. We would continue to take observations (measuring each bridge) until every bridge has been measured and this total number of measured bridges becomes our sample size. It is important to recognize that if we have measured 200 bridges, our sample size is 200 but we have taken only *one sample*. 200 is the number of observations in our *single* sample.

Statistical Inference

Statistical inference can be defined as the process of drawing conclusions from data that is subject to random variation. It is easiest to think of inference as "estimate". Because our population is rarely known, we infer through statistics what the parameters of the population are. For example, because we rarely know the population mean, μ, we infer the value of μ by the estimate $\bar{x}$

Notation

To understand and work with statistical equations, it is important to understand the notation of symbols used:

$$\sum_{n=1}^{n} x_i$$

This notation is read as "The sum (∑) of measurement values (x_i) beginning at the first (1) through the last (n).

Example

For this next section we are going to work with a set of 100 numbers generated from a random number generator. This is our ***population*** (N = 100).

8	5	1	1	6
1	5	4	2	6
4	4	8	7	4
1	7	7	3	1
4	2	3	7	2
2	9	2	7	1
4	1	2	5	7
6	8	4	2	1
2	9	3	5	3
8	2	9	9	1
3	6	6	4	2
2	3	1	6	9
4	4	7	2	4
6	8	7	3	2
2	7	2	5	8
8	5	6	6	7
8	7	9	5	8
8	5	1	8	7
2	2	1	9	2
6	2	4	6	8

If we select 10 data points randomly (numbers in boxes above) from this population (n = 10), we could write this as

$$\sum_{n=1}^{10} x_i$$

If our data points were the following;

2,7,8,3,5,9,2,2,6,5 (n = 10)

Then;

$\sum_{1}^{10} x_i$ = 2+7+8+3+5+9+2+2+6+5 = 49

Normally we would not have any knowledge of the statistics in a population but in this example with a small population (N=100) we can generate population statistics to compare with our sample statistics.

The Mean

The mean ($\bar{x}$) is defined as " the mathematical average of a set of numbers". The average is calculated by adding up two or more scores and dividing the total by the number of scores. It can be calculated by:

$$\bar{x} = \frac{\sum_{n=1}^{n} x_i}{n}$$

For the example above this would be:

$$\bar{x} = \frac{\sum_{n=1}^{10} x_i}{n} = \frac{49}{10} = 4.90$$

For our population, the mean, μ = 4.68

The Standard Deviation

The standard deviation (sd) is defined as "*a measure of dispersion (or differences) of individual observations (measurements) about their mean*." It can be calculated by:

$$sd = \sqrt{\frac{\sum \left(x_i - \bar{x}\right)^2}{n-1}} \quad \text{or} \quad sd = \sqrt{\frac{\sum x_i^{\,2} - \frac{\left(\sum x_i\right)^2}{n}}{n-1}}$$

The first equation better shows what we are measuring - differences of individual observations about their mean or ($y_i - \bar{y}$). The second equation is the one more commonly used because calculation is easier in this form, particularly when using desk calculators.

Note the parentheses carefully;

For our example above;

$$sd = \sqrt{\frac{\sum x_i^{\,2} - \frac{\left(\sum x_i\right)^2}{n}}{n-1}} = sd = \sqrt{\frac{301 - \frac{(49)^2}{10}}{9}} = 2.60$$

In simple terms, the standard deviation is a measure of the variability of the unit being measured. This is the shape of the distribution curve and defines the width of the curve. The more precise the measurement, the narrower the distribution curve.

For our population, the standard deviation $\sigma = 2.74$

Note that the sigma symbol, σ is used for the population mean while the lowercase sd is used for the sample mean.

Central Limit Theorem (CLT)

The statistics that we are dealing with make the assumption of normal distributions of variances about the true mean. In fact, variances are not normally distributed however we can make this assumption based on the CLT. The CLT states that:

As the sample size increases, the distribution of sample means about the true mean becomes more normal.

NOTE: The word "sample" is singular in this statement

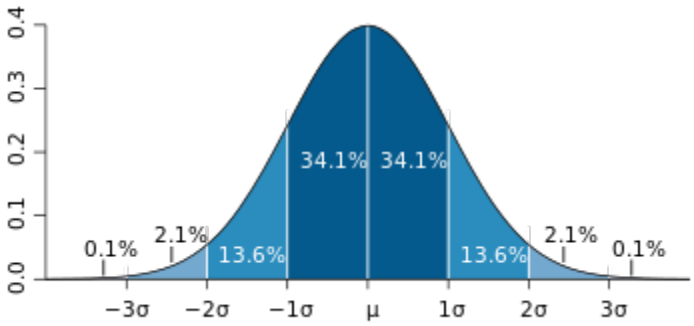
Interpreting the Standard Deviation

The normal distribution of the standard deviation allows us to make the following statements:

1. At $\pm$ one standard deviation, we expect to find approximately 68% of the data points
2. At $\pm$ two standard deviations, we expect to find approximately 95% of the data points
3. At $\pm$ three standard deviations, we expect to find approximately 99.7% of the data points

These values are estimated from the distribution of t-values and are also found in the ***68–95–99.7 rule***, also known as the ***three-sigma rule*** *or* ***empirical rule***. This states that nearly all values lie within 3 standard deviations of the mean in a normal distribution.

About 68.27% of the values lie within 1 standard deviation of the mean. Similarly, about 95.45% of the values lie within 2 standard deviations of the mean. Nearly all (99.73%) of the values lie within 3 standard deviations of the mean.



For the above example, we would expect to find about 68% of the data points between $\bar{x}$ ±1*sd or 4.9 ± 2.60 (2.3-7.5). Does this work for our example?

For the data set, 2,7,8,3,5,9,2,2,6,5 7,3,5,6,5 fall within the approximation. This is slightly off of the approximation (5 vs. 6.8) but can be explained by the small size of the sample (n = 10). In addition, a reasonable question to ask is, "Does this population come from one that is normally distributed?". The data in this case happen to be random. But the beauty of the CLT is that even these random numbers start to approach normal.

* Remember that the distribution becomes more normal as the size of the sample increases

The Standard Error

The standard error of the mean* ($SE_{\bar{y}}$) is defined as *a measure of dispersion (or differences) of sample means about the mean of all possible means*. At first this definition sounds similar to that for standard deviation, but there is a distinct difference.

It must be emphasized that the SD deals with *individual observations* and their mean while the $SE_{\bar{x}}$ deals with *sample means* and the mean of all possible sample means. It can be calculated by:

$$se_{\bar{Y}} = \frac{sd}{\sqrt{n}} = \sqrt{\frac{\sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}}{n(n-1)}}$$

For our example;

$$se_{\bar{Y}} = \frac{sd}{\sqrt{n}} = \frac{2.60}{\sqrt{10}} = 0.82$$

Interpreting the Standard Error

The normal distribution of the standard error allows us to make the following statements:

1. We are approximately 68% confident that the true mean (μ) lies between $\bar{x} \pm 1$ standard error
2. We are approximately 95% confident that the true mean (μ) lies between $\bar{x} \pm 2$ standard error
3. We are approximately 99.7% confident that the true mean (μ) lies between $\bar{x} \pm 3$ standard error

For our example we would interpret this to mean that we are 68% confident that the true mean, μ, lies between 4.90 ± 0.82 or between 4.08 and 5.72. Looking at our population mean, μ = 4.68, this looks pretty good. If we would like to have a higher confidence, we can expand the se to 1.64 (2 * se = 1.64) and interpret the results as being 95% confident that the true mean (μ) lies between 4.90 ± 1.64 or from 3.26 to 6.54

It should appear logical that in order to expand the confidence limit, (68% to 95%), you would need to expand the confidence interval.

Put simply, standard error is an estimate of how close to the population mean your sample mean is likely to be, whereas standard deviation is the degree to which individuals within the sample differ from the sample mean.

Standard error should decrease with larger sample sizes,
as the estimate of the population mean improves.
Standard deviation will be unaffected by sample size.

IMPORTANT POINT!

In order to compute measurement error, we have to have some replication of the measurement

Logically, we cannot have variance without a minimum of 2 values to compare against. It should seem logical then that the more values we have, (the larger the sample size), the better our estimate of variance becomes.

In addition, we can see this concept in the standard deviation equation:

$$sd = \sqrt{\frac{\sum \left(x_i - \bar{x}\right)^2}{n-1}} \quad \text{or} \quad sd = \sqrt{\frac{\sum x_i^{\,2} - \frac{\left(\sum x_i\right)^2}{n}}{n-1}}$$

Without replication we have division by zero

Reading for this Section
Stafford, Susan. A Statistical Primer for Foresters
Kiser, pages 8-13