Lecture 5: Comparisons

Criminology 250

Prof Maria Cuellar

University of Pennslyvania

Relationships between two variables

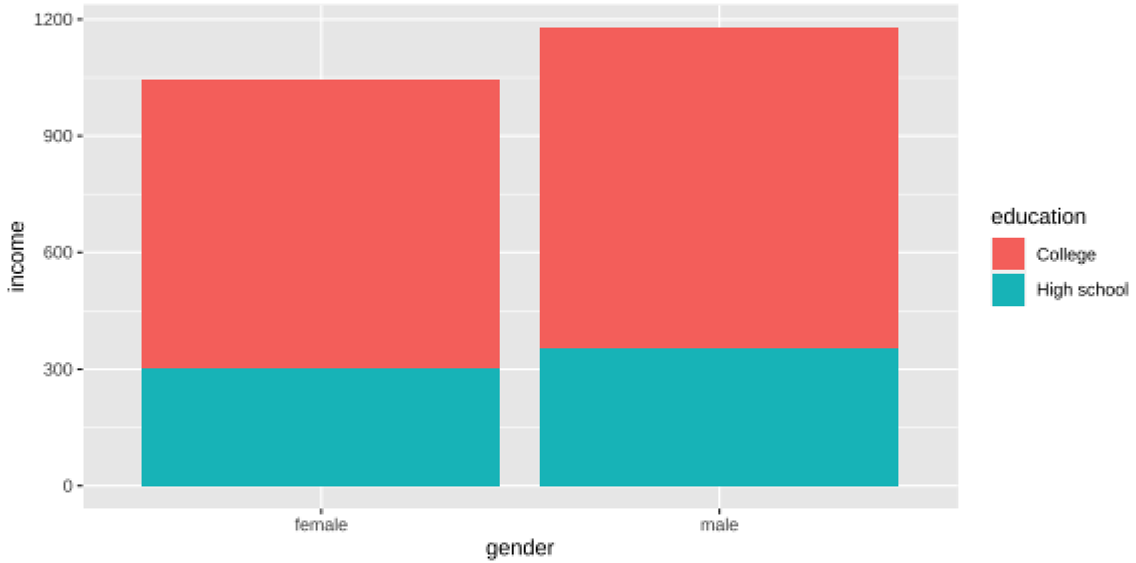
We talked about EDA for comparing two variables. Here are some ideas, but you could use others as well.

- Two categorical variables: stacked barplot.
- One categorical, one quantitative: side-by-side boxplots, side-by-side histograms.
- Two quantitative variables: scatterplot.

Comparing two categorical variables

A stacked bar plot is a good idea.

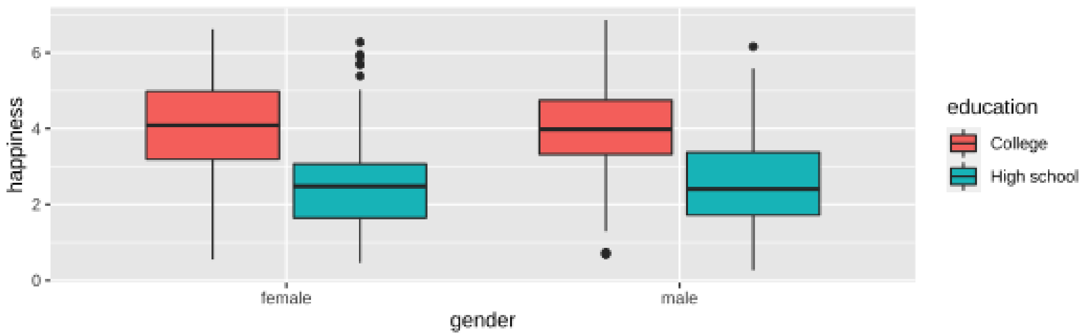
```
library(ggplot2)
dat.income <- read.csv(file = 'income.data.csv')

# Stacked
ggplot(dat.income, aes(fill=education, y=income, x=gender)) +
    geom_bar(position="stack", stat="identity")
```



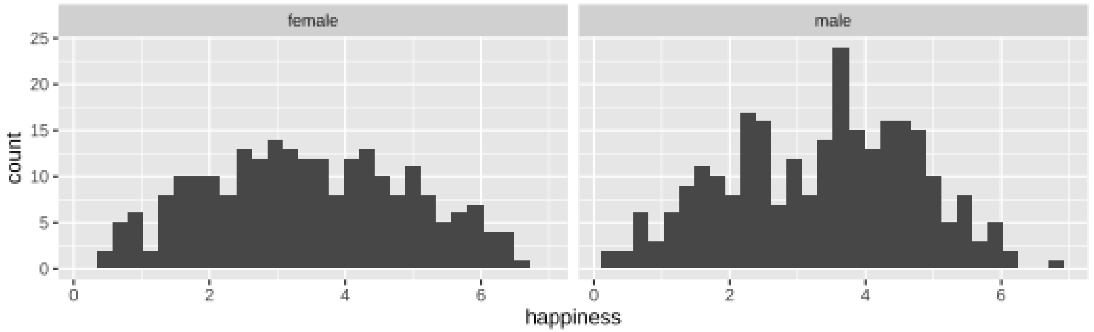
Comparing one categorical variable, one quantitative variable

Two ideas (probably not the only possibilities): side-by-side bar plots, side-by-side histograms.

```
ggplot(dat.income, aes(y=happiness, x=gender
                       , fill=education ) )+
  geom_boxplot()
```



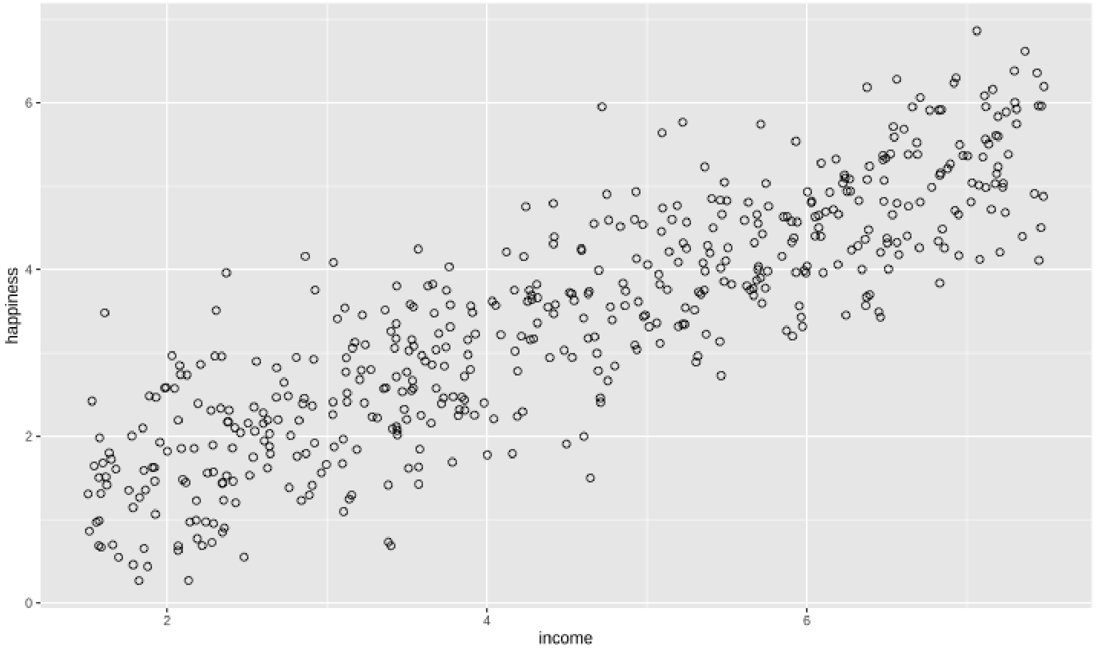
```
ggplot(data=dat.income, aes(x = happiness)) +
  geom_histogram() +
  facet_wrap(vars(gender))
```



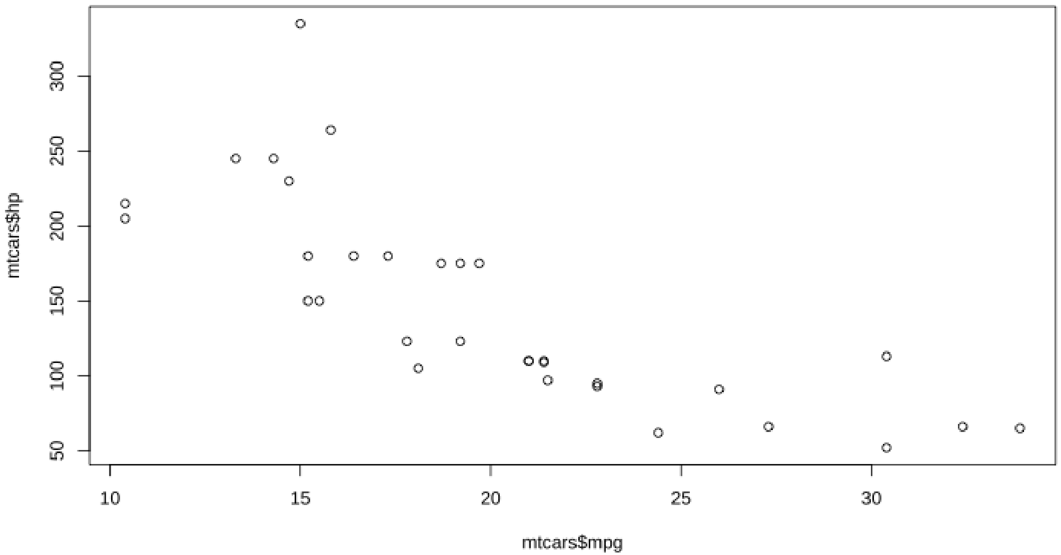
Comparing two quantitative variables

Scatterplots.

```
ggplot(dat.income, aes(x=income, y=happiness)) +
  geom_point(size=2, shape=1)
```



```
plot(mtcars$mpg, mtcars$hp) # mile per gallon vs h
```



3D Scatterplots

```r
library(plotly)

mtcars$am[which(mtcars$am == 0)] <- 'Automatic'
mtcars$am[which(mtcars$am == 1)] <- 'Manual'
mtcars$am <- as.factor(mtcars$am)

fig <- plot_ly(mtcars, x = ~wt, y = ~hp, z = ~qsec
fig <- fig %>% add_markers()
fig <- fig %>% layout(scene = list(xaxis = list(ti
                     yaxis = list(title = 'Gross h
                     zaxis = list(title = '1/4 mil
```

● Automatic
● Manual

Correlation

Correlation gives us a numeric measure of the strength of a relationship, between -1 and 1. It measures the strength of linear association between two quantitative variables. (-1 or 1 is a perfect linear relationship, a correlation of 0 means there is no linear relationship.)

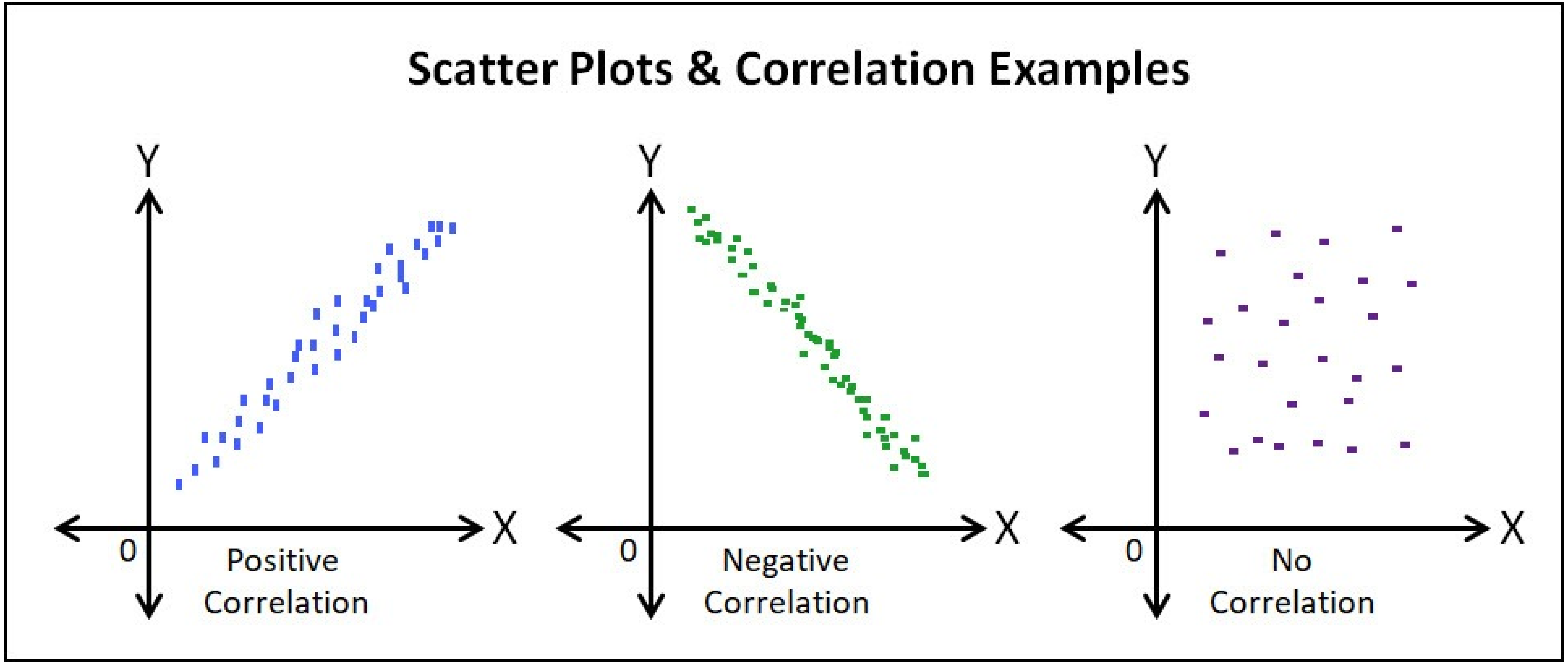
Make sure the two variables are quantitative, that the scatterplot looks "straight", and that there are no extreme outliers.

The *correlation coefficient* is defined as:

$$r = \frac{\sum(x - \overline{x})(y - \overline{y})}{\sqrt{\sum(x - \overline{x})\sum(y - \overline{y})}}.$$

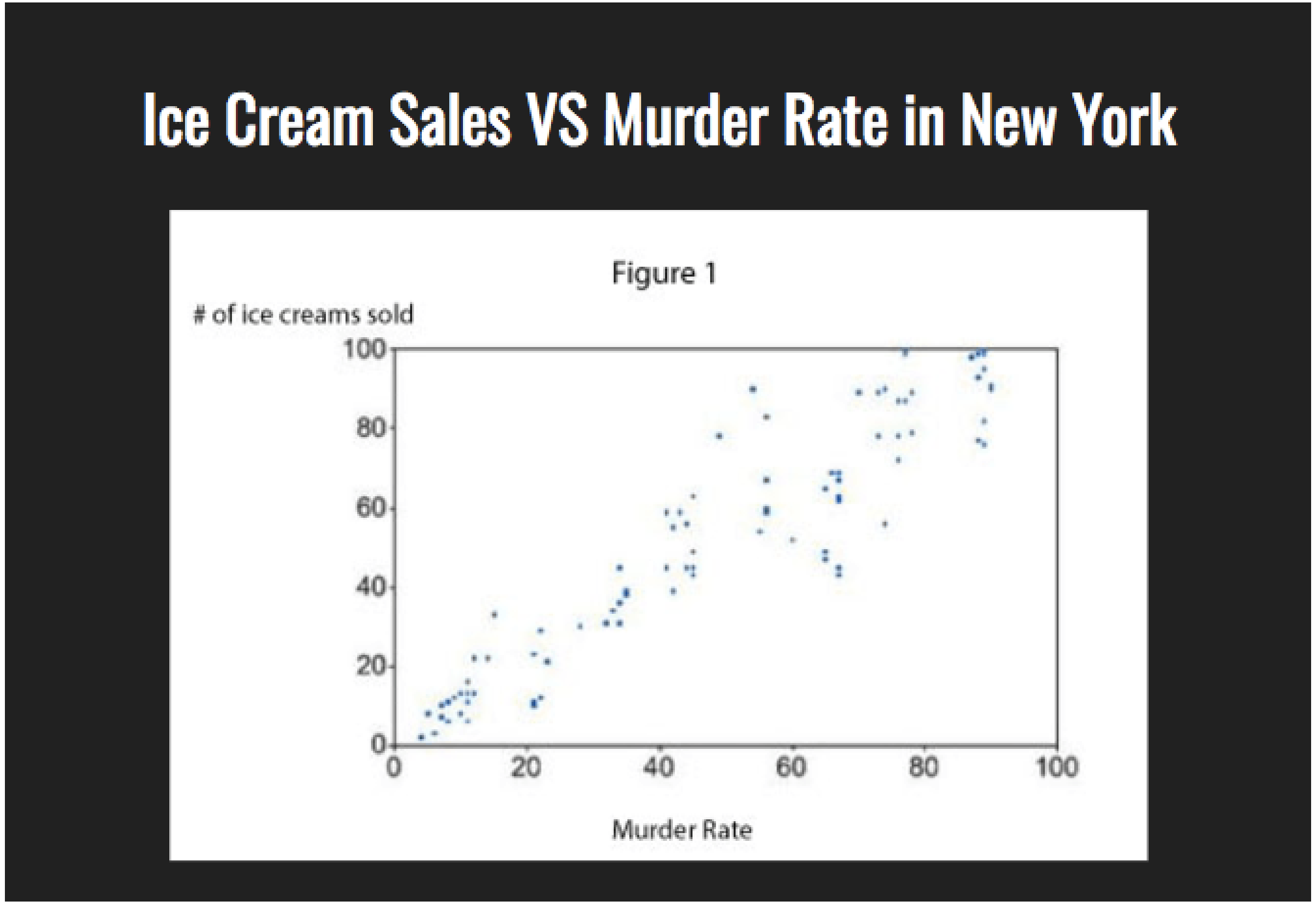
This would be very tedious to calculate by hand, so we use R.

What do correlations look like in a scatterplot?

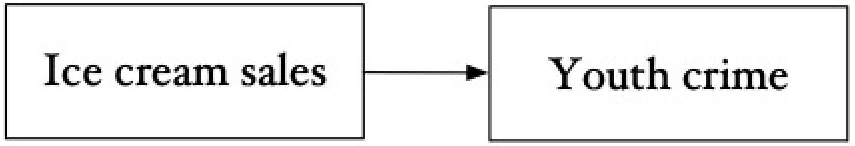


Correlation does not imply causation

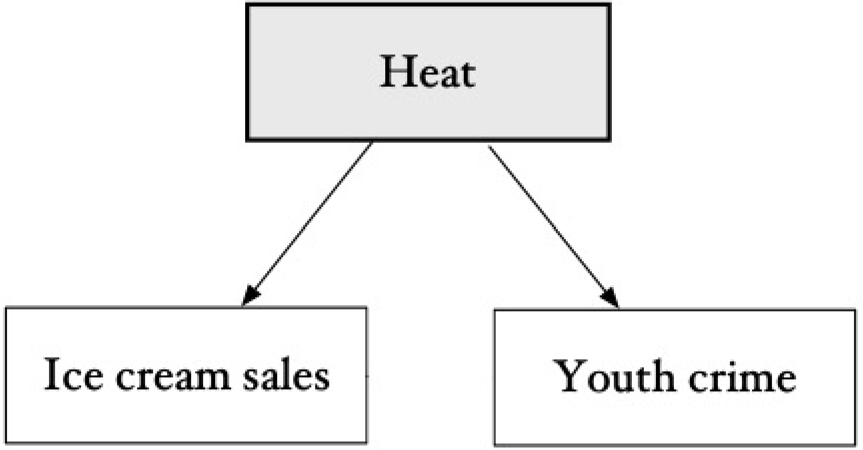
When ice cream sales are high, the murder rate is also high.



Correlation does not imply causation



Correlation does not imply causation



Linear regression

Regression models describe the relationship between variables by fitting a line to the observed data. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change.

A *model* is an equation or formula that simplifies and represents reality.

Simple linear regression is used to estimate the relationship between two quantitative variables. You can use simple linear regression when you want to know:

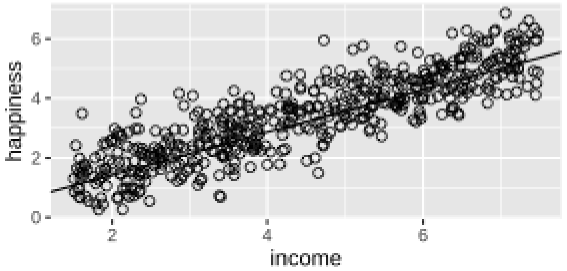
- How strong the relationship is between two variables (e.g. the relationship between rainfall and soil erosion).
- The value of the dependent variable at a certain value of the independent variable (e.g. the amount of soil erosion at a certain level of rainfall).

Linear model

A linear model is an equation of the form

$$\hat{y} = b_0 + b_1 x.$$

The predicted value is the value of $\hat{y}$ found for a given x-value in the data.



Some terminology

x: explanatory, independent, predictor

y: outcome, dependent, response

Assumptions of linear regression

1. Homogeneity of variance (homoscedasticity): the size of the error in our prediction doesn’t change significantly across the values of the independent variable.

2. Independence of observations: the observations in the dataset were collected using statistically valid sampling methods, and there are no hidden relationships among observations.

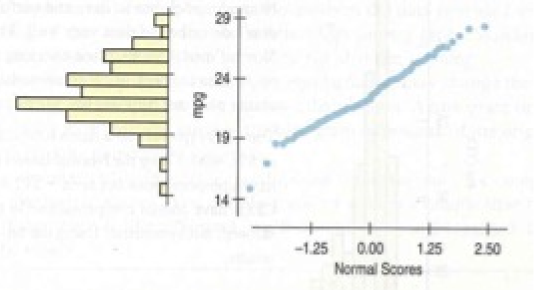
3. Normality: The data follows a normal distribution.

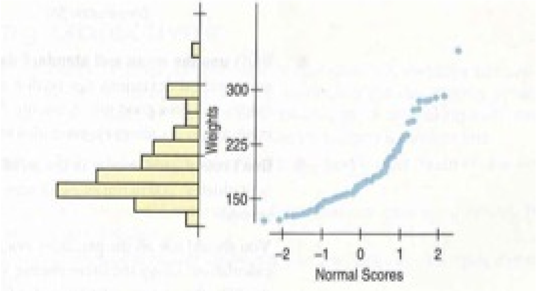
4. The relationship between the independent and dependent variable is linear: the line of best fit through the data points is a straight line (rather than a curve or some sort of grouping factor).

Visual test for normality: Q-Q plot

How can we decide if using a Normal model is appropriate?

If the distribution of the data is roughly Normal, a *quantile-quantile plot* is roughly a diagonal straight line. Deviations from a straight line indicate that the distribution is not Normal.





```
## [1] 0.8656337
```