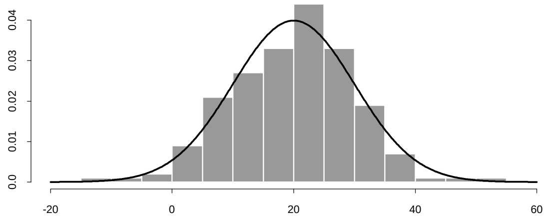
## *The Normal Distributions*

The distribution of data can often be modeled by a "normal" (also called a "Gaussian") distribution.

A probability density histogram for such data will approximately follow the *normal density curve*.

Here is an example, with 200 data points:




There is a normal density curve for every mean ($\mu$) and standard deviation ($\sigma$). The curve above is for $\mu = 20$ and $\sigma = 10$.

## *Mathematics of the Normal Distributions*

The equation of the normal density curve for a variable $x$ with mean $\mu$ and standard deviation $\sigma$ is

$$\textit{density} \;=\; \frac{1}{\sigma\sqrt{2\pi}}\, e^{-(x-\mu)^2/2\sigma^2}$$

These curves all look the same, except for centre and scale.

Why the complicated formula? It can be derived mathematically when the data are the sum of many small, independent influences.

The normal distribution with mean $\mu$ and standard deviation $\sigma$ (ie, variance $\sigma^2$) is sometimes denoted by $N(\mu, \sigma^2)$.

## *Normal Distributions and Reality*

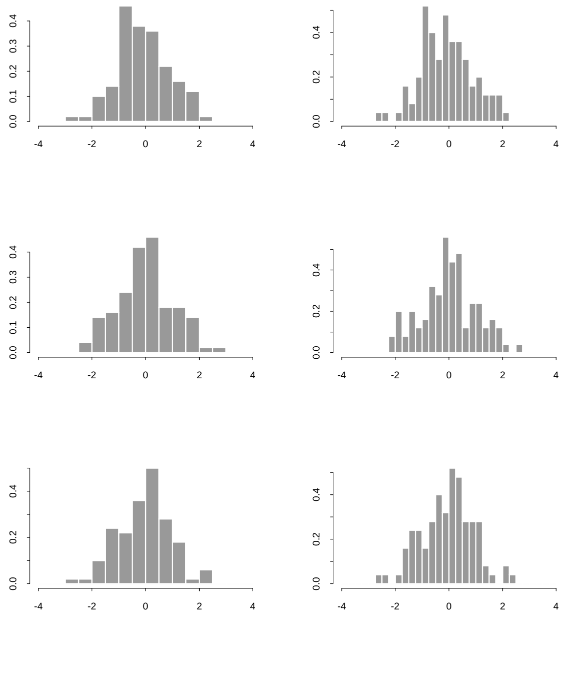
In real life, data seldom has a perfectly normal distribution.

But a normal distribution may be an adequate *model*.

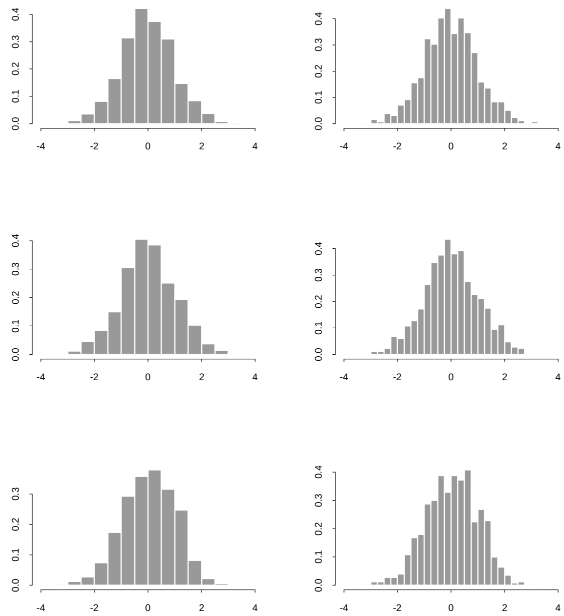
Even if the source of the data is modeled very well by a normal distribution, the actual histogram will show chance variation. What it looks like will also depend on the bin width.

Data simulated with a computer show this...

## *Histograms of N(0,1) Data, 100 Points*



## *Histograms of N(0,1) Data, 1000 Points*



## *The 68 − 95 − 99.7 Rule*

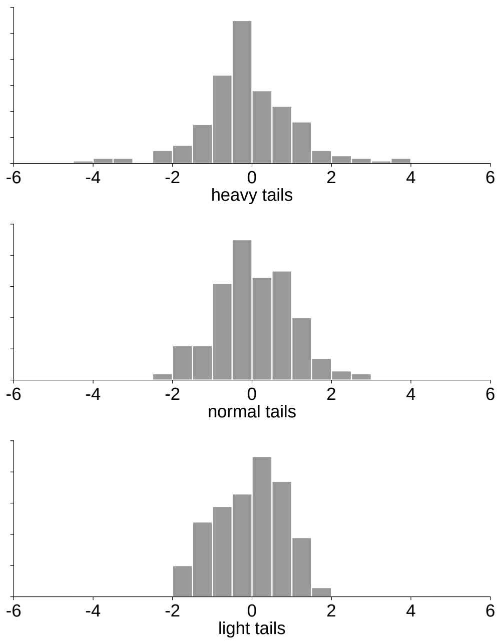
If data comes from the $N(\mu, \sigma^2)$ distribution,

- 68% is within $[\mu - \sigma,\ \mu + \sigma]$
- 95% is within $[\mu - 2\sigma,\ \mu + 2\sigma]$
- 99.7% is within $[\mu - 3\sigma,\ \mu + 3\sigma]$

Even though in theory a normal distribution extends out to $\pm\infty$, data is very unlikely to be more than about three standard deviations from the mean.

**Caution!** This assumes the data is *really* normally distributed. The distribution of data that is just approximately normal may have heavier "tails" than a normal, and so have a higher proportion of extreme observations.

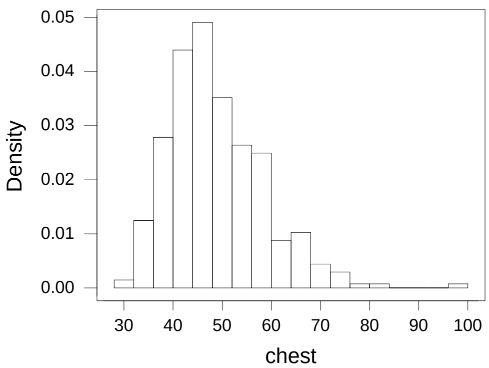
## *Heavy, Normal, and Light Tails*




## *Skewed Distributions*

Some distributions have "extreme" values in only one direction — they are said to be *skewed* in that direction ("left" or "right").

Here is a histogram of the measure of chest injury in the crashtest dummies data set:

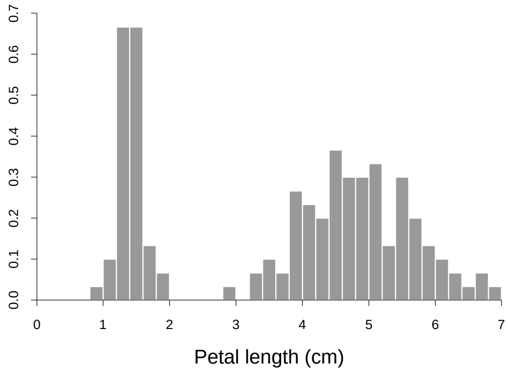



This data is skewed to the right.

## *Bimodal Distributions*

Not all data has a distribution with only one "hump" — called a *mode* in statistics.

Here is a histogram of the lengths of the petals in 150 iris plants:




There are two modes because these plants belong to at least two different species.

## *Exploring Relationships of Variables*

A big part of statistics is concerned with *relationships* between variables. For example:

- How does whether a person smokes relate to whether they get lung cancer?
- How does a person's age relate to how many magazines they buy?
- How does the temperature at which a pot is fired relate to whether or not it cracks?
- How does whether a car has a manual or automatic transmission relate to its fuel efficiency?

Related variables may be categorical or quantitative.

We are often interested in *cause and effect* relationships, but not always.

## *Relationships of Categorical Variables*

The relationship between two categorical variables can be displayed in a table.

For the data on people on the *Titanic*:

```
Rows: class     Columns: survived

           no      yes      All

1st       122      203      325
2nd       167      118      285
3rd       528      178      706
crew      673      212      885

All      1490      711     2201

    Cell Contents -- Count

Rows: class     Columns: survived

           no      yes      All

1st     37.54    62.46   100.00
2nd     58.60    41.40   100.00
3rd     74.79    25.21   100.00
crew    76.05    23.95   100.00

All     67.70    32.30   100.00

    Cell Contents -- % of Row
```

## *Relationships of Categorical and Quantitative Variables*

Side-by-side boxplots can be used to explore the relationship of one categorical and one quantitative variable.

From the crashtest data set (the full version), here are boxplots of the measure of head injury for vehicles in different size classes: