Checking For Normality

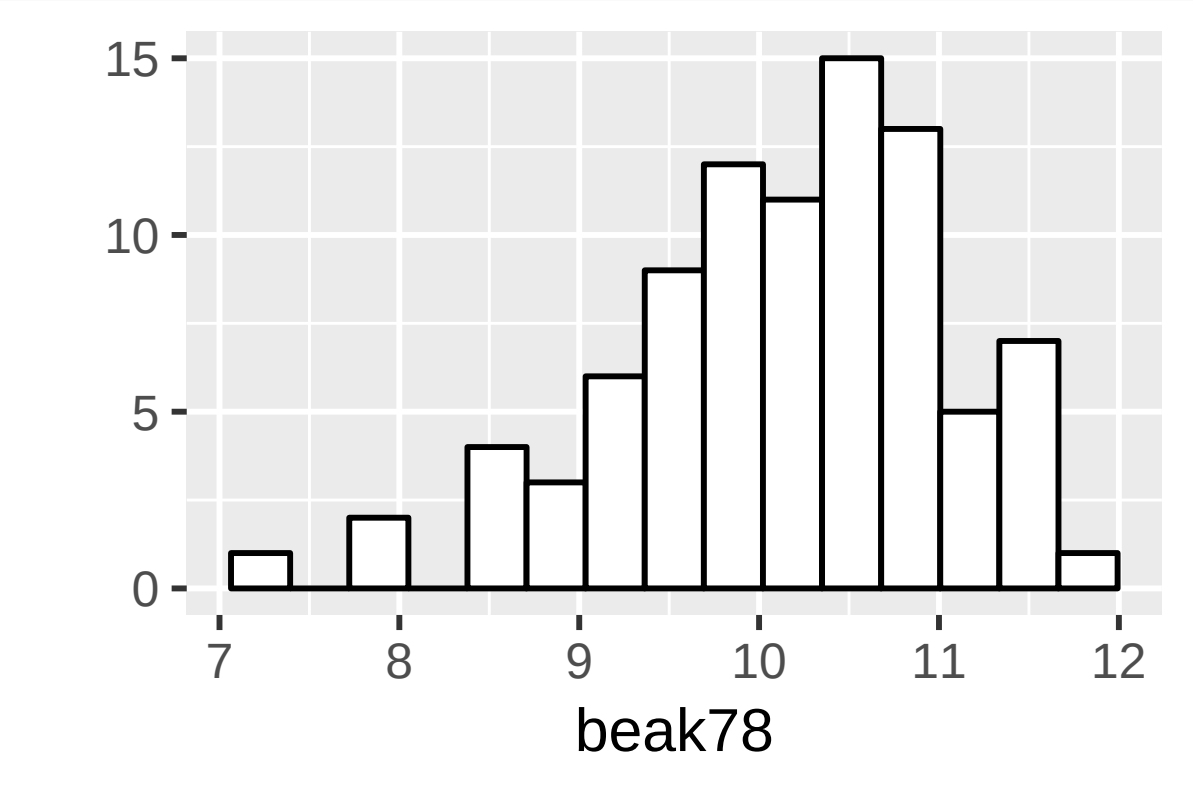
David Gerard

2018-12-07

Checking for normality

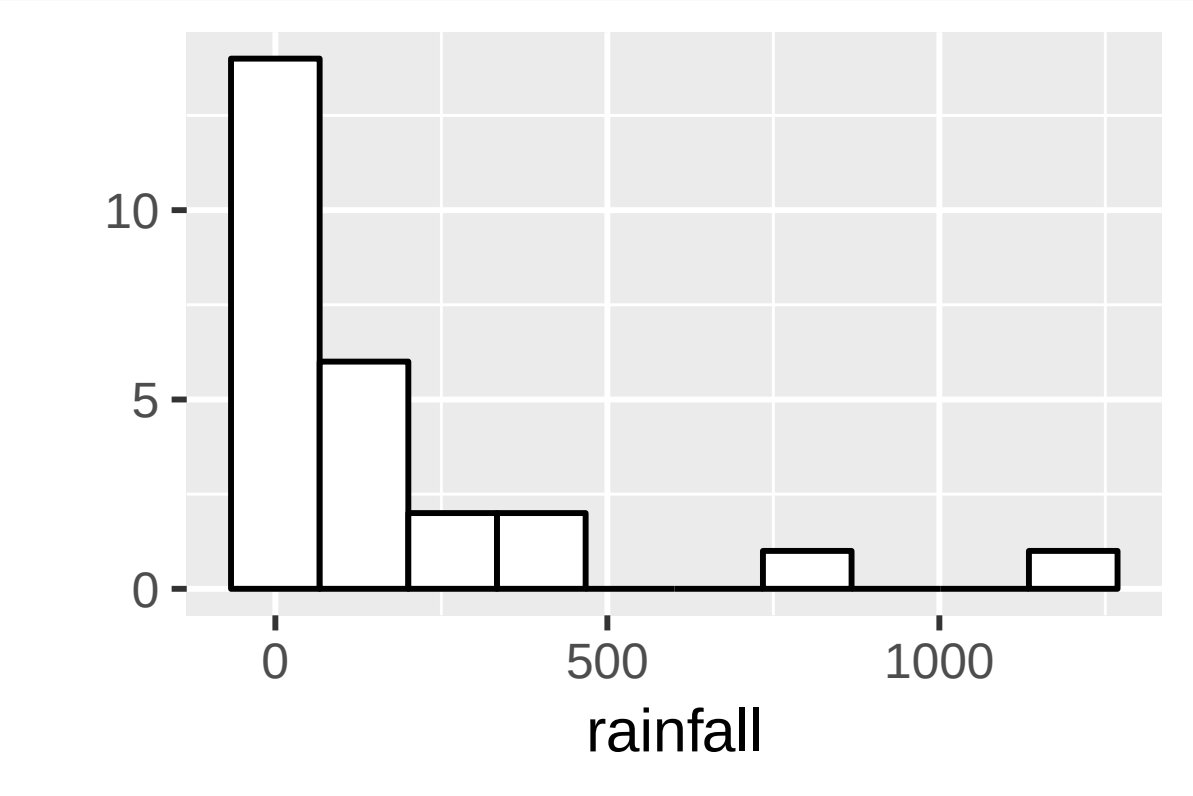
Some distributions look approximately normal

```
library(Sleuth3)
library(ggplot2)
data("case0201")
beak78 <- case0201$Depth[case0201$Year == 1978]
qplot(beak78, bins = 15, color = I("black"), fill = I("whit
```



Clearly not all distributions are normal

```
data("case0301") ## Rainfall data
rainun <- case0301$Rainfall[case0301$Treatment == "Unseeded
qplot(rainun, bins = 10, color = I("black"),
      fill = I("white"), xlab = "rainfall")
```



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

1. Idea: Is the 68-95-99.7 rule approximately correct for a given dataset?
2. More generally, do the percentiles (quantiles) of the data match with the percentiles (quantiles) of the theoretical normal distribution?
3. 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(beak78)
sigma <- sd(beak78)
qnorm(p = 0.2, mean = mu, sd = sigma)
```

```
## [1] 9.375
```

```
quantile(x = beak78, probs = 0.2)
```

```
##  20%
## 9.46
```

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

More quantiles

```
qnorm(p = 0.7, mean = mu, sd = sigma)
```

```
## [1] 10.61
```

```
quantile(x = beak78, probs = 0.7)
```

```
##  70%
## 10.6
```

More quantiles

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

```
## [1] 11.3
```

```
quantile(x = beak78, probs = 0.9)
```

```
##    90%
## 11.14
```

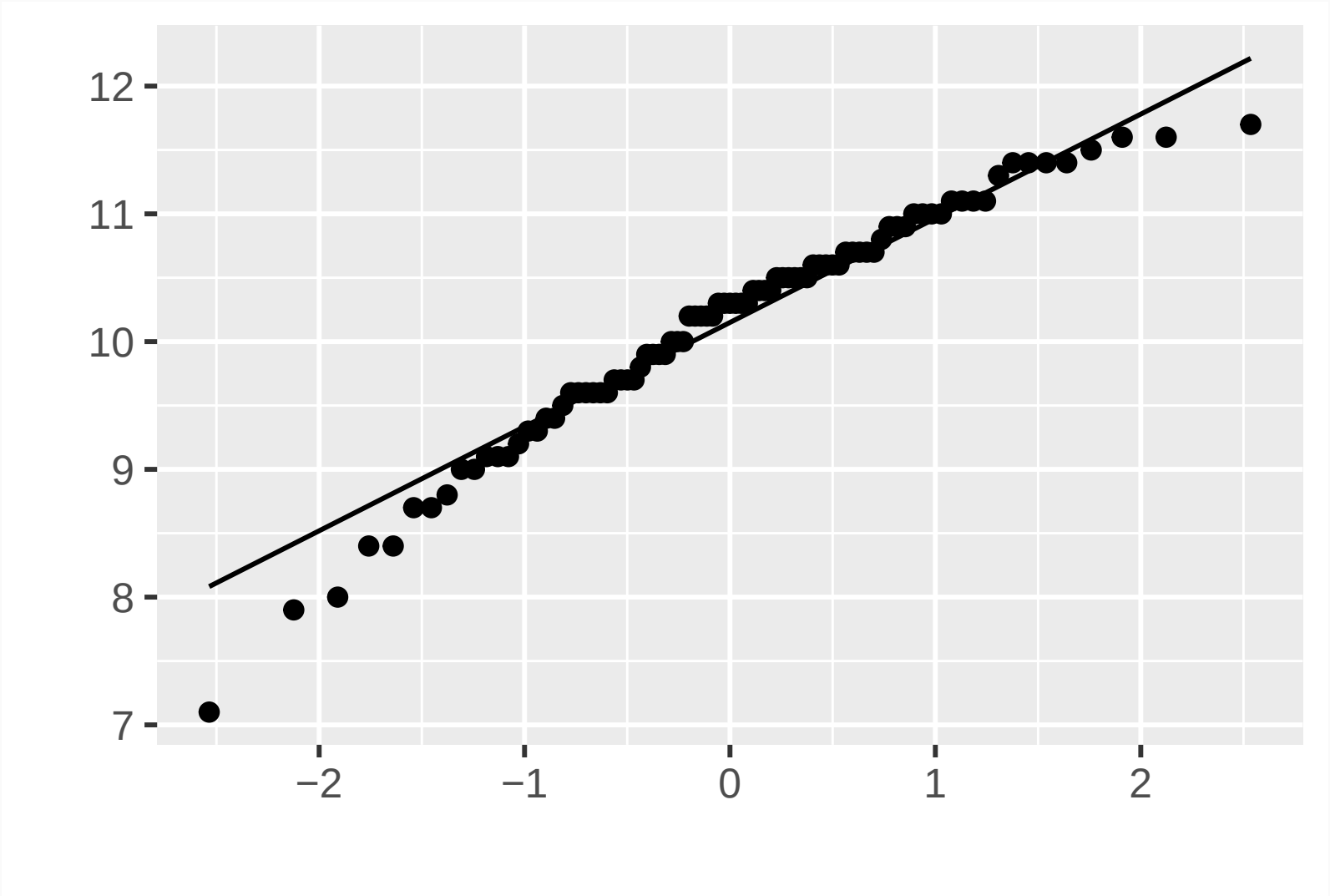
These are all pretty close!

Quantile-quantile plot

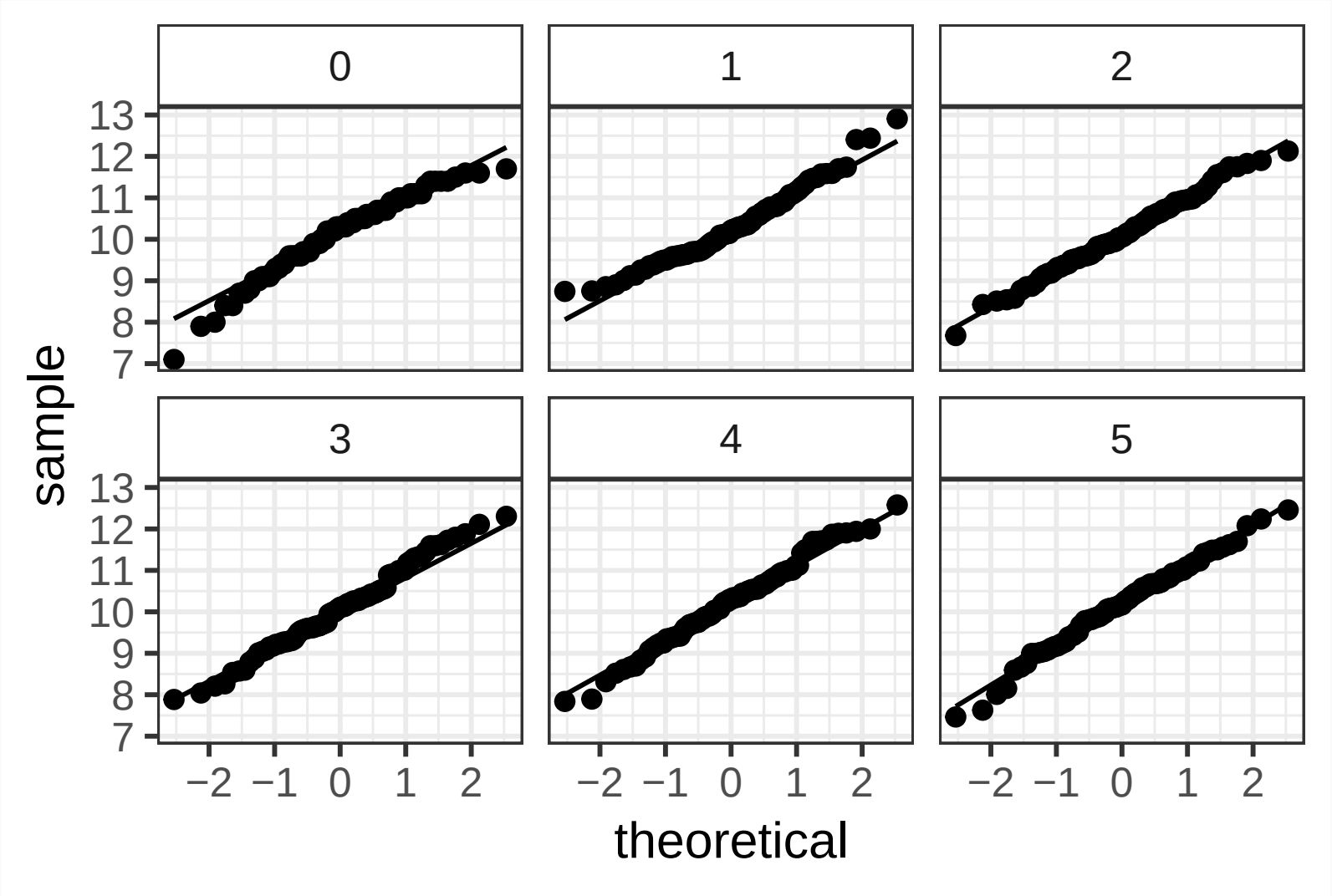
1. Plots the observed quantiles against the quantiles of a $N(\bar{x}, s^2)$ density.
2. If the points lie close to a line, then the normal approximation is approximately correct.
3. Can just plot the observed quantiles against $N(0, 1)$ and look for a straight line (more on why later).

QQplot

```
qplot(sample = beak78, geom = "qq") +
    geom_qq_line()
```

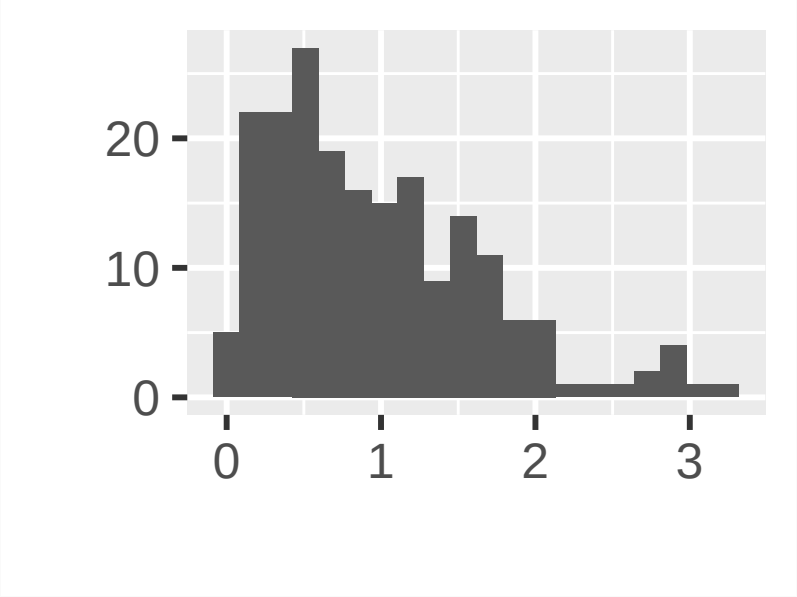


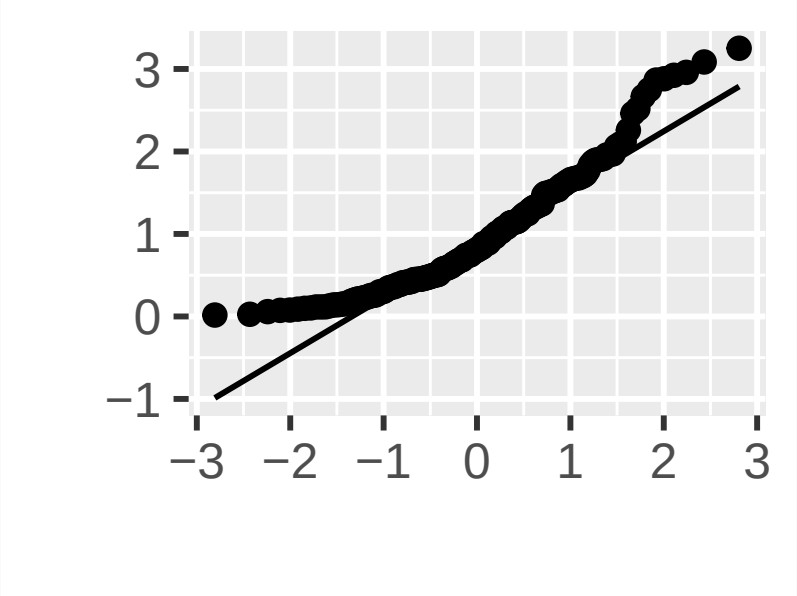
But what does a "good" qqplot look like?



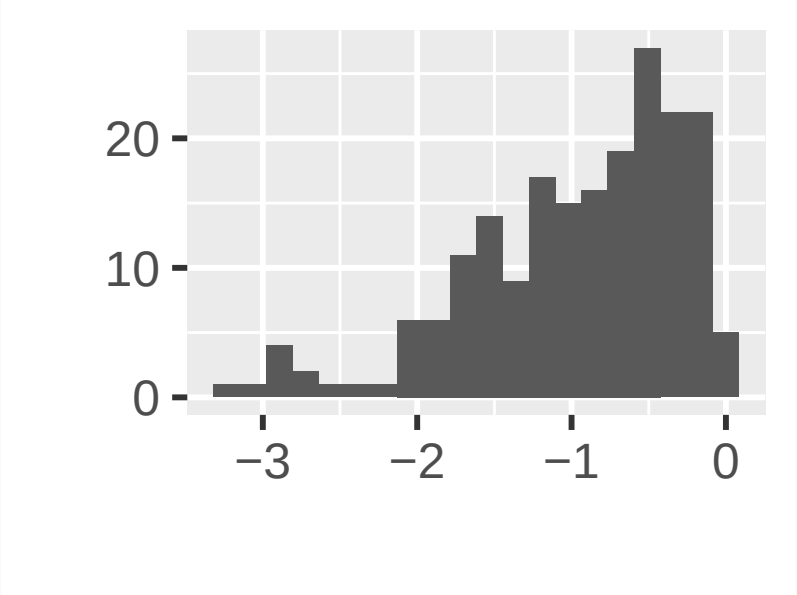
Top left is real data, rest are simulated from $N(\bar{x}, s^2)$ — maybe a little non-normal?

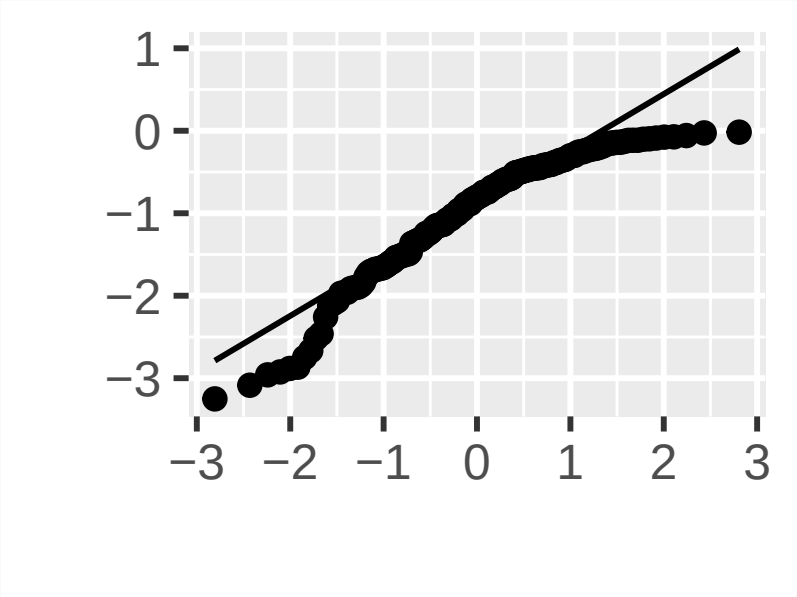
Problem: Skewed right



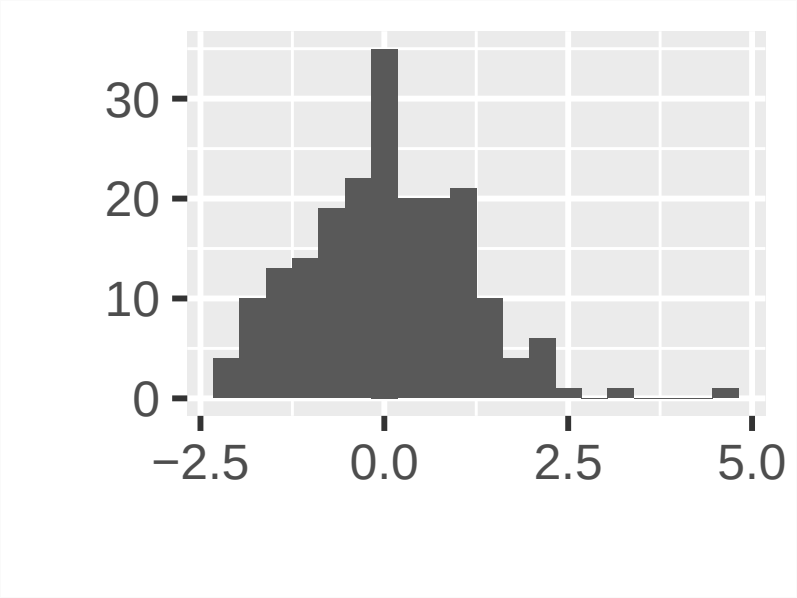


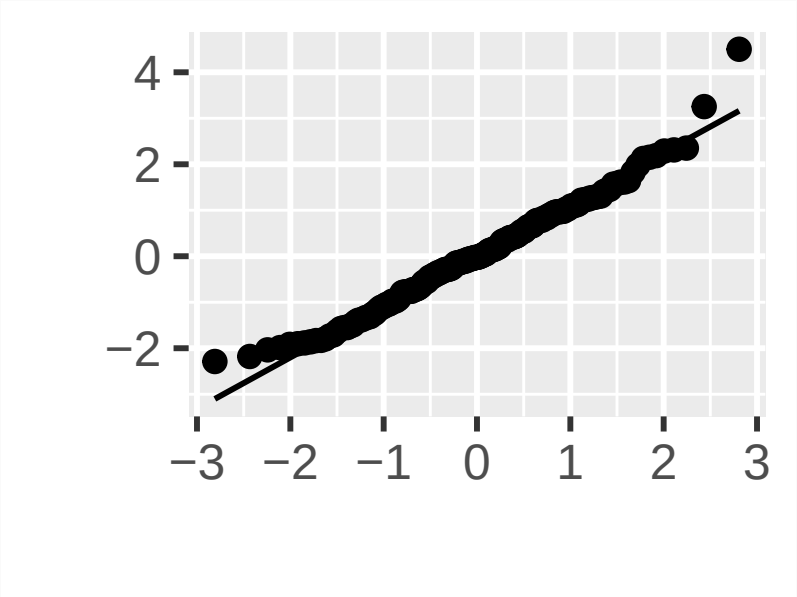
Problem: Skewed left



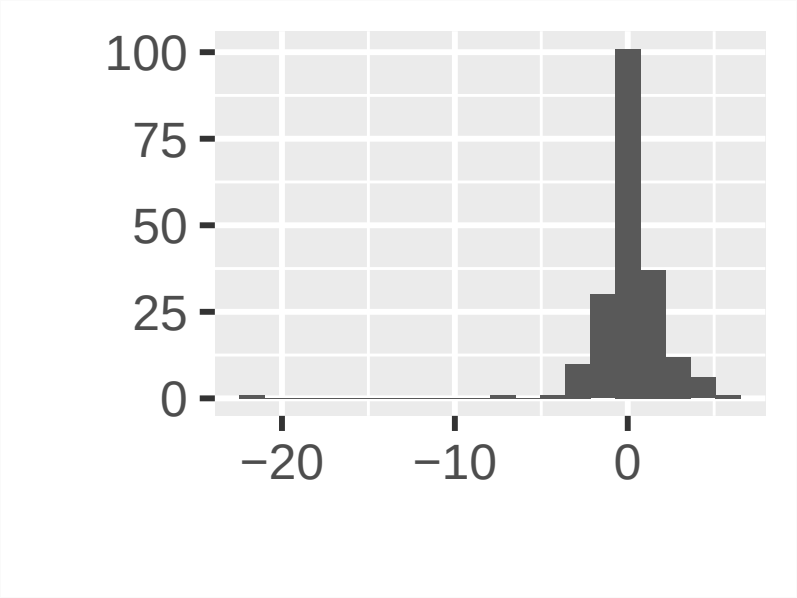


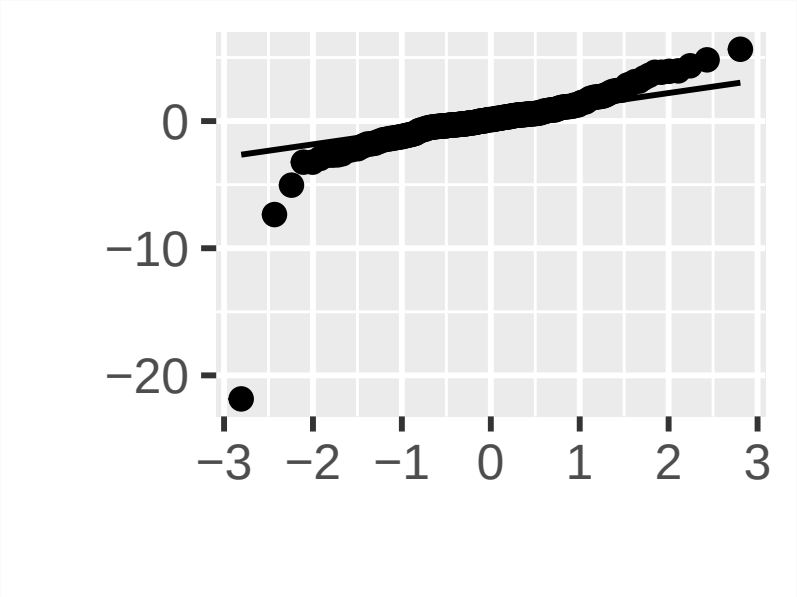
Problem: Outliers



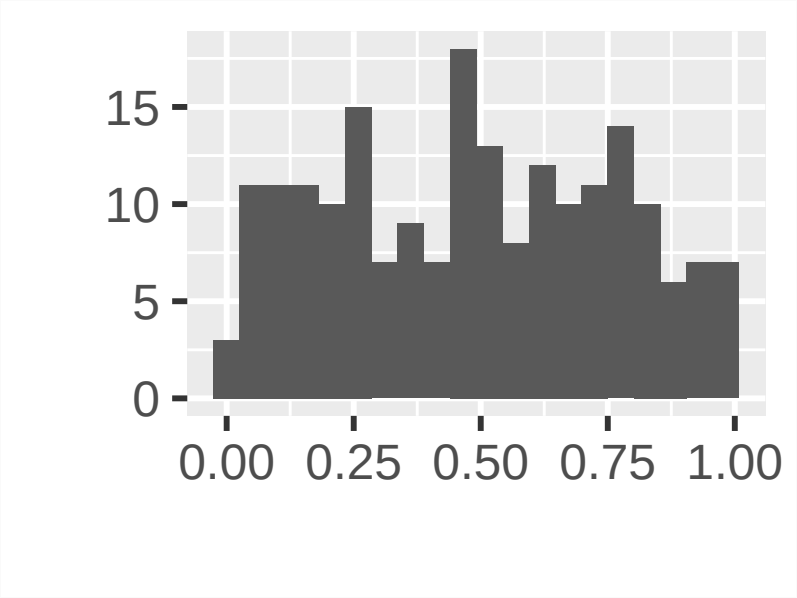


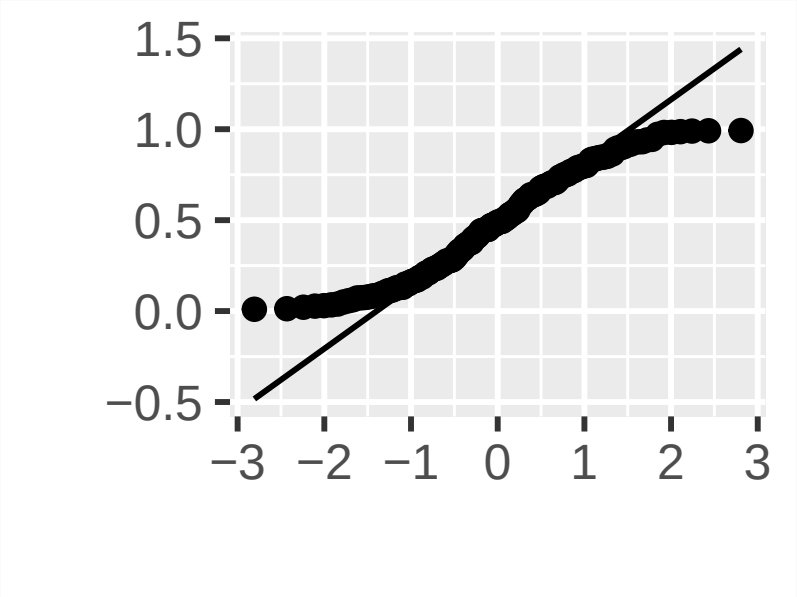
Problem: Heavy tails





Problem: Light tails





The rainfall data

```
qplot(sample = rainun, geom = "qq") +
  geom_qq_line()
```