Most everybody wears shoes most all the time. Sometimes professors write the word “shoes” in the oddest places. Including here. What can we say about them in the context of the binomial and normal distributions?

1 binomial distribution

If a random variable X records the number of successes in n identical trials (iid) then X has the binomial distribution with parameters n and p, the success probability.

For the binomial distribution, the mean is $\mu = np$, the variance $\sigma^2 = npq$ and distribution

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

The R software has four functions which compute with the binomial distribution:

- `dbinom(k,size,prob)` will return the distribution, $P(X = k)$.
- `pbinom(k,size,prob)` will give all values k or less: $P(X \le k)$.
- `qbinom(p,size,prob)` will find the quantile, which k so that $P(X \le k) = p$.
- `rbinom(n,size,prob)` will find a random sample of size n.

For example, it is determined that 30% of shoes sold are laceless. If a shoe store sells 50 shoes, what is the probability that exactly 20 are laceless? What about 12 or fewer sold are laceless. Finally, 13 or more are laceless.

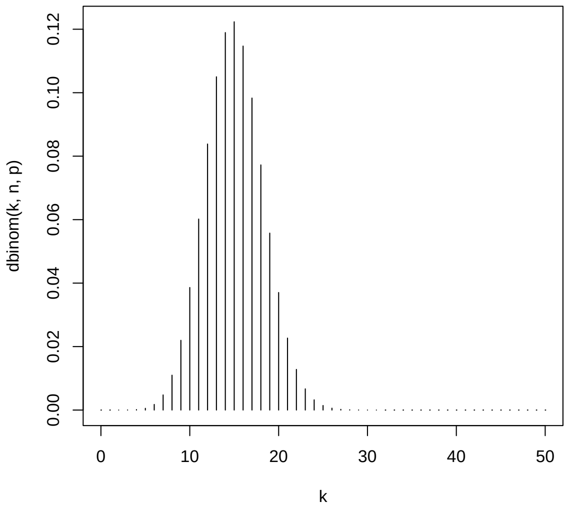
To answer this, we let X be the number of laceless shoes sold. We assume that X is binomial with $n = 50$ and $p = 0.3$. Then we can use `dbinom()` for the first and `pbinom()` for the second

```
> dbinom(20, size = 50, prob = 0.3)

[1] 0.03704

> pbinom(12, size = 50, prob = 0.3)

[1] 0.2229

> 1 - pbinom(12, size = 50, prob = 0.3)

[1] 0.7771
```

The last one, $P(X \ge 13)$ is answered with $1 - P(X \le 12)$. Why?

We see that the probability of exactly 20 is small. Are all the probabilities small, or is this one small? We can plot a graph of the distribution to find out. We draw the plot with vertical lines using `type="h"`. The height of the spike above k is the probability that X is equal to k.

```
> n = 50
> p = 0.3
> k = 0:n
> plot(k, dbinom(k, n, p), type = "h")
```



The plot looks "bell-shaped" despite the fact that it is a discrete distribution.

We see the value for $k = 20$ is about 1/2 the largest, so relatively speaking it isn't small. It is just that the probability is spread out over lots of points so there is not much for any one point.

Question 1: A shoe manufacturer knows that a sports celebrity endorsement is effective about 20% of the time. If they have 12 celebrities, what is the probability that 4 of them will be successful? What is the probability that 2 or fewer are successful?

Question 2: Some shoe manufacturers market themselves to niche markets such as wide feet or narrow feet. Suppose, 15% of the population has wide feet. In a group of 100 people chosen from this population what is the expected number with wide feet? What is the probability 10 or fewer do? What is the probability, 20 or more have wide feet?

Question 3: It is expected that out of every 10 customers to Foot Locker, 2 will buy sneakers. If 30 customers come in, what is the probability 5 or fewer buy sneakers?

Question 4: Make plots of the distribution of X when X is binomial with $n = 10$ and $p = .1$ and when $n = 50$ and $p = .1$. Are either "bell-shaped?"

2 normal distribution

The normal distribution also has four functions for it

- `dnorm(x,mean=mu,sd=sigma)` will give the density for a mean `mu` and standard deviation `sigma`
- `pnorm(x,mean=mu,sd=sigma)` will give $P(X \leq x)$
- `qnorm(p,mean=mu,sd=sigma)` will find b so that $P(X \leq b) = p$
- `rnorm(x,mean=mu,sd=sigma)` will find a random sample of size `x`

For example, if shoe costs are approximately normal with mean \$50 and standard deviation \$15, then what is the probability a shoes is less than \$20? What about more than \$75?

```
> pnorm(20, mean = 50, sd = 15)

[1] 0.02275

> 1 - pnorm(75, mean = 50, sd = 15)

[1] 0.04779
```

2.1 quantiles

In Japan a study was done that found the length of an adult, womans foot has a normal distribution with $\mu = 22.8$ cm and $\sigma = 0.89$ cm. If this is so, what percent of people have feet less than 25cm? What is the cutoff for the top and bottom 5% of foot lengths?

The first question is answered as a probability

```
> pnorm(25, mean = 22.8, sd = 0.89)

[1] 0.9933
```

The second asks for a quantile, so we use `qnorm()`

```
> qnorm(0.05, mean = 22.8, sd = 0.89)

[1] 21.34

> qnorm(0.95, mean = 22.8, sd = 0.89)

[1] 24.26
```

We could do this simultaneously with

```
> qnorm(c(0.05, 0.95), mean = 22.8, sd = 0.89)

[1] 21.34 24.26
```

Question 5: For the foot study, if a shoe manufacturer only made shoes for the middle 50% what would be the range of lengths in cms. made?

Question 6: The same study found the foot circumference is normally distributed with $\mu = 22.6$ and $\sigma = 0.99$.

1. Freehand sketch a plot of the density of this population. Identify 22 on it.
2. What is the probability that a randomly chosen foot has a circumference less than 22cm?
3. If only the bottom 95% of foot circumferences are important to a manufacture, what would be the cutoff for circumference?
4. A company specializes in shoes for large feet. If "large" means the top 10% what would be the smallest circumference shoe they would make?

2.2 The empirical rule

The empirical rule for the normal distribution says that 68%, 95% and 99.7% of the data falls within 1, 2 and 3 standard deviations from the mean.

Let's verify these numbers for a random sample of size 100 with $\mu = 22.6$ and $\sigma = 0.99$.

```
> mu = 22.6
> sigma = 0.99
> x = rnorm(100, mean = mu, sd = sigma)
> z = (x - mu)/sigma
> sum(-1 < z & z < 1)

[1] 68

> sum(-2 < z & z < 2)

[1] 95

> sum(abs(z) < 3)

[1] 99
```

Random numbers were generated with `rnorm()` and then turned into z-scores. For these, we asked how many were in the proper range. The last way of asking uses `abs()` for the absolute value.

The numbers fit pretty well.

Question 7: Do the same simulation above. You should get different, but similar answers. Do you?

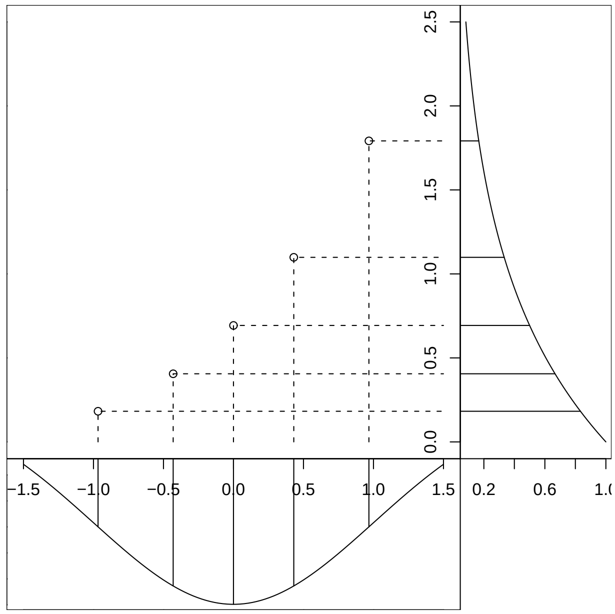
Question 8: You can generate different random numbers with the `rexp()` command. For example, this will find z-scores for a different type of data set

```
> x = rexp(100)
> z = (x - 1)/1
```

For a random sample, make a histogram of the data, and see if the empirical rule fits

2.3 quantile-quantile plots

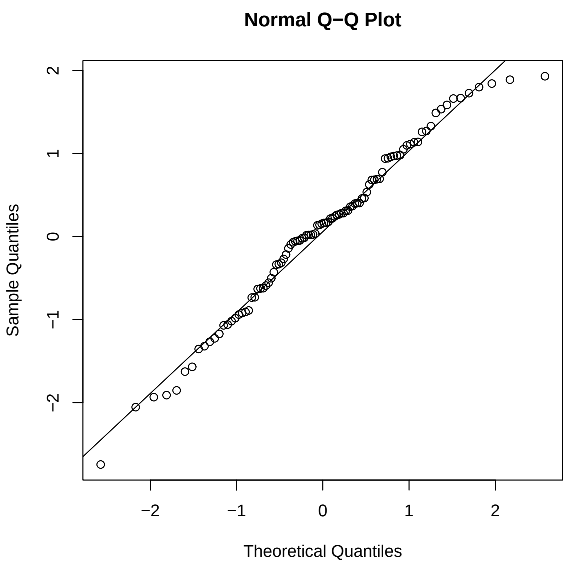
To test if a population is normal, a histogram and the empirical rule can be used as a rough guide. However, a more precise plot made by comparing quantiles is better.



A quantile-normal plot plots the quantiles of a dataset against the quantiles of the normal distribution. If the data set is normally distributed, these points will fall near a straight line. If not, the shape will curve or deviate somehow.

These plots are made with the function `qqnorm()` (two q's).

```
> x = rnorm(100)
> qqnorm(x)
> qqline(x)
```



For the normal data set, the points line up pretty much on a line.

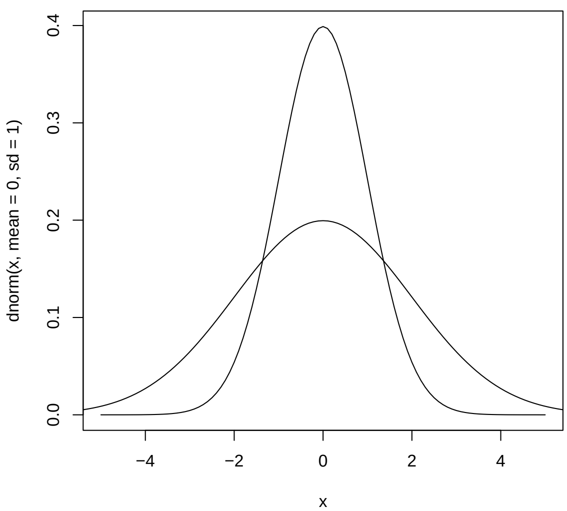
Question 9: Repeat the above with your sample. Do the points have exactly the same shape or generally the same shape?

Question 10: Make another random sample using `rexp()`. Then make a quantile-normal plot. Do these points line up along a line?

2.4 Graphs of the normal distribution

The function `curve()` will plot a function of x for you. When drawing a new plot, you add the limits as arguments. When adding to an existing plot, you use the argument `add=T`. To plot the standard normal density from -5 to 5 and a normal density with mean 0 and standard deviation 2 is done with. done with

```
> curve(dnorm(x, mean = 0, sd = 1), -5, 5)
> curve(dnorm(x, mean = 0, sd = 2), add = T)
```



Question 11: On the same graphic, make plots of a standard normal, a normal with mean 1 and variance 1 and a normal with mean 0 and variance 4.

Question 12: A binomial random variable X, with $n = 30$ and $p = 1/3$, will have mean $\mu = np = 10$ and standard deviation $\sigma = \sqrt{np(1-p)} = \sqrt{20/3}$.

Make a spike plot of the distribution of X. Layer on top a normal density with the same μ and σ as above. Are the two graphs similar? How so?

How might you use the normal density to answer a question like $P(X \leq 12)$?