Densities and the Normal Distribution

David Gerard

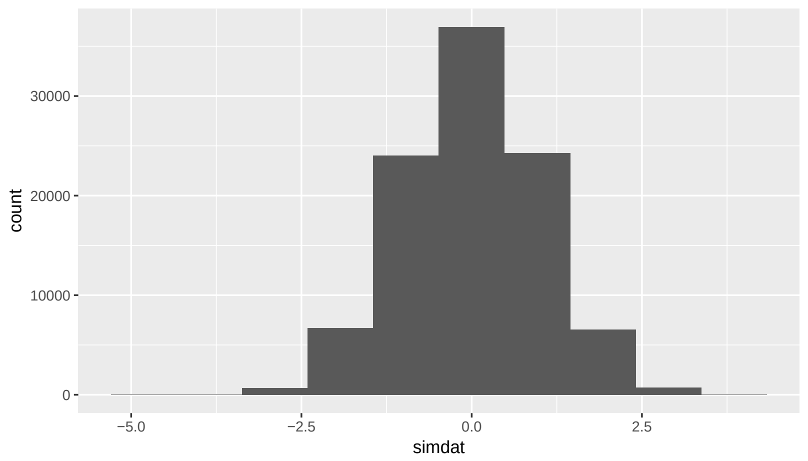
2017-09-18

Learning Objectives

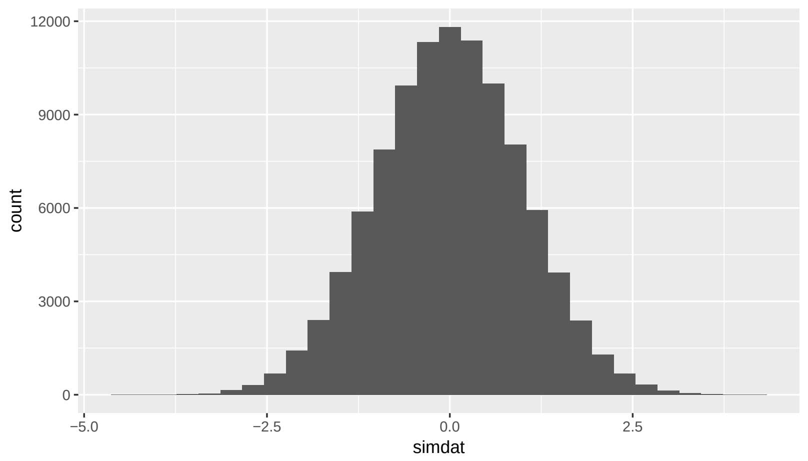
- Density Curves
- Normal curves
- QQ-plots
- Sections 2.5.1, 3.1.1, 3.1.2, 3.1.5, 3.2

Density Curves

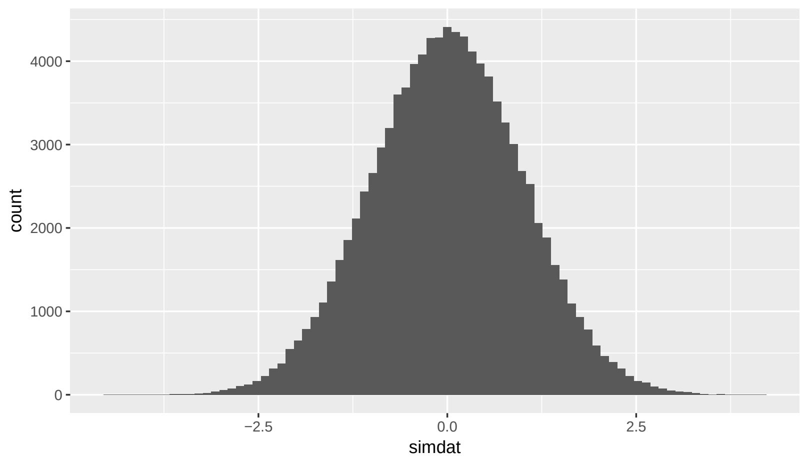
A histogram of simulated data



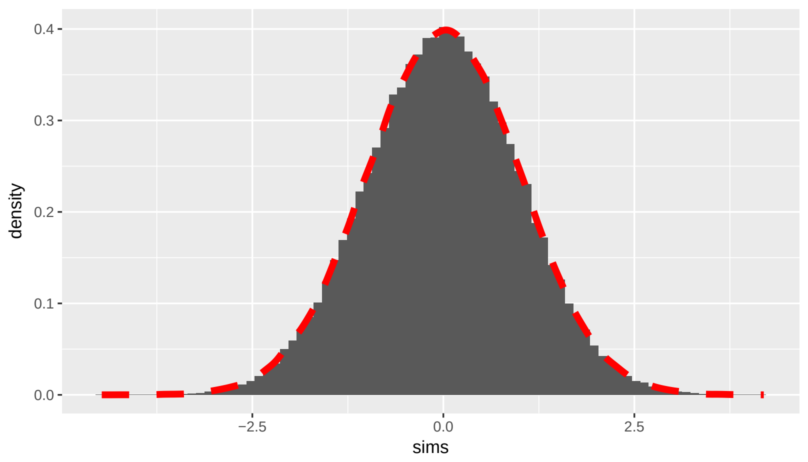
What if we decrease the binwidth?



And more



What do you notice?



Starting to look like a smooth curve!

Density curve

- The distributions of many quantitative variables can be approximated by a density curve

density curve

A density curve describes the overall pattern of a distribution. The area under the curve and above any range of values is the proportion of all observations that fall in that range. A density cuve is a curve that

- Is always on or above the horizontal axis.
- Has area exactly 1 underneath it.

Recall: Movie Scores

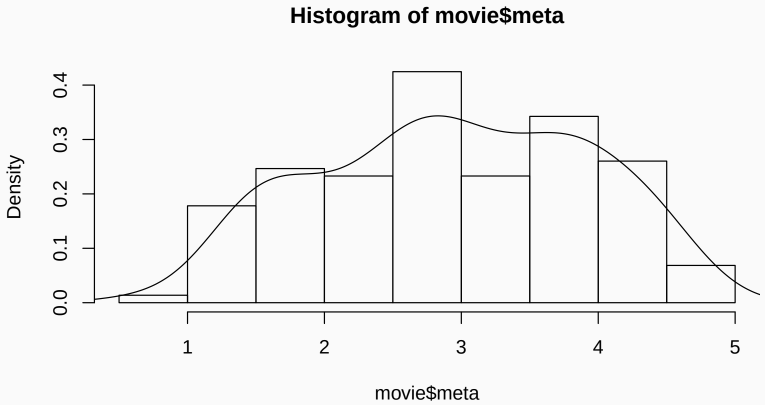
Observational units: Movies that sold tickets in 2015.

Variables:

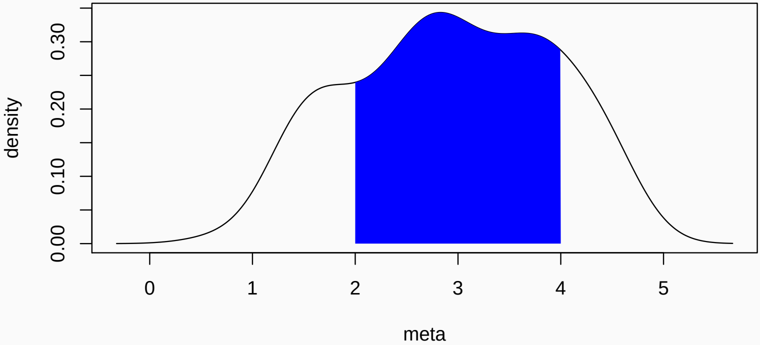
- `rt` Rotten tomatoes score normalized to a 5 point scale.
- `meta` Metacritic score normalized to a 5 point scale.
- `imdb` IMDB score normalized to a 5 point scale.
- `fan` Fandango score.

Density of Metacritic scores

```r
md <- density(movie$meta)
hist(movie$meta, freq = FALSE)
lines(md$x, md$y)
```

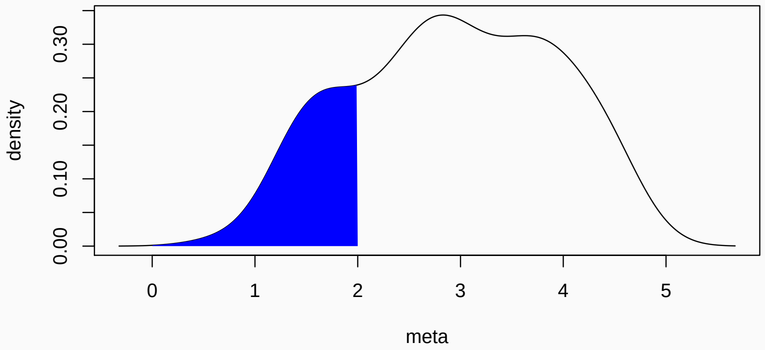


Density example



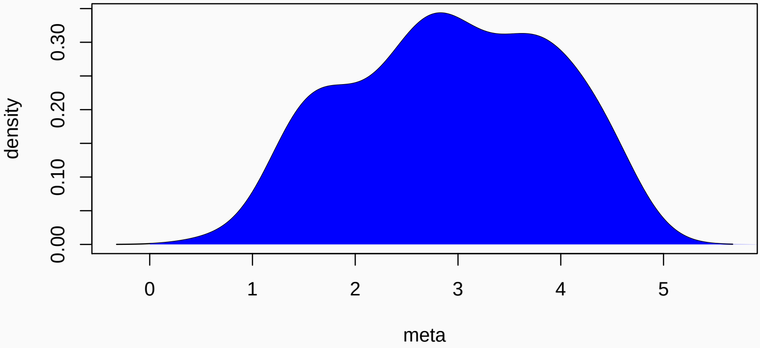
E.g.: Area of shaded region is approximately the proportion of metracritic scores that falls between 2 and 4.

Density example



E.g.: Area of shaded region is approximately the proportion of metracritic scores that are less than 2.

Density example

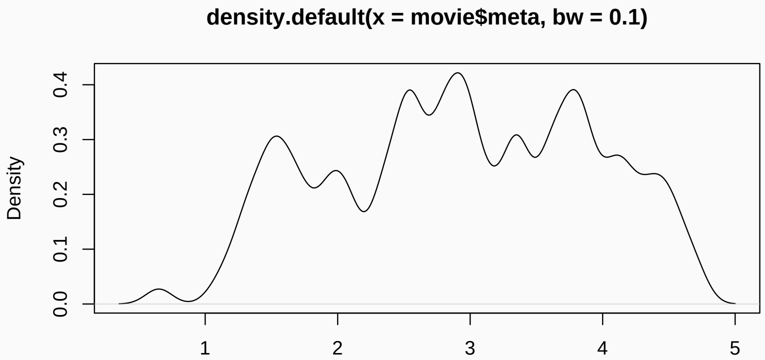


E.g.: Area of shaded region is exactly 1.

Smoothness

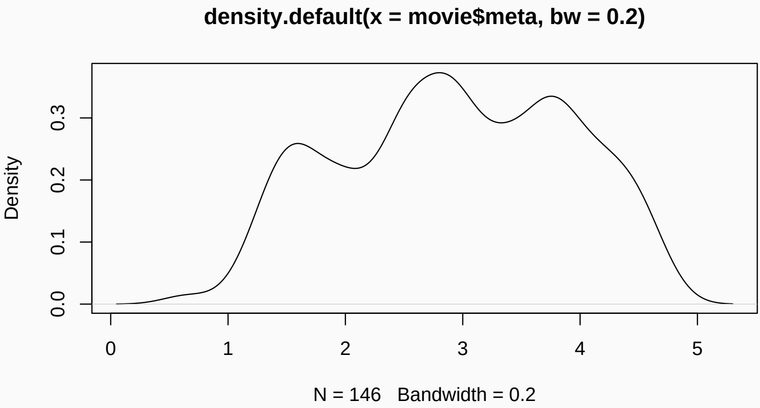
Just as you can control the bin-width of histograms, you can control the smoothness (aka "bandwidth") of density plots.

```
md <- density(movie$meta, bw = 0.1)
plot(md)
```



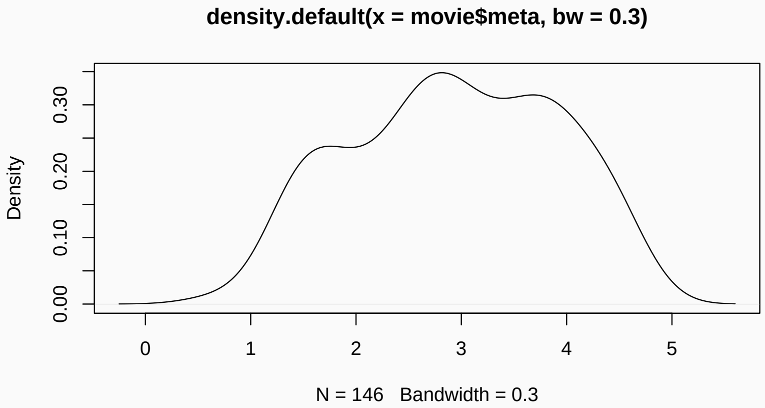
More smooth

```r
md <- density(movie$meta, bw = 0.2)
plot(md)
```



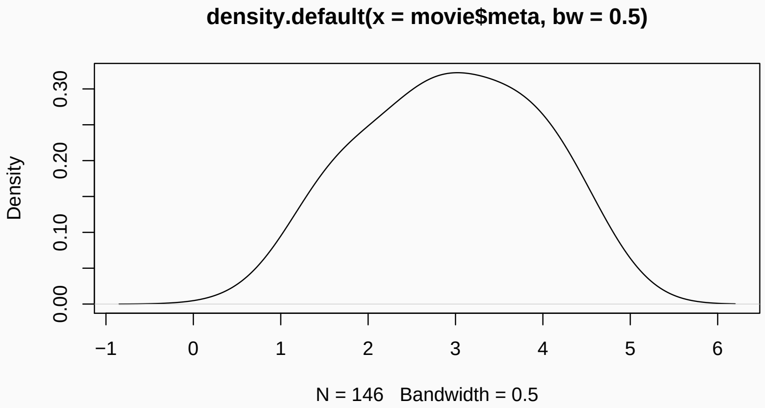
More smooth

```r
md <- density(movie$meta, bw = 0.3)
plot(md)
```



Too smooth!

```
md <- density(movie$meta, bw = 0.5)
plot(md)
```



Mean and median

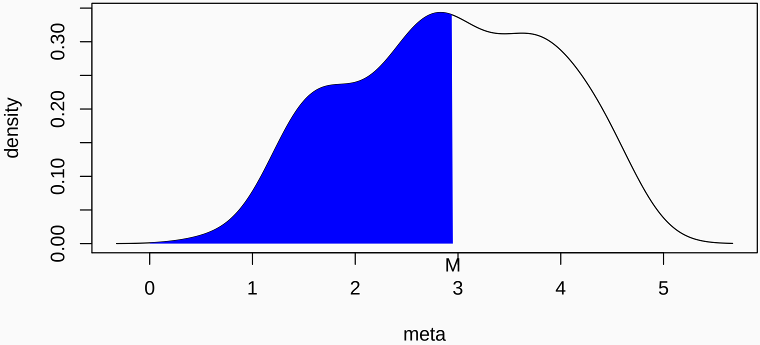
median

The median of a density curve is the equal-areas point, the point that divides the area under the curve in half.

mean

The mean of a density curve is the balance point, at which the curve would balance if made of solid material.

Median



Median M is where half of the area is to the left and to the right of M.

Normal Density Curves

Recall SAT scores

A data frame with 1000 observations on the following 6 variables.

- `sex` Gender of the student.
- `SATV` Verbal SAT percentile.
- `SATM` Math SAT percentile.
- `SATSum` Total of verbal and math SAT percentiles.
- `HSGPA` High school grade point average.
- `FYGPA` First year (college) grade point average.

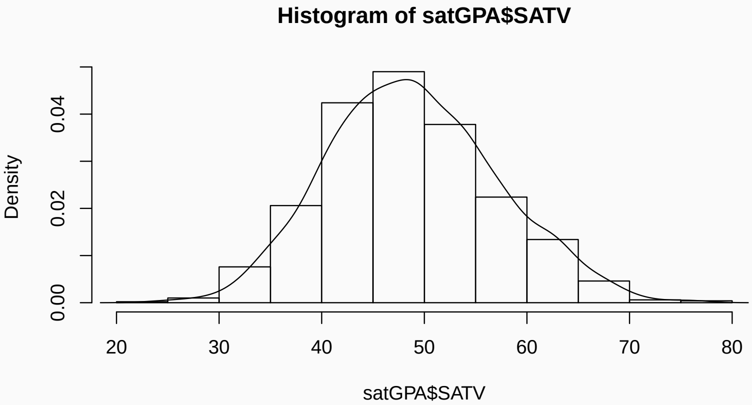
satGPA

```
library(tidyverse)
data(satGPA, package = "openintro")
glimpse(satGPA)
```

```
Observations: 1,000
Variables: 6
$ sex    <int> 1, 2, 2, 1, 1, 2, 1, 1, 2, 1, 1, 2, 2, 2...
$ SATV   <int> 65, 58, 56, 42, 55, 55, 57, 53, 67, 41, ...
$ SATM   <int> 62, 64, 60, 53, 52, 56, 65, 62, 77, 44, ...
$ SATSum <int> 127, 122, 116, 95, 107, 111, 122, 115, 1...
$ HSGPA  <dbl> 3.40, 4.00, 3.75, 3.75, 4.00, 4.00, 2.80...
$ FYGPA  <dbl> 3.18, 3.33, 3.25, 2.42, 2.63, 2.91, 2.83...
```

Bell-shaped curves

```
hist(satGPA$SATV, freq = FALSE)
md <- density(satGPA$SATV)
lines(md$x, md$y)
```



Normal density

One particular bell-shaped density curve is the normal density.

normal curve

The normal curve describes the normal distribution. It is bell-shaped and is defined by the equation:

$$f(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2},$$

where μ is the mean and σ is the standard deviation of the normal distribution.

Facts about the normal density.

- Symmetric, unimodal.
- Completely described by its mean μ and its standard deviation (or variance) σ.
- 1 σ from μ is an inflection point — a point where the 2nd derivative switches from positive to negative (or vice versa). I.e. transition from concave to convex (or vice versa).
- Many variables follow a normal distribution (test scores, physical measurements)
- Many chance processes converge to a normal distribution (more on this later).

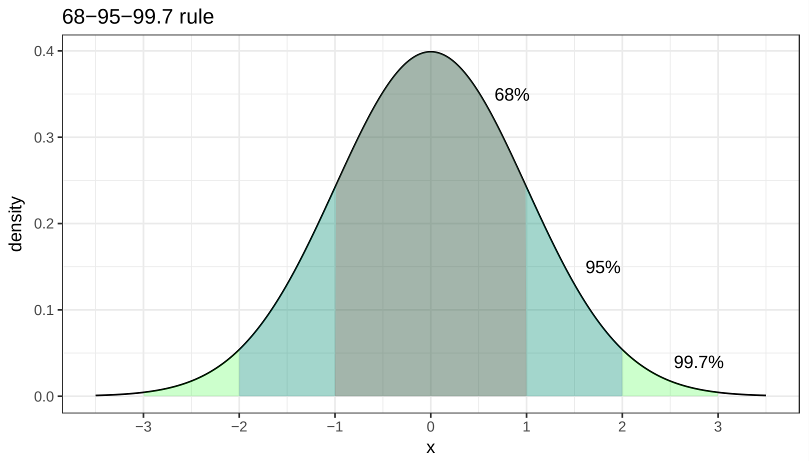
68-95-99.7 rule

In the Normal distribution with mean μ and standard deviation σ

- Approximately 68% of the observations fall within σ of μ
- Approximately 95% of the observations fall within 2σ of μ
- Approximately 99.7% of the observations fall within 3σ of μ

This rule does not depend on the values of μ and σ.

68-95-99.7 rule



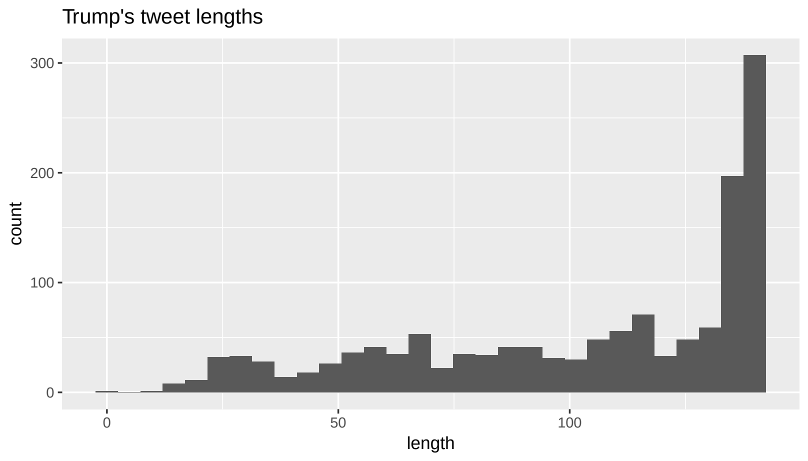
Percentiles

Use the 68-95-99.7 rule to answer these questions.

- What percentile is -3σ? 0.0015
- What percentile is -2σ?
- What percentile is -1σ?
- What percentile is 0σ? 0.5
- What percentile is 1σ?
- What percentile is 2σ? 0.975
- What percentile is 3σ?

Checking for normality

Clearly not all distributions are normal



It's sometimes important to check if normality is a valid approximation.

- Idea: Is the 68-95-99.7 rule approximately correct for the `satGPA` data?
- More generally, do the percentiles (quantiles) of the data match with the percentiles (quantiles) of the theoretical normal distribution?
- Compare the pth percentile (quantile) of the data and the pth percentile (quantile) of a $N(\bar{x}, s^2)$ distribution. If they are pretty close, then normality is a good approximation.

Look at percentiles (quantiles)

```
mu    <- mean(satGPA$SATV)
sigma <- sd(satGPA$SATV)
qnorm(p = 0.2, mean = mu, sd = sigma)
```

```
[1] 42
```

```
quantile(x = satGPA$SATV, probs = 0.2)
```

```
20%
 42
```

That matches almost exactly, what about other percentiles (quantiles)?

More percentiles (quantiles)

```r
qnorm(p = 0.4, mean = mu, sd = sigma)
```

```
[1] 46.85
```

```r
quantile(x = satGPA$SATV, probs = 0.4)
```

```
40%
 46
```

more percentiles(quantiles)

```r
qnorm(p = 0.9, mean = mu, sd = sigma)
```

```
[1] 59.49
```

```r
quantile(x = satGPA$SATV, probs = 0.9)
```

```
90%
 60
```

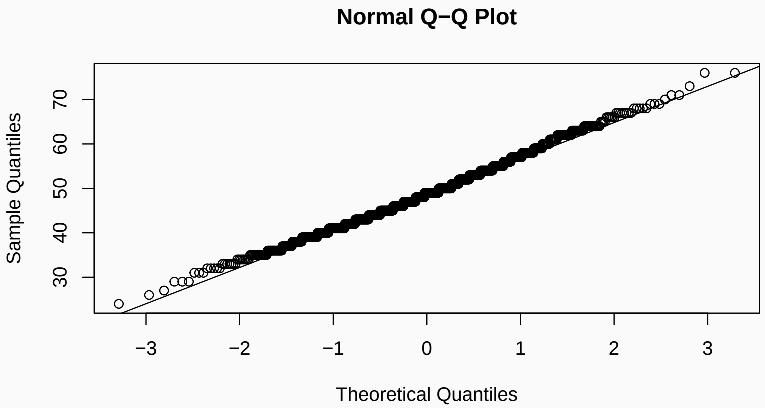
These are all pretty close!

Quantile-quantile plot

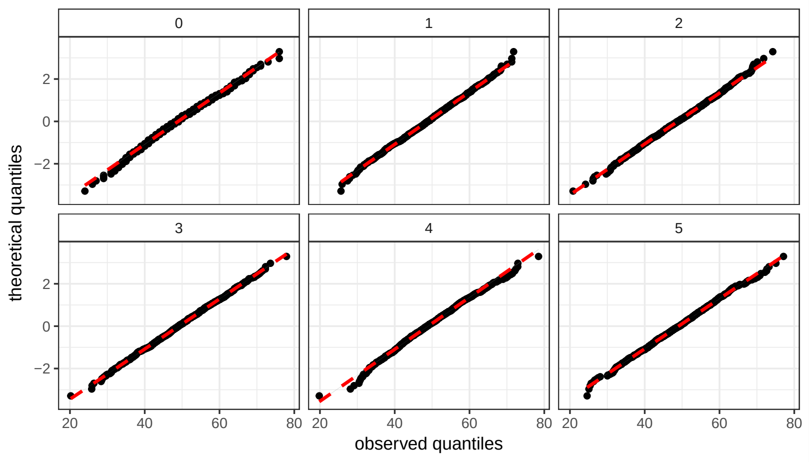
- Plots the observed quantiles against the quantiles of a $N(\bar{x}, s^2)$ density.
- If the points lie close to a line, then the normal approximation is approximately correct.
- Can just plot the observed quantiles against $N(0, 1)$ and look for a straight line (more on why later).

QQplot

```
qqnorm(satGPA$SATV)
qqline(satGPA$SATV)
```

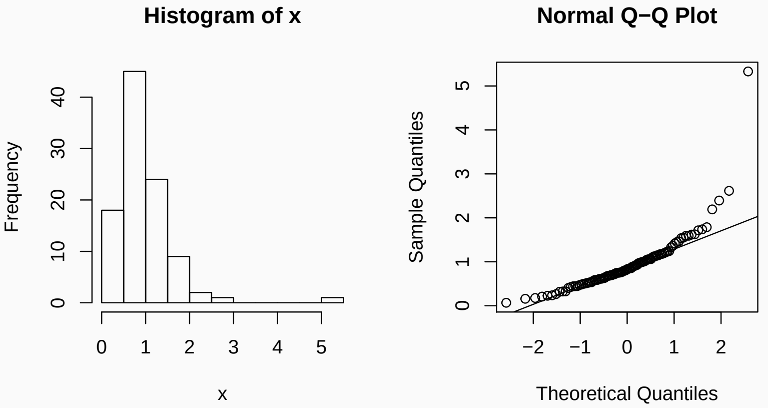


But what does a “good” qqplot look like?

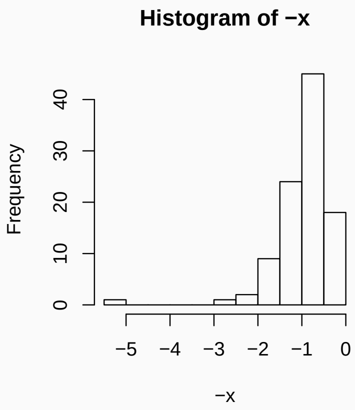


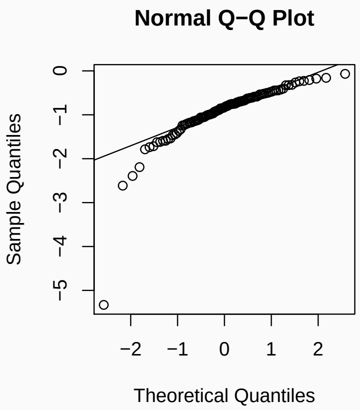
Top left is real data, rest are simulated from $N(\bar{x}, s^2)$ — looks good to me!

Problem: Skewed left

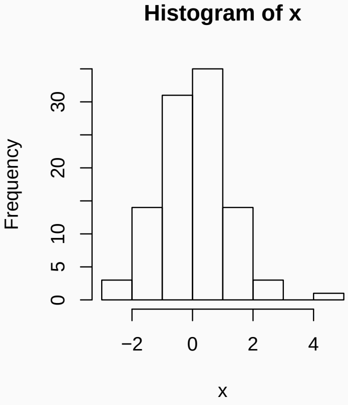


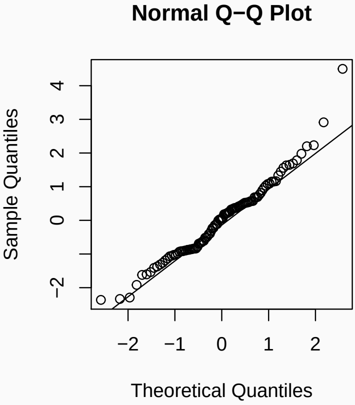
Problem: Skewed right



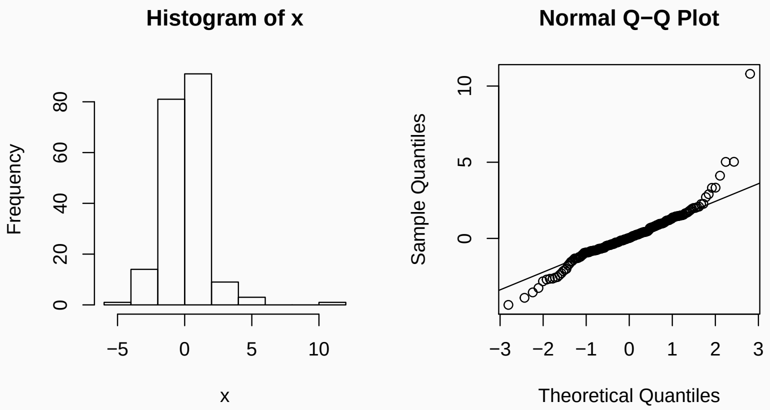


Problem: Outliers





Problem: Heavy tails



Problem: Light tails