**The Standard Normal Distribution and z-scores**

If X is normal with mean  and standard deviation , N(,), then the pdf of X is . If Z is standard normal, N(0,1), then the pdf of Z is .

**Important Fact**: If X is N(,), then  is standard normal.

**As a result**, any normal probability can be calculated using standard normal probabilities.

**Example**: Suppose X is N(3,2). Find P(1 ≤ X ≤ 4).

Using R:

> pnorm(.5,0,1)-pnorm(-1,0,1)

[1] 0.5328072

> pnorm(4,3,2)-pnorm(1,3,2)

[1] 0.5328072

**Using Standard Normal Tables to Compute Normal Probabilities.**

**Example:** If X is N(5,3), find P(-1 ≤ X ≤ 4)

**Z-Scores**

If x is a value from a data set that comes from a distribution that is N(,), then  is the **associated z-score**. The z-score allows us to compare values from different normal distributions.

**Example**: Suppose SAT math scores have (approximately) a normal distribution with mean = 500 and standard deviation = 100 and ACT math scores have (approximately) a normal distribution with mean = 21 and a standard deviation = 5. If Wendy had an SAT math score of 720 and Susan had an ACT math score of 31, who performed better?

Of course, in reality we often do not know the values of  and , so we use the sample mean  and the sample standard deviation s to approximate them. So, in practice, the z-score is given by .

**Probability Plots**

Using data to make a statistical inference about a population requires (in general) that we “know” the type of distribution the population has (normal, exponential, gamma,…). Looking at a histogram of the data often allows us to make a good guess about the kind of distribution the population has. A probability plot gives us a way of evaluating whether our guess is good.

**Normal Probability Plot**

**The Idea** Suppose the random variable X from which the data values come is normal, with mean  and standard deviation . Then, the z-scores  will come from a standard normal distribution (approximately). Hence, the quantiles for the standardized data (the z-scores) should match the quantiles for the standard normal distribution. For example, if we plot the quantiles for the z-scores vs the quantiles for the standard normal distribution, the points should follow the line y=x.

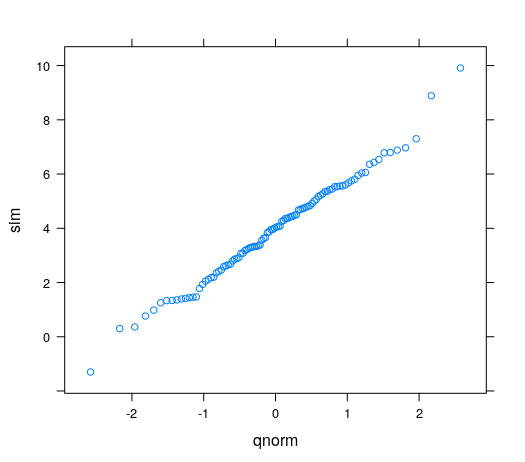
If we plot the quantiles for the original data (not standardized) versus the quantiles for the standard normal distribution, the points won’t line up along y = x, but they will follow a line.

**R has a built-in procedure for generating a normal probability plot.**

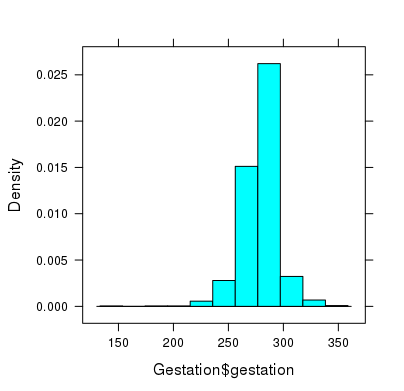
**qqmath(data)**

> sim<-rnorm(100,4,2)

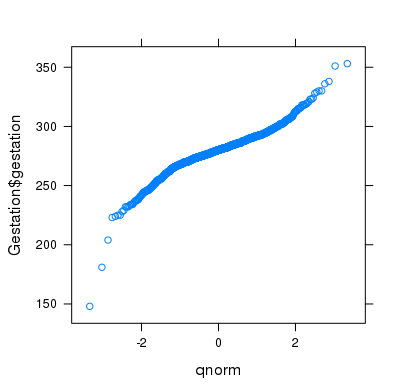
> qqmath(sim)



> histogram(Gestation$gestation)

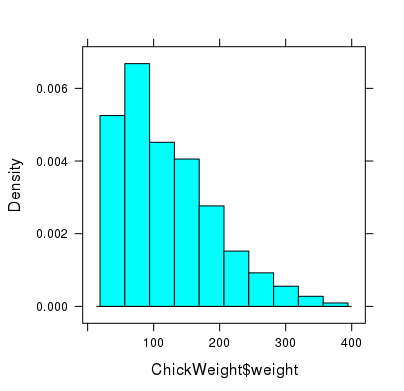
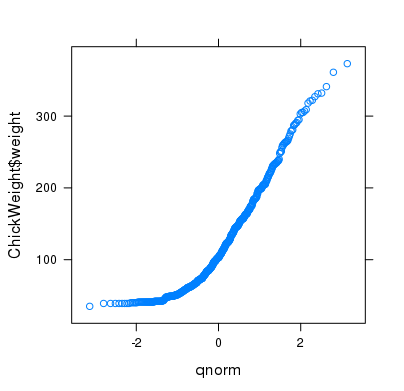


> qqmath(Gestation$gestation)



**What about other kinds of distributions?**

> histogram(ChickWeight$weight) > qqmath(ChickWeight$weight)

** **

**Try an exponential distribution**

> fitdistr(ChickWeight$weight, "exponential")

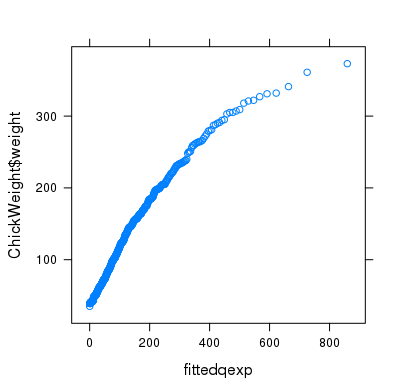
rate

0.0082089446

(0.0003414471)

> fittedqexp<-function(p) qexp(p,.0082089446)

> qqmath(ChickWeight$weight,distribution=fittedqexp)



**Try a gamma distribution.**

> fitdistr(ChickWeight$weight,"gamma")

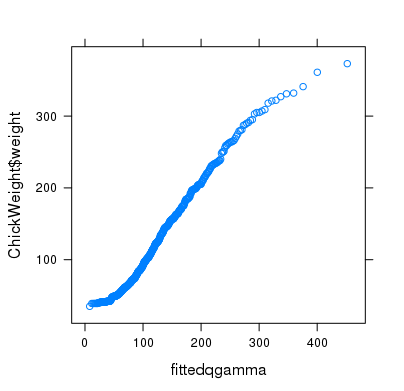
shape rate

3.151334855 0.025869128

(0.175942773) (0.001564504)

> fittedqgamma<-function(p) qgamma(p,3.151334855,.025869128)

> qqmath(ChickWeight$weight,distribution = fittedqgamma)



**Exercises 9**

1. The dataframe **morley** contains 100 measurements that Morley made of the speed of light. These values are contained in the column **Speed**. Use qqmath to create a normal probability plot for this data. Both of the following R commands will work:

> qqmath(~Speed,data=morley)

> qqmath(morley$Speed)

Does this provide strong evidence that a normal distribution fits the data well? (Include the R commands you used, the plot, and your answer to the question.)

2. The column **hincome** in the dataframe **Womenlf** contains household income for 263 Canadian women. (The package **car** must be loaded to access this data.)

1. Use qqmath to create a normal probability plot for this data.
2. Use qqmath to create a qq plot for the gamma distribution.
3. Based on these plots, evaluate which, if either, kind of distribution fits the data well.

(Include the R commands you used and the plots in (a) and (b).)

3. The column **i1** in the data frame **HELPrct** contains the average number of drinks per day for each participant in this study. From the histogram it looks like an exponential distribution might fit this data well. Use qqmath to produce an exponential qq plot for this data. Based on this plot, do you conclude that an exponential distribution does fit the data well? Include the R commands, the plot, and your answer to the question.

4. Another family of continuous distributions is the family of Weibull distributions. R has this family of distributions built in as well. There are two parameters: shape and scale. The usual 4 functions are available: dweibull(x,shape,scale), pweibull(x,shape,scale), qweibull(p,shape,scale), and rweibull(n,shape,scale). For the column **hincome** in the dataframe **Womenlf**, do a qq plot using a Weibull distribution. Include the R commands and the qq plot. How does the Weibull fit compare with the normal and gamma distributions in (2) above?

5. The heights of 18-22 year-olds in the US are approximately normally distributed. For women, mean = 64.3 in. and sd = 2.6 in. For men, the mean = 70 in. and sd = 2.8 in.

1. if a woman is 69 inches tall, what is her z-score?
2. If a man is 69 inches tall, what is his z-score?
3. What is more unusual, a woman who is 70 inches tall or a man who is 77 inches tall?
4. An doorway is to be designed so that 98% of all men 18-22 can pass through the door without bending down to avoid hitting their heads. What is the shortest doorway that meets this condition?

6. If X has a normal distribution with mean = 4 and sd = 2, use the standard normal table to find P( 3 ≤ X ≤ 5).