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 $X_1, X_2, \ldots, X_n$, and it will satisfy the following properties:

1. Each X_i has the same mean μ;
2. Each X_i has the same variance σ^2;
3. Each X_i is statistically independent of any different X_j.

Our goal is to obtain a *representative sample*, that is, a sample distribution that is quantitatively similar to the population distribution. Unless stated otherwise, we will assume that all samples are random samples.

Here's a story that might help with the intuition. Suppose I want to know the average GPA of UC-Davis undergrads; call this μ. 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. 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 tells me their GPA of $x_1 = 3.75$. The second person tells me their GPA of $x_2 = 3.68$. 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. Our sample won't be representative.

If instead I get everyone to answer, then I "expect" each person I ask to have a GPA of μ. This doesn't mean any given student *will* have a GPA equal to μ (in fact they probably won't), but I still "expect" it to be μ in the same way that I "expect" to get five heads and five tails if I flip a fair coin ten times. Mathematically, $E[X_i] = \mu$ for all i. This satisfies condition (a); similar 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. Our sample won't be representative if any X_i depend on X_j.

2 Sample Mean

2.1 Definition and Intuition

For random variables $X_1, \ldots, X_n$ of a random sample, the **sample mean**, $\overline{X}$, is defined as

$$\overline{X} \equiv \frac{X_1 + \ldots + X_n}{n}. \tag{1}$$

We use $\overline{X}$ as the estimate of μ. 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 $\overline{x}$ depends on our random sample: a sum of random variables is itself a random variable, so $\overline{X}$ 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 $\overline{x}$. That's why $\overline{X}$ 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 — even though we have a sample and an estimate of μ, there remains some *uncertainty* about what μ actually is.

2.2 Expected Value and Variance

To analyze the uncertainty about the sample mean, we need to know the expected value and variance of $\overline{X}$. We want to know whether our estimate is correct, on average; and we want to know just how far off our estimate is likely to be.

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

1. $E[X + Y] = E[X] + E[Y]$ for any random variables X and Y;
2. $E[aX] = aE[X]$ for any random variable X and any real number a.

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

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

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

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

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

$$E[\overline{X}] = \frac{1}{n}E\left[X_1\right] + \frac{1}{n}E\left[X_2\right] + \ldots + \frac{1}{n}E\left[X_n\right].$$

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

$$E[\overline{X}] = \frac{1}{n}\Big(E[X_1] + E[X_2] + \ldots + E[X_n]\Big).$$

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

$$\begin{aligned} E[\overline{X}] &= \frac{1}{n}\underbrace{\Big(\mu + \mu + \ldots + \mu\Big)}_{n \text{ times}} \\ &= \frac{1}{n}(n\mu). \end{aligned}$$

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

$$E[\overline{X}] = \mu. \tag{2}$$

In words, the sample mean will be the true mean *on average*. In general, an estimator is said to be **unbiased** when its expected value equals the true parameter. Again, this doesn't mean it *will* equal μ (in fact it probably won't), but we still "expect" it to be μ in the same way that we "expect" to five heads and five tails if we flip a fair coin ten times.

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 $\overline{X}$, we can write

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

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

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

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

$$\text{Var}(\overline{X}) = \frac{1}{n^2}\text{Var}\left(X_1\right) + \frac{1}{n^2}\text{Var}\left(X_2\right) + \ldots + \frac{1}{n^2}\text{Var}\left(X_n\right).$$

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

$$\text{Var}(\overline{X}) = \frac{1}{n^2}\Big[\text{Var}(X_1) + \text{Var}(X_2) + \ldots + \text{Var}(X_n)\Big].$$

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

$$\begin{aligned}\text{Var}(\overline{X}) &= \frac{1}{n^2}\underbrace{\Big[\sigma^2 + \sigma^2 + \ldots + \sigma^2\Big]}_{n \text{ times}} \\ &= \frac{1}{n^2}\Big[n\sigma^2\Big].\end{aligned}$$

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

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

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

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

It is helpful to think of the standard error as being the uncertainty of our estimate $\overline{X}$. 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 $n \to \infty$, the standard error goes to zero. This is the **law of large numbers** at work: as we get more and more data, our estimate is probably getting closer and closer to μ, a property called **consistency**. As we ask more and more people for their GPA, our estimate becomes better. This is illustrated in the next section.

3 Central Limit Theorem

3.1 Intuition

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

When you collect a big enough random sample (say, $n > 30$) and calculate its sample mean, most of the time it's going to be pretty close to μ. 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 $\overline{X}$. 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 μ. If we were to plot a histogram of our estimates, it would look something like the histogram in Figure 1.

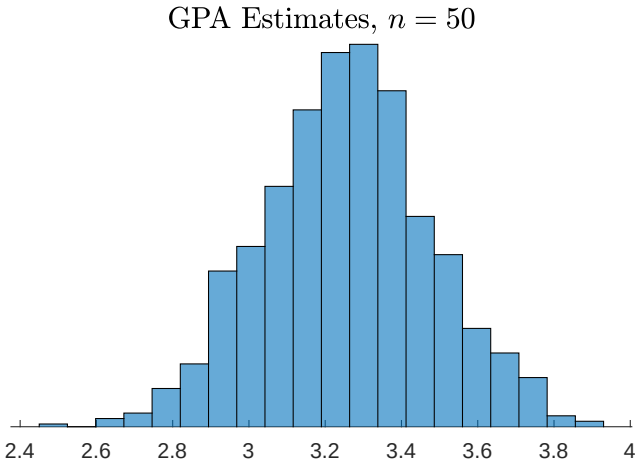

FIGURE 1: Most estimates are fairly close to $\mu = 3.25$, but a small number are pretty far away. It almost kinda looks like a normal distribution; hold that thought.

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 $\overline{x}$. 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 representative and our estimate is in the peak near μ, or whether our sample is weird and our estimate is in a tail far from μ. The person who estimated the average GPA as 2.6 had no idea how bad their estimate was until they looked at the histogram of 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 in Figure 2. 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.

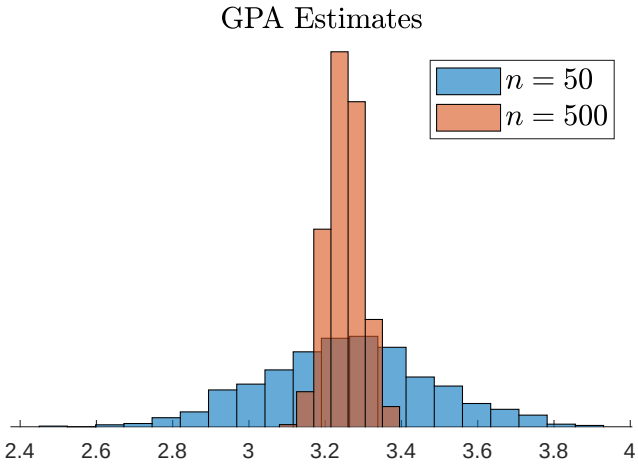

FIGURE 2: The estimates with a larger sample size (in red) are a lot more closely concentrated around $\mu = 3.25$. This is the law of large numbers at work: you're seeing the standard error $\sigma/\sqrt{n}$ get smaller as n gets bigger. As $n \to \infty$, only the bar exactly at μ remains: the estimator is consistent.

3.2 Standardization

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

Forget about estimates for a second. Suppose arbitrary random variable X has mean μ and standard deviation σ. Taking $X - \mu$ will shift the entire distribution so that it is centered at zero. And then dividing by σ 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 $\overline{X}$ as well. It has mean μ and standard deviation $\sigma/\sqrt{n}$, so

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

But we can get more specific. The previous histograms for $\overline{X}$ looked normally distributed. To that end, we can say the following:

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

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

where $\mathcal{N}(0, 1)$ is the **standard normal distribution**.

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

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

In this context, Z is called the ***Z*-statistic**. Proposition (C) is the **central limit theorem**.

3.3 Unknown Variance

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

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

Now we standardize $\overline{X}$ using the sample standard error $S/\sqrt{n}$ to obtain the ***T*-statistic**,

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

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 like the normal distribution but with fatter tails, as illustrated in Figure 3. The idea is that our estimate of σ is itself uncertain, so we're using one uncertain thing to describe another uncertain thing. We have increased uncertainty about $\overline{X}$ from having only an estimate of σ, and we account for that with fatter tails in the distribution of $\overline{X}$. A $T(1)$ distribution (the *Cauchy distribution*) has really fat tails and weird properties. As degrees of freedom increase, tails get thinner until $T(\infty) = \mathcal{N}(0,1)$.

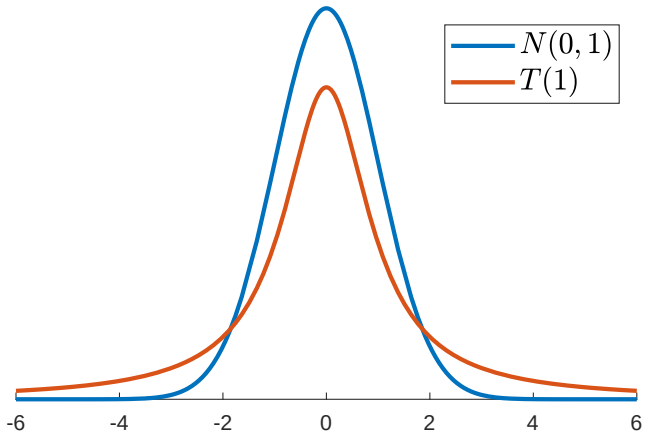

FIGURE 3: A T-distribution has fatter tails than a normal distribution, but the shape is still similar.

Now here's what we can say:

(I) If σ is not known and is estimated with S; and $30 < n < \infty$; then it is approximately true that

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

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

The approximate case become exact as $n \to \infty$. On paper, you can usually use the normal distribution instead of $T(n-1)$ when $n > 30$ because they will be very similar. If you're using R, then just use $T(n-1)$. In practice we will usually have unknown σ 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.