ST 516: Foundations of Data Analytics

Introduction to continuous random variables

Continuous random variables

Until now we've considered **discrete** random variables. The random variables we have seen so far have outcomes that take numeric values in steps.

For example, the number of heads in ten coin flips is discrete. It cannot be any value between 0 and 10, it must be a whole number. You cannot get 4.5 heads!

A **continuous** random variable is one that can take any numeric value in a certain range. Sometimes this range is even infinite.

For example, think about the random variable that takes the value 1 if it rains tomorrow, and zero otherwise.

That is a discrete variable. However, if we think about *how much* it rains tomorrow that is a continuous variable. It could rain any amount that is 0 inches or more.

More examples

Imagine visits to a website, a continuous variable might be how long a visitor spends on the website. This could be any value greater than 0 seconds.

Sometimes the distinction isn't absolutely clear. Consider the amount of money a visitor spends when they make a purchase on a website. This can take on many values as long as they are greater than \$0, but these values are discrete in some sense. If you think of the purchase price in cents, it must be a whole number. Sometimes discrete random variables are treated as continuous for the purpose of data analysis.

For the purposes of data analysis the distinction often doesn't matter. We are more interested in the distributions of a statistic (for example, the sample mean) which are more naturally treated as continuous regardless of nature of the original observations.

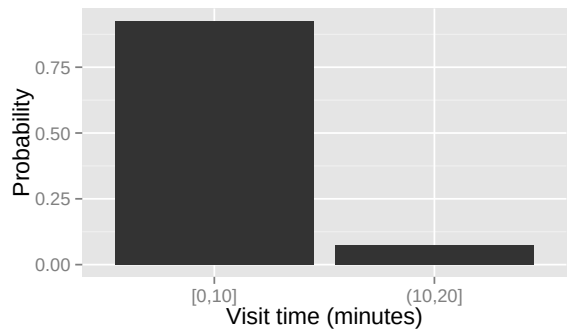
Probability distributions for continuous random variables

When the variable was discrete, we represented the probability distribution as bars. The height of the bar, is the probability that the value occurs.

In our website visit example, the probability distribution for the discrete random variable, X,

$$X = \begin{cases} 0 & \text{a user spends 10 minutes or less at the site} \\ 1 & \text{a user spends more than 10 minutes at the site} \end{cases}$$

might look like this:

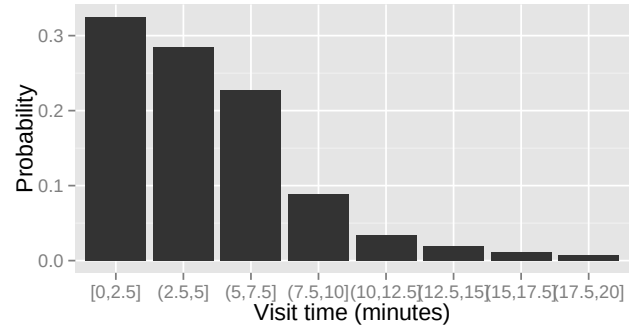


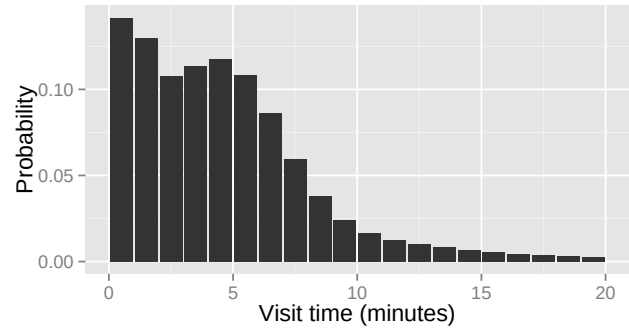
If we are interested in *how long* a user spends on the website, we would be better using more than just two categories,

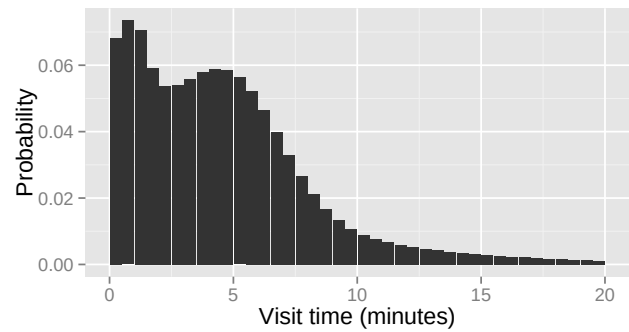
- maybe 8 categories (2.5 min intervals) would be better,
- 20 categories (1 min intervals), or
- 40 categories (30 sec intervals).

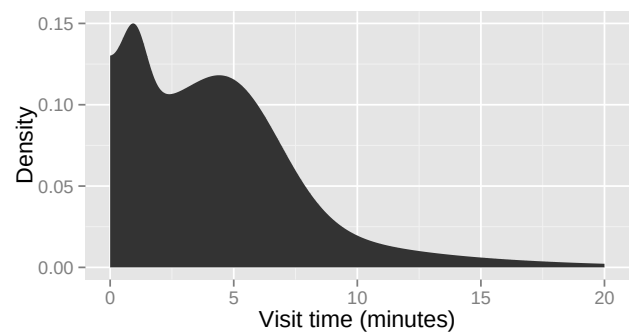
The more categories we have the closer we get to thinking about the time as a continuous variable.

Notice how as the bars get skinnier, the outline of the histogram gets less boxy. This suggests *continuous variables are described by smooth curves* that represent the outline of extremely small bins.









Some properties of the probability distribution for continous random variables

Probability distributions for continuous random variables satisfy the same rules as for discrete variables.

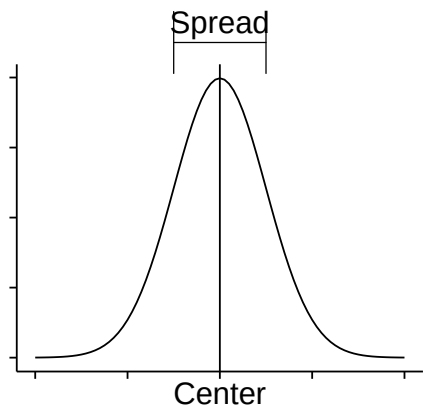
- the curve is never negative, just like probabilities for discrete events are never negative
- the total area under the curve is one, just like the height of the bars in our discrete distribution need to add to one.

We use the probability distribution in the same way as for discrete variables: areas under the curve represent probabilities. More on this in a later lecture.

A useful continuous distribution: the normal distribution

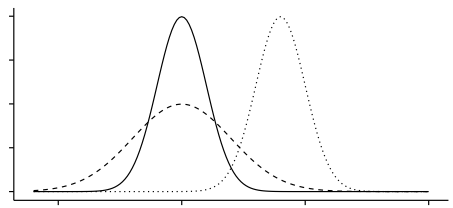
The normal distribution is one example of a probability distribution for a continuous random variable that you have probably seen before.

It is the classic "bell shaped" curve. It is symmetric about its center, and outcomes close to the center are more likely than those far away.



A useful continuous distribution: the normal distribution

The Normal distribution actually encompasses a whole family of curves. To specify a particular one, you must specify a value for its *center*, and a value that controls how *spread* out the outcomes will be. Here are three Normal curves with different centers and spreads.



When we say a variable, X, is Normally distributed (or Normal), we mean it is a random variable whose probability distribution is described by the Normal curve.

We often write $X \sim N(\mu, \sigma^2)$ where μ is the center (or more formally the mean) and σ is the value for the spread (or more specifically the standard deviation).

A useful continuous distribution: the normal distribution

The **standard Normal** distribution, $N(0,1)$, has a mean of zero, and standard deviation of 1

Any Normally distributed random variable, can be transformed to have a standard Normal distribution by subtracting its mean and dividing by its standard deviation.

That is if $X \sim N(\mu, \sigma^2)$ then $\dfrac{X-\mu}{\sigma} \sim N(0,1)$.

Other named distributions

There are many other named continuous distributions. You will see the Gamma in lab, other common ones are the Exponenital, Laplace and lognormal. Take a look at:
`https://en.wikipedia.org/wiki/List_of_probability_distributions#Continuous_distributions` for a whole list.

Simulating continuous random variables

To simulate a continuous random variable we need to know what the probability distribution function looks like.

If our random variable comes from a named distribution there are often built in functions to simulate an outcome.

For example, in R

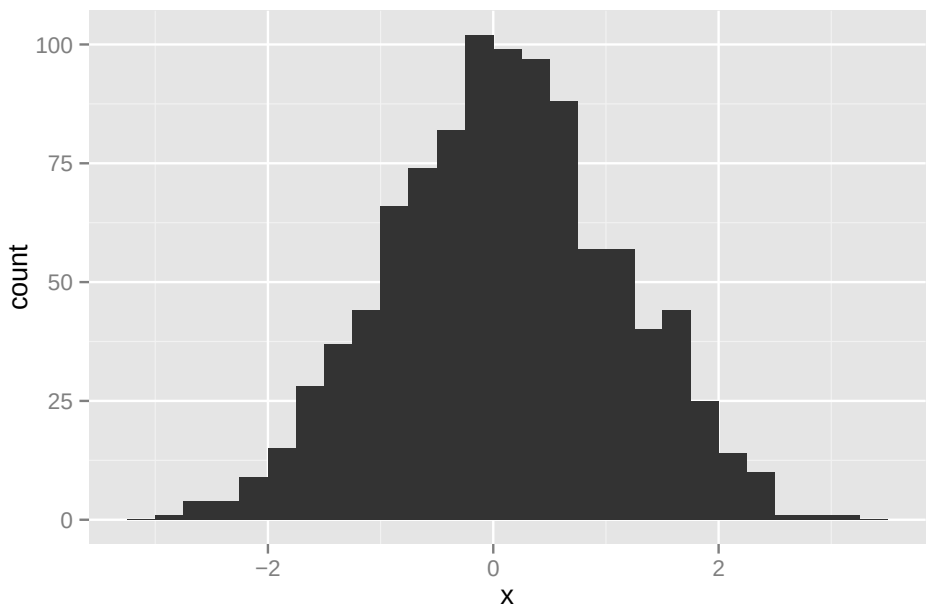
```
rnorm(1)
```

will simulate the value of a single random variable with a Normal distribution centered around zero and with a standard deviation of one (more on standard deviation, next week).

Probability Distribution Functions via Simulation

To approximate the shape of a probability distribution of a continuous variable, we simulate many draws and draw a **histogram** of the outcomes,

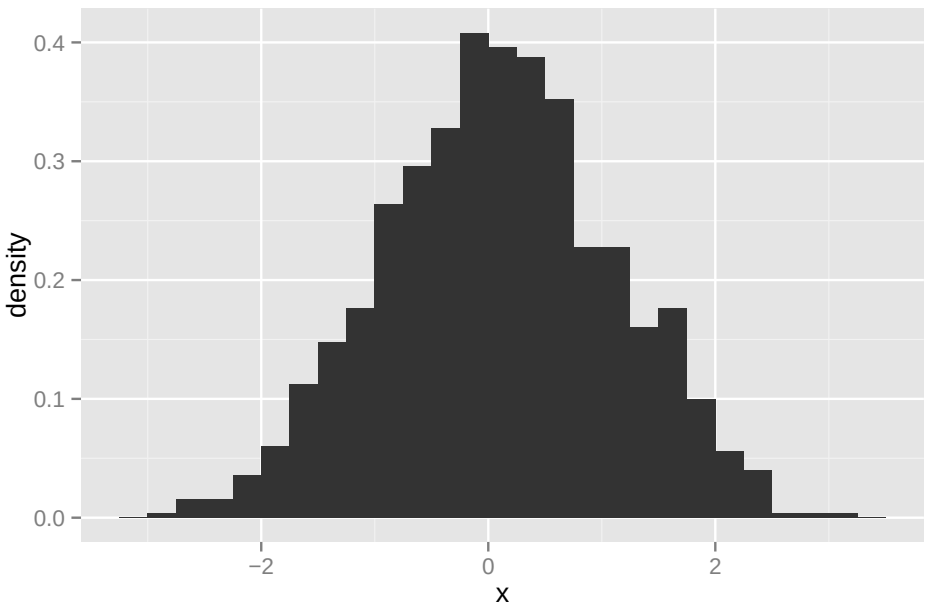
```
x <- replicate(1000, rnorm(1))
qplot(x, binwidth = 0.25)
```



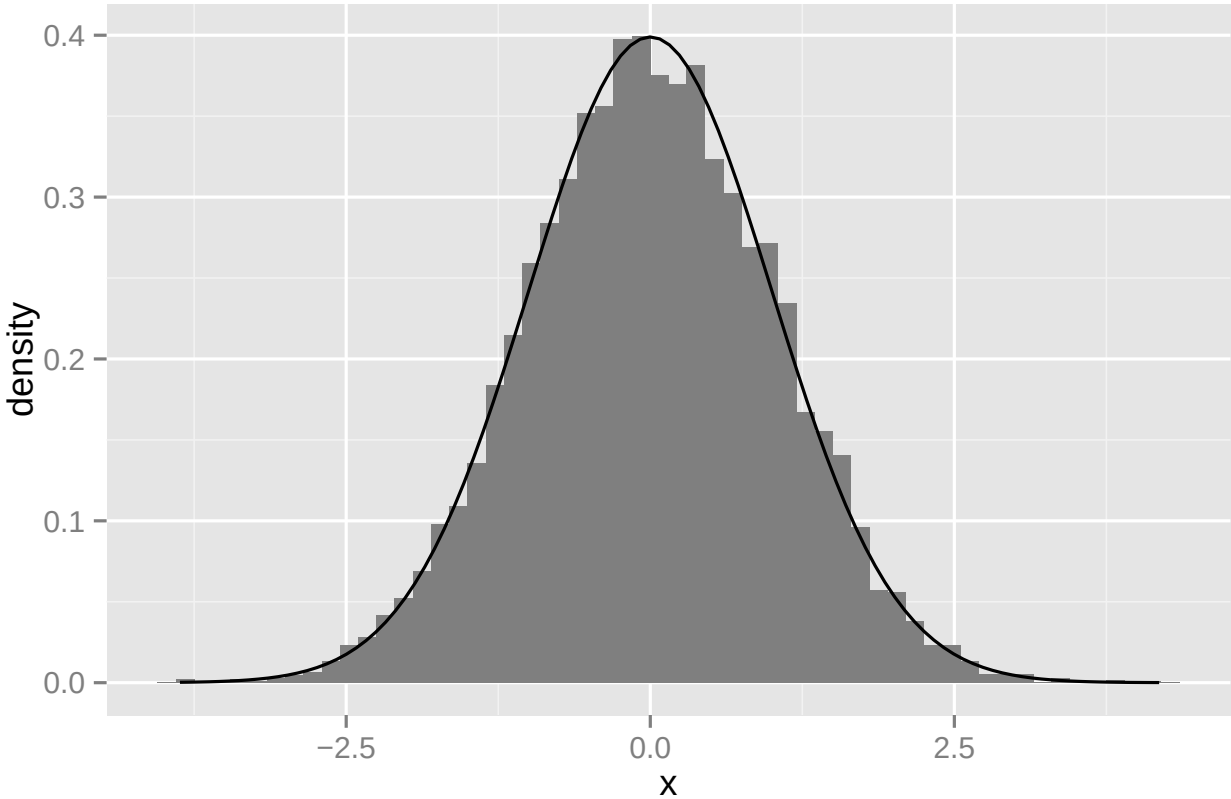
Notice the values on the y-axis are counts. To turn this into something with an area of 1 we need to divide the heights by the number of simulation times the width of each bar ($1000 \times .25 = 250$, in this example).

Or you can get R to do it for you,

```
qplot(x, y = ..density.., binwidth = 0.25, geom = "histogram")
```



The more simulations we do, the smaller our bins should be, and the closer our histogram will be to the true probability distribution curve.



Here I've overlaid the Normal curve on top of my simulation, so you can see that the outline of the histogram is close to the true shape of the Normal curve.

Sampling from a population

Let's return to the idea of sampling from a population from Module 1.

Imagine I'm interested in the heights of US adults. I randomly sample one person from all US adults and measure their height.

We can think of the value I measure as a random variable,

$$X = \text{height of one randomly sampled US adult}$$

X is random because before I do the study I don't know what value I will get for X.

What is this random variable's probability distribution?

Under simple random sampling, the probability distribution is simply the relative frequency of all possible heights in the population. We call this the **population distribution**.