# FA541 Applied Statistics with Applications in Finance

Humphrey De Guzman

10/1/2021

## Contents

# Page 75

## 1.128

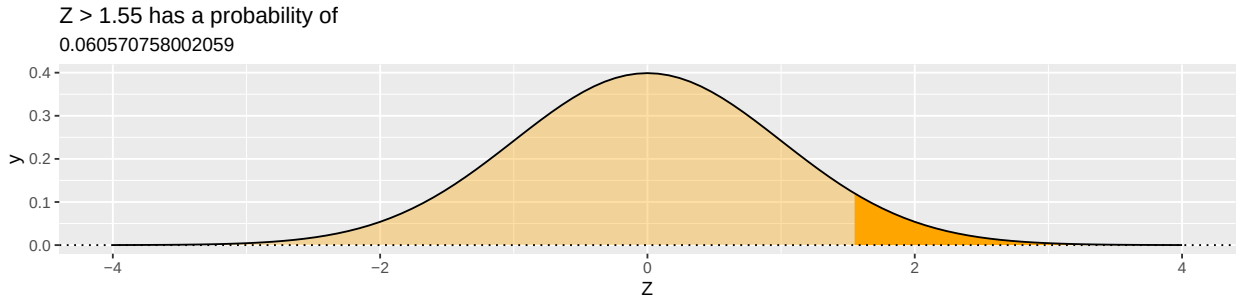
**1.128 Find some proportions.** Using either Table A or your calculator or software, find the proportion of observations from a standard Normal distribution that satisfies each of the following statements. In each case, sketch a standard Normal curve and shade the area under the curve that in the answer to the question.

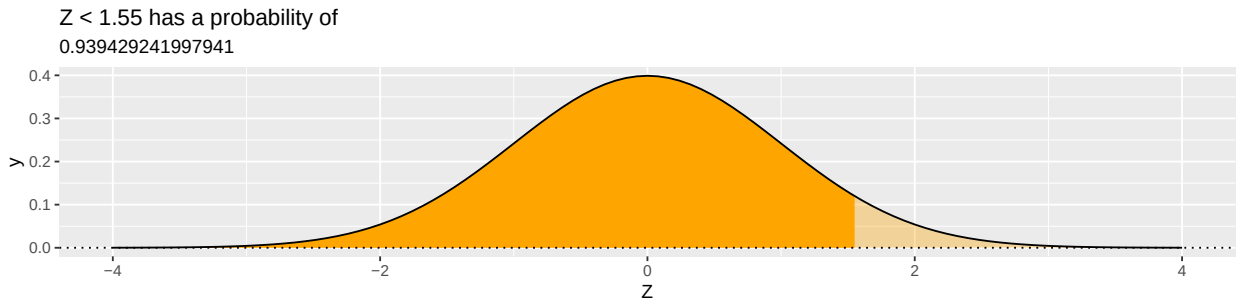
(a) $Z > 1.55$

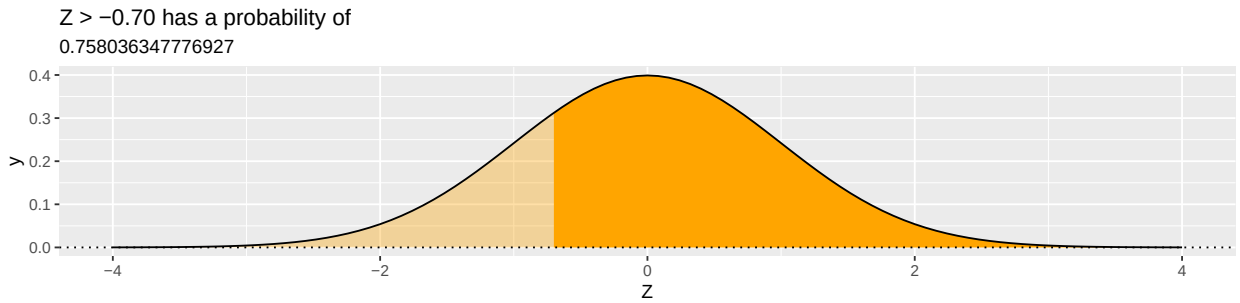
(b) $Z < 1.55$

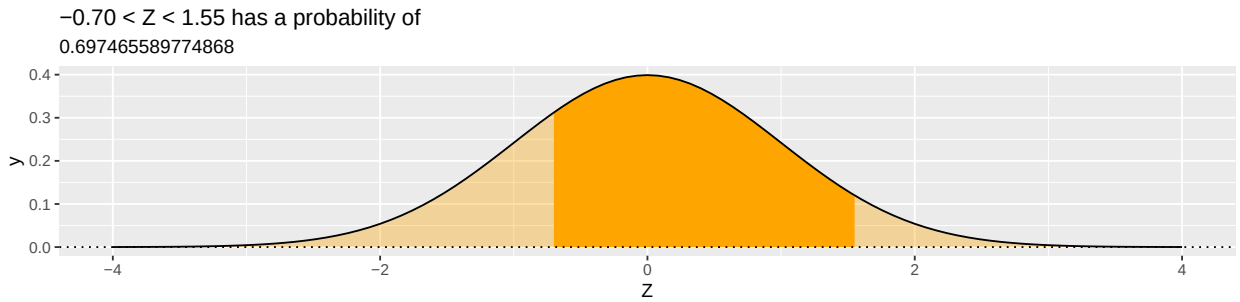
(c) $Z > -0.70$

(d) $-0.70 < Z < 1.55$



## 1.129

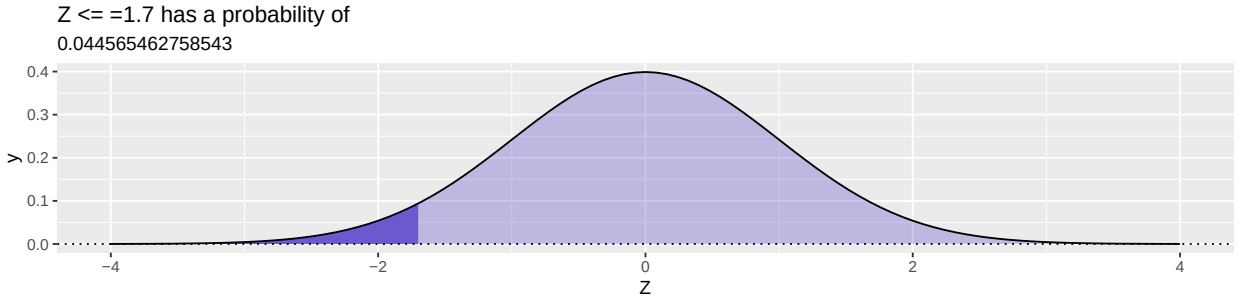
**1.129 Find more proportions.** Using either Table A or your calculator or software, find the proportion of observations from a standard Normal distribution for each of the following events. In each case, sketch a standard Normal curve and shade the area representing the proportion.

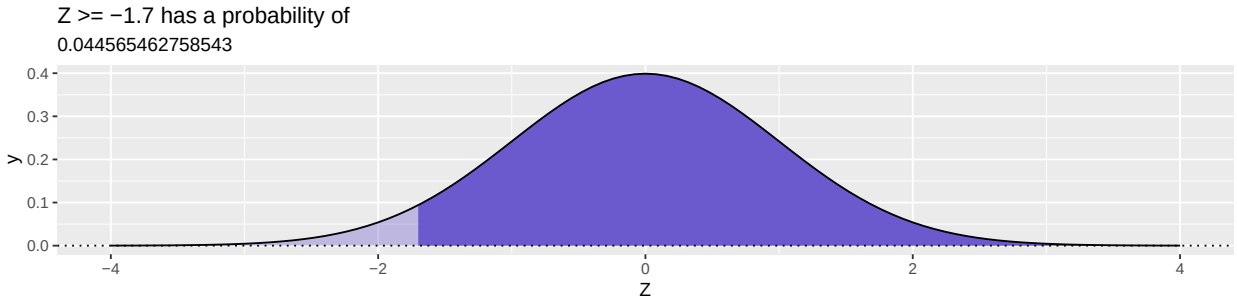
(a) $Z \leq -1.7$

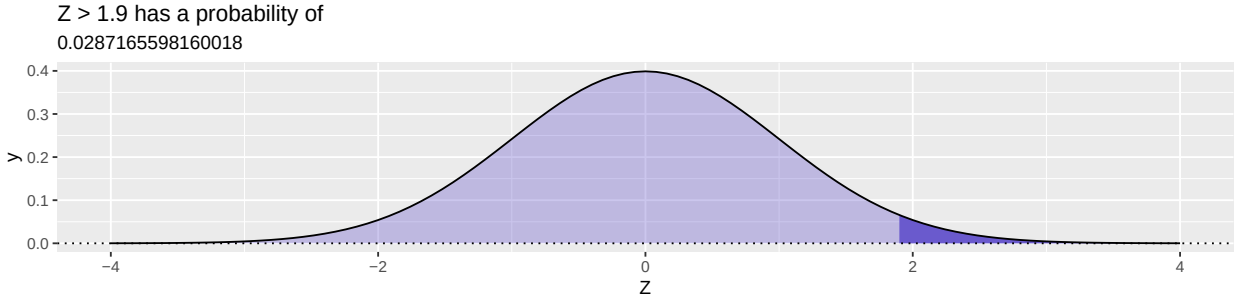
(b) $Z \geq 1.55$

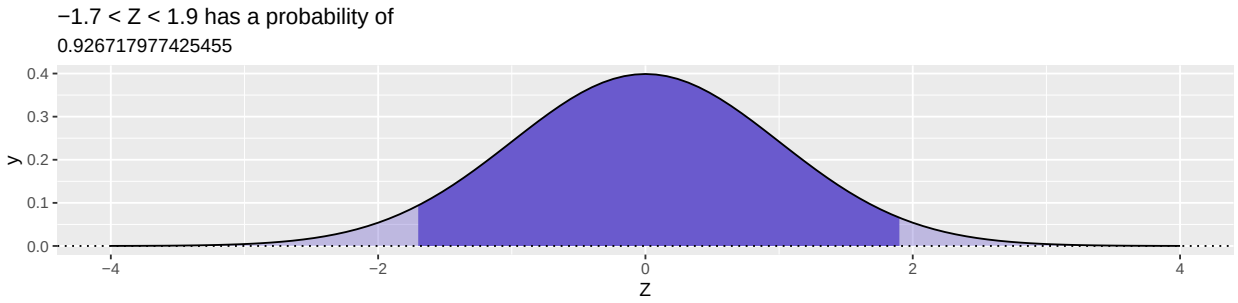
(c) $Z > -0.70$

(d) $-1.7 < Z < 1.9$



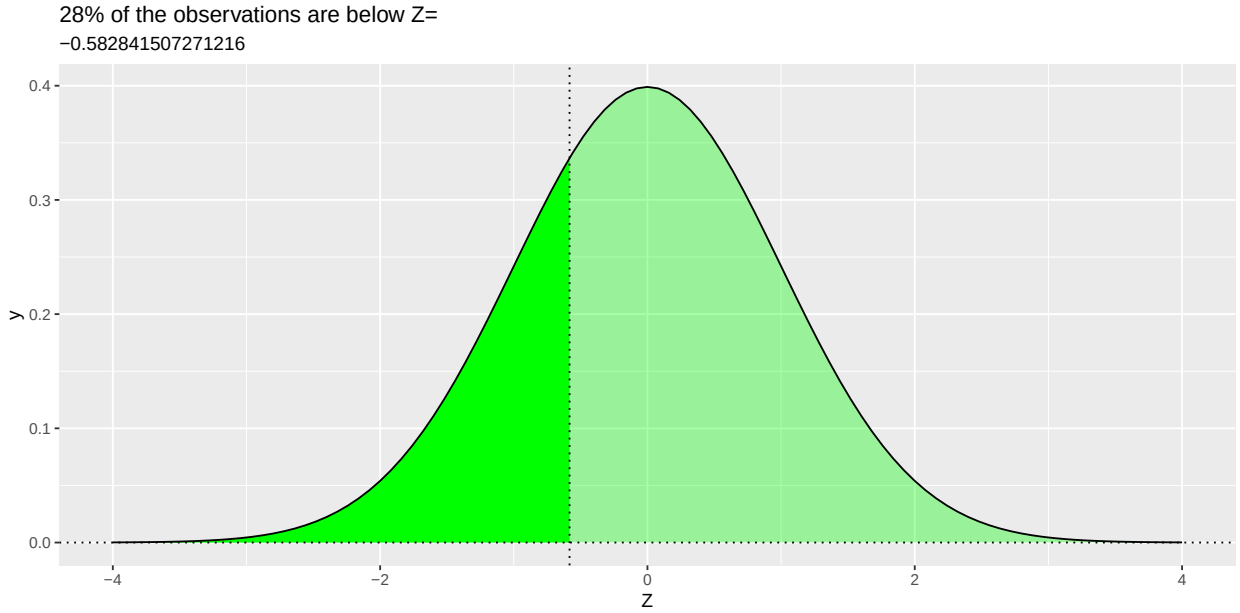



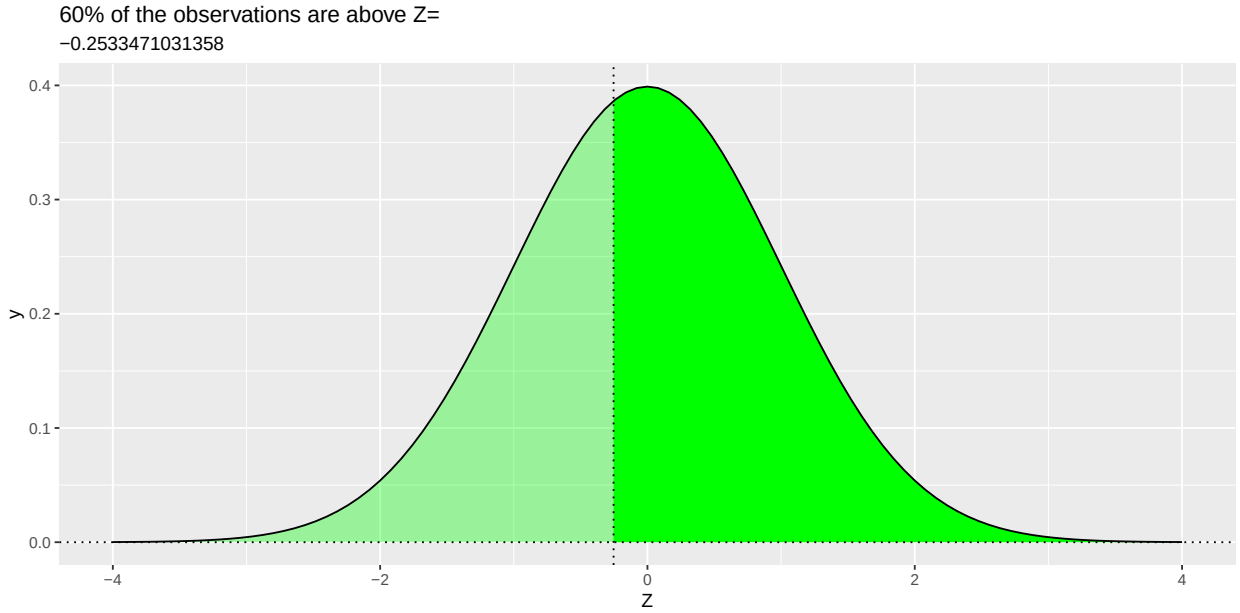
## 1.130

**1.130 Find some values of z.** Find the value z of a standard Normal variable Z that satisfies each of the following conditions. (If you use Table A, report the value of z that comes closest to satisfying the condition.) In each case, sketch a standard Normal curve with your value of z marked on the axis.

(a) 28% of the observations fall below z.

(b) 60% of the observation fall above z.



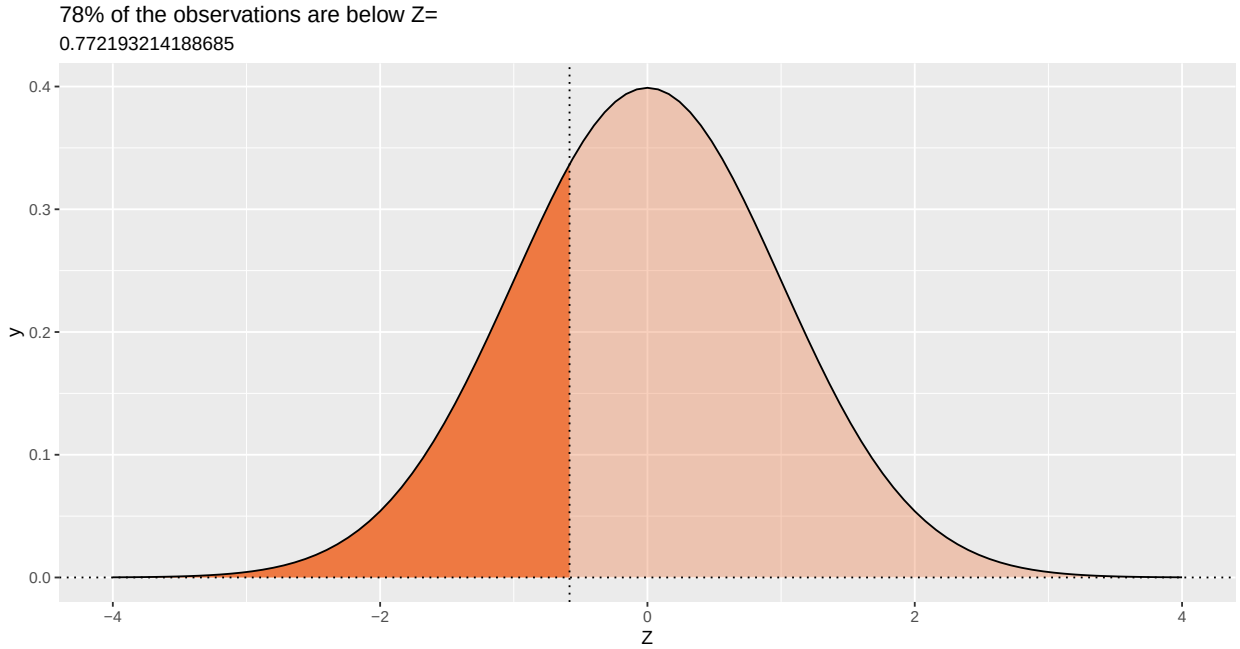



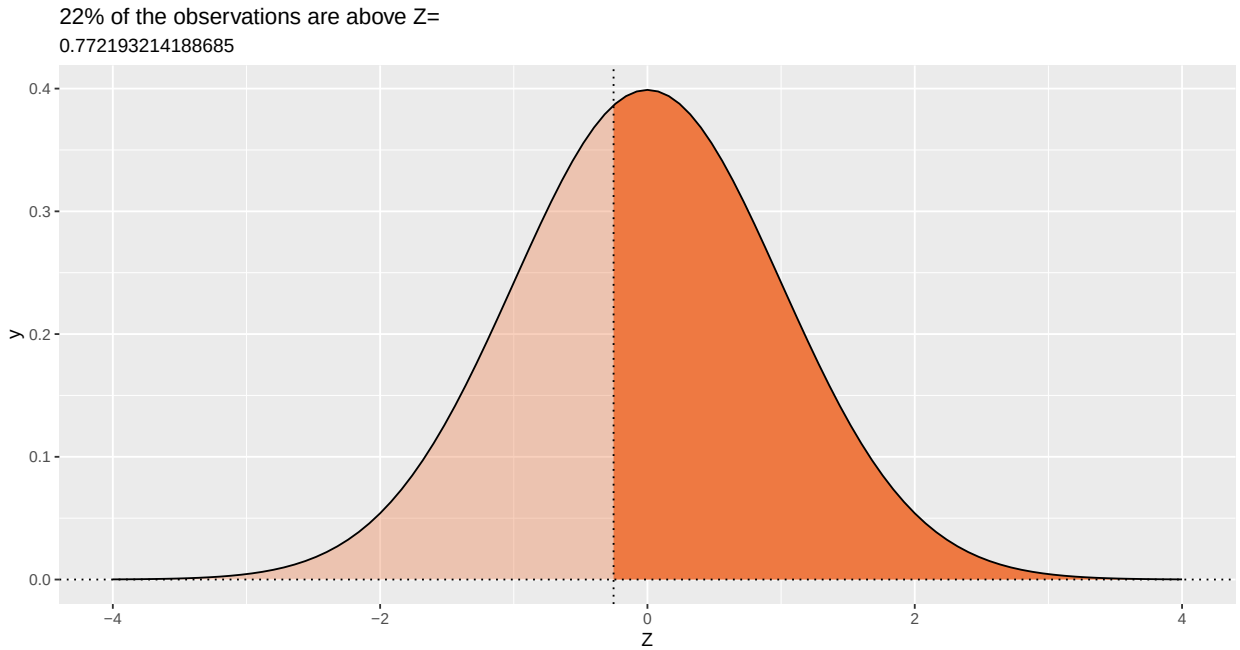
## 1.131

**1.131 Find more values of z.** The variable Z has a standard Normal distribution.

(a) Find the number z that has cumulative proportion 0.78.

(b) Find the number z such that the event Z > z has proportion 0.22.




In this case notice that we have the same z-score, which makes sense as the cumulative proportion of 0.78 should be the same point where the upper proportion 0.22 should meet as they equal the total area of 1.

# Page 79 & 80

## 1.171(a)

**1.171 How much vitamin C do women consume?** To evaluate wether or not the intake of a vitamin or mineral is adequate, comparisons are made between the intake distribution and the requirement distribution. Here is some information about the distribution of vitamin C intake, in milligrams per day, for women aged 19 to 30 years.

| | Percentile (mg/d) | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Mean | 1st | 5th | 19th | 25th | 50th | 75th | 90th | 95th | 99th |
| 84.1 | 31 | 42 | 48 | 61 | 79 | 102 | 126 | 142 | 179 |

(a) Use the 5th, the 50th, and the 95th percentile of this distribution to estimate the mean and standard deviation of this distribution assuming that the distribution is Normal. Explain your method for doing this.

We know that the z-score formula is as follow.

$$Z_{score} = \frac{x - \mu}{\sigma}$$

Where $x$ is our observed value, $\mu$ is our mean or first moment, and $\sigma$ is the root of our second moment, also known as standard deviation. By using what we learned from the previous question, we can easily fit the approriate z-scores found on the space $N(0, 1)$ and solve such system of equations.

For the 5th percentile we see that this value is -1.6449, for the 50th this value is 0, and for the 95th we find 1.6449. This should all make sense intuitively as the 5th and 95th percentile should have the same score value just flipped over our symmetric distribution. The same goes for our 50th percentile where on a $N(0, 1)$ the middle portion of the graph is centered at our mean 0!

This gives us powerful insight given the Percentiles shown are reflective of the true population. We find that then our $\mu = 79$ using the 50th percentile, note that we cannot find $\sigma$ here without the use of the 5th or 95th percentile. We can now set up a system of equations where

$$1.6449 = \frac{142 - \mu}{\sigma}$$
$$-1.6449 = \frac{42 - \mu}{\sigma}$$

Which can also be written as

$$\sigma = \frac{42 - \mu}{-1.6449} = \frac{142 - \mu}{1.6449}$$

Which solving gives us $\mu = 92$ and $\sigma = 30.39784$. Notice how these give us different values depending on if we believe the 50th percentile is indicative of the truth, or the 5th and 95th percentile.

Alternatively if we would like to include all values in our calculation, that is to say we believe all to be true, we set this up.

$$\frac{79 - \mu}{\sigma} = \frac{142 - \mu}{\sigma} + \frac{42 - \mu}{\sigma}$$

This is done as we know that the 5th and 95th percentile are equal to the same value except the negative of the other. Their summation is 0, which is equal to the z-score of our 50th percentile. From here we find $\mu = 105$ and from this point we simply do

$$\frac{142 - 105}{1.6449} \approx 22.4944 = \sigma$$
$$\frac{42 - 105}{-1.6449} \approx 38.3013 = \sigma$$

Again note that our sigma is different, one can take the average of the two which in a sense is a linear probe of our standard deviation, or choose one. In our case I will say that $\mu = 92$ and $\sigma = 30.39784$ and the 50th percentile value found in the data is attributed to variance.

## 1.176

**1.176 Norms for reading scores.** Raw scores on behavioral tests are often transformed for easier comparison. A test of reading ability has mean 70 and standard deviation 10 when given to third-graders. Sixth-graders have mean score 80 and standard deviation 11 on the same test. To provide seperate "norms" for each grade, we want scores in each grade to have mean 100 and standard deviation 20.

(a) What linear transformation will change third-grade scores $x$ into new scores $x_{new} = a + bx$ that have the desired mean and standard deviation? (Use $b > 0$ to preserve the order of the scores.)

(b) Do the same for the sixth-grade scores.

(c) David is a third-grade student who scores 72 on the test. Find David's transformed score. Nancy is a sixth-grade student who scores 78. What is her transformed score? Who scores higher within his or her grade?.

(d) Suppose that the distribution of scores in each grade is Normal. Then both sets of transformed scores have $N(100, 20)$ What percent of third-graders have scores less than 75? What percent of sixth-graders have scores less than 75?

**A** we know our linear transformation takes the form of $bx + a = x_{new}$ and that for the first moment, the mean undergoes the same transformation. For the second moment we find that we simply scale it by $b^2$, though because we are looking for standard deviation, we can just take the roots and see that $\sigma_{new} = b\sigma$. From here we simply find

$$100 = b70 + a$$
$$20 = b10$$
$$x_{new} = -40 + 2x$$

**B** We repeat the same thing

$$100 = b80 + a$$
$$20 = b11$$
$$x_{new} = -45.\overline{45} + \frac{20}{11}x$$

**C** We just show the math here

$$100_{David} = -40 + 2(72)$$
$$96.\overline{36}_{Nancy} = \frac{20(78)}{11} + -45.\overline{45}$$

David scores higher than Nancy in regard to their grade.

**D** This can be done multiple ways, we will check using R's `pnorm()` function for the initial distributions then through a transformed version. In this case 75 transformed for third-graders is 110 and $90.\overline{90}$ for sixth-graders

```
Third = pnorm(75,70,10)
Sixth = pnorm(75,80,11)
cat(Third,"of third-graders have scores less than 75   \n",Sixth,"of sixth-graders have scores less thar
```

0.6914625 of third-graders have scores less than 75
0.3247181 of sixth-graders have scores less than 75

```
nThird = pnorm(110,100,20)
nSixth = pnorm(9000/99,100,20)
cat(nThird,"of third-graders have scores less than 75   \n",nSixth,"of sixth-graders have scores less th
```

0.6914625 of third-graders have scores less than 75
0.3247181 of sixth-graders have scores less than 75

## 1.177

**1.177 use software to generate some data.** Most statistical software packages have routines for generating values of variables having specified distributions. Use your statistical software to generate 30 observations from the $N(25, 8)$ distribution. Compute the mean and standard deviation $\overline{x}$ and $s$ of the 30 values you obtain. How close are $\overline{x}$ and $s$ to the $\mu$ and $\sigma$ of the distribution from which the observations were drawn? Repeat 19 more times the process of generating 30 observations from the N(25,8) distribution and recording $\overline{x}$ and $s$. Make a stemplot of the 20 values of s. Make Normal quantile plots of both sets of data. Briefly describe each of these distributions. Are they symmetric or skewed? Are they roughly Normal? Where are their centers? (The distributions of measures like $\overline{x}$ and $s$ when repeated sets of observations are made from the same theoretical distribution will be very important in later chapters.)

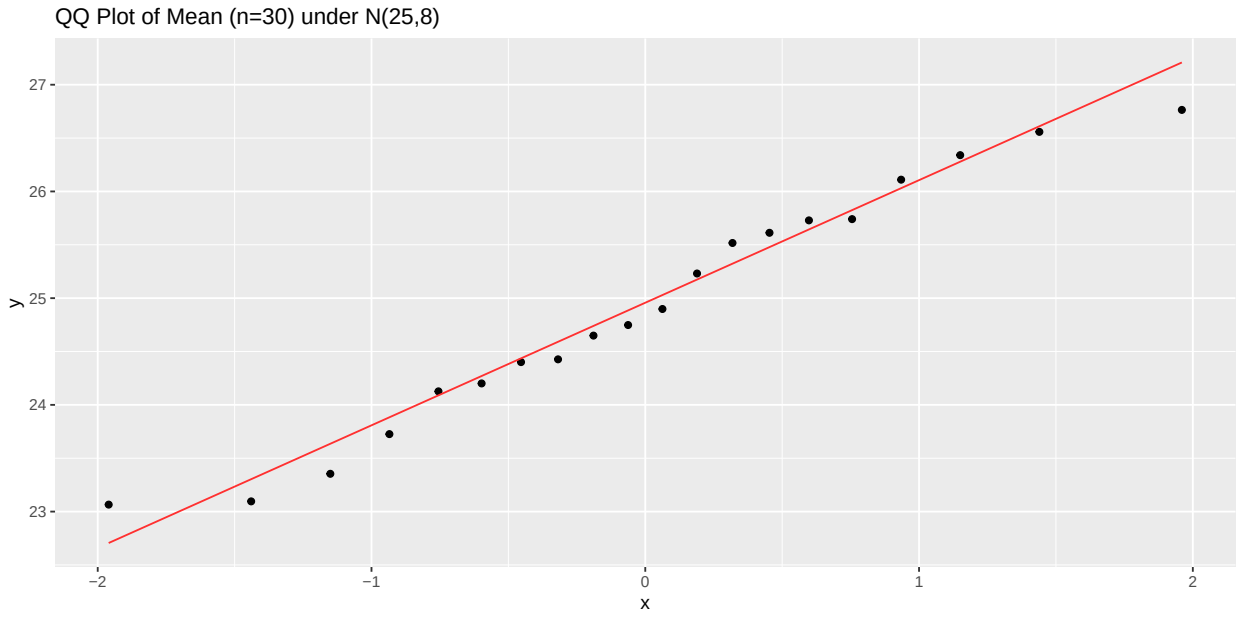
```
set.seed(100)
df = data.frame(Mean=double(),StdDev=double())
for (x in seq(1,20)){
  foo = rnorm(30,25,8)
 df[x,] = c(mean(foo),sd(foo))
}
stem(df$Mean)
```

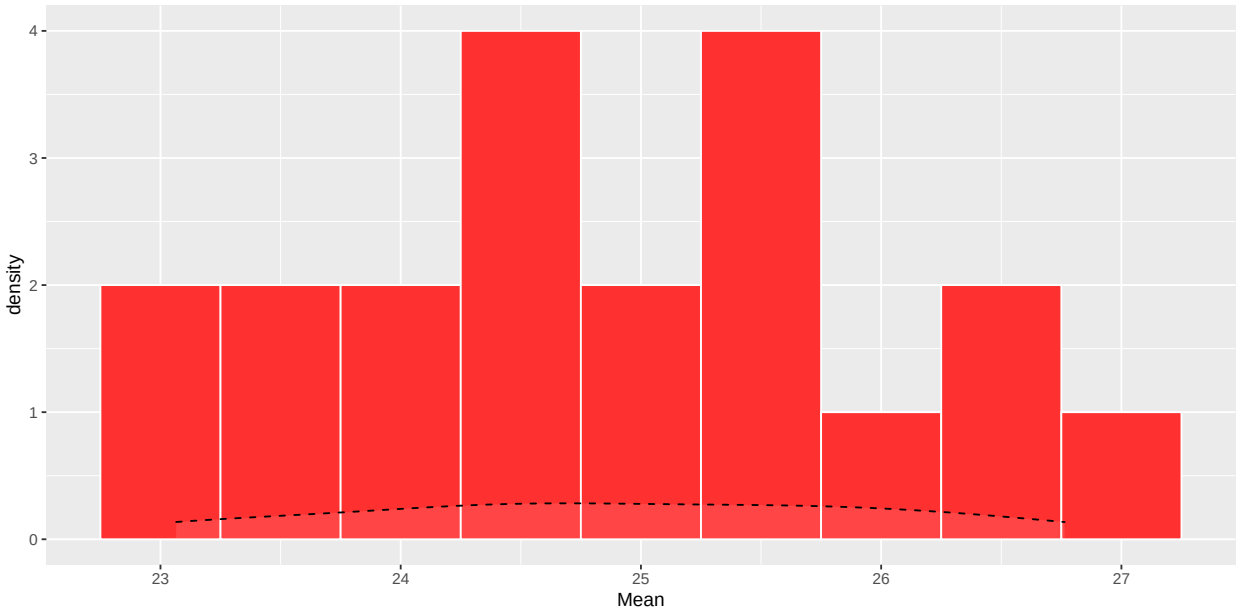
```
##
##   The decimal point is at the |
##
##   23 | 1147
##   24 | 1244679
##   25 | 25677
##   26 | 1368
```

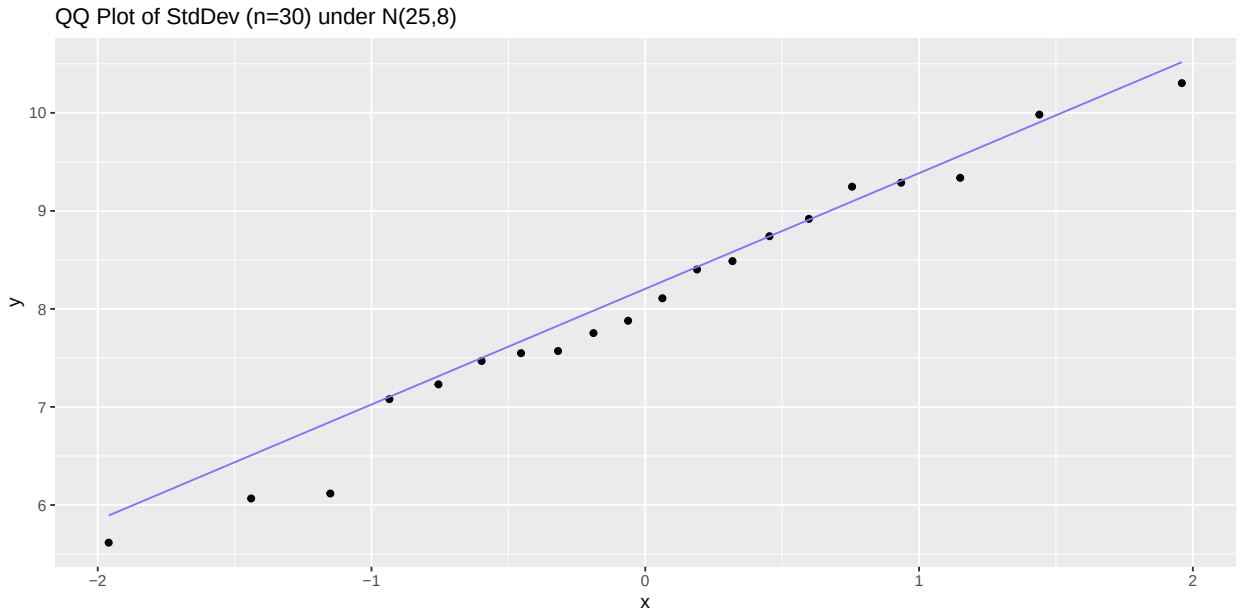
```
stem(df$StdDev)
```

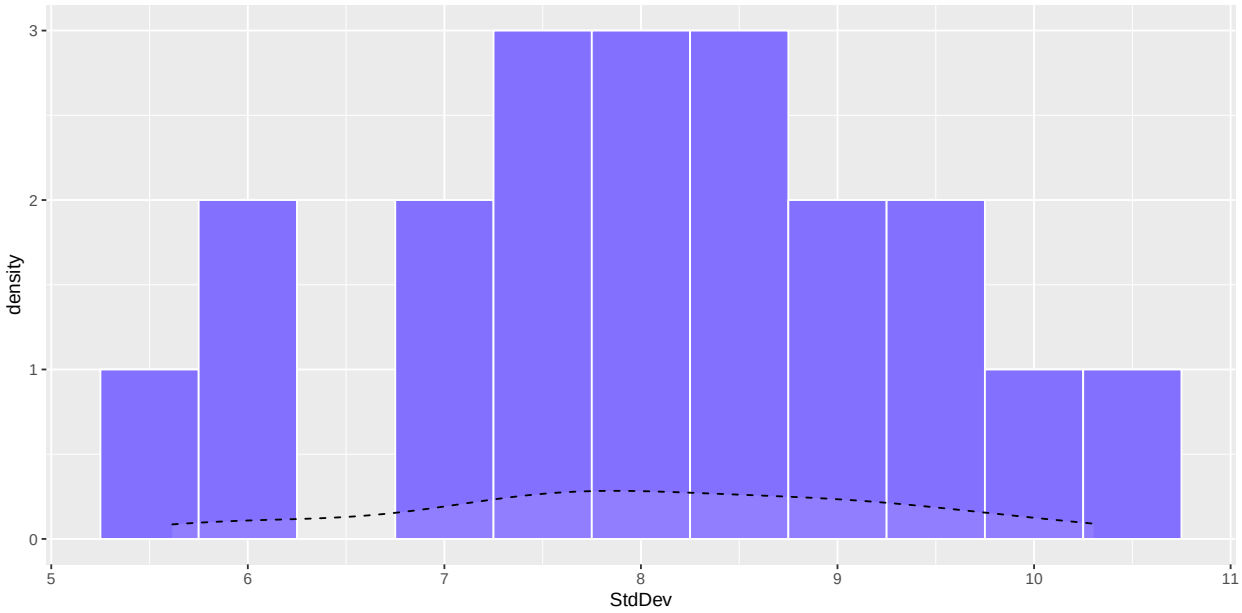
```
##
##   The decimal point is at the |
##
##    5 | 6
##    6 | 11
##    7 | 1255689
##    8 | 14579
##    9 | 233
##   10 | 03
```

My values of $\overline{x}$ and $s$ are $\overline{x} = 24.8791$ and $s = 5.4828$. The mean is quite close however we can see that our standard deviation is not close to our value of 8.






Looking at both the stem plots and QQ-plots we can see that both are typically normal, which is expected of both values. Both have centers near the values of their expected values of respectively 25 and 8. In fact the average of our $\overline{x}$ and $s$ are 25.8266 and 7.8071 respectively. With the QQ-plots we can see that the distributions of both Means and Standard Deviations are quite normal however both look sort of cubic, (at a very small scale), which gives us the inclination that there are fatter tail. Both seem to have some small skewness from what we've seen in the stemplots, though those are quite primitive. Looking at histograms shows us better pictures in which we find the the distribution of both are slightly skewed to the right, though the Standard Deviation distribution looks more even.

# Page 102

## 2.36

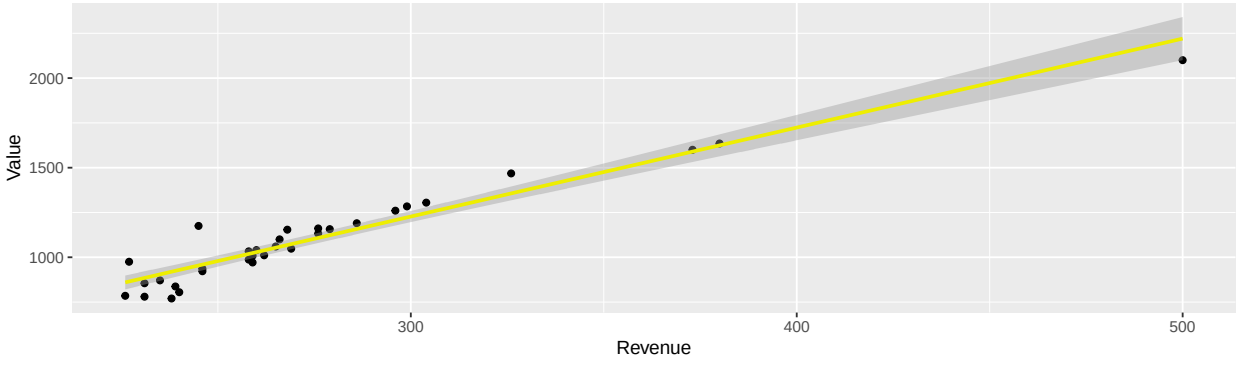
**2.36 Team value in the NFL.** Management theory says that the value of a business should depened on its operating income, the income produced by the business after taxes. (Operating income excludes income from sales of assets and investments, which don't reflect the actual business). Total revenue, which ignores costs, should be less important. Debt includes borrowing for the construction of a new arena. That date file NFL gives the value (in millions of dollars), debt (as percent of value), revenue (in millions of dollars), and operating income (in millions of dollars) of the 32 teams in the National Football League (NFL).

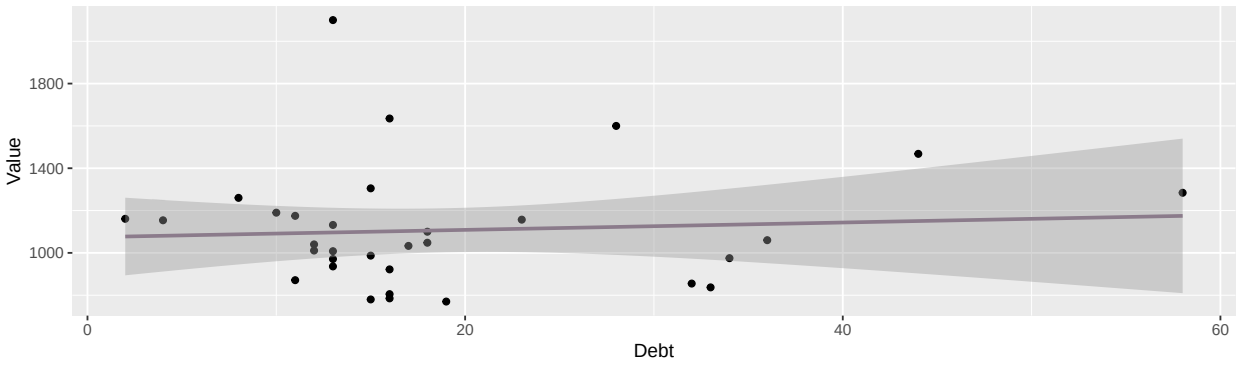
(a) Plot team value against revenue. Describe the relationship.

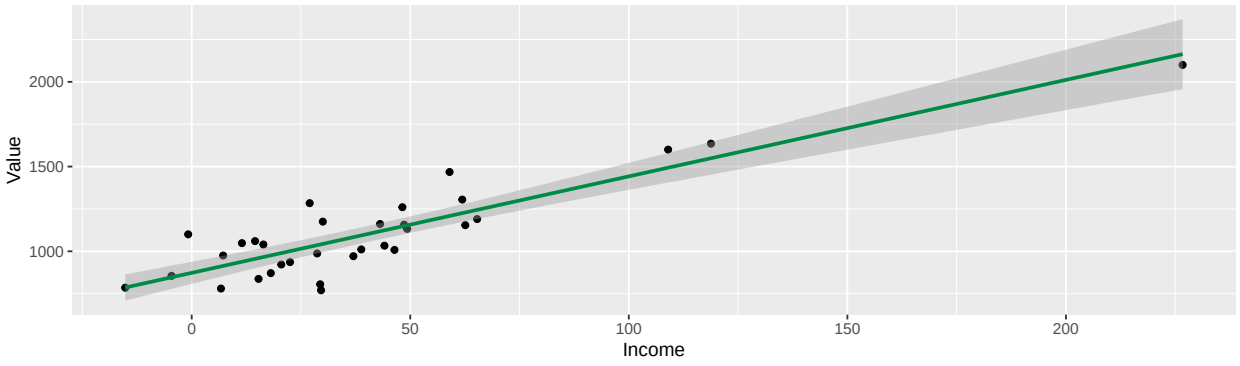
(b) Plot team value against debt. Describe the relationship.

(c) Plot team value against operating income. Describe the relationship.

(d) Write a short summary comparing the relationships that you described in parts (a), (b), and (c) of this exercise.



**A** The relationship looks strong and linear, though I would say that extremely large or small values of Income have some "extra" variation.

**B** There seems to be no relationship between Debt and Value, at least in terms of first glance linear relationships. It may have some quadratic relationship but even that looks like a stretch, espescially at the ends of our Debt. This makes sense as debt structure has more to do with liquidity rather than actual value (it is what the debt is used for that is more reflective of value).

**C** Again we see a linear relationship however, this one is quite varied throughout the graph, which allows us to infer some weaker correlation between the two variable, at least compared to the Revenue vs Value graph. This is quite surprising as operating income is a more "refined" revenue and we assume that operating income should be more reflective of value as this is the money that a business has to work with.

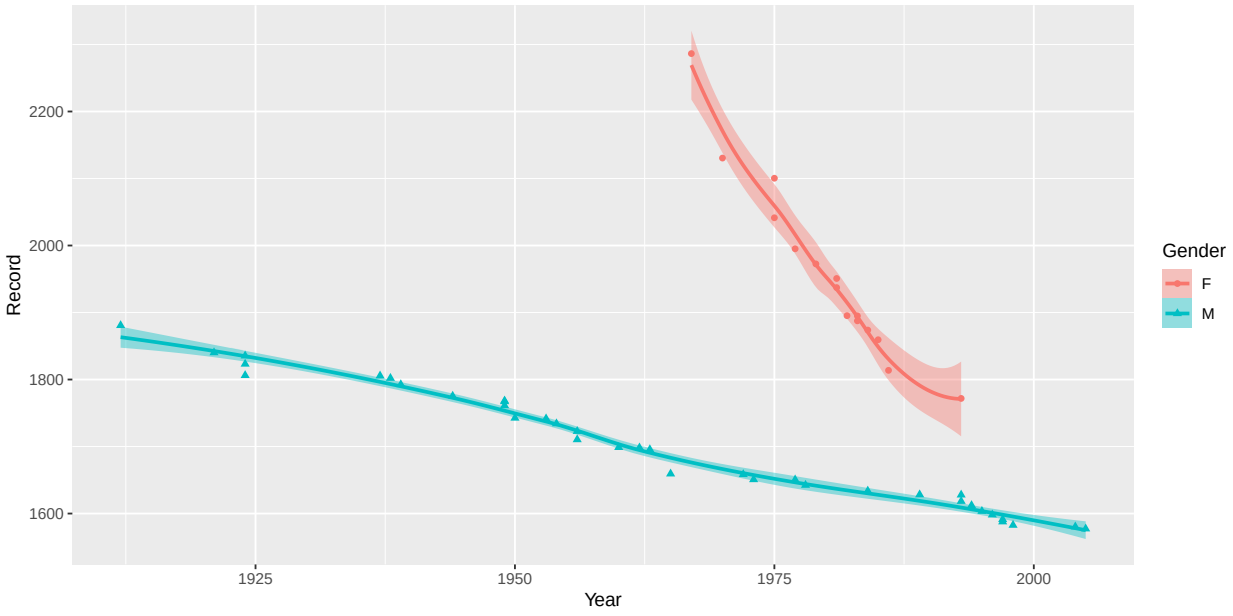
**D** In summary we see linear relationships with Revenue and Income, were the former is stronger than the latter. I cannot think of a logical reason for this other than the fact that the literal costs that teams pay can be abstracted into items that have no real monetary value but provide some sort of "return" or investment opportunity through things such as better performance in games, public opinion, etc. Debt has no linear relationship and it seems to have some weak polynomial relationship, maybe a rotation of $ax - b\ln x + c$, but really a hard relationship to model at that. We agree with this finding as debt structure isn't really indicative of value, ex: Teams that have a lot of debt structure may not invest said debt efficiently than others in which more debt may be required to attain the same Value.

## 2.37

**2.37 Records for men and women in the 10k.** Table 2.1 shows the progress of world record times (in seconds) for the 10,000-meter run for both men and woman.

(a) Make a scatterplot of world record time against year, using seperate symbols for men andd women. Describe the pattern for each sex. Then compare the progress of men and women.

(b) Women began running this long distance later than men, so we might expect their improvement to be more rapid. Moreover, it is often said that men have little advantage over women in distance running as opposed to sprints, where muscular strength plays a greater role. Do the data appear to support these claims?



**A** Both seem to have strong linear relationships with negative correlations. Women seem to have greater absolute correlation, or in real world terms, have improved better than their male counterparts. Though towards the end of the graph we see some asymptotic behavior on the women's plot, though that can be attributed to variance. This can mean that while the difference in the 10K time between genders has been shrinking, if their is some "real" difference between the capabilities of long distance running of men and women, the graphs should reach a cointegration point and the lines become parallel.

**B** I would agree with the first claim that women's improvement is more rapid than men's. Though I am reluctant to take a stance on the second statement. This is due to the fact that while that data does show some significant advantage that men have over women, this can change over time and our system may continue to evolve in a direction we may not necessarily know. It may be possible that the rate of women's improvement continues on their trajectory and eventually cross the men's line. Alternatively, this data set only tracks 10K times, which is fairly long, but not as long as say marathons and half-marathons, which might give us a more conducive picture. In addition, this sample does not show us the whole population of women and men, more so it is a sample of male runners and female runners, which can be messy when talking about the semantics for gender. This is to say while I would agree on the use of female and male runners to reflect running time potential's, from a technical standpoint, some may not agree that this can be translated to all women or all men.