Homework 2 - Key

Math 530/630

1. A coin is tossed three times and the sequence of heads and tails is recorded.
   a. List the sample space.
   b. List the elements that make up the following events:
      1. A = at least two heads;
      2. B = the first two tosses are heads;
      3. C = the last toss is a tail.
   c. List the elements of the following events:
      1. complement of A,
      2. $A \cap B$,
      3. $A \cup C$.

Note, for all answers in this homework that don't explicitly call for R, I don't require or expect that you use R. I use it on most problems just for the sake of convenience and so you can follow the calculations.

Answers

a) hhh, ttt, hht, tth, htt, hth, tht, thh

b)
1. hhh, hht, hth, thh
2. hhh, hht
3. ttt, hht, htt, tht

c)
1. ttt, tth, tht, htt
2. hhh, hht
3. ttt, hht, htt, tht, hhh, hth, thh

2. In a city, 65% of people drink coffee, 50% drink tea, and 25% drink both.
   a. What is the probability that a person chosen at random will drink at least one of coffee or tea?
   b. Will drink neither?

Answers

a) .9

```
coffee <- .65
tea <- .5
both <- .25
```

Recalling that $P(A \cup B) = P(A) + P(B) - P(A \cap B)$:

```
coffee_or_tea <- coffee + tea - both
coffee_or_tea
```

```
## [1] 0.9
```

b) .1

Recalling that $P(A^c) = 1 - P(A)$:

```
1 - coffee_or_tea
```

```
## [1] 0.1
```

3. In this problem, we'll explore quantile-quantile (Q-Q) plots. A quantile is the proportion of cases we find below a certain value, calculated from the inverse of the cumulative distribution function (CDF)

of a random variable, X. The pth quantile of X is the value q_p such that $P(X \leq q_p) = p$. In other words, p is the amount of area under the density curve of X that is to the left of q_p. So the smallest observation in X corresponds to a probability of 0 and the largest to a probability of 1. A Q-Q plot displays quantiles of one distribution against quantiles of another. What this means is that the data are ranked and sorted. A normal Q-Q plot displays quantiles of the normal distribution on the x-axis against quantiles of the empirical (i.e., the observed) distribution on the y-axis. A straight line is typically plotted through the points corresponding to the 1st and 3rd quantiles of each variable. If the empirical data is normally distributed, all the points on the normal Q-Q plot will form a perfectly straight line.

a. Draw a random sample of size $n = 15$ from $N(0, 1)$ and plot both the normal quantile plot and the histogram. Do the points on the quantile plot appear to fall on a straight line? Is the histogram symmetric, unimodal, and mound-shaped? Do this several times.
b. Repeat part(a) for of size n = 30, n = 60, and n = 100.
c. What lesson do you draw about using graphs to assess whether or not a data set follows a normal distribution?

Answers

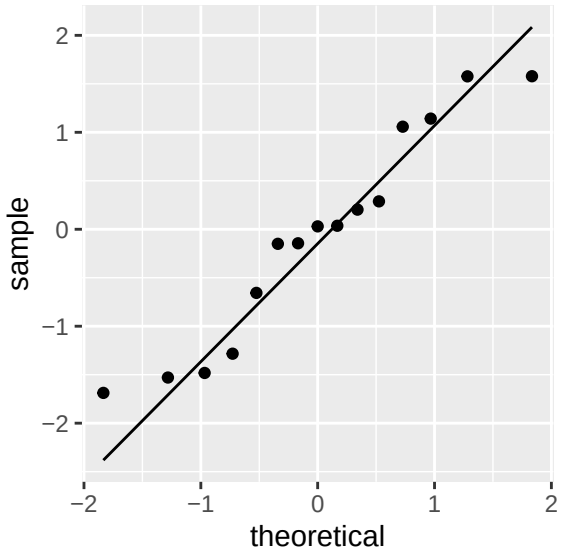
*Note: When plotting histograms, you typically want to reduce the number of bins until there are few/no breaks (empty bins) in your histogram. With small data sets, this can make it hard to visualize the true underlying distribution, so sometimes a density plot can be a bit more helpful here. From `?geom_density`:

> *"Computes and draws kernel density estimate, which is a smoothed version of the histogram. This is a useful alternative to the histogram for continuous data that comes from an underlying smooth distribution."*

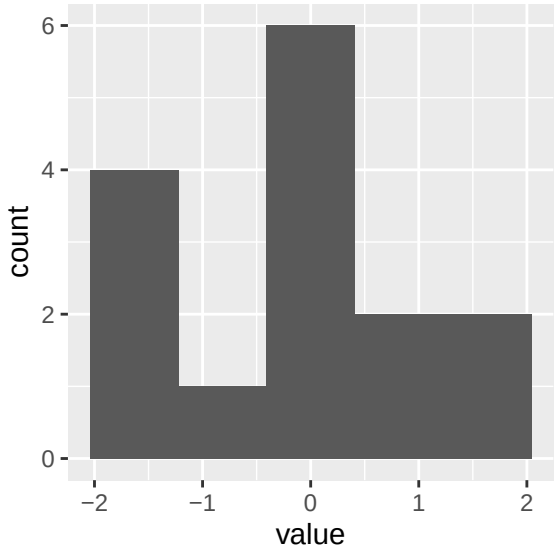
```
library(tibble)
library(ggplot2)

x <- as_tibble(rnorm(15)) # draw random sample of size 15 from N(0,1)
```

```
## Warning: Calling `as_tibble()` on a vector is discouraged, because the behavior is likely to change i
## This warning is displayed once per session.
```

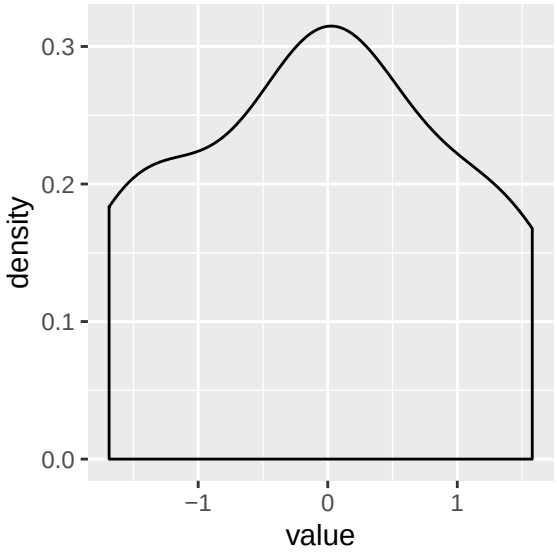
```
ggplot(x, aes(sample = value)) +
  stat_qq() +
  stat_qq_line()
```



```
ggplot(x, aes(x = value)) +
  geom_histogram(bins = 5)
```

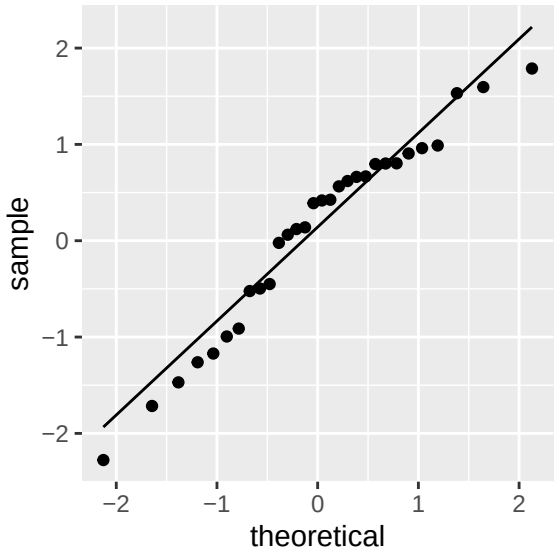



```
ggplot(x, aes(x = value)) +
  geom_density()
```

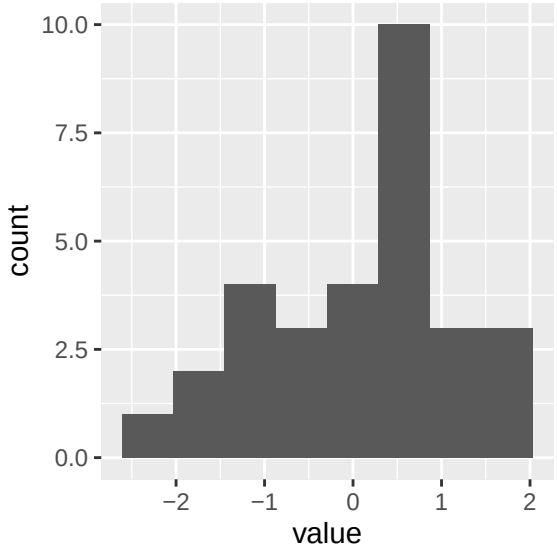


Now here are the results for n = 30, n = 60, and n = 100.

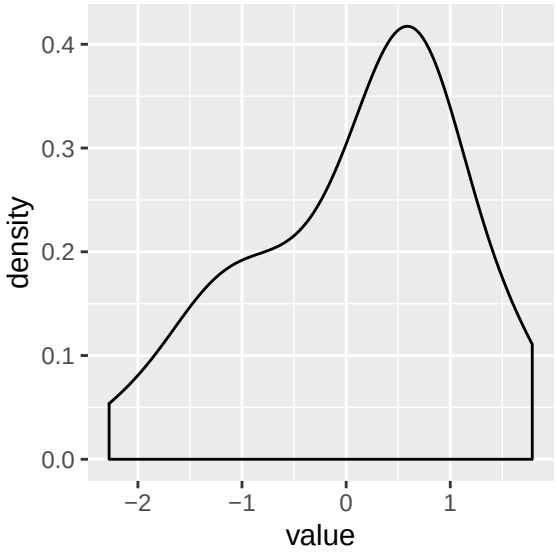
```
x <- as_tibble(rnorm(30)) # draw random sample of size 15 from N(0,1)

ggplot(x, aes(sample = value)) +
  stat_qq() +
  stat_qq_line()
```



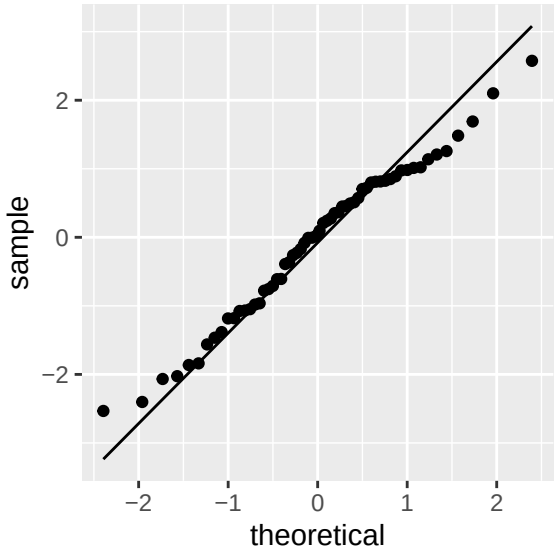

```
ggplot(x, aes(x = value)) +
  geom_histogram(bins = 8)
```



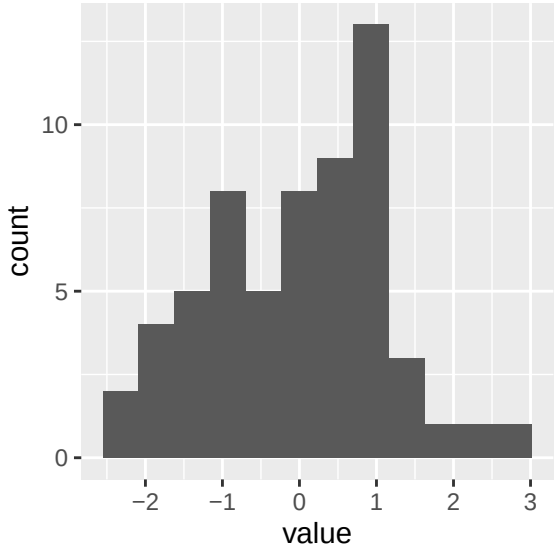

```
ggplot(x, aes(x = value)) +
  geom_density()
```


0.4
0.3
0.2
0.1
0.0
density
−2
−1
0
1
value

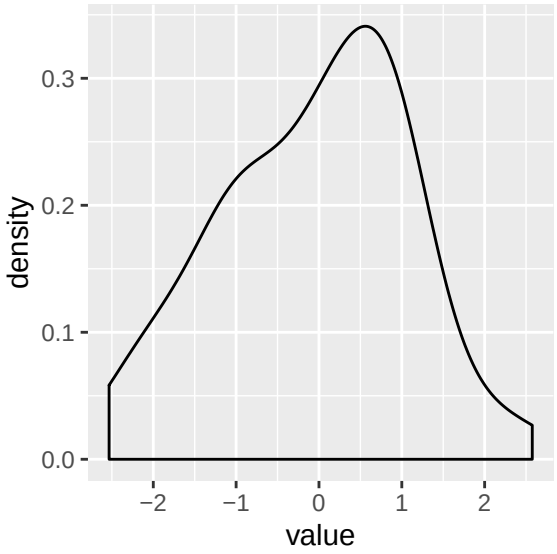
```r
x <- as_tibble(rnorm(60)) # draw random sample of size 15 from N(0,1)

ggplot(x, aes(sample = value)) +
  stat_qq() +
  stat_qq_line()
```



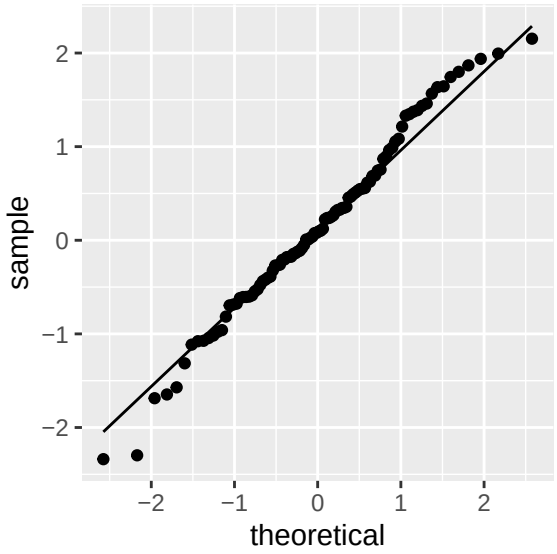

```r
ggplot(x, aes(x = value)) +
  geom_histogram(bins = 12)
```



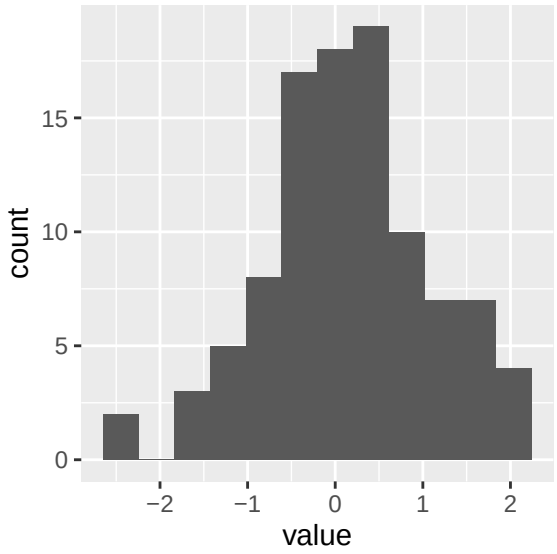

```r
ggplot(x, aes(x = value)) +
  geom_density()
```


0.3
0.2
0.1
0.0
density
−2
−1
0
1
2
value

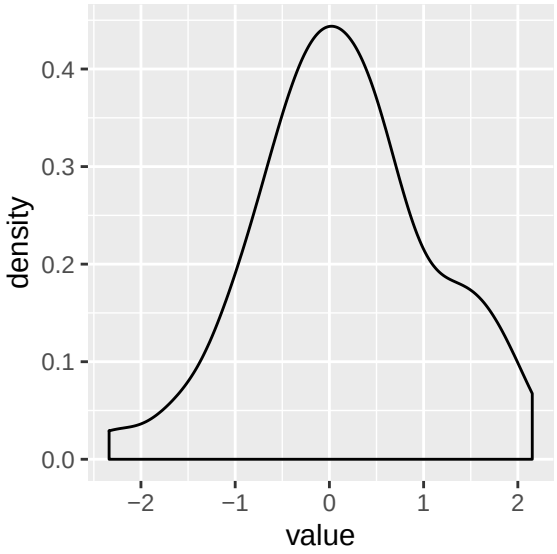
```r
x <- as_tibble(rnorm(100)) # draw random sample of size 15 from N(0,1)

ggplot(x, aes(sample = value)) +
  stat_qq() +
  stat_qq_line()
```




```r
ggplot(x, aes(x = value)) +
  geom_histogram(bins = 12)
```




```r
ggplot(x, aes(x = value)) +
  geom_density()
```

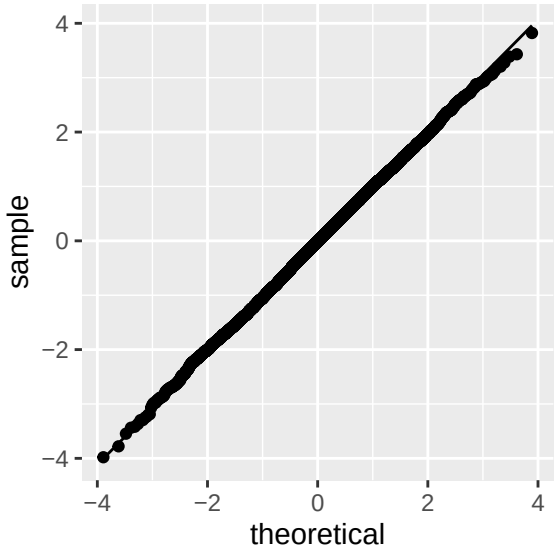

0.4
0.3
0.2
0.1
0.0
density
−2
−1
0
1
2
value

Here's the meat of the problem. Your answer should be something along the lines of:

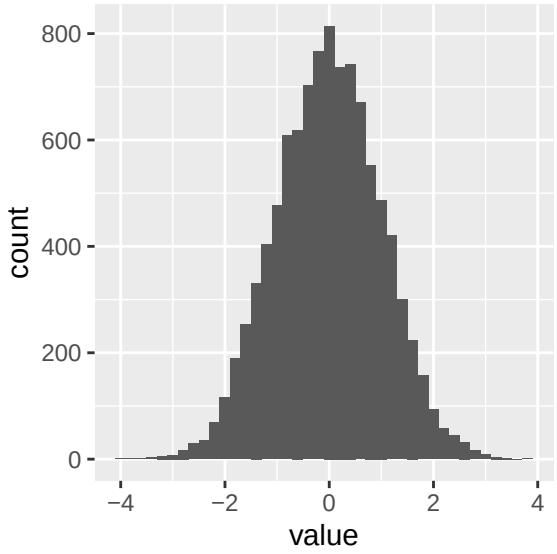
For small values of n, samples drawn from the normal distribution may not have the classic "bell shaped" distribution. In addition, qq-plots may not fall precisely on a straight line.

And an example with a very large n just to drive the point home...

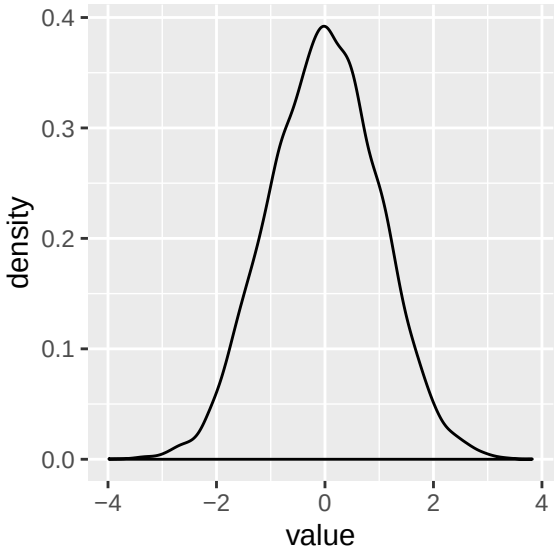
```
x <- as_tibble(rnorm(10000)) # draw random sample of size 15 from N(0,1)

ggplot(x, aes(sample = value)) +
  stat_qq() +
  stat_qq_line()
```




```
ggplot(x, aes(x = value)) +
  geom_histogram(bins = 40)
```




```
ggplot(x, aes(x = value)) +
  geom_density()
```


0.4
0.3
0.2
0.1
0.0
density
−4
−2
0
2
4
value