Statistical Modelling and Distributions

Jason Willwerscheid

2022-10-24

Intro

- Exploratory data analysis (data wrangling, data visualization) asks about the data *as it is*.
- Statistical modelling asks about *where the data came from*: how was it generated?

Example: Coin Tosses

The data is as follows:

$$HHTTHHHTHHTHHHH\ldots$$

or

$$HHHHHHHHHHHHHHH\ldots$$

- What is some exploratory data analysis you might do?

Example: Coin Tosses

Two possible **models** for coin tosses:

- Each flip is independent and has a 50-50 chance of coming up heads or tails.
- Every flip will come up heads.

Example: Coin Tosses

In statistics, we refer to distributions with two possible outcomes as **Bernoulli** distributions. We assign one of the outcomes (e.g., tails) a value of zero and the other (e.g., heads) a value of one. We then specify the probability p that the outcome is a one.

Example: Coin Tosses

Two possible **models** for coin tosses:

- Each flip is independent and has a 50-50 chance of coming up heads or tails: this is a Bernoulli($p = 0.5$) distribution.
- Every flip will come up heads: this is a Bernoulli($p = 1$) distribution.

Example: Coin Tosses

- We have very strong **prior beliefs** that coin flips have a 50-50 chance of coming up heads or tails (i.e., they are **unbiased**).
- Unless there is very strong evidence to the contrary we (usually unconsciously) model the coin flips as Bernoulli($p = 0.5$).

Example: Free Throws

Writing 0 for misses and 1 for successes, the data might be:

$$100111100101100$$

or

$$011111111001111$$

We will assume that all free throws are independent (i.e., there is no such thing as a “hot hand”).

Example: Free Throws

- The key difference in this example is that *we don't know the "true" probability that a given player will make his or her free throw*.
- In other words, we know that the free throws are Bernoulli but we don't know the value of the parameter p.

Example: Free Throws

- Assume that $p = 0.9$ (think Kevin Durant or Elena Delle Donne). We can **simulate** data using R:

$$11011101110101111111 = 16/20$$

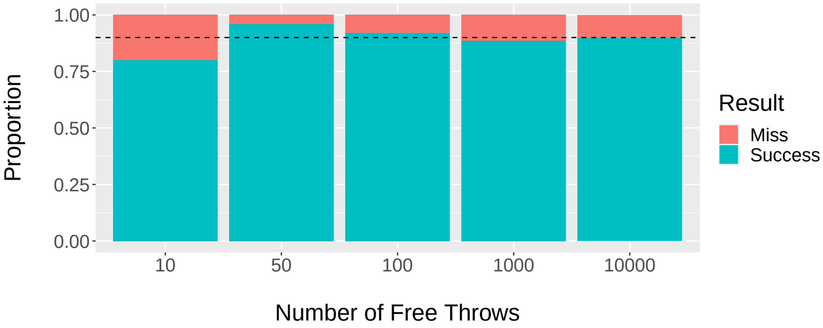
$$11111111111111111111 = 20/20$$

$$10111011111110111111 = 17/20$$

- In no case was *exactly* 90% of free throws made!

Example: Free Throws

One of the key results of probability, however, is that as more and more free throws are attempted, the proportion that are made will get closer and closer to the true value of p. This is known as the **law of large numbers**:



Example: Free Throws

Again, we usually have the data but not the value of p. In other words, we have to deal with **uncertainty**:

$$00010101011101100111 = 11/20$$

- What is your best guess for p here?
- If you had to give a range of p that you are 90% certain is correct, what would you choose?

Example: Free Throws

Again, we usually have the data but not the value of p. In other words, we have to deal with **uncertainty**:

$$00010101011101100111 = 11/20$$

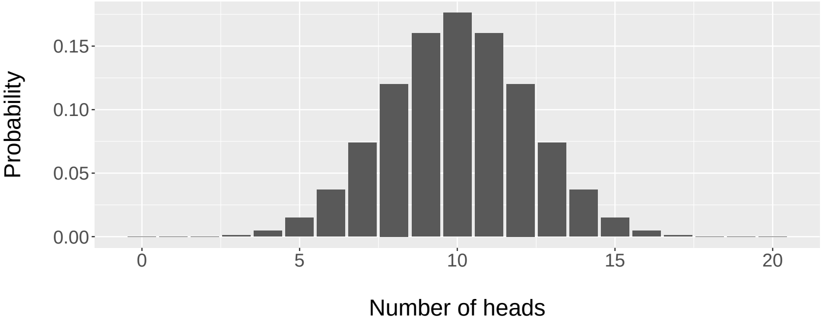
- Actual value used to generate this data: $p = 0.8$.

Example: Free Throws

- How do we deal with this uncertainty? That is, how can we give mathematical expression to our confidence in the value of p?
- I will return to this question, but first it will be useful to talk about some other distributions you will encounter.

Other Useful Distributions

A *binomial* distribution is what we get when we count up the results of independent Bernoulli trials. For example, if we count the number of heads out of 20 coin flips, we get a Binomial($n = 20$, $p = 0.5$) distribution:

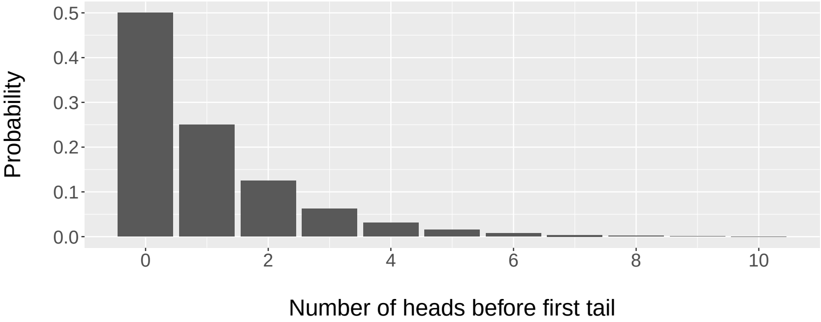


Other Useful Distributions

Notice that with the Bernoulli distribution, we only had to specify one **parameter** p. With the binomial distribution, we have to specify two parameters: the number of trials n and the probability of success p. However, n is always known so we only have to estimate p.

Other Useful Distributions

Another coin-flip-related distribution is the **negative binomial** distribution, which counts up the number of heads before the first tail (or the second tail, or the third, etc.). Now we have the two parameters p and r (the number of failures needed to stop flipping).



Other Useful Distributions

Recall that we began our discussion of Rosencrantz and Guildenstern by talking about **outcome spaces**. The outcome space of a Bernoulli trial is $\{0, 1\}$.

- What is the outcome space of a Binomial($n = 20$, $p = 0.5$) trial?
- What about a NB($r = 1$, $p = 0.5$) trial?

Other Useful Distributions

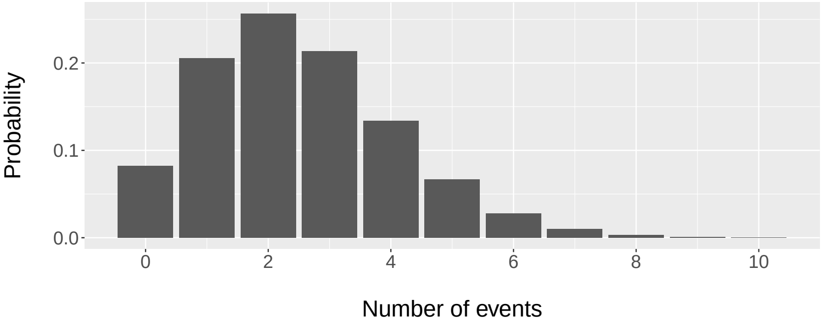
Another distribution whose outcome space is the entire range of nonnegative integers ($\{0, 1, 2, 3, \ldots\}$) is the **Poisson** distribution. Some applications (courtesy of Wikipedia):

- The number of stars found in a unit of space.
- The number of chewing gums on a tile of sidewalk:

- The number of soldiers in the Prussian army accidentally killed by horse kicks in a given year.

Other Useful Distributions

- A Poisson distribution arises when you are counting up events that can occur at any time, and whether or not they occur does not depend on the last time they occurred in the past (e.g., getting kicked by a horse).
- It is specified by a single parameter, the mean μ of the distribution. Here is a Poisson($\mu = 2.5$) distribution:

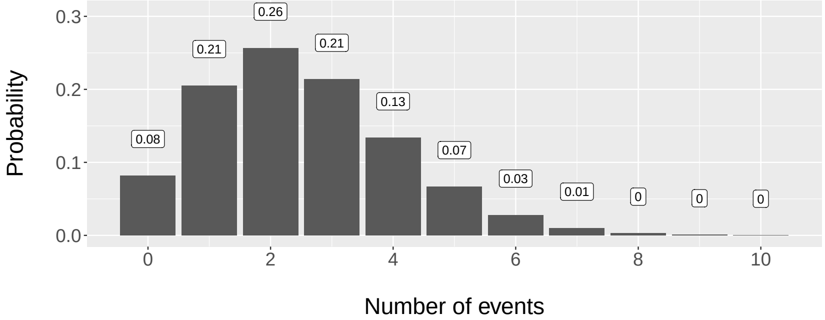


Measures of Location

- The **mode** of a distribution is the value that has the largest probability of occurring.
- The **median** of a distribution is such that 50% of events are to the left and 50% are to the right.
- The **mean** of a distribution is the "average" result. I like to think of it as the center of mass (where would you put your finger to balance the distribution on your hand).

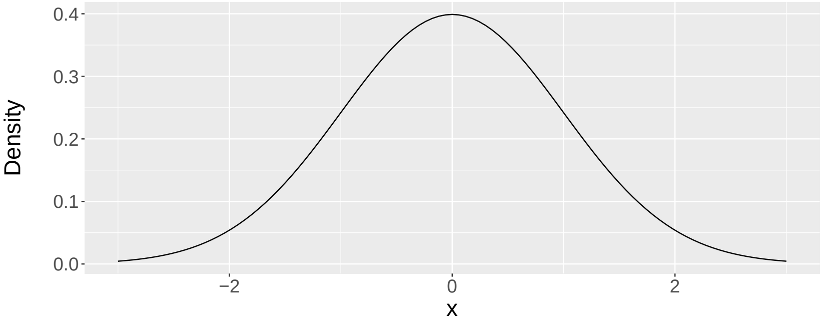
Measures of Location

- What are the mean, median, and mode of the Poisson($\mu = 2.5$) distribution?



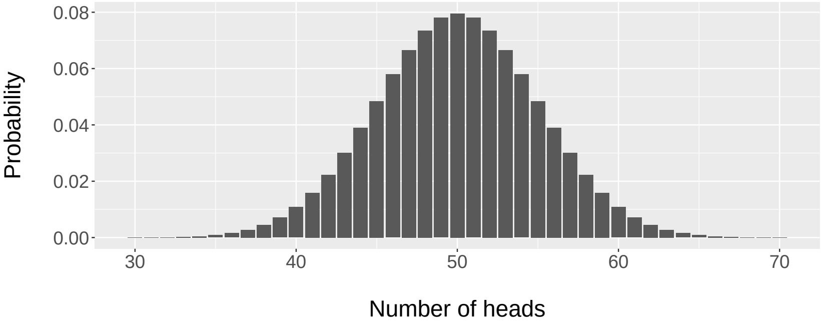
Other Useful Distributions

- The above distributions are all **discrete** (you can list the possible outcomes). A normal distribution is an example of a **continuous** distribution.



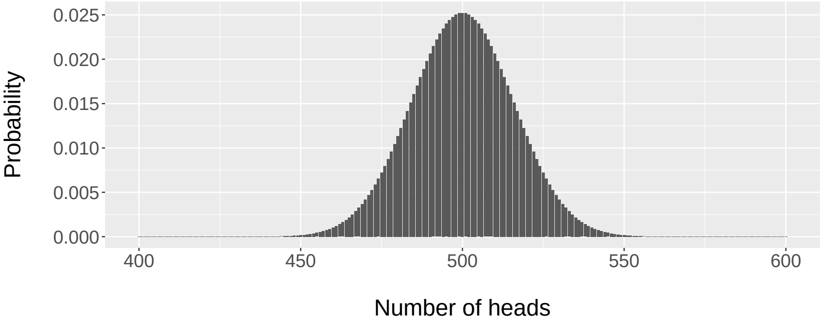
Other Useful Distributions

- Normal distributions occur naturally, so they're extremely useful. We won't get too far into the statistical weeds, but one example is that with more and more trials, the binomial distribution gets closer and closer to a normal distribution:



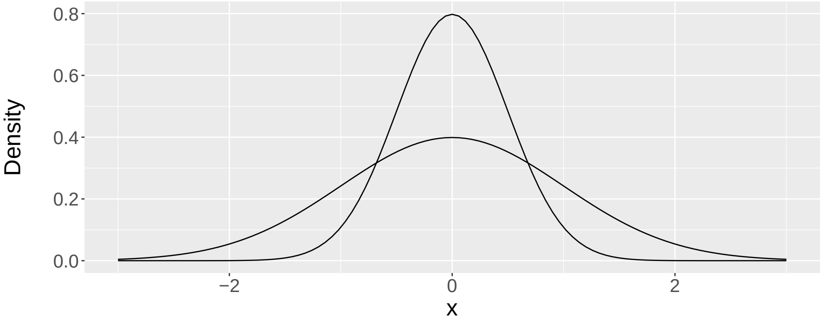
Other Useful Distributions

- Normal distributions occurs naturally, so they're extremely useful. We won't get too far into the statistical weeds, but one example is that with more and more trials, the binomial distribution gets closer and closer to a normal distribution:



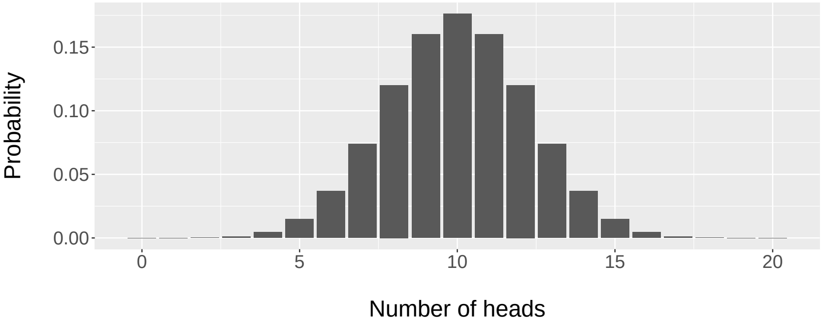
Other Useful Distributions

- The normal distribution has two parameters, the mean μ and the **standard deviation** σ. The standard deviation of a distribution is a measure of **spread**; i.e., how wide is the distribution? Here are two normal distributions with the same mean but different standard deviations:



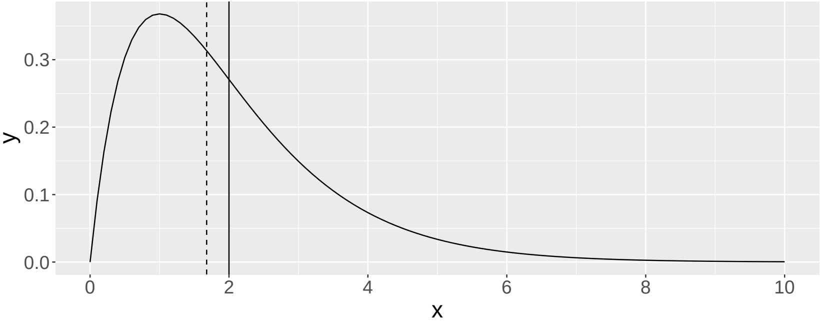
Other Properties of Distributions

- Note that normal distributions are **symmetric**. As a result, the mean, median, and mode are identical.



Other Properties of Distributions

- If a distribution is asymmetric, we say that it is either **left-skewed** or **right-skewed**. To remember the difference, think about which way the distribution would fall if you put your finger at the median instead of at the mean.



Properties of Distributions

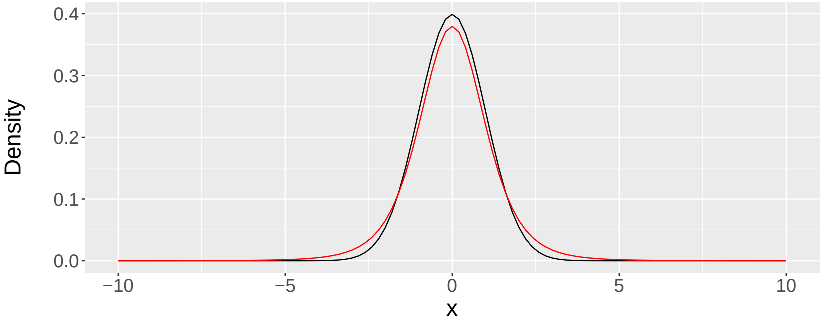
- When a distribution is left-skewed, the center of mass (mean) is to the left of (less than) the median.
- When a distribution is right-skewed, the center of mass (mean) is to the right of (greater than) the median.

Properties of Distributions

- We've talked about measures of location (mean, median, mode), spread (standard deviation), and skewness.
- A final important property concerns the **tails** of distributions. Two distributions can be identical in terms of location, spread, and skewness, but can differ in how fast the distribution decays at the extremes.

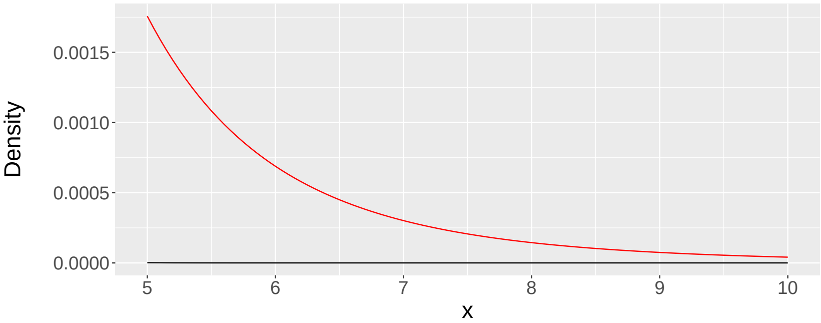
Properties of Distributions

- The t distribution is commonly used to model continuous data when **heavy tails** are needed:



Properties of Distributions

- The *t* distribution is commonly used to model continuous data when **heavy tails** are needed:

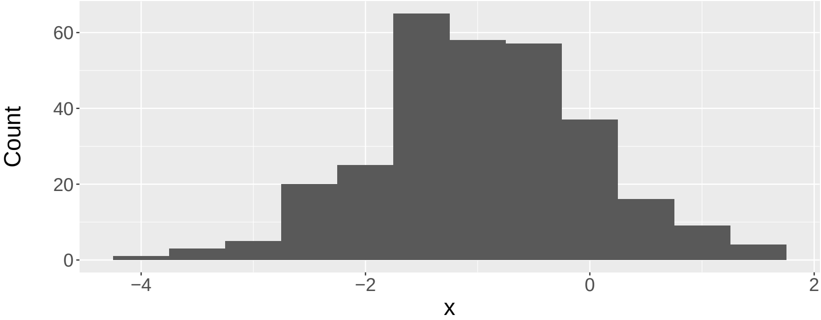


Summary

- Distributions: Bernoulli, Binomial, Negative Binomial, Poisson, Normal, t
- Properties: location (mode, mean, median), spread (variance), skewness (left- or right-), tails (heavy or light)

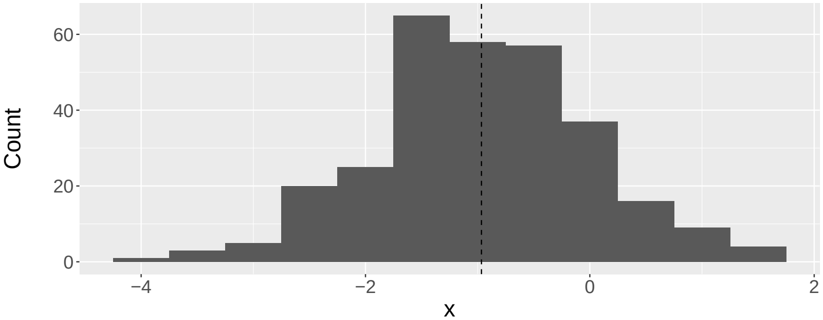
Uncertainty

Again, we rarely have the “true” distribution of the data, so we often need to estimate it. If we have data that we can assume is normally distributed, for example, we need to estimate the mean and variance. How can we estimate the mean here?



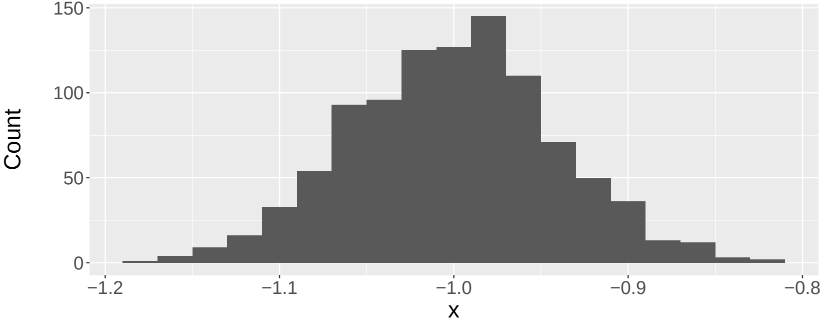
Uncertainty

We can use the **sample mean** (here, -0.966).



Uncertainty

It's crucial to remember that the sample mean is in general not the same as the "true" mean or **population mean**. In the previous example, the mean I used to simulate the data (consisting of 300 observations) was $\mu = -1$. If I run multiple simulations, I get different sample means. Let's do 1000 simulations and plot the sample means:



Most of the samples are concentrated in an area about 0.2 units wide.

Uncertainty

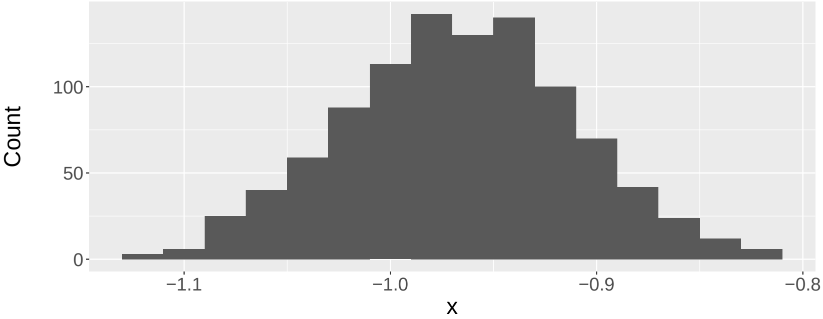
But simulating data means knowing the true values of the parameters. In other words, if we have the true mean, we can get a distribution of sample means, but how can we go the other way? That is, if we have a sample mean, can we say anything about the "distribution" of true means?

Uncertainty

One clever way to do this is called the **bootstrap**. The idea is to use the distribution of the data (also called the **empirical distribution**) as the true distribution and sample from that.

Uncertainty

In the normally distributed example above, I have 300 observations. To get an additional sample mean, I take 300 draws from those observations (with replacement!) to get a new dataset, and I repeat as many times as I like.



The samples are again concentrated in an area about 0.2 units wide, so the bootstrap works pretty well here!

Uncertainty

Notice that the bootstrap can work even when the data is not a standard, known type of distribution like the normal distribution! We'll explore an example of this kind of bootstrap in the next demo. . . .