Biostats week 2: Descriptive statistics for one variable at a time

Outline

This packet reviews statistics for working with one variable at a time:

1. Categorical vs. continuous
2. Frequency distribution
3. Central tendency
4. Spread
5. Probability distributions
6. Visualization

We will be using the `RNHANES`, `car`, and `ggplot2` packages. Install any of these packages you need before we begin. The `RNHANES` package takes a while to install.

1. Categorical vs. continuous

While R allows for many data types (e.g., numeric, character, factor), arguably for the purposes of analyzing data the most important distinction is between categorical and continuous data.

categorical variables: Variables with *categories* like marital status, color, sex, alma mater, religion, etc.

continuous variables: Variables with values that can take *any* value along some *continuum* like age, height, weight, distance, blood pressure, temperature, etc.

discrete variables: Variables that either have categories or are on a continuum but can only have certain values. Examples include the number of siblings or pets you have, the result of rolling dice, cars on a street, people in a state, etc. *Discrete* variables that do not contain categories are often treated as continuous variables.

2. Frequency distribution

A frequency distribution shows the number of times each unique value in a data set occurs. Let's examine a couple variables from the NHANES data set. The NHANES *Drug Use* questionnaire includes information on use of different types of drugs. The codebook for the Drug Use module is here: https://wwwn.cdc.gov/Nchs/Nhanes/2013-2014/DUQ_H.htm.

In addition to the special topics in the questionnaires, NHANES also includes demographics like age, sex, marital status, and other characteristics.

The `RNHANES` package allows easy access to all the online NHANES data through 2013-2014. Open the `RNHANES` package and load data using the `nhanes_load_data` command.

```
# open the RNHANES library
library(RNHANES)

# download 2013-2014 NHANES data
# use the name of the data you want (DUQ_H for drug use)
# add demographic data using demographics = TRUE
nhanes2013 <- nhanes_load_data(file_name = "DUQ_H",
                               year = "2013-2014",
                               demographics = TRUE)
```

One of the variables in this data set is `DUQ200` which is **Ever used marijuana or hashish**. Looking in the codebook for this variable, it appears to be categorical with four categories:

- Yes
- No
- Refused
- Don't know

If we were interested in the frequency distribution of this categorical variable for marijuana use, there are two primary ways to show this: (1) in a table, and (2) in a bar graph.

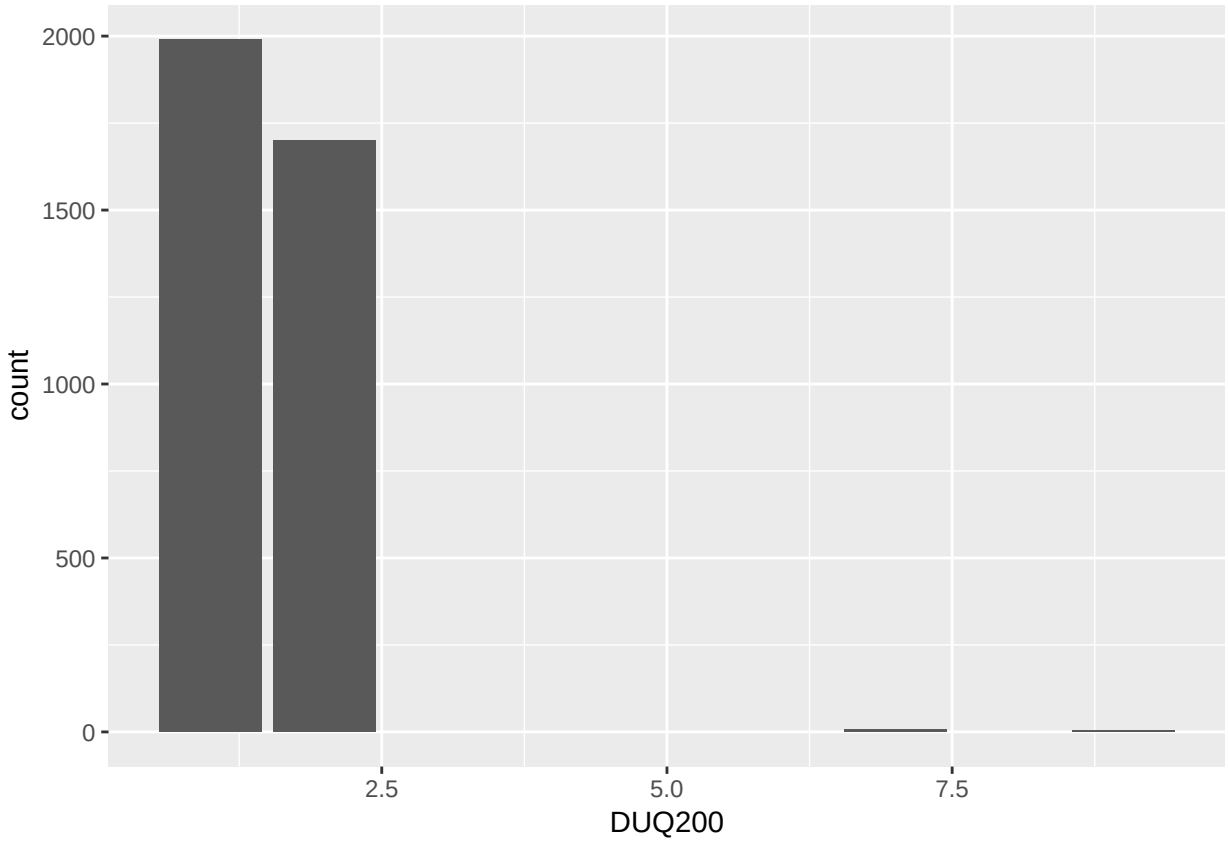
```
# frequency table of marijuana use
table(nhanes2013$DUQ200)
```

```
##
##    1    2    7    9
## 1991 1699    6    5
```

The numbers in the table match the frequencies in the codebook for the DUQ200 variable. However, the categories are not labeled and so it is not easy to understand the table without having the codebook right in front of you.

Would a graph be better?

```
# bar graph of marijuana use
# need ggplot open if it is not already open
library(ggplot2)
ggplot(data = nhanes2013, aes(x = DUQ200)) +
  geom_bar()
```



The graph is not useful at all. There is no way to tell the meaning of the bars. We can make a few changes to the variable to make it more easy to understand:

- Give the variable a better name
- Add labels to the categories (Yes, No, Refused, Don't know)

There are many ways to recode in different packages. One of the most straightforward is the `car` package written by John Fox.

Install the `car` package and open it to recode:

```
# open the car library
library(car)
```

```
## Warning: package 'car' was built under R version 3.5.1
```

```
## Loading required package: carData
```

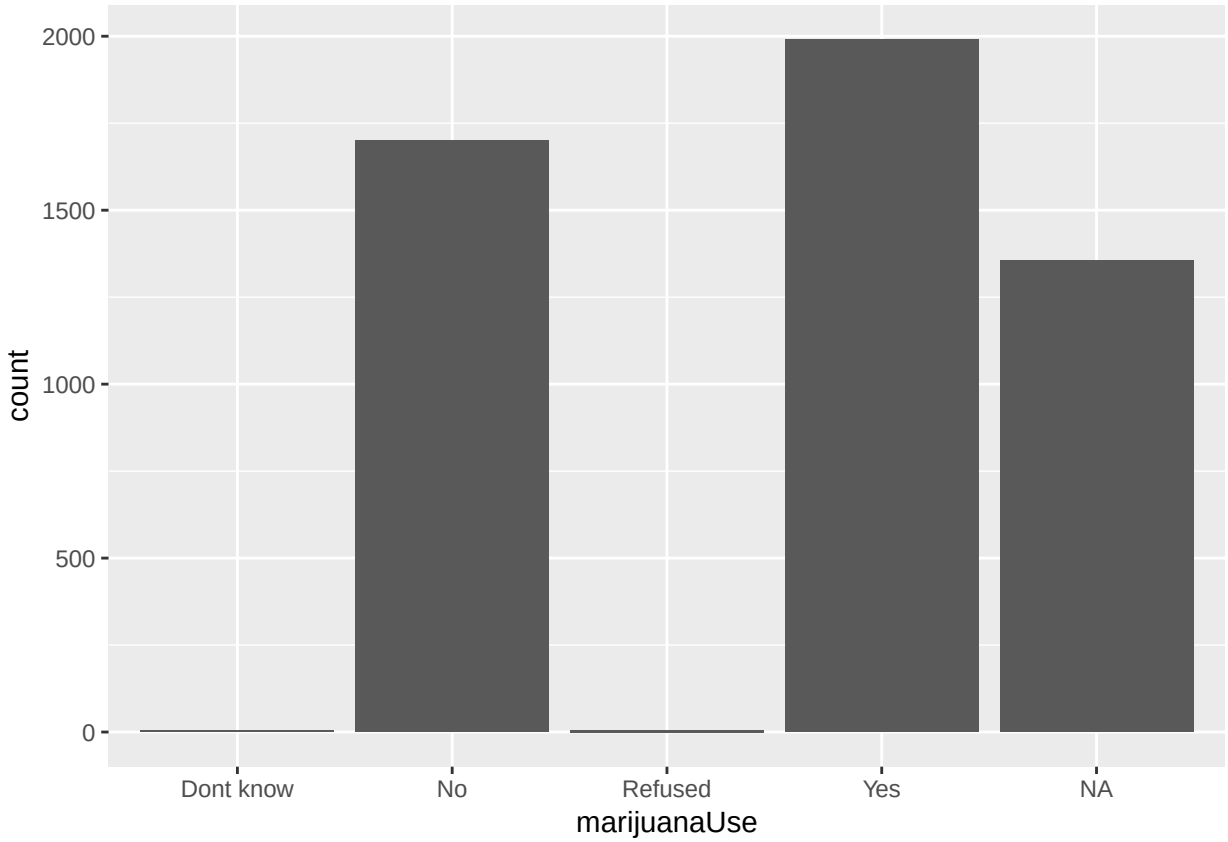
```
# recode into a new marijuanaUse variable
nhanes2013$marijuanaUse <- recode(nhanes2013$DUQ200,
                                  "1 = 'Yes';
                                  2 = 'No';
                                  7 = 'Refused';
                                  9 = 'Dont know'")

# try the table and graph again
table(nhanes2013$marijuanaUse)
```

```
##
## Dont know        No   Refused       Yes
```

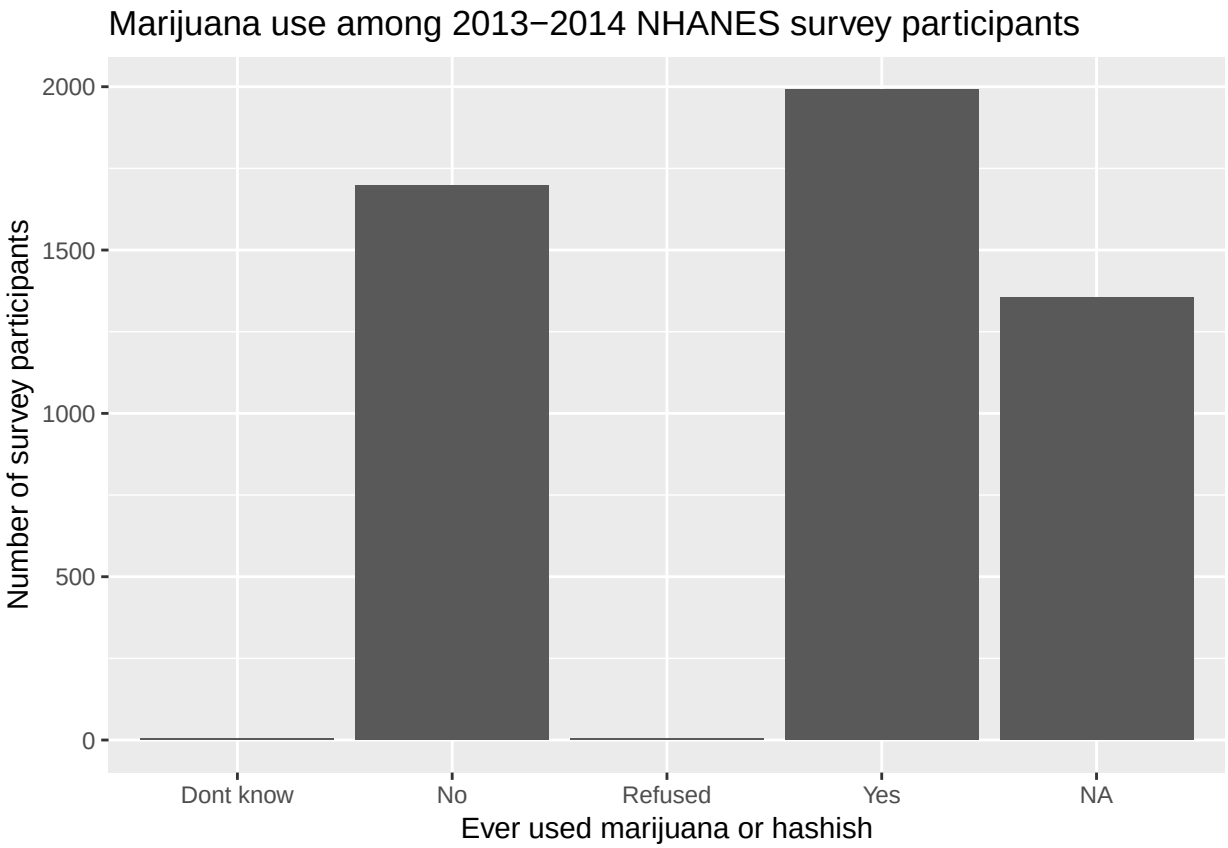
```
##         5      1699         6      1991
```

```
ggplot(data = nhanes2013, aes(x = marijuanaUse)) +
  geom_bar()
```



We can also improve the the frequency distribution graph by adding titles on each axis and a main title:

```
# adding titles to the plot
ggplot(data = nhanes2013, aes(x = marijuanaUse)) +
  geom_bar() +
  xlab("Ever used marijuana or hashish") +
  ylab("Number of survey participants") +
  ggtitle("Marijuana use among 2013-2014 NHANES survey participants")
```



The frequency distribution is now easier to understand.

There are so few people in the refused and don't know categories that excluding them from analysis might be good strategy.

One way to do this is to combine these groups with one of the larger groups depending on what makes sense. In this situation, these people seem to fit better with `missing` than with `Yes` or `No`.

In R, `missing` is coded as `NA` (not available).

We can use the same recode command structure from above to recode, this time assigning 7 and 9 to `NA` instead of refused and don't know:

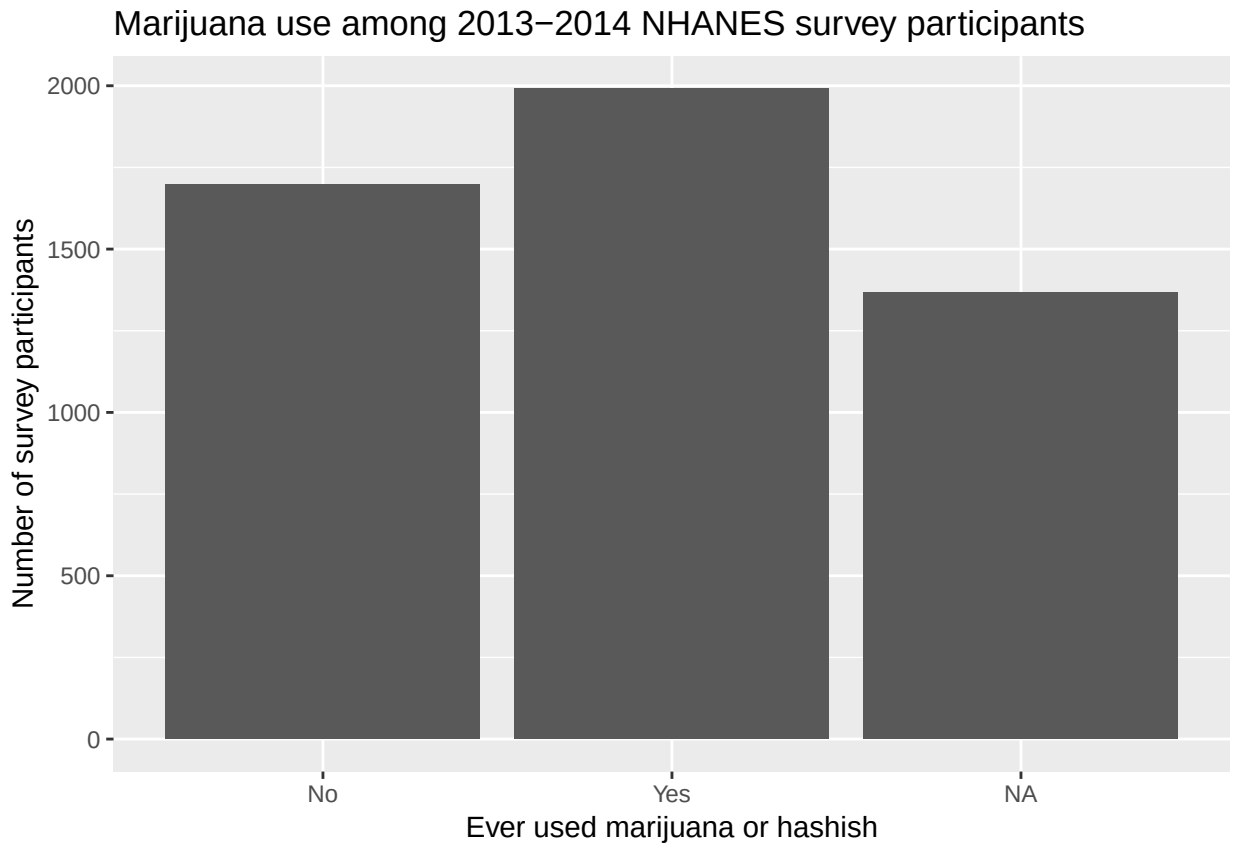
```
# add labels and recode
# refused and don't know to missing NA
nhanes2013$marijuanaUse <- recode(nhanes2013$DUQ200,
                                  "1 = 'Yes';
                                  2 = 'No';
                                  7 = NA;
                                  9 = NA")
```

Once the recoding is done, check the table and graph to see if it looks correct and is easier to read:

```
# try the table and graph again
table(nhanes2013$marijuanaUse)
```

```
##
##   No  Yes
## 1699 1991
```

```r
ggplot(data = nhanes2013, aes(x = marijuanaUse)) +
  geom_bar() +
  xlab("Ever used marijuana or hashish") +
  ylab("Number of survey participants") +
  ggtitle("Marijuana use among 2013-2014 NHANES survey participants")
```



That looks better. There is more work we can do to format the table and graph, but we will get to that later.

Continuous variables are not well described using frequency tables or bar charts. Try a table of the *age when first tried cocaine* variable, `DUQ260` from the NHANES data to see why. Before creating the table, examine the variable in the codebook. From the codebook, it looks like people who tried cocaine were in the 12 to 47 age range when they first tried cocaine.

What does a table show?

```
# first tried cocaine
# frequencies with continuous data
table(nhanes2013$DUQ260)
```

```
##
## 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36
##  1  5  5 15 37 46 81 43 76 56 35 24 16 43  8  8 12  6 24  3  6  3  1  5  2
## 37 38 39 40 42 45 47
##  2  2  1  2  1  4  1
```

This table is really useless for understanding the variable. Instead, to learn more about the distribution of continuous variables, divide the variable up into intervals and count how many people fall into each interval.

For example, divide the age of first cocaine use into 5 intervals using a frequency table with 5 breaks:

```
# first tried cocaine
# frequency table with 5 breaks
table(cut(x = nhanes2013$DUQ260, breaks = 5))
```

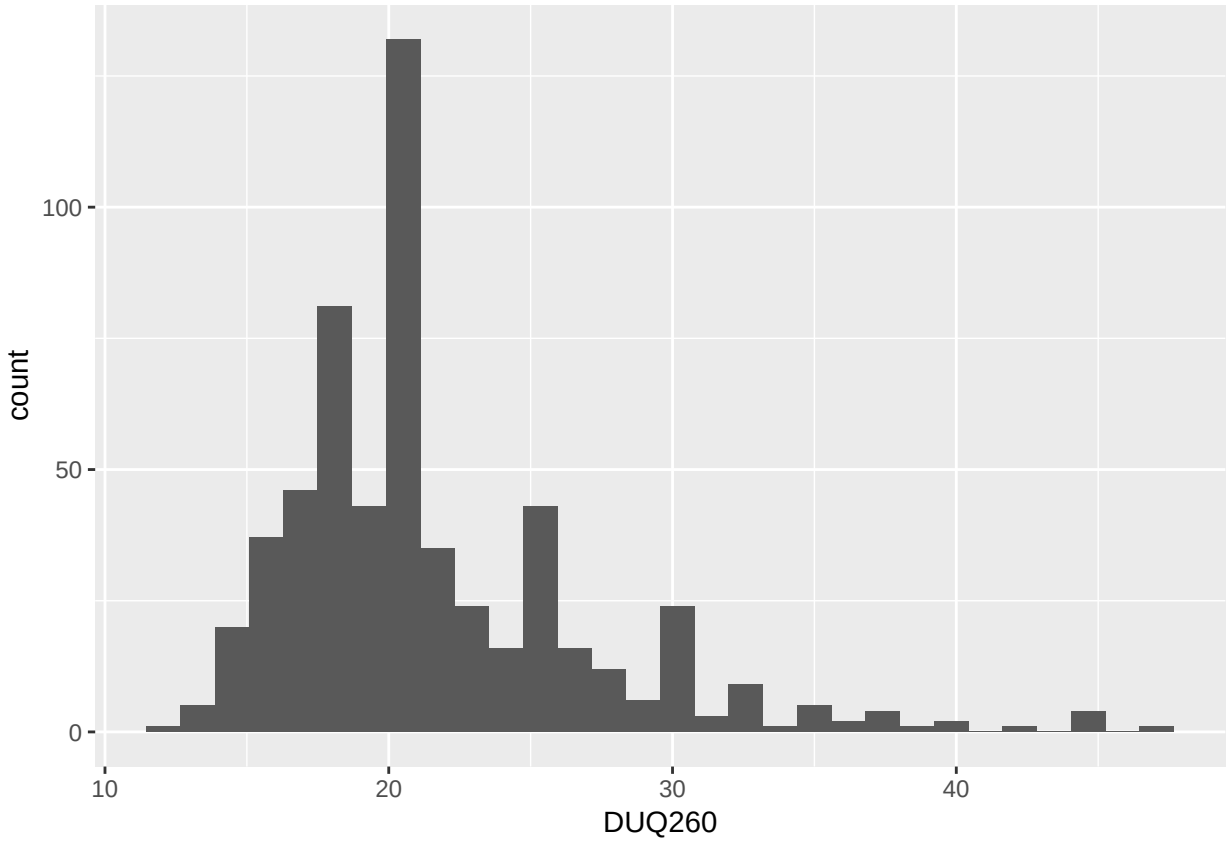
```
##
## (12,19] (19,26] (26,33] (33,40] (40,47]
##     233     258      62      15       6
```

It looks like 233 people were in the first interval, which is 12 to 19 years old.

Note that, in this output, square brackets [and] exclude the value, so (12, 19] means values from 12 up to–but not including–19.

The graphic way of cutting a continuous variable up into intervals is to use a histogram. Histograms show the shape of the data for continuous variables, like this:

```
# histogram of date first tried cocaine
ggplot(data = nhanes2013, aes(x = DUQ260)) +
  geom_histogram()
```

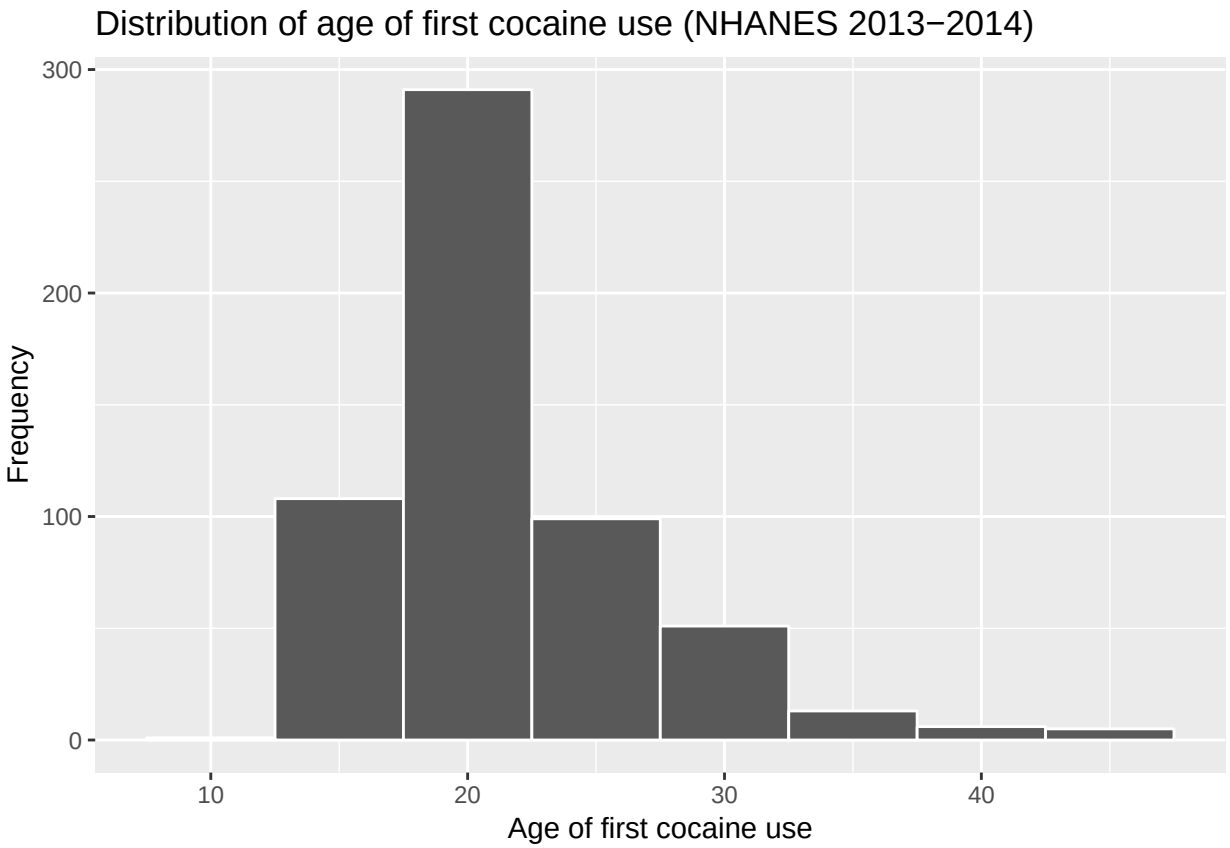


Notice one of the main differences between the histogram and the bar graph is that there are no gaps between bars in a histogram, which is consistent with the idea of a continuous variable being any value within a range.

There are several options to improve this visualization. First, the binwidth option changes how the data are divided up, which may make the shape of the data easier to see. Second, the default in R is to not have any visual distinction between the bins. To more clearly see each bin, you can add a color outline to each bin.

Here is a histogram with a binwidth of 5 and white outlines around the bins. Also, all figures should have titles on the axes and overall:

```
ggplot(data = nhanes2013, aes(x = DUQ260)) +
  geom_histogram(binwidth = 5, color = I("white")) +
  xlab("Age of first cocaine use") +
  ylab("Frequency") +
  ggtitle("Distribution of age of first cocaine use (NHANES 2013-2014)")
```



This gives a little different perspective on the distribution compared to the original histogram. When the overall shape of the distribution is unclear, trying a few different binwidths might help to better understand the data.

3. Central tendency

Central tendency is a measure of the center, or the typical value, of a variable. There are three possible measures of central tendency: mean, median, and mode.

- The mode is the most common value
- The median is the middle value
- The mean is the sum of the values divided by the number of values

To choose the measure of central tendency most appropriate for your data, the first thing to consider is whether your variable is *categorical* or *continuous*. If the variable is categorical, the only appropriate measure of central tendency is the mode.

Consider, for example, the marijuana use variable. It does not make logical sense, nor is it possible, to find a mean by adding up the no and yes responses. Likewise, finding the median, or middle value, is not possible since there is no inherent order to the categories.

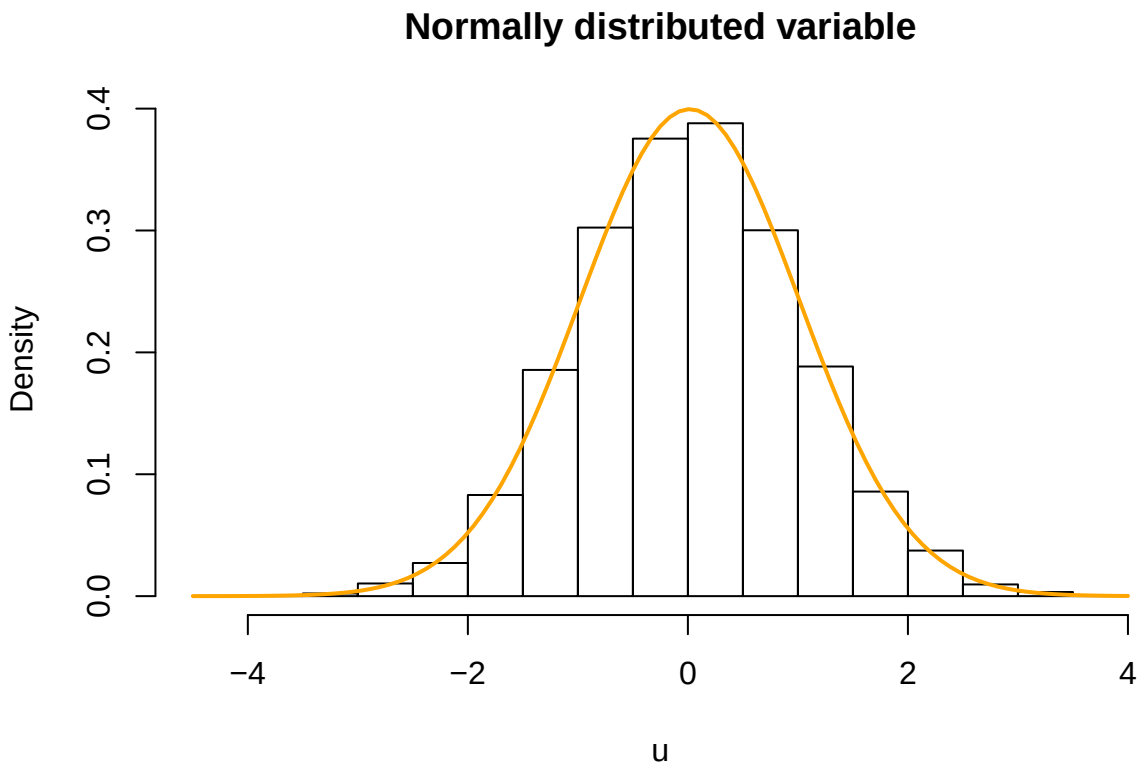
Unfortunately, there is no command *mode* to find the mode of a variable. One way to get this value is to make a frequency distribution table and put it in order from largest to smallest. The largest value is the mode since it happens most often.

```
# find the mode for marijuana use
sort(table(nhanes2013$marijuanaUse), decreasing = T)
```

```
##
##  Yes   No
## 1991 1699
```

The mode for marijuana or hashish use among NHANES participants is *Yes*.

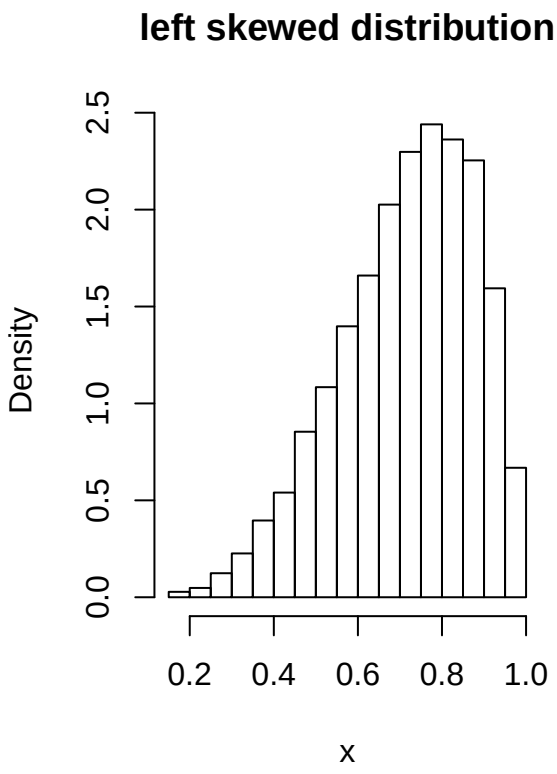
If the variable is continuous (or discrete) the next decision to make is between the mean and the median. The most well-understood and widely used value is the mean, but it is not always appropriate. To decide between these two, check the distribution of the variable. The mean is appropriate if a histogram representing variable distribution looks normal or close to normal, like this:

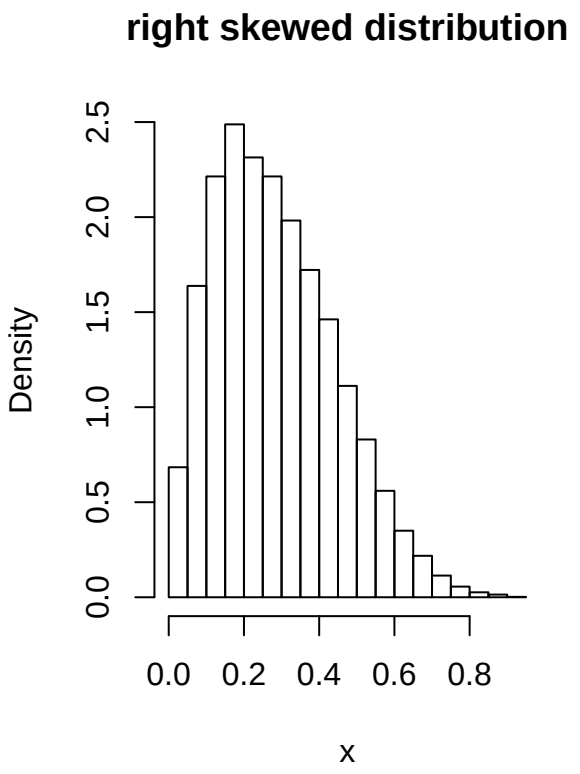


If the variable distribution has a normal, or near normal, distribution shape, compute the mean:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

However, if the histogram looks skewed, with some very large values on the right (right skewed) or left (left skewed), the median is more appropriate.





When you add together a set of values that includes a few very large or very small values like those on the far left or right of the skewed distribution graphs, the mean may be influenced and not be a good representation of a typical value or the middle of the data.

One example of this often used is income in the US. Say we took the mean of the incomes of 5 people you know plus Bill Gates. Since Bill Gates made $11.5 Billion this year (and so would be far on the right of the income distribution), it would look like the average income was millions of dollars. Instead, if we use the median, we end up with the middle number.

Two things of note:

- When there are an even number of numbers, the median is the mean of the middle two numbers
- In a perfect normal distribution, *the mean, median, and mode would be the same.*

Finding the mean and median in R for the DUQ260 variable:

```
# mean and median age first used cocaine
mean(nhanes2013$DUQ260, na.rm = TRUE)
```

```
## [1] 21.52439
```

```
median(nhanes2013$DUQ260, na.rm = TRUE)
```

```
## [1] 20
```

Interpretation: The mean age of first cocaine use is 21.52 and the median age of first cocaine use is 20.

The `na.rm = TRUE` added to the command removes any missing values ("NA") before computing the mean or median. It is needed for these commands if there is missing data.

4. Spread

In addition to using central tendency to characterize a variable, a corresponding measure of how spread out the values are around the central value is important. Each measure of central tendency has one or more corresponding measures of spread.

- variance (goes with the mean)
- standard deviation (goes with the mean)
- range (goes with the median)
- interquartile range or IQR (goes with the median)

The *variance* is the average of the squared differences between each value of a variable and the mean of the variable, computed like this:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

```
# variance of age of first cocaine use
var(nhanes2013$DUQ260, na.rm = TRUE)
```

```
## [1] 29.05263
```

There is no direct interpretation of the variance. It is a general measure of how much variation there is in your data.

The *standard deviation* is the square root of the variance, computed like this:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

```
# standard deviation for age of first use
sd(nhanes2013$DUQ260, na.rm = TRUE)
```

```
## [1] 5.390049
```

The standard deviation is sometimes interpreted as the average amount an observation differs from the mean. This is conceptually close and a good way to think about it, but as you can see in the formula above, it is not 100% accurate.

The *range* is the span between the largest and smallest values of a variable

```
# range of age first use of cocaine
range(nhanes2013$DUQ260, na.rm = TRUE)
```

```
## [1] 12 47
```

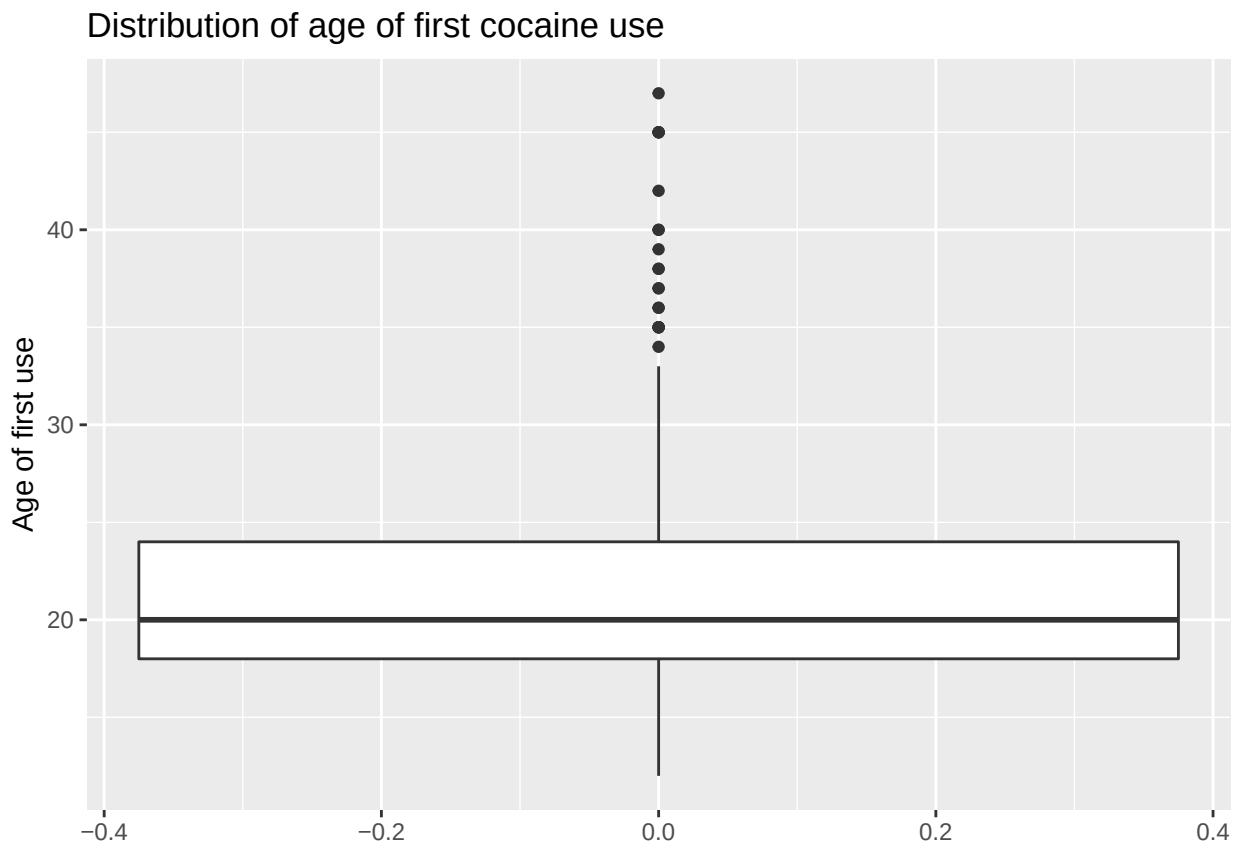
Finally, values we can interpret: The range of ages for first cocaine use is between 12 and 47 years old.

The *IQR* is the difference between the first and third quartiles, or the middle 50% of the data

```
# interquartile range of age of first use
IQR(nhanes2013$DUQ260, na.rm = TRUE)
```

```
## [1] 6
```

While the value is logical, there is a 6 year difference between the first and third quartiles, it is not a straightforward statistic and is often just interpreted like this: IQR = 6. In a preview of later material, the IQR is often shown visually when a boxplot is used to display continuous variables. *The distance between the top and bottom of the box is the IQR.*



The middle 50% looks like it goes from age 15 to 23 or so. So, half of the participants in the NHANES data who had ever used cocaine were between 15 and 23 years old when they first used it.

Looking at the codebook, it appears that the Refused and Don't know categories for this variable are coded as 777 and 999. Luckily for us, there are no people in these categories. If there were people in these categories, the distribution of age would appear to go all the way to 999 years old!

Before we move on, let's clean up this variable with a better name and recode the values of 777 and 999 to `NA`, which is the R indicator for Missing.

```
# rename and recode continuous variable
nhanes2013$ageFirstUse <- recode(nhanes2013$DUQ260,
                                 "777 = NA;
                                 999 = NA")
summary(nhanes2013$ageFirstUse)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   12.00   18.00   20.00   21.52   24.00   47.00    4483
```

5. Probability distributions

Probability is a measure of how likely it is that something happens. For example, if I flip a coin, there is 1 in 2 chance that it will be heads. The probability that I will flip heads is 50%.

Probability distributions show the likelihood of choosing a certain value at random for categorical or discrete variables. For example, we could calculate the probability of picking a person at random who had used marijuana from all the NHANES participants who answered the marijuana use question. If we chose a person at random. We could use the equation:

$$P(use) = \frac{Number.of.marijuana.users}{total.who.answered.question}$$

Where P(use) is the probability of marijuana use.

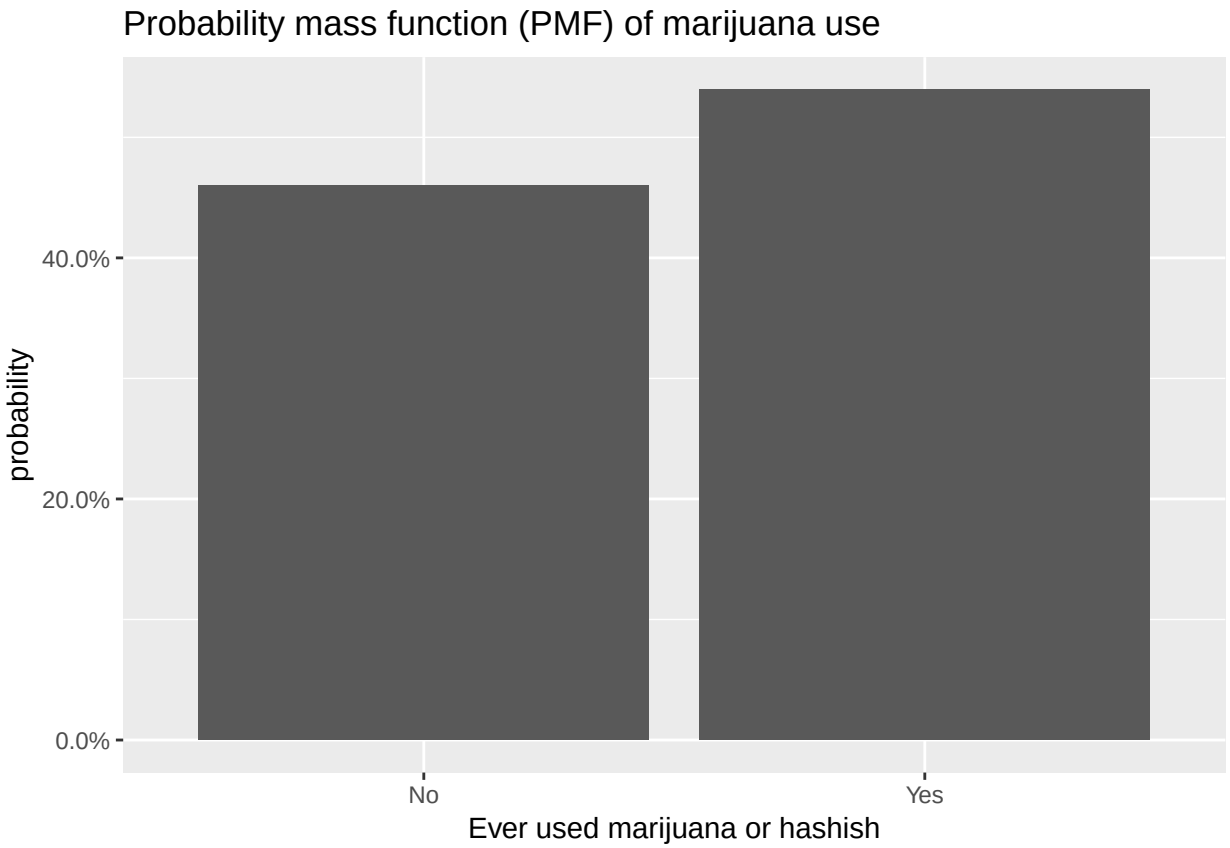
Like frequency distributions, a probability distribution is typically shown as a table or graph. Find the probability of Yes and No responses to the marijuana use question:

```
# table with probability of marijuana use
prop.table(table(nhanes2013$marijuanaUse))
```

```
##
##        No       Yes
## 0.4604336 0.5395664
```

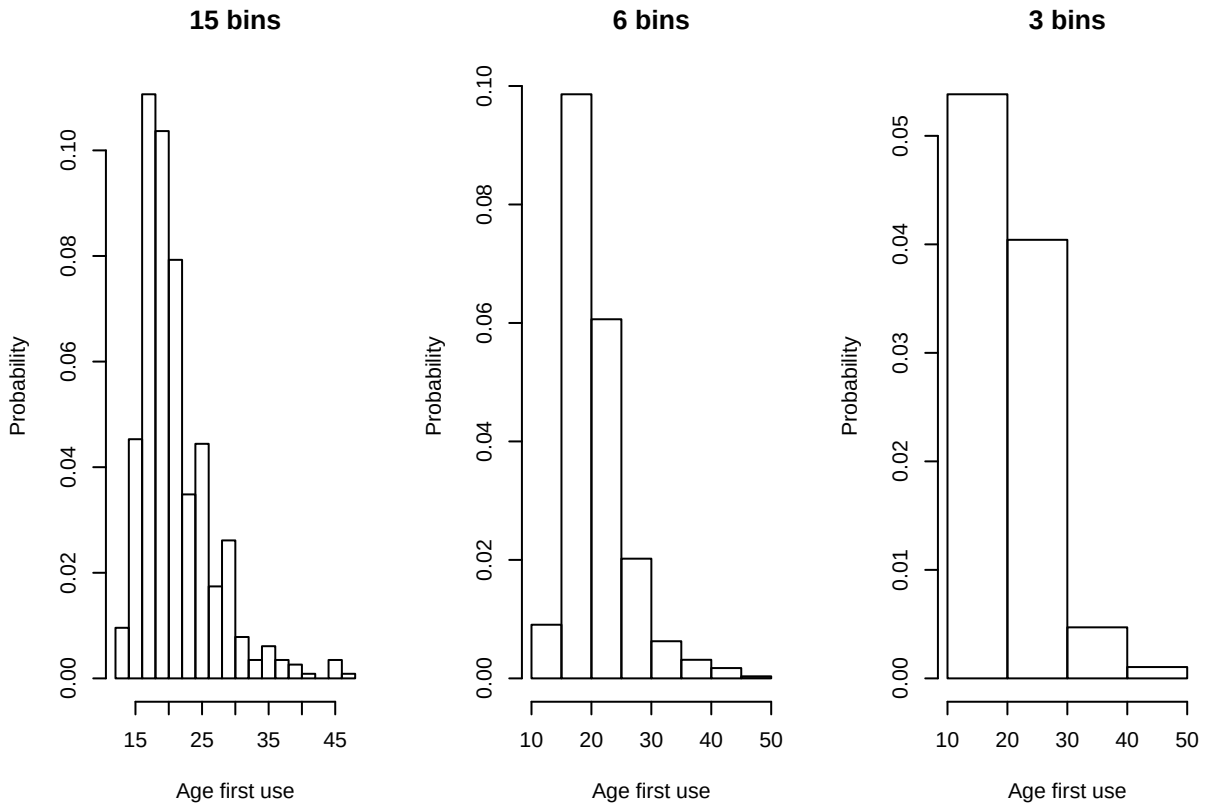
The probability of picking a person at random that *used* marijuana is NA%. Likewise, the probability of picking a person at random that did not use is 53.96%.

Advanced plotting features of R can be used to make a bar graph showing probabilities instead of frequencies. This is called a probability mass function (PMF) graph:

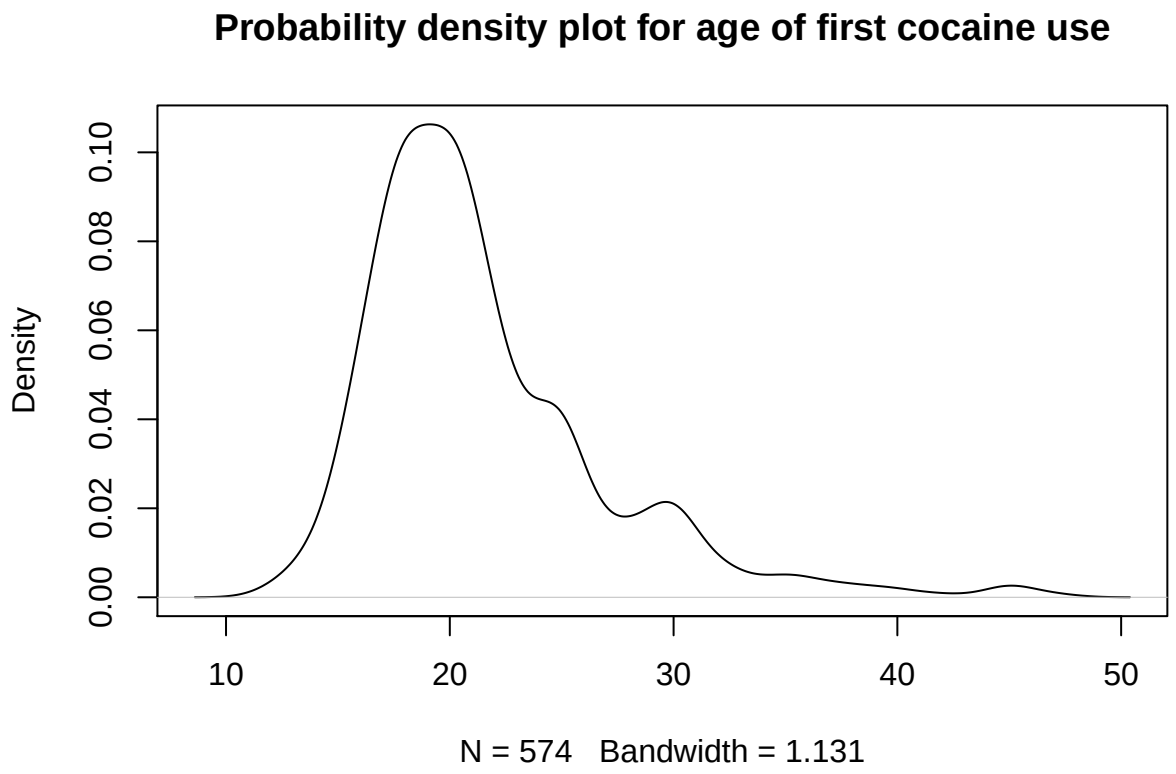


With the graph or table it is clear that there is a greater chance of picking a person who has used marijuana.

Using the same strategy to find the probability of selecting a value in a range for a continuous variable we run into trouble. Remember, continuous variable distributions are shown with histograms. Check out three histograms showing probability for the age of first cocaine use variable. Note that the probability for any given value of calories per ounce can vary widely based on the number of bins displayed:



Instead we can graph what is called a probability density function (PDF) to show the probability of a value being selected at random from a continuous variable. In this case it shows the probability of a person being of a given age when first using cocaine.

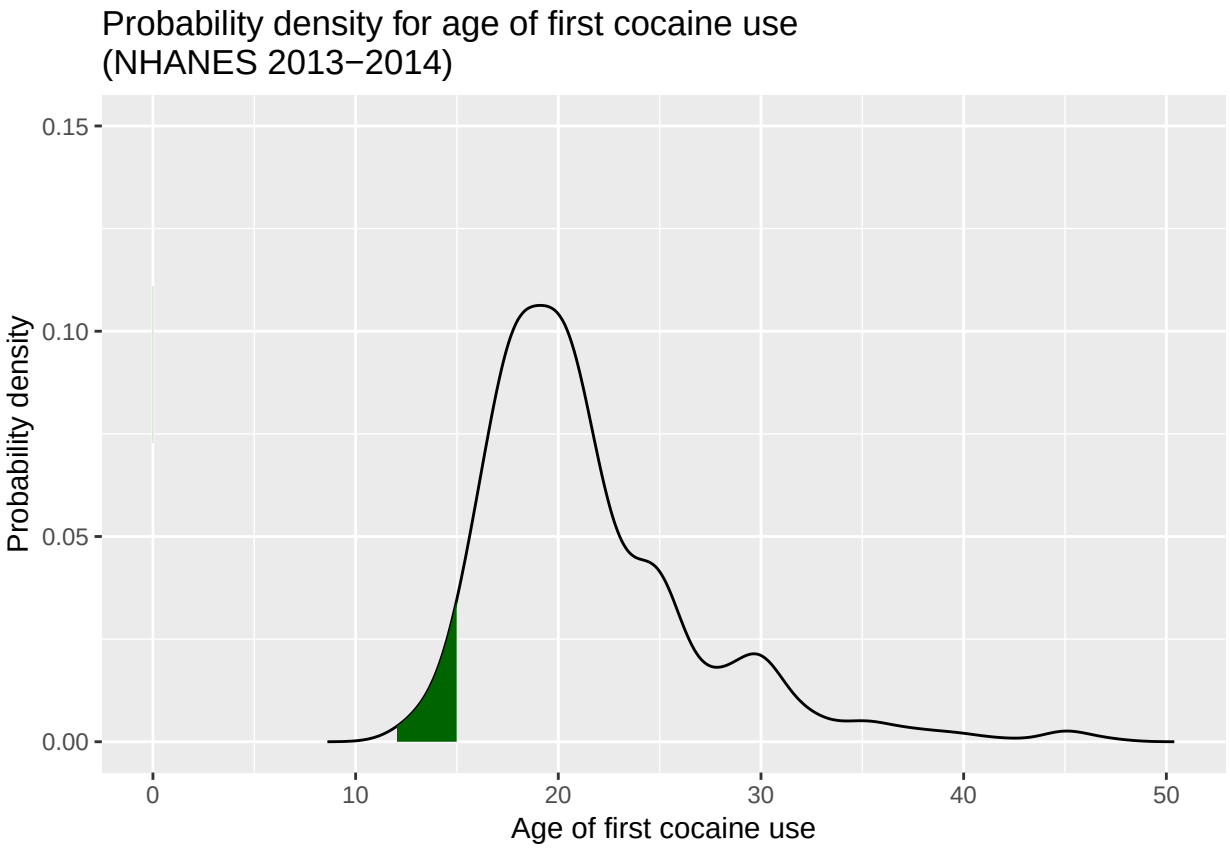


The interpretation of this graph is still a little tricky. To find out what the likelihood of selecting a person at random who used cocaine for the first time between ages 12 and 15, we can use calculus to find the area under the curve between 12 and 15:

First, visualize it using some advanced ggplotting (don't worry if this code looks complicated! it is...):

```
# create a density object and make it a data frame to graph
ageUseDensity <- density(nhanes2013$ageFirstUse, na.rm=TRUE)
ageUseDensityData <- data.frame(x = ageUseDensity$x, y = ageUseDensity$y)

# graph density data
ggplot(data = ageUseDensityData, mapping = aes(x = x, y = y)) +
    geom_line()+
    geom_area(mapping = aes(x = ifelse(x>12 & x< 15 , x, 0)), fill = "darkgreen") +
  ylim(0, .15) +
  xlab("Age of first cocaine use") +
  ylab("Probability density") +
  ggtitle("Probability density for age of first cocaine use\n(NHANES 2013-2014)")
```



Then use some calculus to find the area under the curve shown in green (again, the code will look complex for now, but over time you will get used to the R language). You absolutely do not need to remember calculus!

```
# compute the probability density function
# rule 2 finds the probability in a given range
pdf <- approxfun(ageUseDensity$x, ageUseDensity$y, rule=2)

# find the area under the curve between 12 and 15
area <- integrate(pdf, 12, 15)
area
```

```
## 0.04318919 with absolute error < 2.3e-05
```

So, the probability of choosing a person who used cocaine at random and the person being between 12 and 15 years old on first use (being in the green shaded part under the curve) is 4.32%.

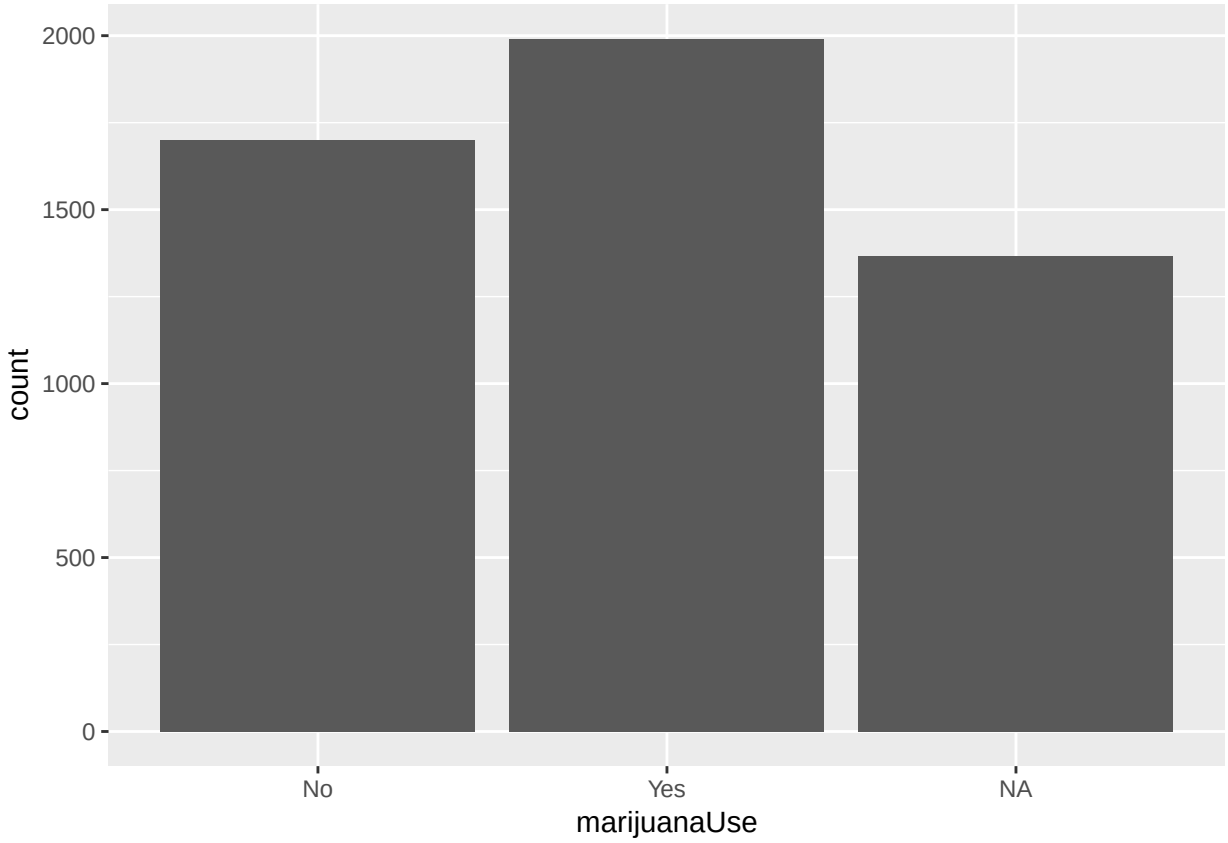
#6. Visualization

There are several different R packages useful for making graphs. The two most widely used are the default graphs that come with the base R installation and ggplot2, both of which were used for several graphics so far.

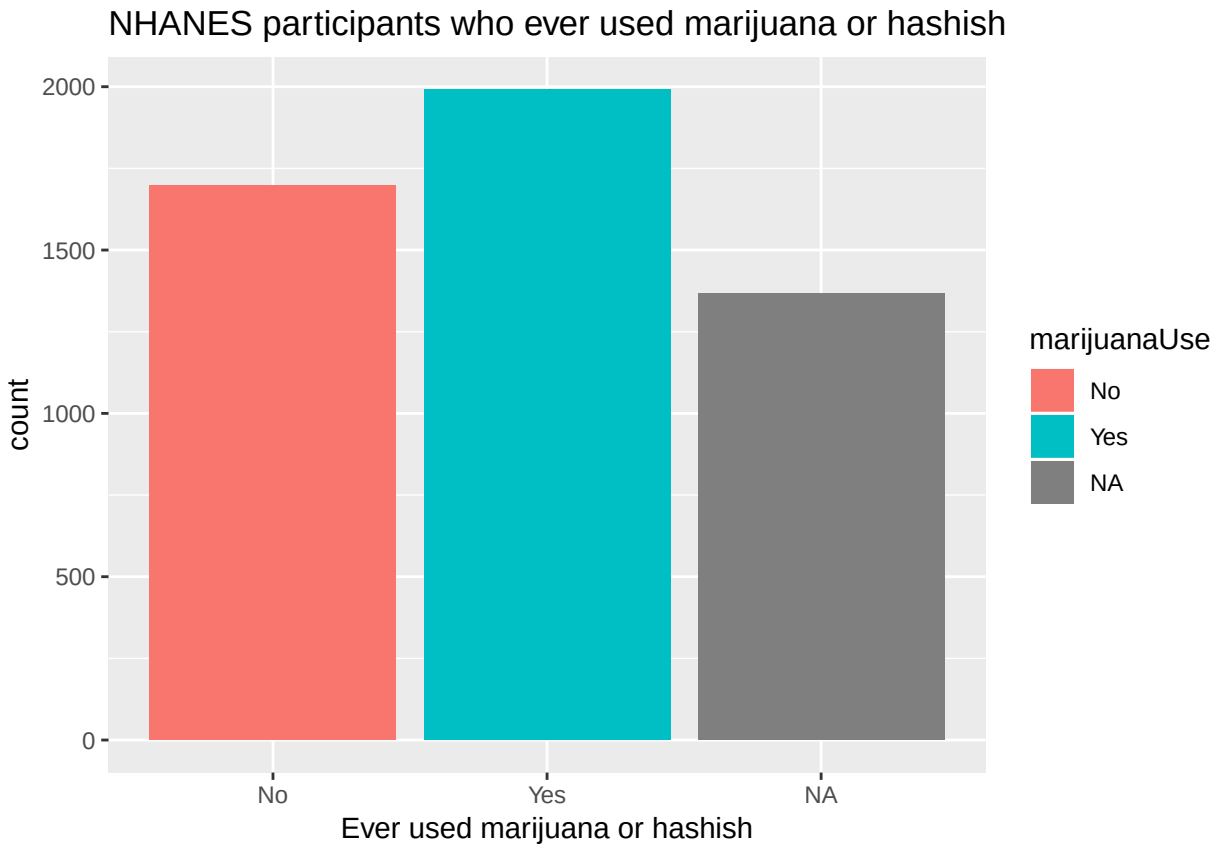
Just like certain descriptive statistics are useful for certain types of variables, different plots are useful in different situations:

bar graph: Useful for showing frequencies or percentages of values in each category for categorical variables

```
# basic bar plot
ggplot(nhanes2013, aes(x = marijuanaUse)) +
  geom_bar()
```




```
# add color, labels, title
ggplot(nhanes2013, aes(x = marijuanaUse, fill = marijuanaUse)) +
  geom_bar() +
  xlab("Ever used marijuana or hashish") +
  ggtitle("NHANES participants who ever used marijuana or hashish")
```

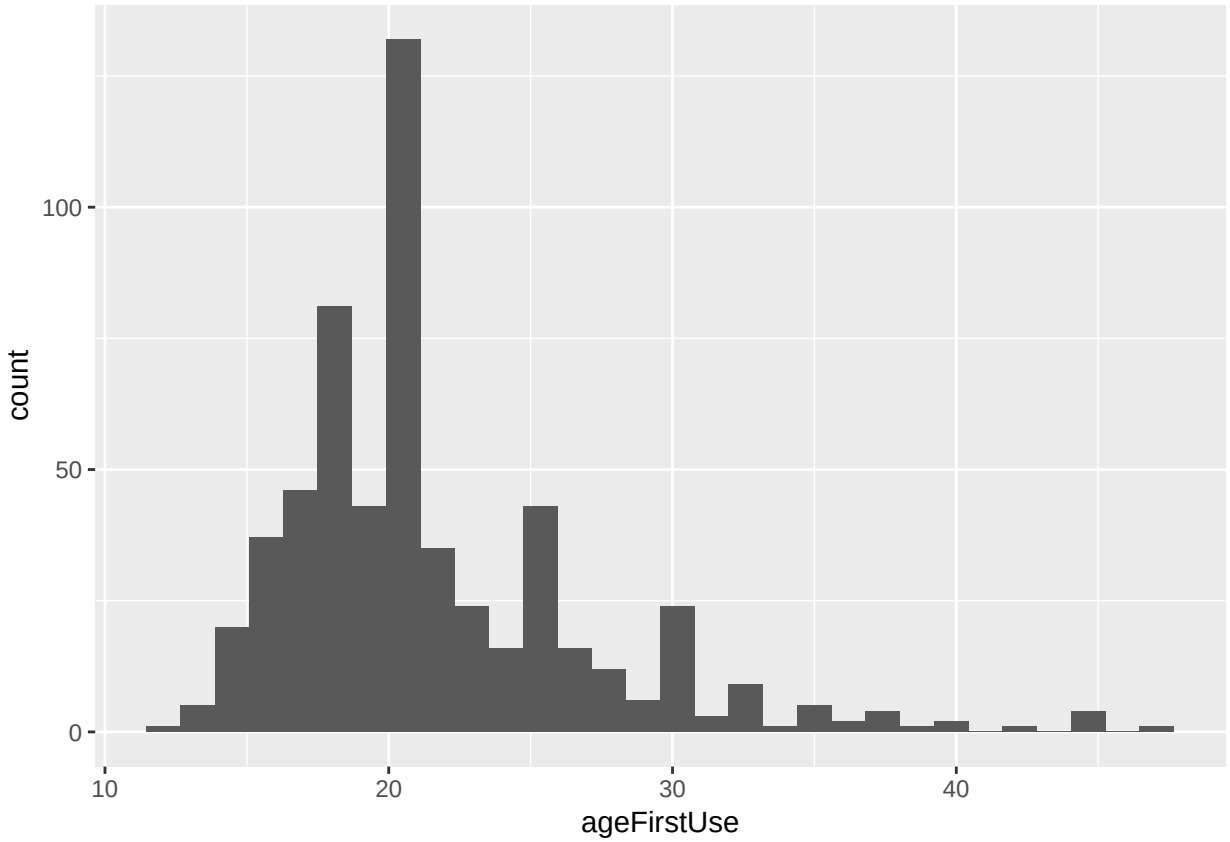


histogram: Useful for showing the distribution of values for a continuous, or near continuous, variable

```
# basic histogram
ggplot(nhanes2013, aes(x = ageFirstUse)) +
  geom_histogram()
```

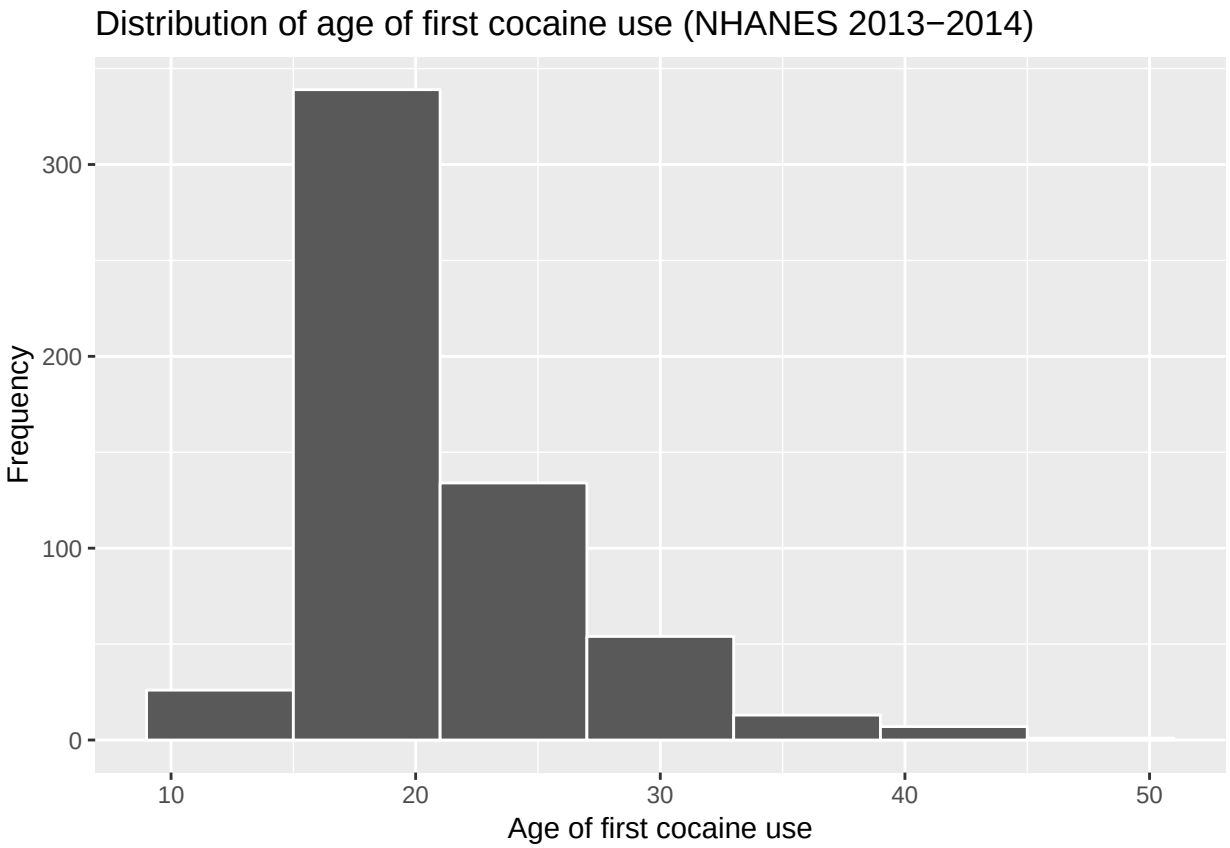
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
## Warning: Removed 4483 rows containing non-finite values (stat_bin).
```




```
# histogram with binwidth of 6 and white dividers
# added labels and titles
ggplot(nhanes2013, aes(x = ageFirstUse)) +
  geom_histogram(binwidth = 6, color = I("white")) +
  xlab("Age of first cocaine use") +
  ylab("Frequency") +
  ggtitle("Distribution of age of first cocaine use (NHANES 2013-2014)")
```

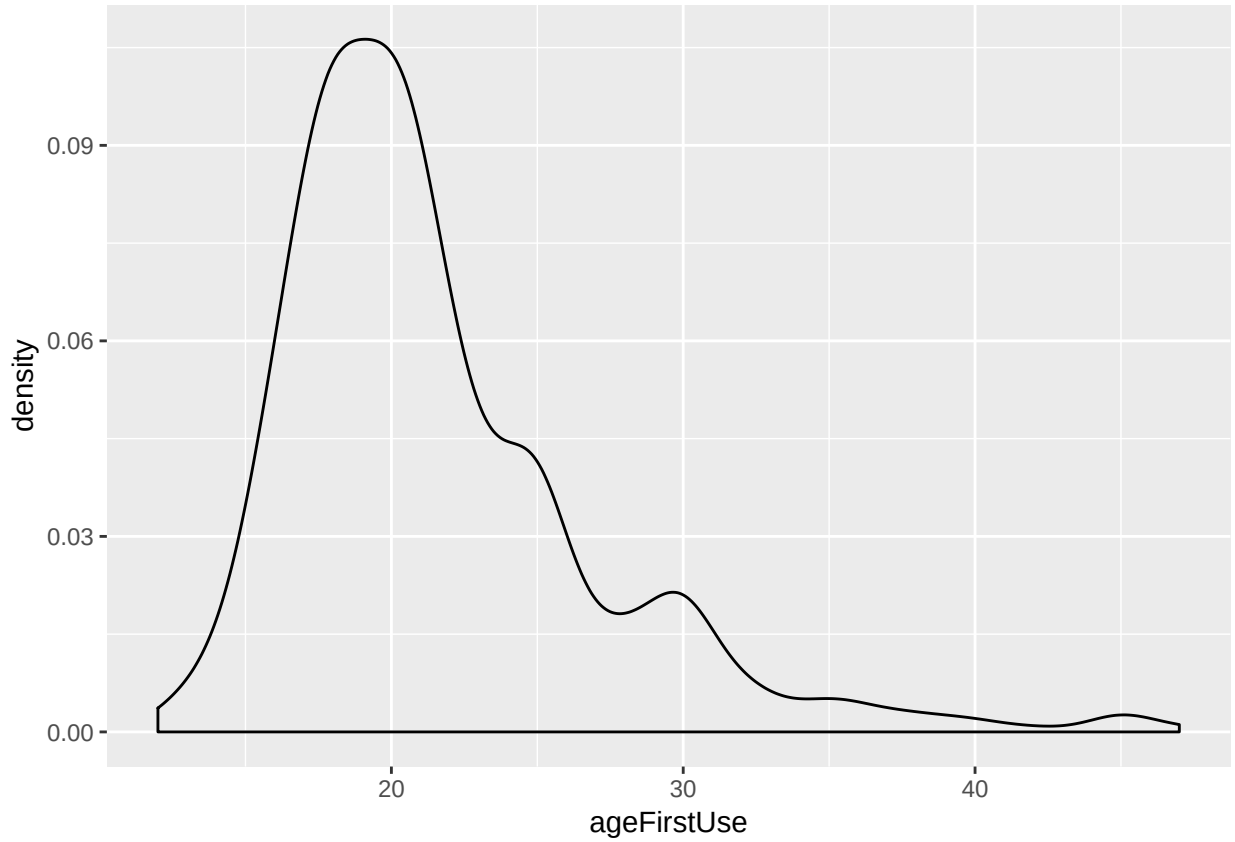
```
## Warning: Removed 4483 rows containing non-finite values (stat_bin).
```



probability density function (PDF): useful for a more precise look at the distribution of continuous variables

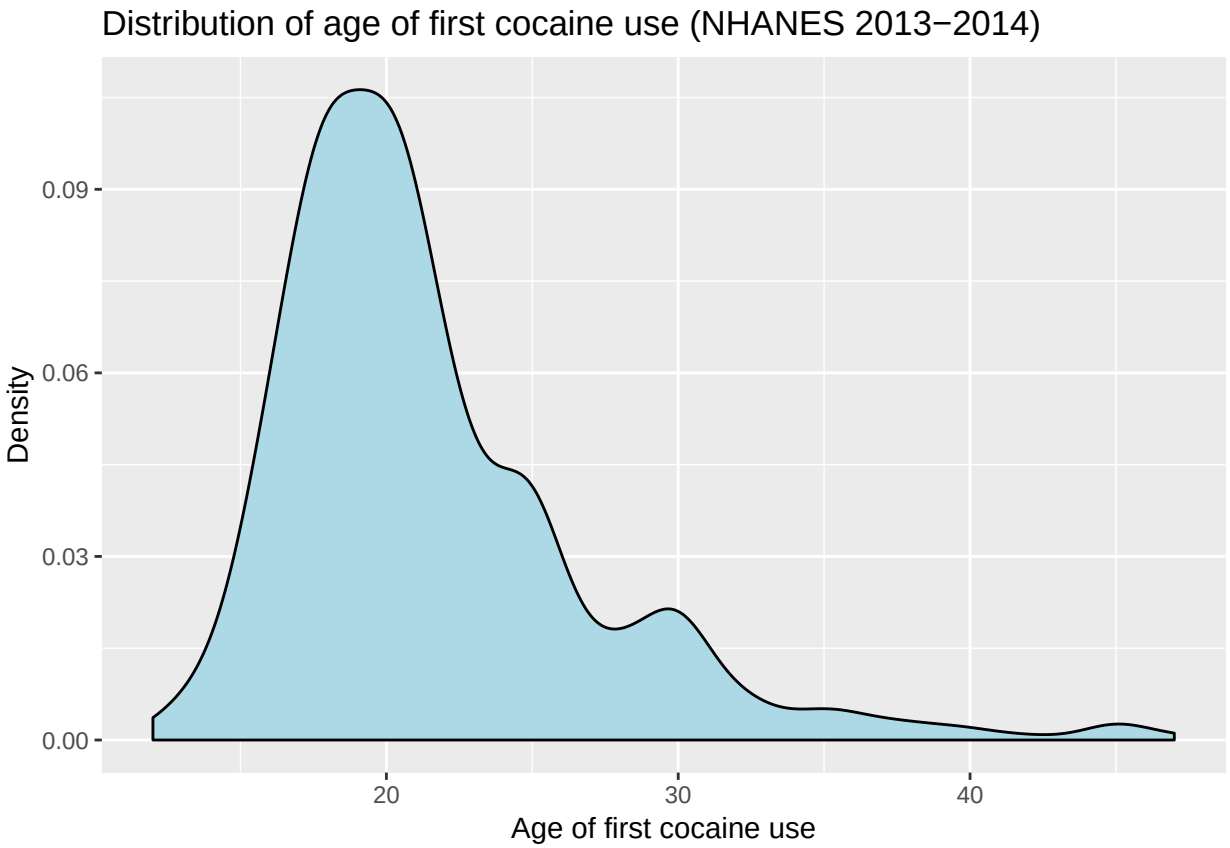
```
# basic density plot
ggplot(nhanes2013, aes(x = ageFirstUse)) +
  geom_density()
```

```
## Warning: Removed 4483 rows containing non-finite values (stat_density).
```




```
# density plot
# added labels and titles
ggplot(nhanes2013, aes(x = ageFirstUse)) +
  geom_density(fill = I("lightblue")) +
  xlab("Age of first cocaine use") +
  ylab("Density") +
  ggtitle("Distribution of age of first cocaine use (NHANES 2013-2014)")
```

```
## Warning: Removed 4483 rows containing non-finite values (stat_density).
```



FoB2 Challenge

Continue to use the 2013-2014 NHANES data set with the Cigarette Use module (SMQ_H) rather than the drug use module examined in this packet. The codebook is available here: https://wwwn.cdc.gov/Nchs/Nhanes/2013-2014/SMQ_H.htm. Open a new R Script file and:

- Use R commands to bring the data into R
- Write R commands to complete the tasks below
- Annotate the R commands so it is clear what you did and which questions you were answering

Submit your *R commands with annotation* to Blackboard before the next class meeting. If you see a standard edition and hacker edition below, you may choose either one. The standard edition will evaluate your application of the commands learned in this packet (and earlier packets) to similar scenarios; the hacker edition will evaluate your use of the procedures from this packet in new scenarios or may ask you to figure out a new command on your own. Both editions are worth the same number of points.

Standard edition

1. Recode the SMQ020 variable to have a logical variable name and to have labels for each category (e.g., Yes, No, Every day) rather than numbers.
2. Recode the SMD030 variable to have a logical variable name and so that values outside the main range are all interpreted by R as `Missing`.

3. Use an appropriate table and an appropriate graph to show the distribution of the variables you recoded in #1 and #2.
4. Compute the central tendency for the variables examined in questions #1 and #2. Use the most appropriate measure of central tendency for each variable.
5. Report the spread for the variable from question 2. Use the most appropriate measure of spread.
6. Determine the probability of selecting person at random from the data who smoked at least 100 cigarettes in their life.
7. Determine the probability of starting smoking regularly between ages 7 and 12.

Hacker edition

Complete # 1 - 6 from standard edition

7. Create a density plot age started smoking cigarettes regularly.
8. Determine the probability of starting smoking regularly between ages 7 and 12.
9. Shade the density plot to highlight the area under the curve representing the probability of starting smoking regularly between ages 7 and 12. Comment on whether the probability you found in #8 seems consistent with the amount of shading under the curve. Be sure you can see the curve and the shaded area well (you may have to adjust some of the values in the plot commands!).