# 1

# Introduction to Statistical Inference

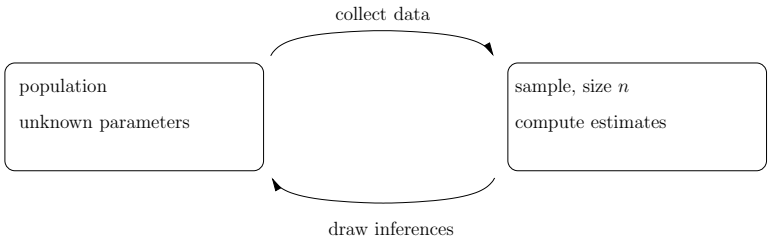
**Question.** Where does Stat 252 fit in the statistics curriculum?

**Answer.** This course provides an introduction to the fundamental and theoretical bases of the subject of statistics. In particular, we will focus on inferential statistics and the theory of estimation. We will study sampling distribution theory and the central limit theorem. We will discuss a number of different methods of estimation, including the method of moments and the method of maximum likelihood. We will learn how to construct confidence intervals using the pivotal method, and how to construct hypothesis tests via the likelihood ratio.

**Remark.** There are two definitions of statistics with which you are familiar.

Numbers computed from data are called *statistics*. The subject of *Statistics* (capitalized for emphasis) is that branch of science that deals with the interpretation of information and the analysis of data.

Computing statistics is easy, but doing statistics is hard!




In Stat 160, you were introduced to the computational aspects of doing statistics, and to the primary tools of inference, namely confidence intervals and hypothesis tests.

In Stat 251, you were taught the language of statistics, namely probability. The "correct" way to view data is as a realization of a random variable.

In Stat 252, we use the tools from probability to rigorously develop the concepts of confidence intervals and hypothesis tests.

In other Stat classes, you study different aspects of this diagram. For example, in Stat 357 you study the "collection of data" step, and in Stat 354 you study inference when there is more than one variate—the so-called "multiple regression."

**Remark.** Because of its central importance to the subject of statistics (and especially to those subjects that employ theoretical statistical techniques such as **actuarial science**), Stat 252 is often seen by students as a *hard class*. For most, the material will not come easily and will require intensive study.

### A Word about Notation

One of the general goals of this course is to teach students how to correctly express mathematical ideas. Symbols have a specific meaning, and must be used correctly if one is to be understood. The equal sign = in particular is often misused by students *even at the second year level*. As such, at this level, I like to introduce the symbols := and $\doteq$ to try and help students with their precision.

The symbol $A := B$ means $A$ *is defined to equal* $B$, whereas $C = D$ by itself means simply that $C$ and $D$ are equal. This is an important distinction because if you write $A := B$, then there is no need to verify the equality of $A$ and $B$. They are equal by definition. However, if $C = D$, then there *IS* something that needs to be proved, namely the equality of $C$ and $D$ (which might not be obvious). The symbol $A \doteq B$ means that $A$ and $B$ are approximately equal. For instance, when writing a decimal approximation to an irrational number such as $\sqrt{2}$ it is incorrect to write $\sqrt{2} = 1.41$. These two numbers are NOT equal. However, it is correct to write $\sqrt{2} \doteq 1.41$.

### The Basic Idea of Statistics: Estimating Parameters

As you are no doubt aware from your previous statistics courses, the language of statistics is probability. That is to say, although it is trivial and **BORING** to compute the summary statistics of a collection of numbers, it is fascinating to know that quite **INTERESTING** ideas such as confidence intervals and hypothesis tests develop when the sampled data is viewed as realizations of a collection of independent and identically distributed random variables.

Recall that the primary goal of inferential statistics is to estimate population *parameters*. This is done by calculating *statistics*, which are simply numbers computed from data, and using them as *point estimates* of the appropriate parameter.

As you learned in Stat 160, if you have a population with an unknown mean $\mu$ and unknown variance $\sigma^2$, and you collect a sample of data $y_1, y_2, \ldots, y_n$, then one way to estimate $\mu$ is using the sample mean

$$\overline{y} := \frac{y_1 + \cdots + y_n}{n},$$

and one way to estimate $\sigma^2$ is using the sample variance

$$s^2 := \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \overline{y})^2.$$

You will also recall that using only a single number (such as $\overline{y}$ or $s^2$) to estimate a parameter (such as $\mu$ or $\sigma^2$) is not that beneficial because it is unlikely that the computed point estimate will equal the parameter exactly. Instead, you learned that much more information is provided by a *confidence interval* or a *hypothesis test*.

Most confidence intervals that you have encountered are based on the normal distribution. For example, you are told in Stat 160 that if $\alpha \in (0,1)$ and $y_1, y_2, \ldots, y_n$ are a random sample of data from a normally distributed population with unknown mean $\mu$ and known variance $\sigma^2$, then a $(1-\alpha)100\%$ confidence interval for $\mu$ is given by

$$\overline{y} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \text{ or, equivalently, } \left[\overline{y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\ ,\ \overline{y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \tag{1.1}$$

where $z_{\alpha/2}$ is the "$z$-test critical value with tail area $\alpha/2$." In order to test the null hypothesis $H_0 : \mu = \mu_0$ against the alternative $H_A : \mu \neq \mu_0$, we compute the test statistic

$$z = \frac{\overline{y} - \mu_0}{\sigma/\sqrt{n}},$$

and then reject the null hypothesis at the significance level $\alpha$ if and only if $|z| > z_{\alpha/2}$.

You are also told that if $\sigma^2$ is unknown, then you need to replace $z_{\alpha/2}$ with $t_{\alpha/2,n-1}$, namely the "$t$-test critical value with $n-1$ degrees of freedom and tail area $\alpha/2$," in which case a $(1-\alpha)100\%$ confidence interval for $\mu$ is given by

$$\overline{y} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \text{ or, equivalently, } \left[\overline{y} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\ ,\ \overline{y} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right] \tag{1.2}$$

where $s = \sqrt{s^2} \geq 0$ denotes the *sample standard deviation*. In order to test the null hypothesis $H_0 : \mu = \mu_0$ against the alternative $H_A : \mu \neq \mu_0$, we compute the test statistic

$$t = \frac{\overline{y} - \mu_0}{s/\sqrt{n}},$$

and then reject the null hypothesis at the significance level $\alpha$ if and only if $|t| > t_{\alpha/2,n-1}$.

**Example 1.1.** The following $n = 10$ observations are a random sample from a normal population:

5.1, 2.6, 3.4, 6.3, 2.8, 4.5, 5.9, 1.7, 3.7, 8.0.

If $y_i$ denotes the $i$th observation, then the sample mean is

$$\overline{y} = \frac{5.1 + 2.6 + 3.4 + 6.3 + 2.8 + 4.5 + 5.9 + 1.7 + 3.7 + 8.0}{10} = \frac{44}{10} = 4.4,$$

and the sample variance is

$$s^2 = \frac{(5.1 - 4.4)^2 + (2.6 - 4.4)^2 + \cdots + (3.7 - 4.4)^2 + (8.0 - 4.4)^2}{9} = \frac{113}{30}$$

so that the sample standard deviation is $s = \sqrt{113/30} \doteq 1.941$. Because the population variance is unknown, a 99% confidence interval for the mean is given by

$$\overline{y} \pm t_{0.005,n-1}\frac{s}{\sqrt{n}} \text{ or, equivalently, } 4.4 \ \pm \ 3.250 \cdot \frac{\sqrt{113/30}}{\sqrt{10}}$$
$$\text{or, equivalently, } [2.405, 6.395]$$

since $t_{0.005,9} = 3.250$. In order to test the null hypothesis $H_0 : \mu = 4$ against the alternative hypothesis $H_A : \mu \neq 4$ at the $\alpha = 0.01$ significance level, we compute the test statistic

$$t = \frac{\overline{y} - \mu_0}{s/\sqrt{n}} = \frac{4.4 - 4}{\sqrt{113/30}/\sqrt{10}} \doteq 0.652,$$

and note that since the absolute value of the test statistic is less than the critical value $t_{0.005,9} = 3.250$, we do not reject the null hypothesis.

However, as we partially discussed in Stat 251, and which will be covered in much more detail later in this course, the point of view of statistics is that the data are realizations of random variables. Therefore, we "pretend" that we have not seen the numerical data and assume that our sample consists of random variables $Y_1, Y_2, \ldots, Y_n$. In Stat 252, we will often assume that we have a *random sample* which is synonymous with *independent and identically distributed random variables*. Therefore, suppose that $Y_1, Y_2, \ldots, Y_n$ are a random sample from a population with common mean $\mu$ and common variance $\sigma^2$. The sample mean $\overline{Y}$ is the random variable

$$\overline{Y} := \frac{Y_1 + \cdots + Y_n}{n},$$

and the sample variance is the random variable

$$S^2 := \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

We say that $\overline{Y}$ and $S^2$ are *(point) estimators* of the population mean $\mu$ and population variance $\sigma^2$, respectively. In Stat 252, we will study many of the properties of these (point) estimators including the reason that they are so

frequently used; namely, we will prove later in this course that $\overline{Y}$ is an unbiased estimator of $\mu$, and $S^2$ is an unbiased estimator of $\sigma^2$.

In the particular case that $Y_1, Y_2, \ldots, Y_n$ are a random sample of normal random variables with common mean $\mu$ and common variance $\sigma^2$, we know from Stat 251 that $\overline{Y} \sim \mathcal{N}(\mu, \sigma^2/n)$ and

$$\frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1).$$

Moreover, we know that if $Z \sim \mathcal{N}(0,1)$ and $\alpha \in (0,1)$, then there exists a unique number $z_{\alpha/2}$ with the property that $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$. This implies that

$$P\left(-z_{\alpha/2} \leq \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

and so isolating $\mu$ implies

$$P\left(\overline{Y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \overline{Y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha. \tag{1.3}$$

Hence, we see from (1.3) that this is precisely the interpretation of the confidence interval given by (1.1). Before the data are collected, the interval

$$\left[\overline{Y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\ ,\ \overline{Y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right],$$

both of whose endpoints are random variables, has probability $1 - \alpha$ of containing the parameter $\mu$. After the data are collected, no such probabilistic statement is valid: the interval

$$\left[\overline{y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\ ,\ \overline{y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right],$$

both of whose endpoints are numbers, may or may not contain $\mu$.

Furthermore, it will be shown in Stat 351 that

$$\frac{\overline{Y} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

Here $T \sim t(n-1)$ indicates that $T$ has a $t$-distribution with $n-1$ degrees of freedom. Moreover, we know that if $T \sim t(n-1)$ and $\alpha \in (0,1)$, then there exists a unique number $t_{\alpha/2,n-1}$ with the property that

$$P(-t_{\alpha/2,n-1} \leq T \leq t_{\alpha/2,n-1}) = 1 - \alpha$$

which implies that

$$P\left(-t_{\alpha/2,n-1} \leq \frac{\overline{Y}-\mu}{S/\sqrt{n}} \leq t_{\alpha/2,n-1}\right) = 1-\alpha.$$

Isolating $\mu$ yields

$$P\left(\overline{Y} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{Y} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right) = 1-\alpha. \tag{1.4}$$

As before, we see from (1.4) that this is precisely the interpretation of the confidence interval given by (1.2). Before the data are collected, the interval

$$\left[\overline{Y} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\ ,\ \overline{Y} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right],$$

both of whose endpoints are random variables, has probability $1-\alpha$ of containing the parameter $\mu$. After the data are collected, no such probabilistic statement is valid: the interval

$$\left[\overline{y} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\ ,\ \overline{y} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right], \tag{1.5}$$

both of whose endpoints are numbers, may or may not contain[1] $\mu$.

One of the things that you will do in Stat 252 is **prove** these formulæ, and derive other, more general, formulæ for confidence intervals. It should also be noted that there are, of course, theoretical interpretations for hypothesis tests. These will certainly be discussed later in this course.

[1] This is really one of the subtle points about confidence intervals. Since the interval (1.5) is deterministic given the data, the event $\left\{\overline{y} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \leq \mu \leq \overline{y} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}\right\}$ occurs with probability equal to either 0 or 1. For instance, in Example 1.1 we have $P(2.405 \leq \mu \leq 6.395) \in \{0,1\}$. Without knowing the true value of the parameter $\mu$, there is no way for us to determine if this probability is 0 or if it is 1.