

# 1 Random Samples

We rarely have access to data for an entire population. To do any meaningful analysis, we instead work with a sample of data taken from the population. Ideally, we will obtain a **random sample** of data. The random sample will have  $n$  observations consisting of random variables  $X_1, X_2, \dots, X_n$ , and it will satisfy the following properties:

- (a) Each  $X_i$  has the same mean  $\mu$ .
- (b) Each  $X_i$  has the same variance  $\sigma^2$ ;
- (c) Each  $X_i$  is statistically independent of any different  $X_j$ .

Here's a story that might help with the intuition.

Suppose I want to know something about the average GPA of UC-Davis undergrads; call this  $\mu$ . I don't have access to the registrar or anything like that, and I'm not about to ask every single UC-Davis undergrad what their GPA is, so I don't know what  $\mu$  is. Instead, I'll estimate it by asking  $n = 50$  undergrads. The first undergrad I will eventually ask is represented by the random variable  $X_1$ , the second by  $X_2$ , and so on.

Now I go out and start asking people. Suppose the first person I ask says their GPA is  $x_1 = 3.75$ . The second person says their GPA is  $x_2 = 3.68$ . The third person I ask has a GPA of 1.12 but doesn't give me an answer because they're kinda embarrassed by it, so it doesn't get added to my sample. This pattern repeats itself: only people with high GPAs are comfortable answering, so I end up collecting mostly above-average GPAs. In this case, I would expect the GPA of those who answer to be *higher* than the true GPA. In other words,  $E[X_i] > \mu$  because of a selection bias. This violates condition (a), and we want to avoid things like this in our samples.

If instead I get everyone to answer, then I expect the first person I ask to have a GPA of  $\mu$ . This doesn't mean it *will* be equal to  $\mu$  (and in fact it probably won't be), but I still expect it to be  $\mu$  in the same way that I expect to get five heads and five tails if I flip a fair coin ten times. This is also true for the second person and third person and so forth. Mathematically,  $E[X_i] = \mu$  for all  $i$ . This satisfies condition (a), and the same chain of logic holds for condition (b) as well so that  $\text{Var}(X_i) = \sigma^2$  for all  $i$ .

Condition (c) means that the GPA of the first person I ask doesn't have any influence on the GPA of the second person I ask. The first person I ask might tell me their GPA is  $x_1 = 4.0$ . The second person hears this, feels insecure, and lies by telling me that their GPA is  $x_2 = 3.90$ , even though it's really 3.40. In this case,  $X_1$  and  $X_2$  are not statistically independent: a high answer for  $X_1$  makes it more likely that I'll get another high answer for  $X_2$ . This is something we want to avoid in our samples;  $X_2$  should not depend on  $X_1$ .

Unless stated otherwise, we will assume that all samples are random samples.

## 2 Sample Mean

### 2.1 Definition and Intuition

For random variables  $X_1, \dots, X_n$  of a random sample, the **sample mean** is defined as

$$\bar{X} \equiv \frac{X_1 + \dots + X_n}{n}. \quad (1)$$

We use  $\bar{X}$  as the estimate of  $\mu$ . But because it's just an estimate, it's pretty much guaranteed to not be exactly right.

It is of critical importance to recognize that the number we calculate for  $\bar{x}$  depends on our random sample: a sum of random variables is itself a random variable, so  $\bar{X}$  will have an expected value and a variance. In other words, if I go out and ask 50 random people what their GPA is, and you go out and ask 50 random people what their GPA is, chances are we're each going to ask a different group of 50 people and therefore we'll each calculate a different  $\bar{x}$ . That's why  $\bar{X}$  is a random variable with an expected value and a variance: it's totally possible (unlikely, but possible) that I get unlucky and randomly sample a group of 50 weird people and get a bunch of weird answers that aren't representative of the population, and I (unknowingly) get a lousy estimate because of that. We have to account for the possibility that our sample is a bad one — that even though we have a sample, there remains some *uncertainty* about what  $\mu$  actually is.

### 2.2 Expected Value and Variance

To analyze the uncertainty about our sample mean, we need to know the expected value and variance of  $\bar{X}$ .

The expected value operator is a *linear operator*, which means it satisfies the two following properties:

- (a)  $E[X + Y] = E[X] + E[Y]$  for any random variables  $X$  and  $Y$ ;
- (b)  $E[aX] = aE[X]$  for any random variable  $X$  and any real number  $a$ .

Using the definition of the sample mean  $\bar{X}$ , we can write

$$E[\bar{X}] = E\left[\frac{X_1 + X_2 + \dots + X_n}{n}\right].$$

Property (a) says we can write the expectation of a sum as the sum of expectations, so

$$\begin{aligned} E[\bar{X}] &= E\left[\frac{X_1 + X_2 + \dots + X_n}{n}\right] \\ &= E\left[\frac{X_1}{n}\right] + E\left[\frac{X_2}{n}\right] + \dots + E\left[\frac{X_n}{n}\right]. \end{aligned}$$

Property (b) says that we can take the scalar  $1/n$  out of each expectation so that

$$E[\bar{X}] = \frac{1}{n}E[X_1] + \frac{1}{n}E[X_2] + \dots + \frac{1}{n}E[X_n].$$

We have a common factor of  $1/n$  in each term, so we can factor it out and write

$$E[\bar{X}] = \frac{1}{n}\left(E[X_1] + E[X_2] + \dots + E[X_n]\right).$$

By the assumptions of a random sample,  $E[X_i] = \mu$  for all  $X_i$ . This substitution gives

$$\begin{aligned} E[\bar{X}] &= \frac{1}{n}\underbrace{\left(\mu + \mu + \dots + \mu\right)}_{n \text{ times}} \\ &= \frac{1}{n}(n\mu). \end{aligned}$$

Cancel out an  $n$  in the numerator and denominator and we are left with the result:

$$E[\bar{X}] = \mu. \quad (2)$$

In words, we expect our sample mean to be the true mean. Again, this doesn't mean it *will* be equal to  $\mu$  (and in fact it probably won't be), but we still expect it to be  $\mu$  in the same way that we expect to get five heads and five tails if we flip a fair coin ten times. The sample mean will be correct *on average*.

Let's do the same thing for variance. Variance is not a linear operator, but as long as all  $X_i$  are independent from each other (which we are assuming as part of the random sample assumptions), then it will be similar. Specifically,

- (i)  $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$  for independent random variables  $X$  and  $Y$ ;
- (ii)  $\text{Var}(aX) = a^2 \text{Var}(X)$  for any random variable  $X$  and any real number  $a$ .

Now let's do the whole rigmarole for variance.

Using the definition of the sample mean  $\bar{X}$ , we can write

$$\text{Var}(\bar{X}) = \text{Var}\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right).$$

Property (i) says that we can write the variance of an independent sum as the sum of variances, which yields

$$\begin{aligned}\text{Var}(\bar{X}) &= \text{Var}\left(\frac{X_1 + X_2 + \dots + X_n}{n}\right) \\ &= \text{Var}\left(\frac{X_1}{n}\right) + \text{Var}\left(\frac{X_2}{n}\right) + \dots + \text{Var}\left(\frac{X_n}{n}\right).\end{aligned}$$

Property (ii) says that we can take the scalar  $1/n$  out of each variance *and square it* so that

$$\text{Var}(\bar{X}) = \frac{1}{n^2} \text{Var}(X_1) + \frac{1}{n^2} \text{Var}(X_2) + \dots + \frac{1}{n^2} \text{Var}(X_n).$$

We have a common factor of  $1/n^2$  in each term, so we can factor it out and write

$$\text{Var}(\bar{X}) = \frac{1}{n^2} \left[ \text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_n) \right].$$

By the assumptions of a random sample,  $\text{Var}(X_i) = \sigma^2$  for all  $X_i$ . This substitution gives

$$\begin{aligned}\text{Var}(\bar{X}) &= \frac{1}{n^2} \underbrace{\left[ \sigma^2 + \sigma^2 + \dots + \sigma^2 \right]}_{n \text{ times}} \\ &= \frac{1}{n^2} \left[ n\sigma^2 \right].\end{aligned}$$

Cancel out the  $n$  in the numerator and denominator and we are left with the result:

$$\text{Var}(\bar{X}) = \frac{\sigma^2}{n}. \quad (3)$$

In practice we will more often be using the standard deviation of  $\bar{X}$ , sometimes called the **standard error of the sample mean**, denoted  $\text{se}(\bar{X})$  and given by

$$\text{se}(\bar{X}) \equiv \frac{\sigma}{\sqrt{n}}. \quad (4)$$

It is helpful to think of the standard error as being the uncertainty of our estimate  $\bar{X}$ . I rec-

ommend becoming comfortable with the term *standard error* because it will be appearing frequently once we get into regression analysis.

Notice that as  $n$  gets bigger and bigger, the standard error (and variance) becomes smaller and smaller. In fact, as  $n \rightarrow \infty$ , the standard error goes to zero. This is the **law of large numbers** at work: as we get more and more data, our estimates get better and better, less and less uncertain. As we ask more and more people for their GPA, our estimate becomes better.

## 3 Central Limit Theorem

### 3.1 Intuition

It's good that we know the expected value and variance of  $\bar{X}$ . We can go further.

When you collect a random sample and calculate its sample mean, most of the time it's going to be pretty close to  $\mu$ . But every now and then, you'll collect a weird sample and your sample mean will be pretty far off.

Suppose a bunch of us go out and randomly ask 50 people what their GPA is, and we each come back with our own  $\bar{X}$ . The true mean is  $\mu = 3.25$ , and most of us get an estimate that's pretty close to that. One or two of us, just by chance, ended up asking a group of weird people and obtained estimates as far away as 2.6 or 3.8. But still, most of our estimates are fairly close to  $\mu$ . If we were to plot a histogram of our estimates, it would look something like the histogram below.

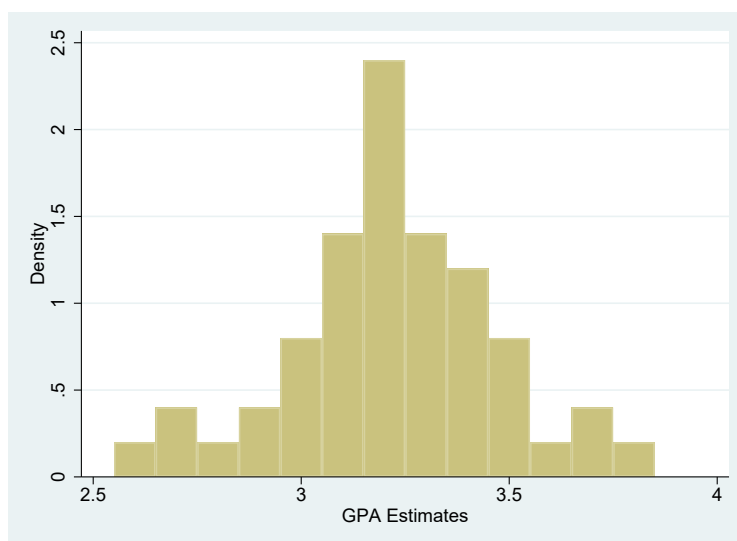


FIGURE 1: Most estimates are fairly close to  $\mu = 3.25$ , but a small number are pretty far away. It almost kinda looks like a normal distribution; hold that thought.

This was just a thought experiment, though. In reality we only collect one sample, as large of a sample as we can, and calculate one  $\bar{X}$ . So in practice, we could be anywhere on such a histogram *and we don't know where for sure*. We have no idea whether our sample is good and our estimate is in the peak near  $\mu$ , or whether our sample is weird and our estimate is in the tail far from  $\mu$ . The person who estimated GPA as 2.6 had no idea how bad their estimate was until they looked at the other estimates — in reality, we don't have the other estimates to compare ours to.

Note that if instead we each ask  $n = 500$  people what their GPA is, then we'll get a histogram that looks more like the green one below. In this case, we might not be so worried about whether our estimate is a bad one or not. That's why large sample sizes are nice.

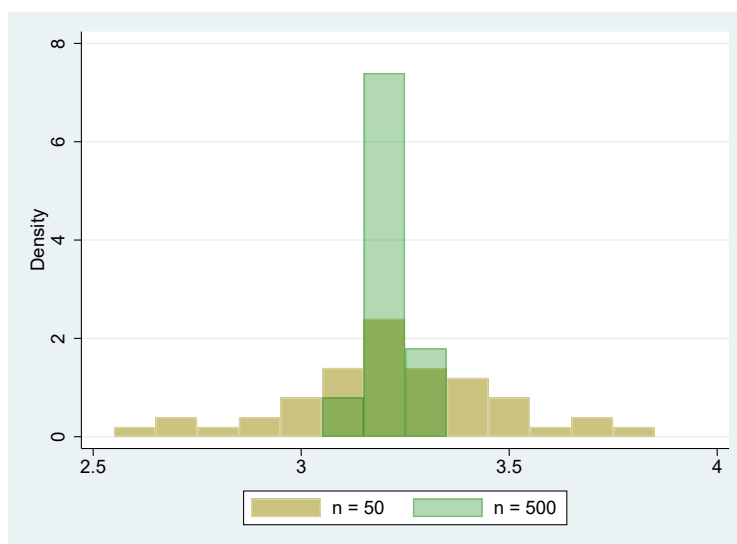


FIGURE 2: The estimates with a larger sample size are a lot more closely concentrated around  $\mu = 3.25$ . Again, that is the law of large numbers at work: you are seeing  $\sigma / \sqrt{n}$  get smaller as  $n$  gets bigger.

### 3.2 Standardization

Okay, great, grand, wonderful. The problem is that every time we want to estimate something, we'll probably be estimating something with a different mean and a different variance. Estimating GPA and estimating student debt will have different means and different standard deviations and therefore different-looking estimate histograms. That's why we will *standardize* the procedure.

Suppose arbitrary random variable  $X$  has mean  $\mu$  and standard deviation  $\sigma$ . Taking  $X - \mu$  will shift the entire distribution so that it is centered at zero. And then dividing by  $\sigma$  will re-shape the distribution so that it has variance of one. In other words, the

**standardization** of random variable  $X$ , here denoted by random variable  $Z$  is given by

$$Z = \frac{X - \mu}{\sigma} \sim (0, 1),$$

where notation  $\sim (0, 1)$  reads as “is distributed with mean 0 and variance 1.”

We can do this with  $\bar{X}$  as well. It has mean  $\mu$  and standard deviation  $\sigma / \sqrt{n}$ , so

$$Z \equiv \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim (0, 1). \quad (5)$$

Sometimes in this context,  $Z$  is called the **Z-statistic**.

We can get even more specific. The above histograms for  $\bar{X}$  looked like they were normally distributed. To that end, here’s what we can say:

(A) If  $\sigma$  is known and  $30 < n < \infty$ , then it is approximately true that

$$Z \equiv \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim \mathcal{N}(0, 1).$$

(B) If  $\sigma$  is known and the underlying distribution is normal, then  $Z \sim \mathcal{N}(0, 1)$  exactly, even for small  $n$ .

(C) If  $n \rightarrow \infty$ , then  $Z \sim \mathcal{N}(0, 1)$  exactly, regardless of the underlying distribution.

Proposition (C) is known as the **central limit theorem**.

### 3.3 Unknown Variance

Again, this is all nice and everything, but we’ve been making a not-so-innocent assumption throughout. When we standardized  $\bar{X}$ , we did so as if we knew what the population standard deviation  $\sigma$  was. In practice, we won’t know what  $\sigma$  is and we have to use an estimate instead. The estimate for  $\sigma$  is given by

$$S \equiv \sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}. \quad (6)$$

Now we standardize  $\bar{X}$  to obtain the **T-statistic**,

$$T \equiv \frac{\bar{X} - \mu}{S / \sqrt{n}} \sim T(n-1), \quad (7)$$

where  $T(n - 1)$  is the ***T-distribution*** with  $n - 1$  *degrees of freedom*.

A  $T$ -distribution is centered at zero and has a bell shape much like the normal distribution, but it has fatter tails. The idea is that our estimate of  $\sigma$  is itself uncertain, so we're using an uncertain thing to describe an uncertain thing. That is, we have increased uncertainty about  $\bar{X}$  from having estimated  $\sigma$ , and we need to account for that with fatter tails in the distribution of  $\bar{X}$ . Note that  $T(\infty) = \mathcal{N}(0, 1)$ . Things get weird with  $T(1)$ .

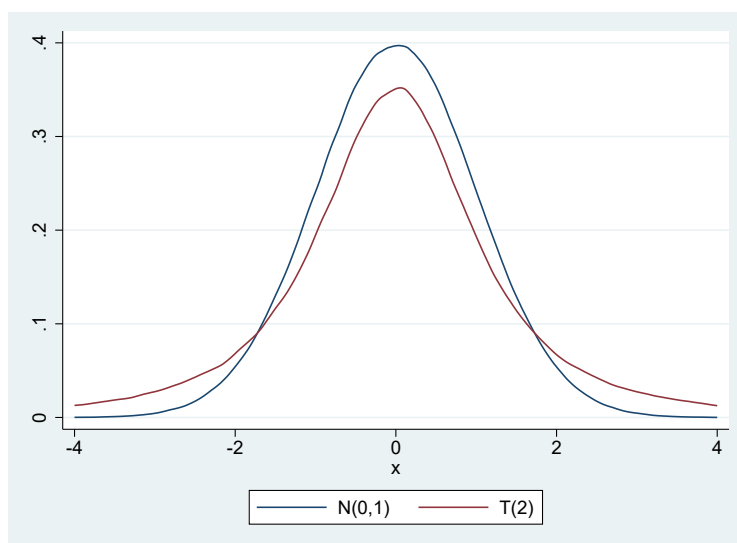


FIGURE 3: A  $T$ -distribution has fatter tails than a normal distribution, but the shape is still similar.

Now here's what we can say:

- (I) If  $\sigma$  is not known and  $30 < n < \infty$ , then it is approximately true that

$$T \equiv \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim T(n - 1).$$

- (II) If  $\sigma$  is not known and the underlying distribution is normal, then  $T \sim T(n - 1)$  exactly, even for small  $n$ .

The approximate case become exact as  $n \rightarrow \infty$ . On paper, if  $n > 30$ , then you can usually use the normal distribution instead of  $T(n - 1)$  because they will be very similar. (In fact,  $T(\infty)$  is exactly standard normal.) If you're using Stata, then just use  $T(n - 1)$ . In practice, we will usually have unknown  $\sigma$  and  $n > 30$ , so the  $T(n - 1)$  distribution is used heavily.

Note that if  $n \leq 30$ , then we can only do "reliable" inference if we have reason to believe that the underlying data is normally distributed. Accordingly, you should be skeptical of inference on small sample sizes.