STAT 515 Lec 7

The Normal Distribution

Karl Gregory

The Normal distribution

Figure 1: Carl Friedrich Gauß (1777 in Braunschweig – 1855 in Göttingen)

The Normal distribution is sometimes called the Gaussian distribution after him who first suggested it: Carl Friedrich Gauß.

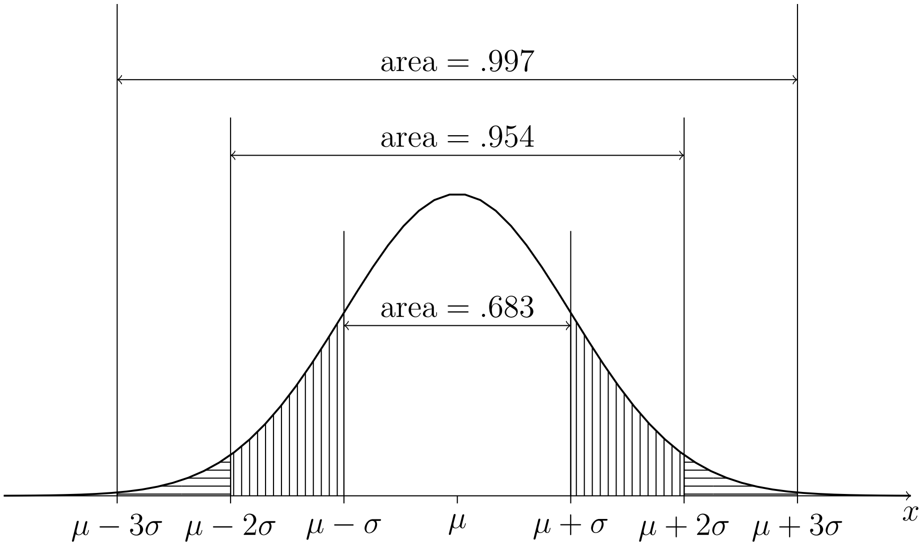
Definition 1: Probability density function of the Normal distribution

The Normal distribution with mean μ and variance σ^2 has probability density function

$$f(x) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sigma}\exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right].$$

If a random variable X has this distribution, we write $X \sim \text{Normal}(\mu, \sigma^2)$.

The pdf of the Normal distribution looks like this:



Result 1: Mean and variance of Normal distribution

If $X \sim \text{Normal}(\mu, \sigma^2)$, then $\mathbb{E}X = \mu$ and $\operatorname{Var} X = \sigma^2$.

Note that the probability density function is centered at μ and that its spread is controlled by σ.

When we use the Normal distribution, we often think in terms of *the number of standard deviations from the mean*, where by standard deviation, we refer to σ, the square root of the variance σ^2. We see that if X has a Normal distribution, then it will lie within 1 standard deviation of its mean 68% of the time; it will lie within 2 standard deviations of its mean about 95% of the time; and it will lie within 3 standard deviations of its mean about 98% of the time.

We sometimes assume that the values in a population are Normally distributed in the sense that if we were to take a census of the population—that is if we were to observe every single value—and build a histogram of all the values, the histogram would take on the shape of the Normal probability density function. When this is true of a population, we may regard each sampled value X from the population as a random variable having a Normal probability distribution. We can then use the Normal probability density function to compute probabilities concerning X.

Exercise. Suppose the growth in height of Loblolly pines from age three to age five is Normally distributed with a mean of 6 feet and a standard deviation of 0.5 feet. Let X be the growth from age three to age five of a randomly selected Loblolly pine.

1. What is the probability that the growth in height of the tree from age three to five is

in between 5.5 and 6.5 feet?

Answer: A growth of 5.5 feet is 1 standard deviation below the mean and 6.5 is one standard deviation above the mean. Since the growth in height of Loblolly pines from age three to age five is Normally distributed, we can say $P(5.5 \leq X \leq 6.5) = .683$.

2. What is the proportion of Loblolly pines that grow more than 7 feet in height from age three to age five?

   Answer: Since the growth in height of Loblolly pines from age three to five has mean 6 and standard deviation 0.5, 7 feet is 2 standard deviations above the mean. Since the growth in height of Loblolly pines from age three to five is Normally distributed, the proportion of Loblolly pines whose growth in height from age three to five is at least 2 standard deviations *away* from the mean is $1 - .954 = .046$. However, we are only interested in the proportion whose whose growth in height from age three to five is at least 2 standard deviations *above* the mean. Since the Normal probability density function is symmetric, it places the same "probability mass" to the right of 2 standard deviations above the mean as it places to the left of 2 standard deviations below the mean (refer to the picture). Therefore, we must divide the proportion 0.046 by 2, getting $P(X > 7) = .023$.

Sometimes the "population" is not so easy to define, as in the next example about miles-per-gallon on each tank of gas. In this case, the population is rather a hypothetical one: if we were to fill the tank many many times and record the gas mileage many many times, than we might assume that these gas mileages would produce a histogram conforming in shape to the Normal probability density function.

Exercise. Suppose your mpg follows a Normal distribution with mean 28 and standard deviation 2. What is an interval within which your mpg should lie about 95% of the time?

Answer: If the mpg is Normal, then it should lie within 2 standard deviations of its mean about 95% of the time. Therefore, the mpg should lie in the interval

$$(28 - 2 \times 2, 28 + 2 \times 2) = (24, 32)$$

about 95% of the time.

The standard Normal distribution

Recall that for a continuous random variable with probability density function f, we may compute, for any $a \leq b$ the probability of the event $a \leq X \leq b$ as

$$\begin{aligned} P(a \leq X \leq b) &= \text{Area between } f \text{ and the horizontal axis on the interval } [a, b] \\ &= \int_a^b f(x)dx \quad \text{(for those who have taken some calculus).} \end{aligned}$$

If $X \sim \text{Normal}(\mu, \sigma^2)$, then

$$\begin{aligned} P(a \leq X \leq b) &= \text{Area beneath Normal}(\mu, \sigma^2) \text{ pdf on the interval } [a, b] \\ &= \int_a^b \frac{1}{\sqrt{2\pi}} \frac{1}{\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx. \end{aligned}$$

Even if you are a calculus master, you cannot evaluate this integral by hand exactly; it must be approximated, and the calculations are very tedious. Fancy calculators can compute (that is they can calculate very close approximations to) these integrals. The software R has built-in functions to do this. But people have been using the Normal or Gaussian distribution since before these tools were available. How?

In early days, the idea was to consider a *standard Normal* random variable having mean 0 and standard deviation 1 and to publish a table containing many already-computed integrals of the function

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right],$$

which is the Normal probability density function with $\mu = 0$ and $\sigma = 1$. Then, practitioners could shift and scale their random variable X to give it a mean of 0 and a standard deviation of 1 and use the tables of already-computed integrals of this function to get their probabilities of interest.

We find that if we subtract from X the mean μ and then divide by the standard deviation σ, the result is a random variable, call it Z, which has the *standard Normal* distribution—that is, the Normal distribution with mean 0 and standard deviation 1. Precisely:

Result 2: Transformation to the Standard Normal distribution

If $X \sim \text{Normal}(\mu, \sigma^2)$, then

$$Z = \frac{X - \mu}{\sigma} \sim \text{Normal}(0, 1).$$

A probability concerning X can thus be obtained by looking up the corresponding probability concerning $Z = (X - \mu)/\sigma$ in the table of integrals of the standard Normal probability density function.

Note that if we want to move from the "standardized" or Z world back into the "unstandardized" or X world, we do the inverse transformation

$$X = \mu + \sigma Z.$$

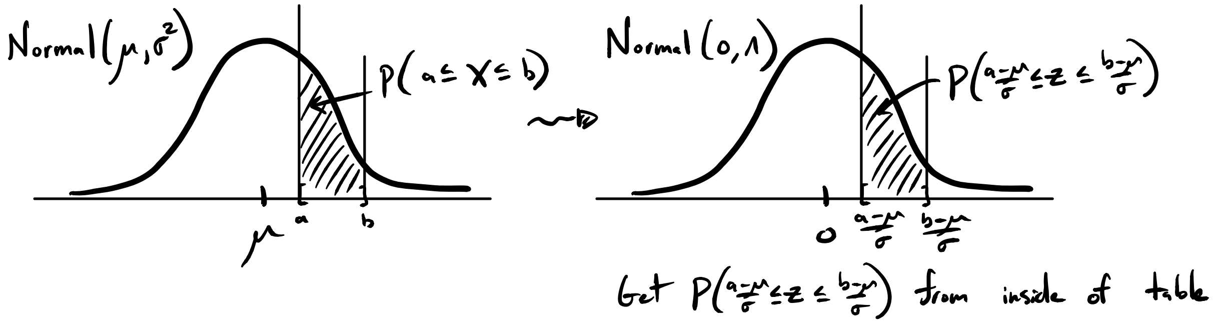
Important: We interpret Z as the number of standard deviations X is from the mean.

Finding probabilities for Normal random variables

If X has the Normal distribution with mean μ and standard deviation σ, we may find $P(a \leq X \leq b)$ as follows:

1. Transform a and b from the X world to the Z world (the number-of-standard-deviations world) by
$$a \mapsto \frac{a-\mu}{\sigma} \quad \text{and} \quad b \mapsto \frac{b-\mu}{\sigma},$$
since
$$\begin{aligned} P(a \leq X \leq b) &= P\left(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) \\ &= P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right). \end{aligned}$$
Note that if $a = \pm\infty$ then $(a-\mu)/\sigma = \pm\infty$ and if $b = \pm\infty$ then $(b-\mu)/\sigma = \pm\infty$.
2. Look up probabilities for Z on the z-table.

We start in the X world and go to the Z world to get a probability. It looks like this:


Normal(μ, σ^2)
$P(a \leq X \leq b)$
μ
a
b
Normal$(0,1)$
$P(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma})$
0
$\frac{a-\mu}{\sigma}$
$\frac{b-\mu}{\sigma}$
Get $P(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma})$ from inside of table

Finding quantiles of Normal random variables

A *quantile* is a percentile: Recall that if your height is at the 90th percentile, then 90% of people are your same height or shorter. We refer to the 90th percentile as the 0.90 quantile.

To be precise, the θth quantile q_θ of a continuous random variable X is the value such that

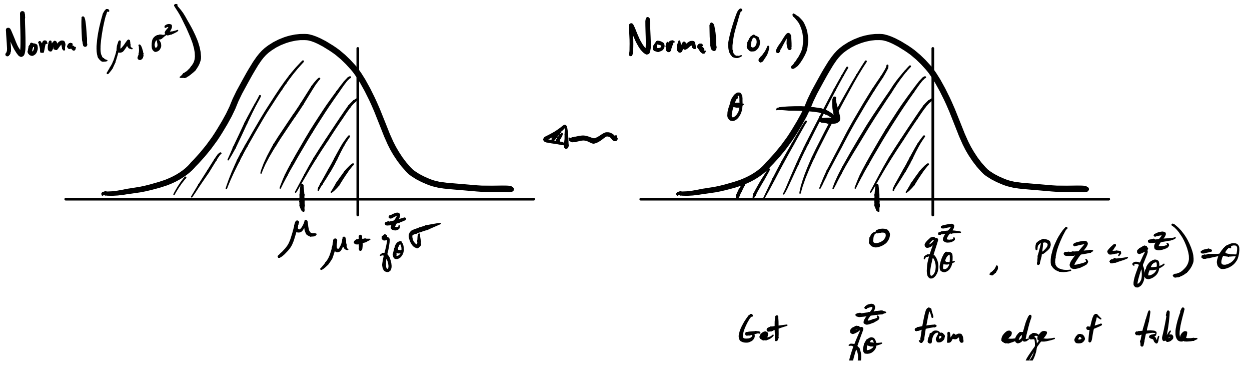
$$P(X \leq q_\theta) = \theta.$$

We may find the θth quantile of a Normal random variable X with mean μ and standard deviation σ as follows:

1. Find the θth quantile q_θ^Z of the standard Normal distribution. That is, find the value q_θ^Z such that
$$P(Z \leq q_\theta^Z) = \theta.$$

2. Transform this value from the Z-world into the X-world by
$$q_\theta^Z \mapsto \mu + q_\theta^Z \sigma$$

We start in the Z world and go to the X world to get a quantile. It looks like this:


Normal(μ, σ^2)
μ
$\mu + q_\theta^z \sigma$
Normal$(0,1)$
θ
0
q_θ^z , $P(Z \leq q_\theta^z) = \theta$
Get q_θ^z from edge of table

Exercise. Suppose your mpg follows a Normal distribution with mean 28 and standard deviation 2. What is the 0.90 quantile of your mpg distribution?

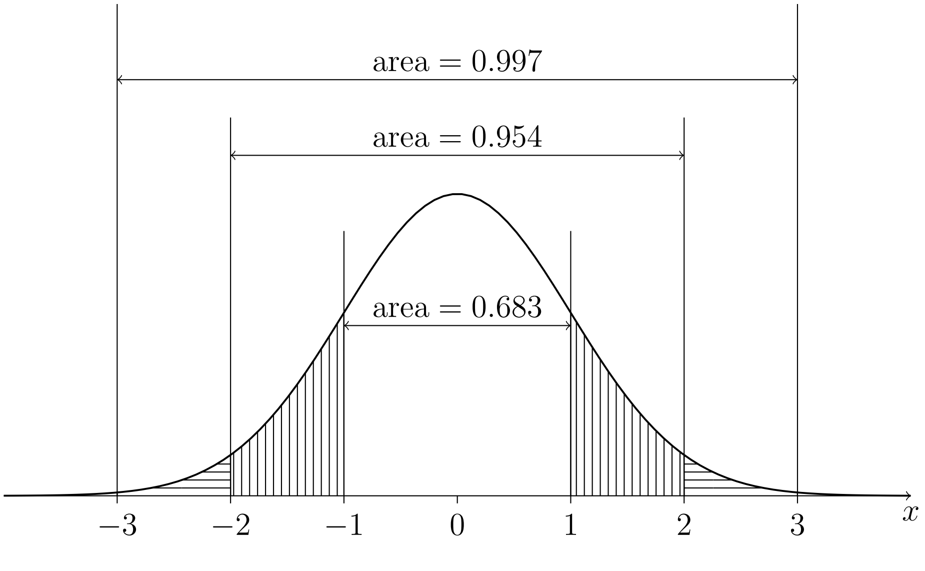
Answer: Since the mpg is Normal, and since the 0.90 quantile of the standard Normal distribution is 1.282, we know that the 0.90 quantile of the mpg is 1.282 standard deviations above the mean. This is $28 + 1.282(2) = 30.564$. We expect the mpg to be less than or equal to 30.654 90% of the time.

Some plots of the standard normal distribution

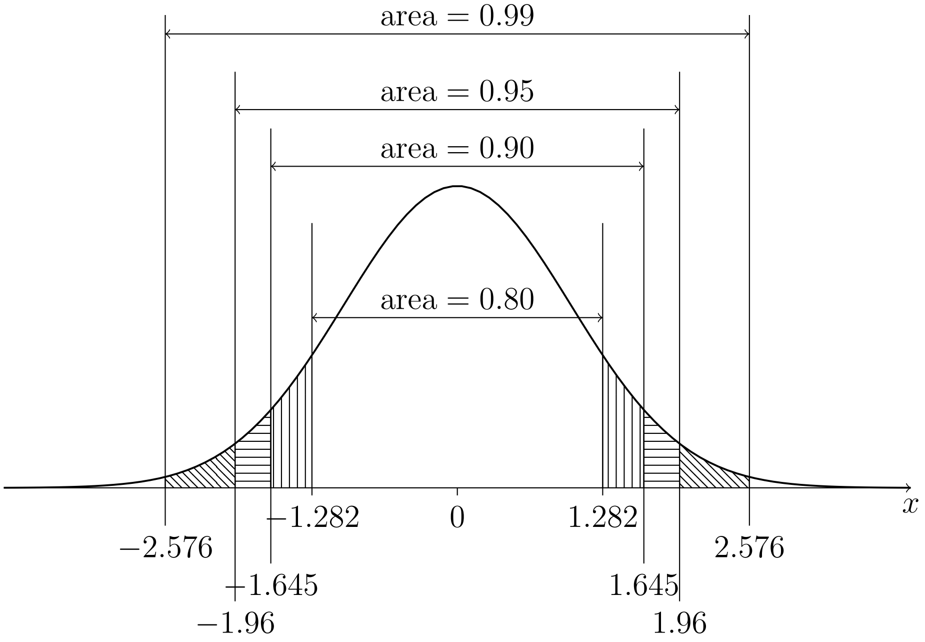
The *standard Normal probability density function*

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{z^2}{2}\right],$$

looks like this:



The next plot shows some commonly used quantiles of the standard Normal distribution.



From the above plot we can see that the 0.005, 0.025, 0.05, 0.95, 0.975, and 0.995 quantiles of the standard Normal distribution are −2.576, −1.96, −1.645, 1.645, 1.96, and 2.576, respectively.

Getting standard Normal quantiles/probabilities

It is not possible (even if you have taken Calculus), to compute areas under the curve

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{x^2}{2}\right]$$

on paper. We need to use a computer or a table which lists a bunch of areas under the curve for us which someone else has already computed using a computer. Alas! On tests, we will need to use the paper table method!

The easy way (with R)

If the random variable Z has the standard Normal distribution, then probabilities of the form $P(Z \leq z)$ for any z can be computed with the R function `pnorm()`. For example, we can find the probabilities $P(Z \leq 1.282)$, $P(Z \leq 1.645)$, and $P(Z \leq 1.96)$ with

```
> pnorm(1.282)
[1] 0.9000787
> pnorm(1.645)
[1] 0.9500151
> pnorm(1.96)
[1] 0.9750021
> pnorm(1.645)
[1] 0.9500151
```

Conversely, if we wish to get a quantile, we can use the R function `qnorm()`, for example

```
> qnorm(.025)
[1] -1.959964
> qnorm(.05)
[1] -1.644854
> qnorm(.95)
[1] 1.644854
> qnorm(.975)
[1] 1.959964
```

Note that these numbers correspond to those in the figures in the previous sections.

The cumbersome way (with tables)

Discuss in class. Find table on page 816 of MCS13.

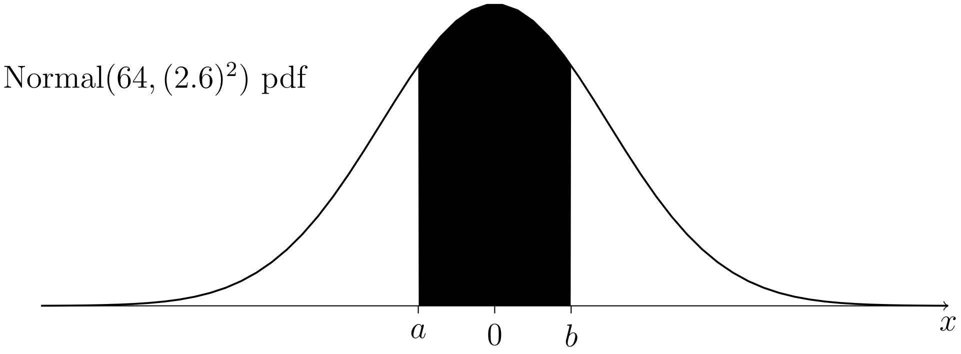
Exercise. Suppose the heights of American females who are over 20 years old follow a Normal distribution with mean 64 inches and standard deviation 2.6 inches.

1. Find an interval within which 50% of the heights lie.

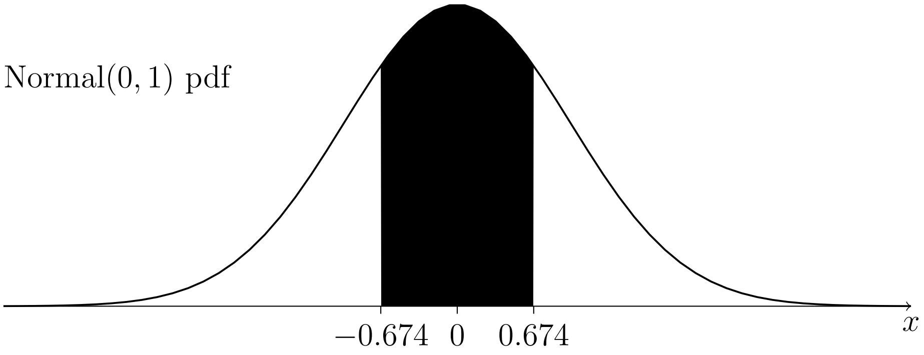
   Answer: We can find such an interval immediately by noting that 50% of the heights lie above the mean, since the Normal distribution is symmetric. So 50% of the heights lie in the interval $(64, \infty)$.

2. Find an interval centered at the mean within which 50% of the heights lie.

   Answer: The plot below depicts the desired interval, over which 50% of the area under the curve lies:



The corresponding interval for the standard Normal distribution can be found by looking at the z-table. It is depicted below:



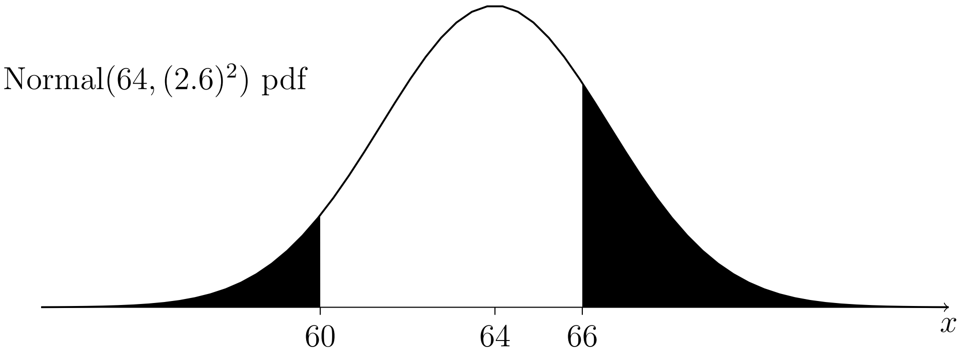
So we get a and b from the transformations

$$a = 64 + 2.6(-0.674) = 62.2476 \quad \text{and} \quad b = 64 + 2.6(0.674) = 65.7524.$$

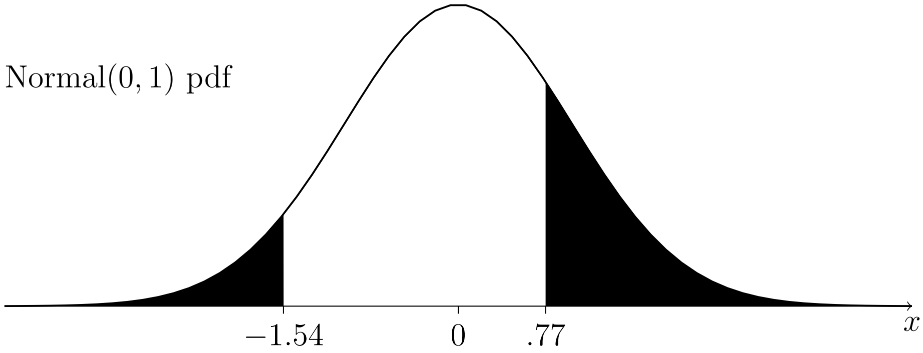
So the interval is $(62.2476, 65.7524)$.

3. Let X be the height of a randomly selected American female over the age of 20. Find $P(X < 60 \text{ or } X > 66)$.

   Answer: The probability is equal to the area of the shaded region in the figure below:



   The total area of the shaded regions in the above figure is equal to that of the shaded regions under the standard Normal distribution in the figure below:



   where we have obtained the values -1.54 and 0.77 via

$$-1.54 = (60 - 64)/2.6 \quad \text{and} \quad 0.77 = (66 - 64)/2.6.$$

   Using the z-table on pg. 816 of MCS13, the area on the left is given by $0.5 - 0.4382 = 0.0618$ and the area on the right is given by $0.5 - 0.2794 = 0.2206$. So we have

$$P(X < 60 \text{ or } X > 66) = P(X < 60) + P(X > 66) = 0.0618 + 0.2206 = 0.2824.$$

Exercise. You sell jars of baby food labelled as weighing 4oz $\approx$ 113g. Suppose your process results in jar weights with the Normal$(\mu = 120, \sigma^2 = 4^2)$ distribution and that you will be fined if more than 2% of your jars weigh less than 113g.

1. What proportion of your jars weigh less than 113g?

   Answer: Letting $X \sim \text{Normal}(\mu = 120, \sigma^2 = 4^2)$, we want to find $P(X < 113)$. We have

   $$P(X < 113) = P((X - 120)/4 < (113 - 120)/4) = P(Z < -1.75) = 0.0401,$$

   so the 0.0401 is the proportion of jars that weigh less than 113 grams. If a jar is selected at random, its weight will be less than 113 with probability 0.0401.

2. To what must you increase μ to avoid being fined?

   Answer: Now let $X \sim \text{Normal}(\mu, \sigma^2 = 4^2)$, where we do not yet specify a value for μ. We want to have $P(X < 113) \leq 0.02$. So we write

   $$P(X < 113) = P((X - \mu)/4 < (113 - \mu)/4) = P(Z < (113 - \mu)/4) \leq 0.02.$$

   By drawing a picture, we can see that the above implies

   $$(113 - \mu)/4 \leq q^Z_{0.02} \iff \mu \geq 113 - 4q^Z_{0.02}.$$

   Plugging in $q^Z_{0.02} = \texttt{qnorm(0.02)} = -2.053749$, we obtain

   $$\mu \geq 113 - 4(-2.053749) = 121.215.$$

   So, in order that the proportion of jars with weights less than 113g does not exceed 0.02, we need to increase μ to at least 121.215.

3. Keeping $\mu = 120$g, to what must you reduce σ to avoid being fined?

   Answer: Now we let $X \sim \text{Normal}(\mu = 120, \sigma^2)$, where we do not yet specify a value for σ. We want to have $P(X < 113) \leq 0.02$. So we write

   $$P(X < 113) = P((X - 120)/\sigma < (113 - 120)/\sigma) = P(Z < -7/\sigma) \leq 0.02.$$

   By drawing a picture, we can see that the above implies

   $$-7/\sigma \leq q^Z_{0.02} \iff \sigma \leq -7/q^Z_{0.02}.$$

   Note that the inequality changes direction when we divide both sides by $q^Z_{0.02}$, as this is a negative number. Plugging in $q^Z_{0.02} = \texttt{qnorm(0.02)} = -2.053749$, we obtain

   $$\sigma \leq -7/(-2.053749) = 3.408401.$$

   So, in order that the proportion of jar weights less than 113 grams be no greater than 0.02, we can allow a value of σ of at most 3.408401.

Checking if a random variable is Normal

It is not always reasonable to assume that a random variable X has a Normal distribution. If we want to use the Normal distribution to make predictions or probabilistic statements about X, we must make sure that X has, at least approximately, a Normal distribution. One way to do this is to take a sample of values of X and see if their distribution resembles a Normal distribution.

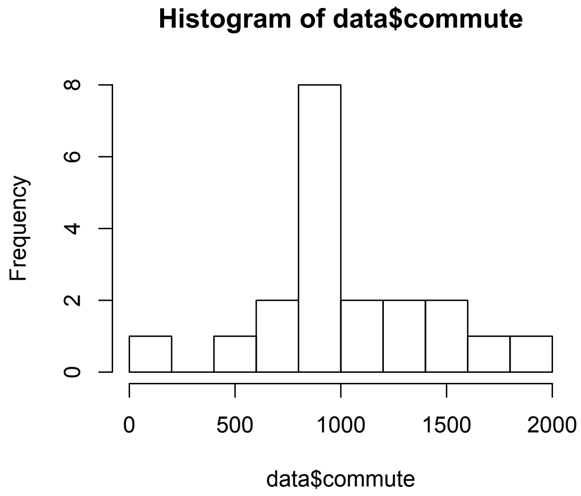
Example. A previous STAT 515 class was polled about the time it takes to commute to class on Monday morning. Each student in the class reported the number of seconds it took to get to his or her first class, resulting in 20 values which the following R command stores in an object called `commute`.

```
commute <- c(1832,1380,1440,913,1620,654,1362,878,577,172,
                   934,773,928,1171,998,1574,1062,900,900,900)
```

Do the commute times of USC undergraduates follow a Normal distribution? Although the class does not represent a random sample of USC students, let us pretend for the moment that it does. A histogram of the commute times can be generated by the command

```
hist(data$commute)
```

which produces



A very crude way of assessing Normality is to compare the shape of the histogram with that of the Normal probability density function. However, since our sample is very small ($n = 20$), the histogram does not give us a good picture of the distribution.

A much better way of assessing Normality is to construct what is called a Normal quantile-quantile, or Normal QQ plot. First we compute the mean $\bar{X}_n$ and standard deviation S_n of the sample. Then we see how close the quantiles of the sample are to the quantiles of the Normal distribution which has mean $\bar{X}_n$ and standard deviation S_n. Recall that S_n is the square root of the sample variance

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

For the commute data, we have

```
> mean(commute)
[1] 1048.4
> sd(commute)
[1] 394.6439
```

Now we ask questions like the following: Is the sample median close to the median of the Normal$(1{,}048.4, 394.6439^2)$ distribution? According to the Normal$(1{,}048.4, 394.6439^2)$ distribution, the median is 1048.4, as the mean of the Normal distribution is also its median. The sample median is 931, so we compare the values 1,048.4 and 931. Now we choose another quantile, say, the 0.10 quantile, at or below which 10% of the observations should lie. For the Normal$(1{,}048.4, 394.6439^2)$ distribution, this number is

$$1{,}048.4 + 394.6439(-1.281552) = 542.6433,$$

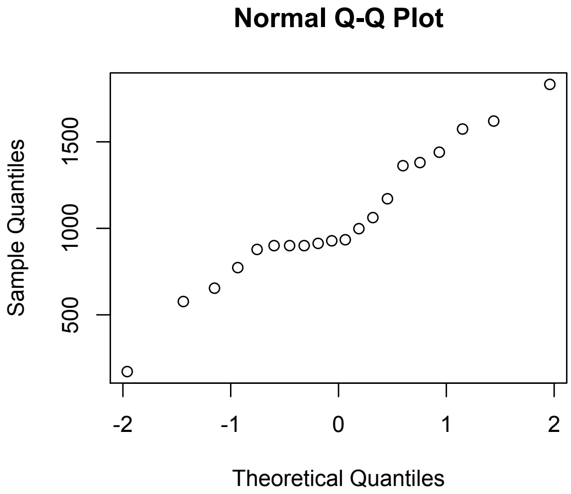
where -1.281552 is the 0.10 quantile of the standard Normal distribution. The 0.10 quantile of the sample is 577 (this is the second smallest value). So we compare the value 577 to 542.6433.

We do this for several quantiles and in the end, we plot the sample quantiles against the corresponding quantiles of the Normal$(1{,}048.4, 394.6439^2)$ distribution. If the points in this plot, which we call a quantile-quantile or QQ plot, fall roughly around a straight line, then we can assume that the distribution of X is Normal.

We can produce a QQ plot to assess the Normality of the commute times with the R command

```
qqnorm(data$commute),
```

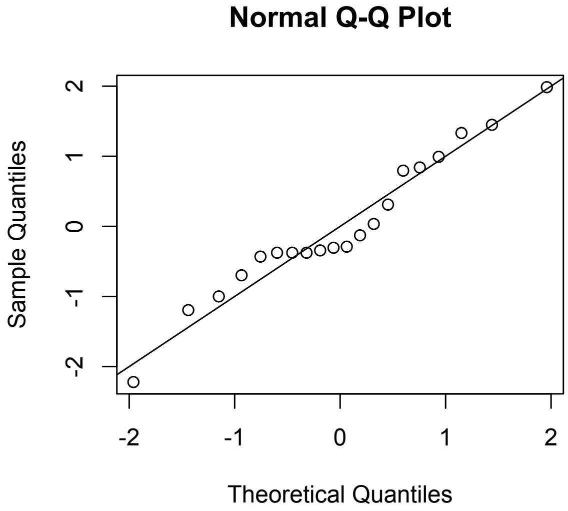
which gives the plot



Note that the quantiles on the horizontal axis are those of the standard Normal distribution. This is okay, we just want to see whether the points fall around a straight line. What I prefer to do is the following:

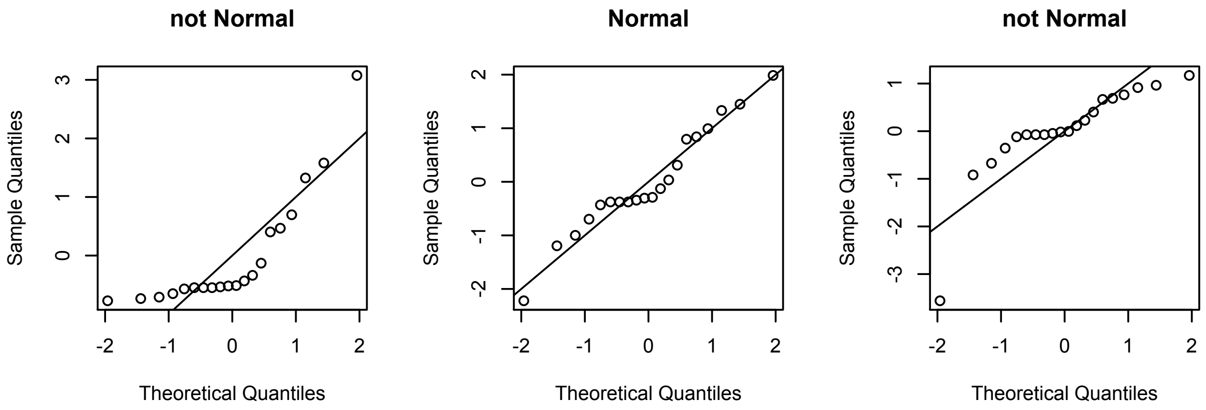
```
qqnorm(scale(commute))
abline(0,1)
```

which produces the plot



in which the axes are on the same scale, allowing us to draw the line $y = x$ with the command `abline(0,1)`. This gives us a visual aid when assessing whether the points fall around a straight line.

It takes some practice looking at QQ plots to know what is Normal and what is not. This passes for Normal; the points stay fairly close to a straight line. Here are some examples of what non-Normal data might look like:



See textbook MCS13 pg. 258 for some examples of QQ plots.

Sums of independent Normal random variables

Result 3: Sum of independent Normal random variables

If $X_1 \sim \text{Normal}(\mu_1, \sigma_1^2), \ldots, X_n \sim \text{Normal}(\mu_1, \sigma_n^2)$ are independent random variables, then

$$\sum_{i=1}^{n} X_i \sim \text{Normal}\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right).$$

In the above, *independent* means that the values of the random variables do not affect one other.

Exercise. Consider boxes containing 25 jars of baby food (from previous).

1. Give the expected value and variance of the box weights.

   Answer: Let $X_1, \ldots, X_{25}$ be the weights of the jars in the box. Then the sum of the box weights $Y = \sum_{i=1}^{n} X_i$ must have the Normal distribution with mean $\mu = 25 \cdot 120 = 3000$ and a variance of $25 \cdot 16 = 400$.

2. Give the probability that the box weighs less than 2,975 grams.

   Answer: Noting that $\sqrt{400} = 20$, we have

$$\begin{aligned} P(Y \leq 2975) &= P((Y - 3000)/20 \leq (2975 - 3000)/20) \\ &= P(Z \leq -5/4) \\ &= \texttt{pnorm(-1.25)} \\ &= 0.1056498. \end{aligned}$$