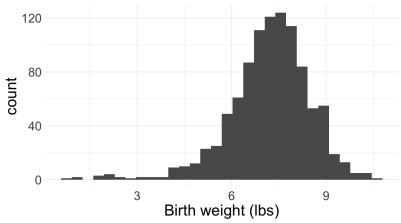
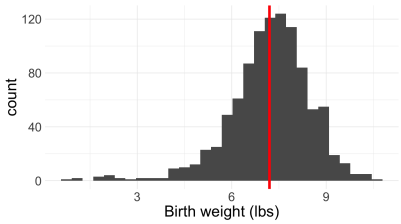
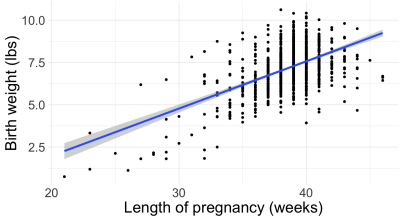
1 Introduction

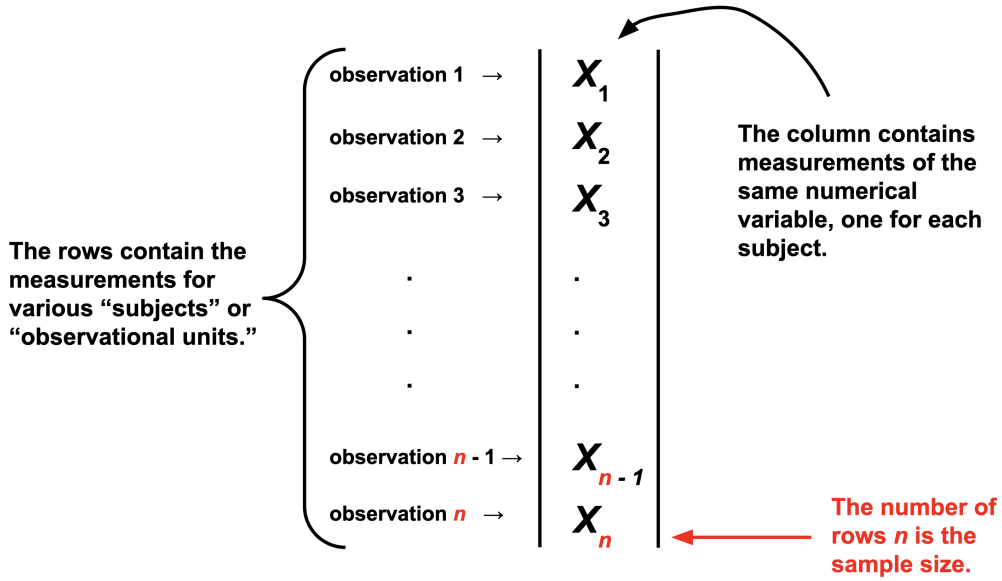
Imagine you work at a hospital. Dozens of babies are delivered in the hospital each week. A lot of information is collected about each birth, and hiding inside of these data is a lot of useful information about the health and future of the community. How can we extract it? If you take classes like STA 101 or 199, you learn various methods:

Question	**Method**	**Cartoon**
How are birth weights distributed?	histogram	
What's the typical birth weight?	sample average	
How is birth weight related to gestational age?	linear regression	

In general, the goal of *data science* is to convert data into knowledge, and we perform this "conversion" by applying statistical methods: histogram, sample average, line of best fit, etc. A *mathematical statistician* looks at these methods and wonders: do they actually work? Do they behave the way we expect? Do they deliver what they promise? Are they reliable? How reliable? Under what conditions? And before we can answer these questions, we have to back up and ask a more fundamental one: what does it even *mean* for a statistical method to "work"? These are all theoretical questions, and mathematical statisticians answer them using the tools of probability theory that we have studied for the last twelve weeks.

1.1 Baby's first dataset

For us, a data set will be a spreadsheet with one column of numbers.



Of course, in the modern era, datasets are huge. There could be millions of columns. There could be so many columns and rows that you cannot fit the entire dataset on a single computer (so-called "big data"). Furthermore, modern datasets are weird. It is not just a box of numbers anymore. Text is data. Images are data. Video is data. Sometimes all at once. If you continue studying statistics, you will get there eventually. But for now, let us keep it simple.

1.2 The implicit assumption of all statistics

Observed data are the result of a random process.

By the time the data arrive in our spreadsheet, they are a fixed set of numbers. But how did those numbers get there in the first place? Plenty of seemingly random forces might have had their way before we ever got to observe anything:

- (**nature's randomness**) many of the phenomena we study have an intrinsic random component that is simply irreducible. Think mutations during DNA replication, the quantum behavior of subatomic particles, or the "random walk" in stock prices;
- (**human error**) data are collected by humans, and humans make mistakes;
- (**measurement error**) the laboratory devices we use to collect measurements in the sciences are far from perfect;
- (**study design**) the gold standard for teasing out cause and effect is the randomized controlled trial (RCT) where, *by design*, the researcher randomly divides subjects into treatment and control groups;

- (**survey non-response**) much of our economic data on unemployment, inflation, and the behavior of firms is collected by survey. Political polls are a type of survey. Course evaluations are a survey. Say you issue a survey to a population of interest, and only 10% respond. Who are they? Why did they respond? Why did the other 90% abstain? There is a lot of randomness going on here, and you need to get your arms around it if you want to interpret the survey results correctly.

For all of these reasons and more, it is sensible to regard the numbers in our spreadsheet as the end result of a complex random process. One of the statistician's goals is to model this process and explain how the data turned out the way they did. Part of this job involves modeling the true underlying science, and another part involves modeling the errors introduced in the measurement process. Tricky stuff!

1.3 Classical statistical inference

Since the data in our spreadsheet are the result of a random process, we model them on the blackboard as realizations of random variables. To keep it simple, on a first pass we model the data as **independent and identically distributed** (**iid**) from some shared distribution P:

$$X_1, X_2, X_3, \ldots, X_{n-1}, X_n \overset{\text{iid}}{\sim} P.$$

If you keep studying statistics, you learn how to relax the iid assumption, which is often bogus.

The main idea of statistics is that we have no clue what the distribution P is that governs the data. All we have access to are realizations of the X_i, and we need to use these to try to reverse engineer P. Contrast this with probability. When we were doing probability, P was always known. Every problem thus far has begun with a statement equivalent to: "Assume P is…" You were *told* what the distribution was, and then you did some math that reasoned from that assumption to a conclusion about the behavior of X: its mean, median, variance, mgf, how to sample it, etc. In statistics, the situation is exactly reversed: given examples of data generated from P, can you figure out what P is?

1.4 Parametric statistics

In principle, P could be literally any probability distribution under the sun. That makes searching for the right one quite daunting. To simplify life, we often make the *extra* assumption that the unknown P belongs to some convenient **parametric family** of distributions, such as the binomial, Poisson, normal, etc:

$$X_1, X_2, X_3, \ldots, X_{n-1}, X_n \overset{\text{iid}}{\sim} f_\theta.$$

Here, f is generic notation for *either* a PMF or PDF, and θ is generic notation for the *finite* set of parameters that govern the distribution. Here are some of the examples we have seen:

- The Bernoulli family has the probability of success $\theta = p$;
- The Poisson family has the rate $\theta = \lambda$;
- The normal family has the mean and variance $\theta = (\mu, \sigma^2)$;

- The gamma family has the shape and rate $\theta = (\alpha, \beta)$.

If you are willing to swallow the parametric assumption, the goal of statistical inference becomes "use the data to learn the value of the unknown parameter θ." As soon as you know the parameters, you know the entire distribution. Furthermore, statisticians want to *quantify uncertainty* about what we have learned from the data, not just produce a good guess and call it a day.

To be clear from the very outset, the "true" data-generating distribution is seldom if ever literally a member of one of our special parametric families, but as a simplifying assumption, this leap of faith may be "close enough" for practical purposes. If you remain skeptical, I encourage you to dive into the beautiful area of *nonparametric statistics* where we do *not* make strong (and often unrealistic) parametric assumptions about P.

Example 1.1 (Bernoulli data (coin flip))**.** Imagine we have *binary* data:

$$X_i = \begin{cases} 1 & \text{something happened} \\ 0 & \text{something didn't happen.} \end{cases}$$

Assuming the X_i are independent and identically distributed (iid), there is really only one distribution that could have generated these data: Bernoulli! This is a special case where we don't have to worry about whether our model for the distribution is appropriate, because there is only one. So we know

$$X_1, X_2, X_3, \ldots, X_{n-1}, X_n \overset{\text{iid}}{\sim} \text{Bernoulli}(\theta).$$

This serves as a model for the archetypal statistics problem: given a mystery coin with an unknown probability of heads $0 \leq \theta \leq 1$, you flip the coin n times and use the outcomes X_i to guess what the underlying probability is. The goal of statistics is to use the data to learn what θ is *and also* to quantify our uncertainty about it.

1.5 Three's Company

The Catholics have the Father, the Son, and the Holy Spirit. American government has legislative, judicial, and executive branches. Music has melody, harmony, and rhythm. Similarly, we organize classical parametric statistical theory into three related areas of inquiry:

1. (**point estimation**) what is our single number best guess at the unknown parameter?
2. (**interval estimation**) what is a range of likely values for the unknown parameter?
3. (**hypothesis testing**) can the data distinguish between competing claims about the unknown parameter?

With the short time we have left, we will discuss the first two and defer the third to STA 332. Hypothesis testing turns out to be a surprisingly controversial topic. Spicy!