

## Drawing graphs

## Our data

- ▶ To illustrate making graphs, we need some data.
- ▶ Data on 202 male and female athletes at the Australian Institute of Sport.
- ▶ Variables:
  - ▶ categorical: Sex of athlete, sport they play
  - ▶ quantitative: height (cm), weight (kg), lean body mass, red and white blood cell counts, haematocrit and haemoglobin (blood), ferritin concentration, body mass index, percent body fat.
- ▶ Values separated by tabs (which impacts reading in).

## Packages for this section

```
library(tidyverse)
```

## Reading data into R

- ▶ Use `read_tsv` (“tab-separated values”), like `read_csv`.
- ▶ Data in `ais.txt`:

```
my_url <- "http://ritsokiguess.site/datafiles/ais.txt"  
athletes <- read_tsv(my_url)
```

## The data (some)

```
athletes
```

```
# A tibble: 202 x 13
```

	Sex	Sport	RCC	WCC	Hc	Hg	Ferr	BMI	SSB
	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	female	Netball	4.56	13.3	42.2	13.6	20	19.2	49
2	female	Netball	4.15	6	38	12.7	59	21.2	110.
3	female	Netball	4.16	7.6	37.5	12.3	22	21.4	89
4	female	Netball	4.32	6.4	37.7	12.3	30	21.0	98.3
5	female	Netball	4.06	5.8	38.7	12.8	78	21.8	122.
6	female	Netball	4.12	6.1	36.6	11.8	21	21.4	90.4
7	female	Netball	4.17	5	37.4	12.7	109	21.5	107.
8	female	Netball	3.8	6.6	36.5	12.4	102	24.4	157.
9	female	Netball	3.96	5.5	36.3	12.4	71	22.6	101.
10	female	Netball	4.44	9.7	41.4	14.1	64	22.8	126.

```
# i 192 more rows  
# i 1 more variable: Wt <dbl>
```

# Types of graph

Depends on number and type of variables:

Categorical	Quantitative	Graph
1	0	bar chart
0	1	histogram
2	0	grouped bar charts
1	1	side-by-side boxplots
0	2	scatterplot
2	1	grouped boxplots
1	2	scatterplot with points identified by group (eg. by colour)

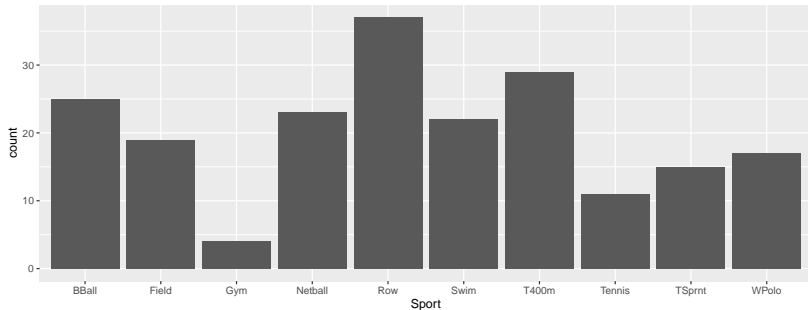
With more (categorical) variables, might want *separate plots by groups*. This is called *facetting* in R.

# ggplot

- ▶ R has a standard graphing procedure `ggplot`, that we use for all our graphs.
- ▶ Use in different ways to get precise graph we want.
- ▶ Let's start with bar chart of the sports played by the athletes.

# Bar chart

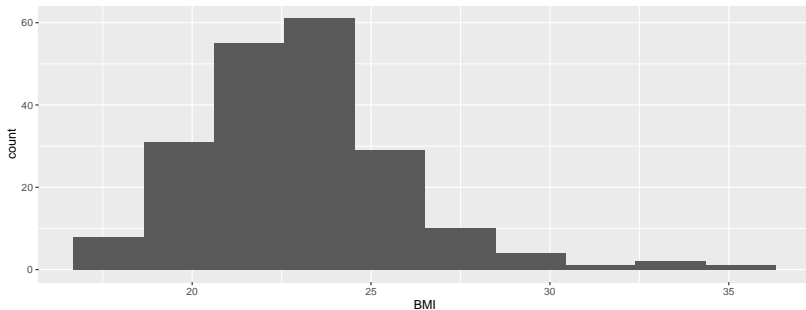
```
ggplot(athletes, aes(x = Sport)) + geom_bar()
```





# Histogram of body mass index

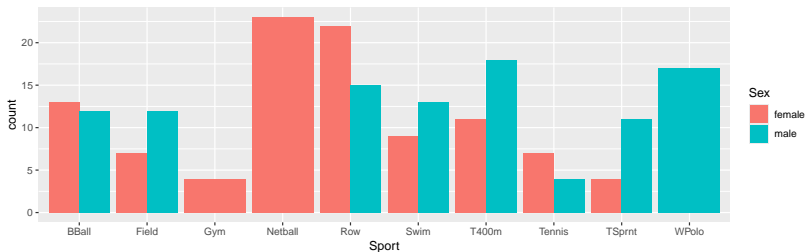
```
ggplot(athletes, aes(x = BMI)) + geom_histogram(bins = 10)
```



# Which sports are played by males and females?

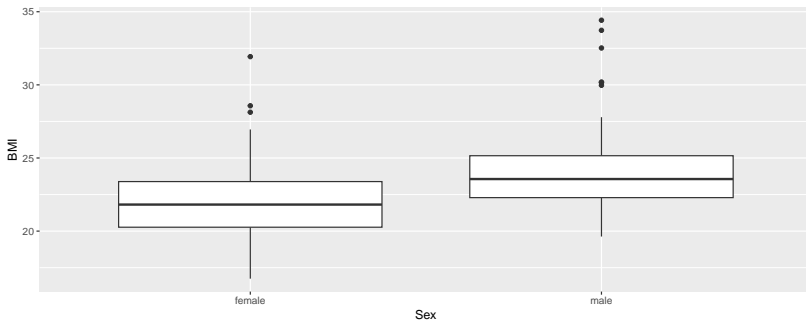
Grouped bar chart:

```
ggplot(athletes, aes(x = Sport, fill = Sex)) +  
  geom_bar(position = "dodge")
```



# BMI by gender

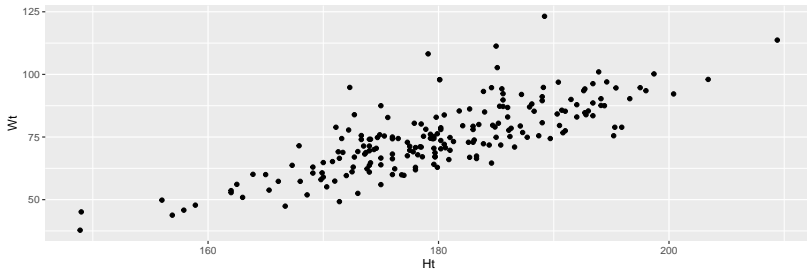
```
ggplot(athletes, aes(x = Sex, y = BMI)) + geom_boxplot()
```



# Height vs. weight

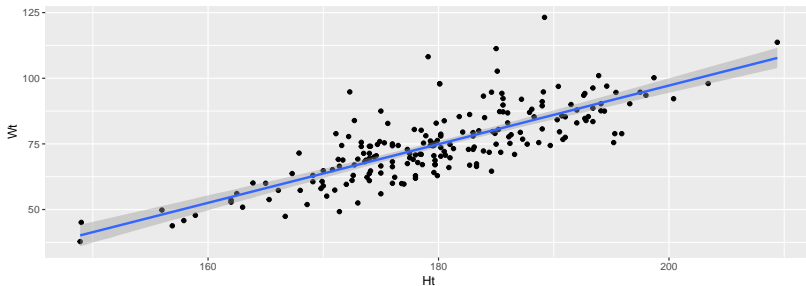
Scatterplot:

```
ggplot(athletes, aes(x = Ht, y = Wt)) + geom_point()
```



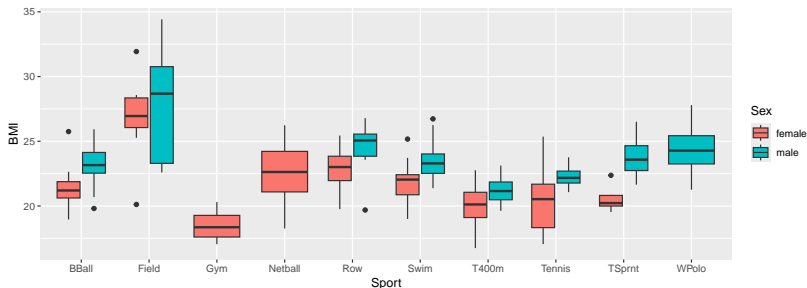
## With regression line

```
ggplot(athletes, aes(x = Ht, y = Wt)) +  
  geom_point() + geom_smooth(method = "lm")
```



# BMI by sport and gender

```
ggplot(athletes, aes(x = Sport, y = BMI, fill = Sex)) +  
  geom_boxplot()
```



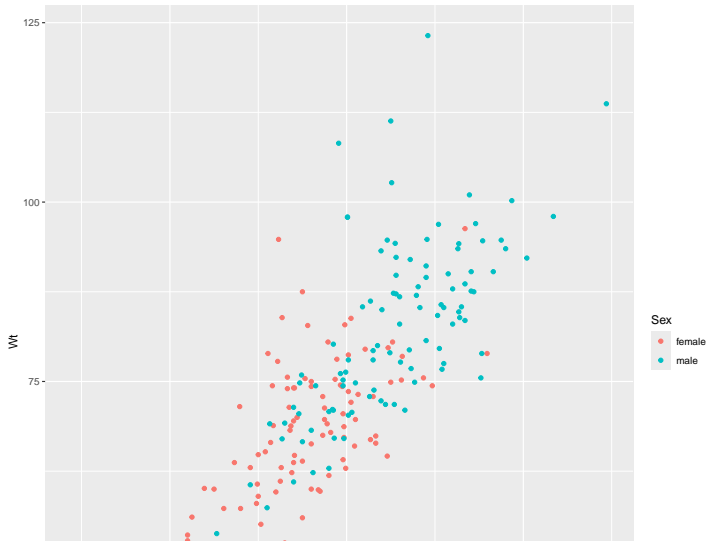
A variation that uses colour instead of fill:

```
ggplot(athletes, aes(x = Sport, y = BMI, colour = Sex)) +  
  geom_boxplot()
```



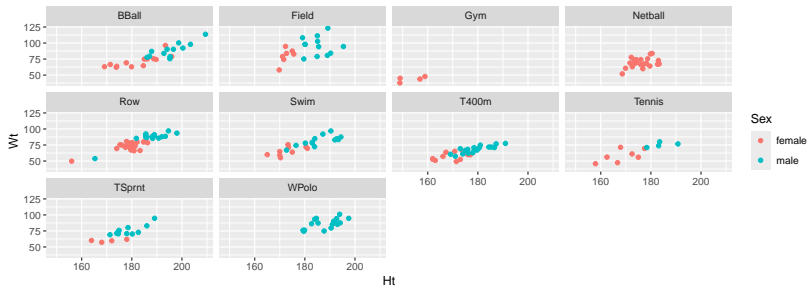
# Height and weight by gender

```
ggplot(athletes, aes(x = Ht, y = Wt, colour = Sex)) +  
  geom_point()
```



# Height by weight by gender for each sport, with facets

```
ggplot(athletes, aes(x = Ht, y = Wt, colour = Sex)) +  
  geom_point() + facet_wrap(~Sport)
```

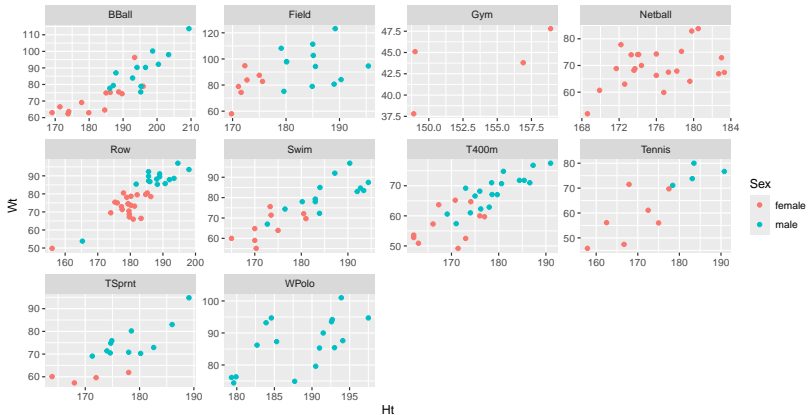




## Filling each facet

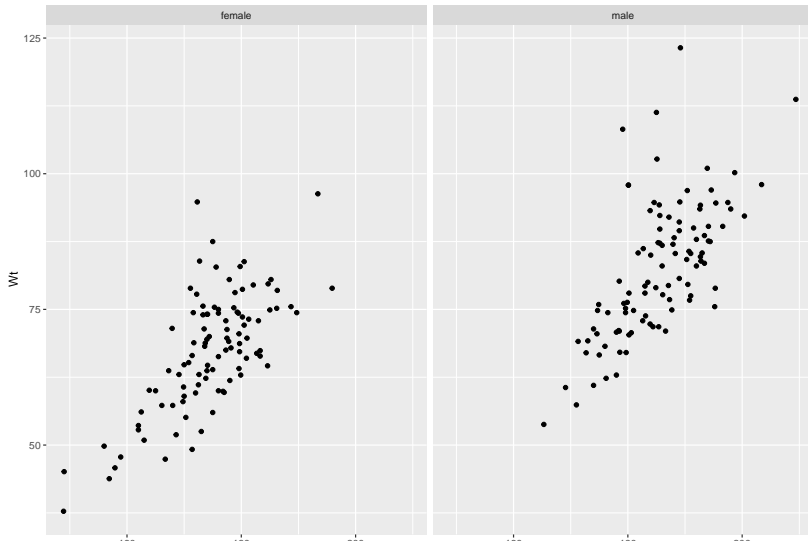
Default uses same scale for each facet. To use different scales for each facet, this:

```
ggplot(athletes, aes(x = Ht, y = Wt, colour = Sex)) +  
  geom_point() + facet_wrap(~Sport, scales = "free")
```



## Another view of height vs weight

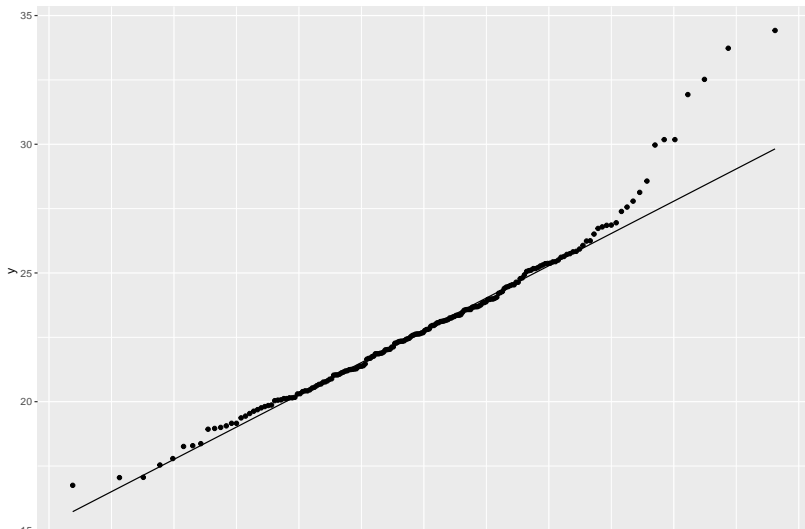
```
ggplot(athletes, aes(x = Ht, y = Wt)) +  
  geom_point() + facet_wrap(~ Sex)
```



## Normal quantile plot

For assessing whether a column has a normal distribution or not:

```
ggplot(athletes, aes(sample = BMI)) + stat_qq() + stat_qq_
```



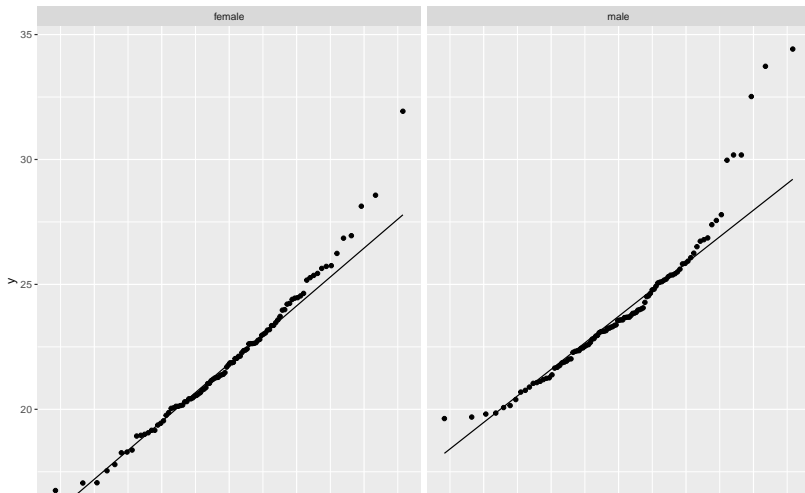
## Comments

- ▶ Data on  $y$ -axis
- ▶ on  $x$ -axis, the  $z$ -scores you would expect if normal distribution correct
- ▶ if the points follow the line, distribution is normal
- ▶ the way in which the points *don't* follow line tell you about how the distribution is not normal
- ▶ in this case, the highest values are too high (long upper tail).

# Facetting

Male and female athletes' BMI separately:

```
ggplot(athletes, aes(sample = BMI)) + stat_qq() + stat_qq_line() +  
  facet_wrap(~ Sex)
```



## Comments

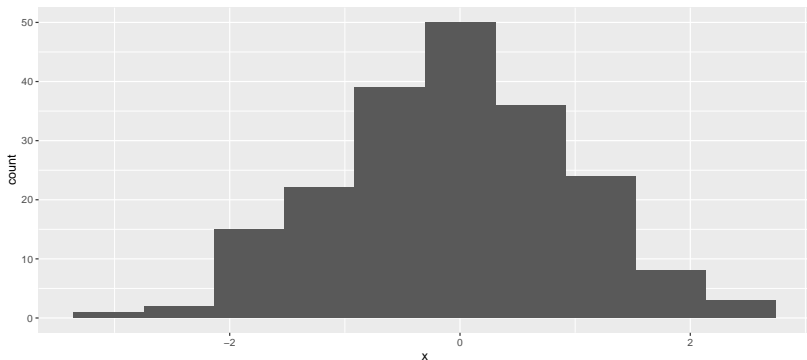
- ▶ The distribution of BMI for females is closer to normal, with only the highest few values being too high
- ▶ The distribution of BMI values for males might even be right-skewed: not only are the upper values too high, but some of the lowest ones are not low enough.

## More normal quantile plots

- ▶ How straight does a normal quantile plot have to be?
- ▶ There is randomness in real data, so even a normal quantile plot from normal data won't look perfectly straight.
- ▶ With a small sample, can look not very straight even from normal data.
- ▶ Looking for systematic departure from a straight line; random wiggles ought not to concern us.
- ▶ Look at some examples where we know the answer, so that we can see what to expect.

## Normal data, large sample

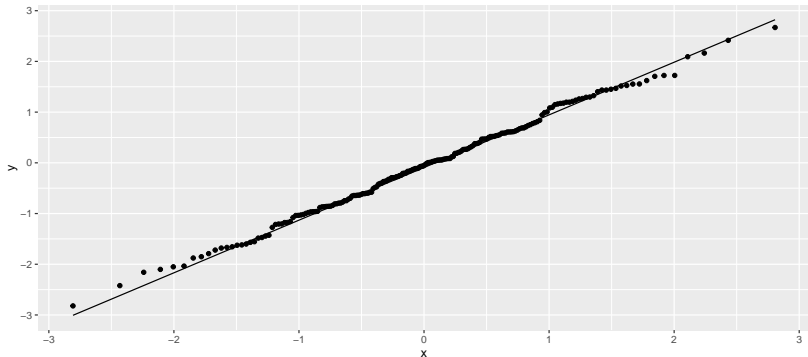
```
d <- tibble(x=rnorm(200))  
ggplot(d, aes(x=x)) + geom_histogram(bins=10)
```





# The normal quantile plot

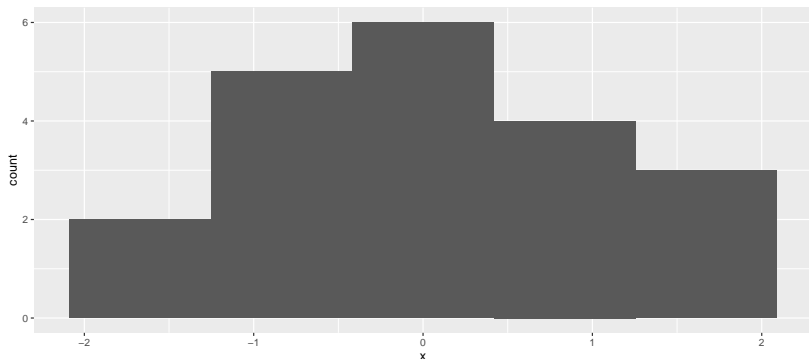
```
ggplot(d,aes(sample=x))+stat_qq()+stat_qq_line()
```



## Normal data, small sample

- Not so convincingly normal, but not obviously skewed:

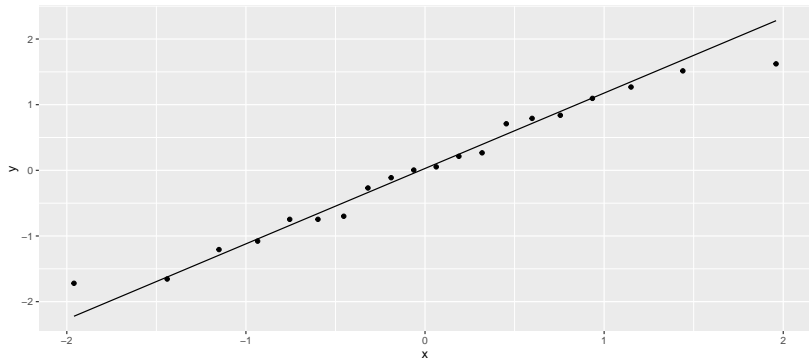
```
d <- tibble(x=rnorm(20))  
ggplot(d, aes(x=x)) + geom_histogram(bins=5)
```



# The normal quantile plot

Good, apart from the highest and lowest points being slightly off.  
I'd call this good:

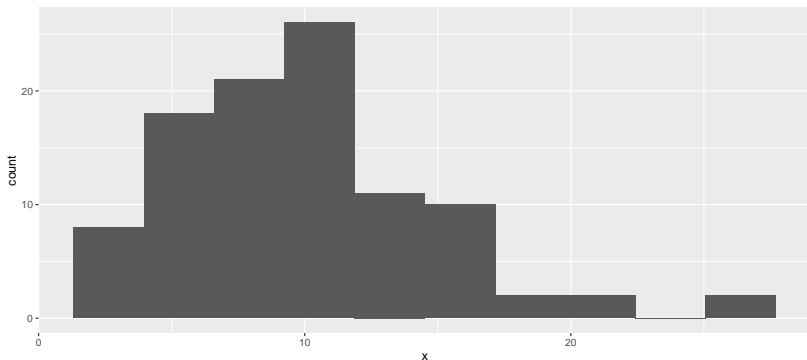
```
ggplot(d, aes(sample=x)) + stat_qq() + stat_qq_line()
```



## Chi-squared data, $df = 10$

Somewhat skewed to right:

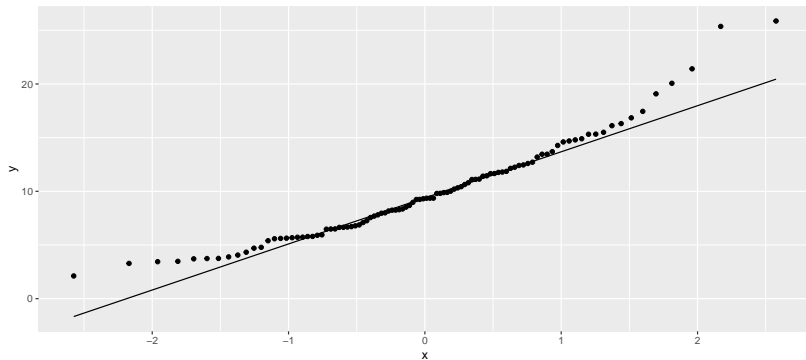
```
d <- tibble(x=rchisq(100, 10))  
ggplot(d,aes(x=x)) + geom_histogram(bins=10)
```



# The normal quantile plot

Somewhat opening-up curve:

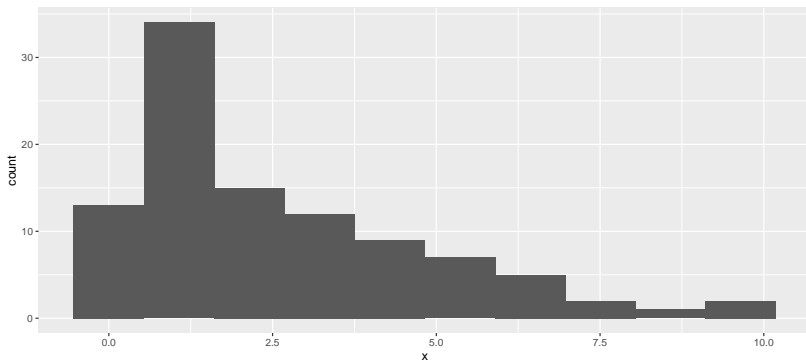
```
ggplot(d,aes(sample=x))+stat_qq()+stat_qq_line()
```



## Chi-squared data, $df = 3$

Definitely skewed to right:

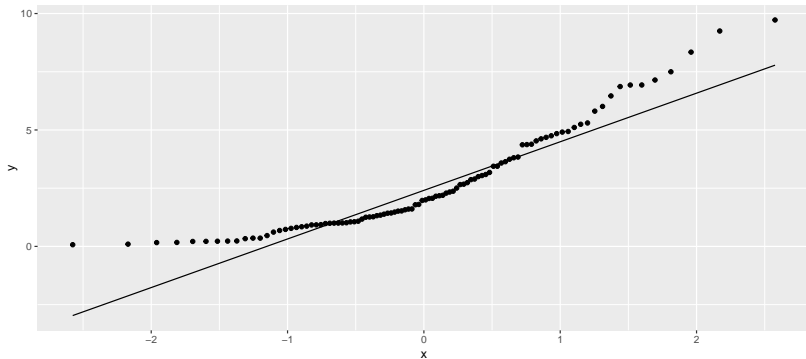
```
d <- tibble(x=rchisq(100, 3))  
ggplot(d, aes(x=x)) + geom_histogram(bins=10)
```



# The normal quantile plot

Clear upward-opening curve:

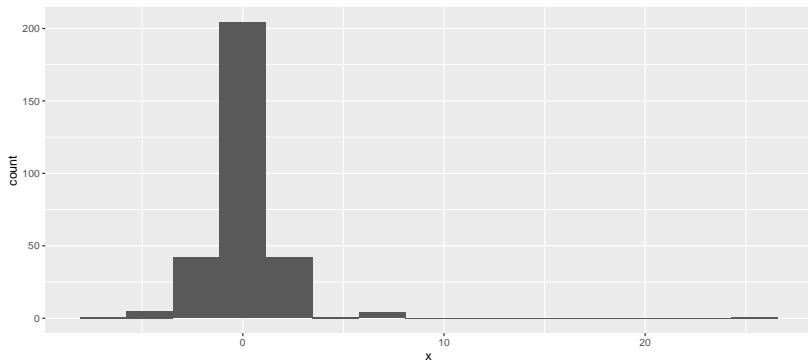
```
ggplot(d,aes(sample=x))+stat_qq()+stat_qq_line()
```



## t-distributed data, $df = 3$

Long tails (or a very sharp peak):

```
d <- tibble(x=rt(300, 3))  
ggplot(d, aes(x=x)) + geom_histogram(bins=15)
```

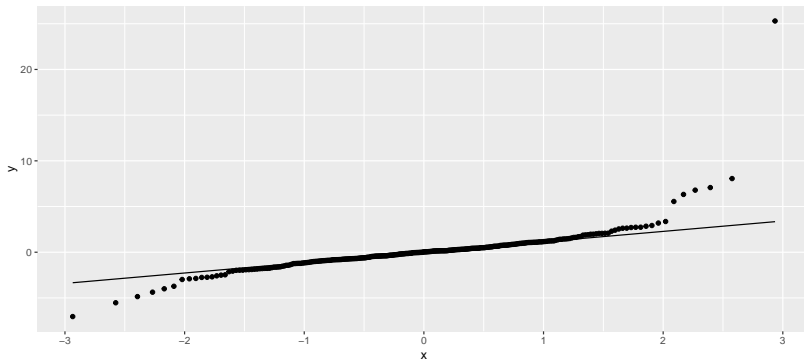




# The normal quantile plot

Low values too low and high values too high for normal.

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



# Summary

On a normal quantile plot:

- ▶ points following line (with some small wiggles): normal.
- ▶ kind of deviation from a straight line indicates kind of nonnormality:
  - ▶ a few highest point(s) too high and/or lowest too low: outliers
  - ▶ else see how points at each end off the line:

High points		
Low points	Too low	Too high
Too low	Skewed left	Long tails
Too high	Short tails	Skewed right