Frequency Distributions

January 4, 2020

Contents

We've all taken a standardized test and received a percentile rank. For example, a SAT score of 1940 corresponds to a percentile of 90. This means that 90% of test takers received a score of 1940 or below. Percentile ranks are a way of converting any set of scores to a standard number, which allows for the comparison of scores from test-to-test or year-to-year.

A common example of the use of percentile ranks is when a professor curves scores from a class to compute the class grades. Here we'll work through a concrete example from an example data set to curve scores for a class.

Suppose you're a professor who wants to convert final grades to a course grades of A, B, C, D and F. (we could also convert to the finer scale of grade points but let's keep things simple).

More specifically, you want to assign a grade of A to the top 10% of students, B's to the next 10%, C's to the next 10%, D's to the next 20%, and F's to the last 50%. Don't worry, I won't fail half of our class!

In your class of 20 students, you obtain the following final scores, which reflect a combination of homework, midterm and final exam grades, sorted from lowest to highest:

You can download the csv file containing these scores here: ExampleGrades.csv

Score
55
56
56
57
60
60
61
61
62
64
72
72
76
76
76
77
77
77
79
79

Frequency histograms

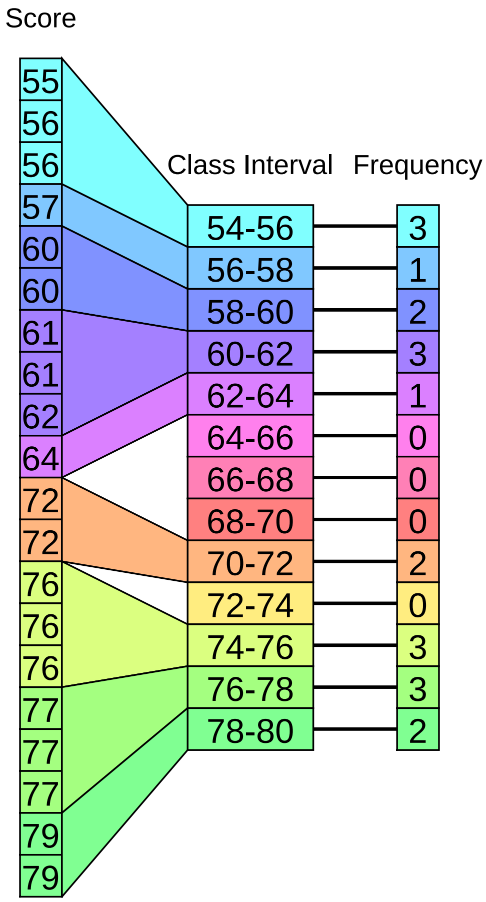
First we'll explore this data set by visualizing the distribution of scores as a histogram. A histogram shows the frequency of scores that fall within specific ranges, called **class intervals**.

The choice of your class intervals is somewhat arbitrary, but there are some general guidelines.

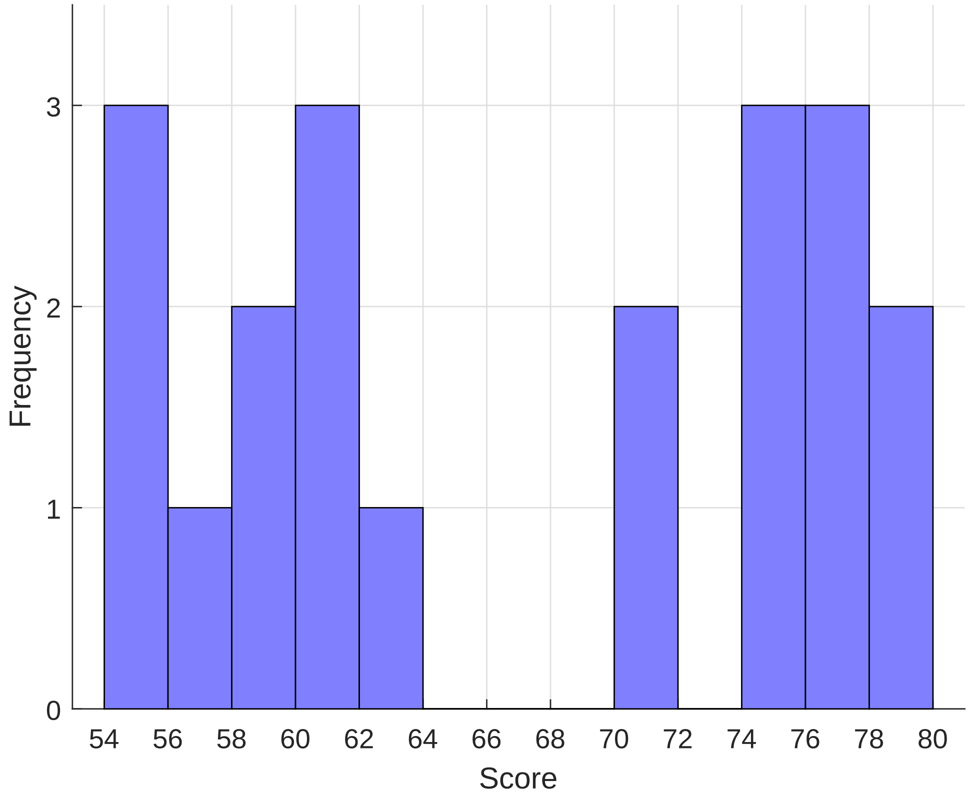
First, choose a sensible number and width for the class intervals. It's good to have something around 10 intervals. Our scores cover a range between 55 and 79, which is 24 points. This means that a width of 2 should be about right.

Second, choose a sensible lowest range of the lowest class interval. A good choice is a multiple of the interval width. Since our lowest score is 55, the lowest factor of 2 below this is 54 . We'll use the rule that if a score lies on the border between two class intervals, the score will be placed in the **lower** class interval. Our first class interval will therefore include the scores greater than or equal to 54 and less than 56.

This figure should help you see how the scores are assigned to each class interval:



We can visualize the distribution of scores with a graph of the **frequency histogram**, which is just a bar graph of the frequencies for the class intervals:



I've labeled the x-axis for the class intervals at the borders. Alternatively you can label the centers of the intervals or the range for each interval. It's up to you.

Take a look at the frequency histogram. What does it tell you about the distribution of scores? Can you see where you might choose the cutoffs for the different grades?

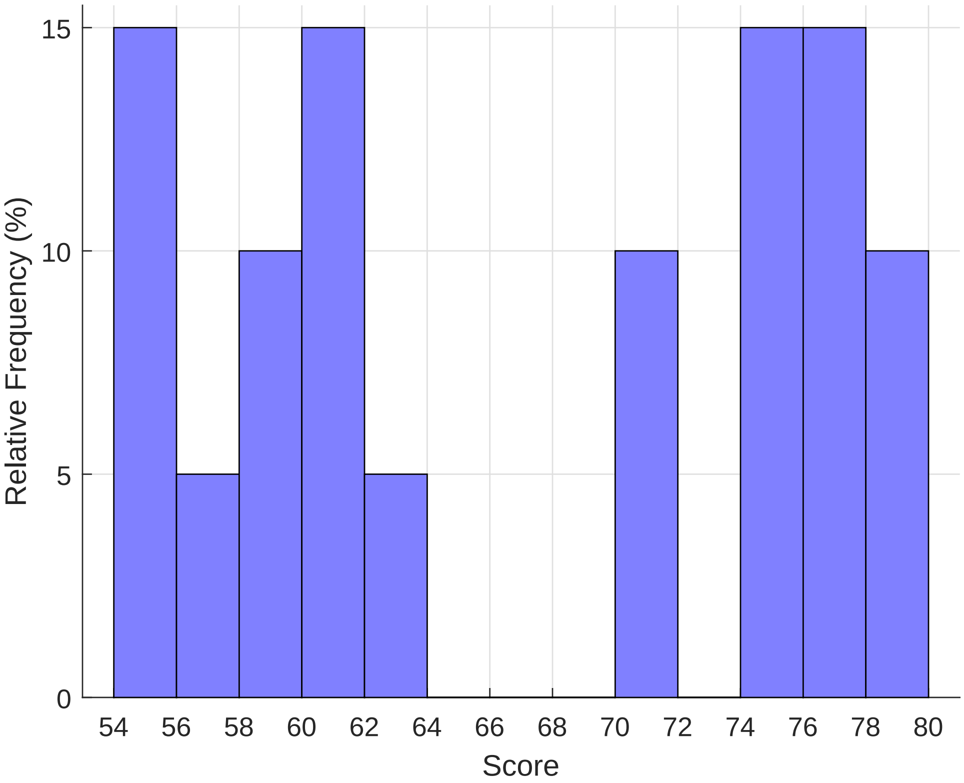
Relative Frequency Histograms

Another way to plot the distribution is to change the y-axis to represent the **relative frequency** in percent of the total number of scores. This is done by adding a third column to the table which is the percent of scores for each interval. This is simply calculated by dividing each frequency by the total number of scores and multiplying by 100. For example, the first class interval contains 3 scores, so the relative frequency is $100\frac{3}{20} = 15\%$.

This means that 15% of the scores fall below 56.

Class Interval	frequency	Relative frequency
54-56	3	15
56-58	1	5
58-60	2	10
60-62	3	15
62-64	1	5
64-66	0	0
66-68	0	0
68-70	0	0
70-72	2	10
72-74	0	0
74-76	3	15
76-78	3	15
78-80	2	10

Here's a graph of the relative frequency distribution. It looks just like the regular frequency distribution but with a different Y-axis:



We're now getting somewhere toward assigning scores to grades. You can see now that for example 10% of the scores fall in the highest class interval. This means that 100-10 = 90% fall below a score of 78. More formally, the score of 78 is called the **percentile point** and the corresponding rank of 90% is called the **percentile rank**, sometimes written as P_{90}. In shorthand, we write:

$P_{90} = 78$.

Looking at the first class interval at the other end of the distribution, you can see that 15% of the scores fall below a score of 56. In other words (or symbols):

$P_{15} = 56$.

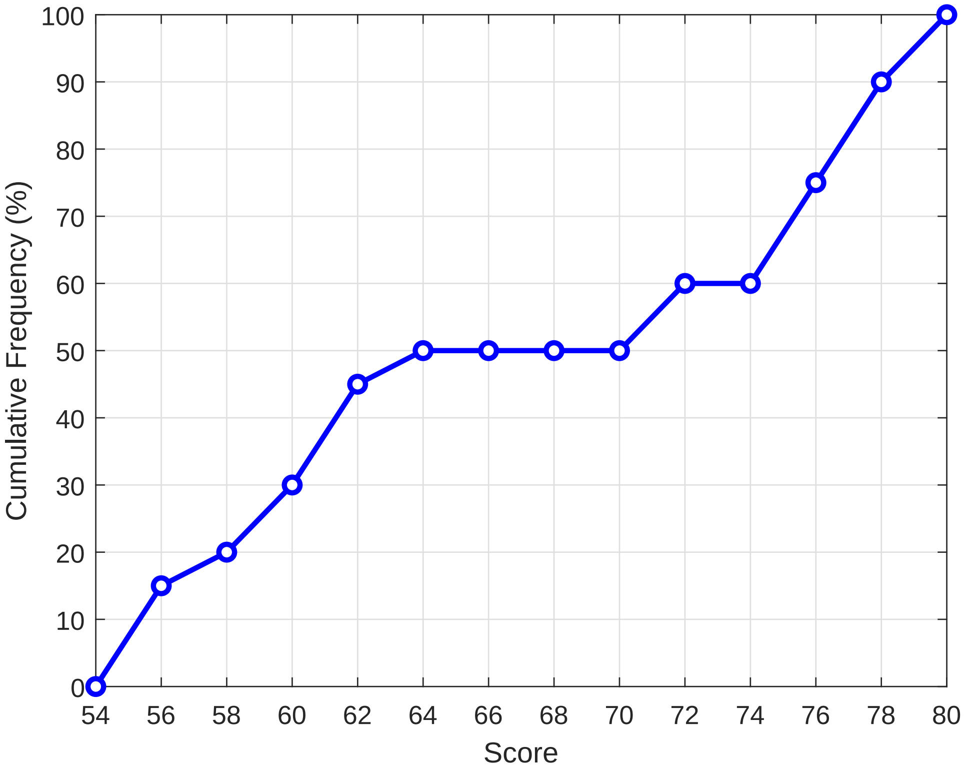
Cumulative Frequency Graph

By adding cumulatively along the class intervals, we can find out what percent of scores fall below the upper end of each class interval. Here's the result in a table:

Class Interval	frequency	Relative frequency	Cumulative frequency
54-56	3	15	15
56-58	1	5	20
58-60	2	10	30
60-62	3	15	45
62-64	1	5	50
64-66	0	0	50
66-68	0	0	50
68-70	0	0	50
70-72	2	10	60
72-74	0	0	60
74-76	3	15	75
76-78	3	15	90
78-80	2	10	100

You should see how this table shows the relationship between percentile points (upper end of each class interval) to percentile ranks (Cumulative frequency).

The cumulative relative frequency can be plotted as a line graph like this:



Frequency Histograms in R

Making histograms in R is pretty easy. As in most programming languages, there are many ways of doing the same thing. The simplest way is using R's 'hist' command.

The R commands shown below can be found here: HistogramExample.R

```
# Clear the workspace:
rm(list = ls())
# The .csv file containing the grades can be found at:
# http://www.courses.washington.edu/psy315/datasets/ExampleGrades.csv
#
# If you open up the .csv file you'll see that it contains a
# single column of numbers with the name 'Grades' as a column
# header.

# Load in the grades from the .csv file on the course website
mydata <-read.csv("http://www.courses.washington.edu/psy315/datasets/ExampleGrades.csv")
# The command 'mydata <- read.csv' loads the data into variable
# called 'mydata'.
#
# The grades are in a field defined by the column header, 'Grades'.
# We access fields of variable with the dollar sign.
# We can use 'head' to show just the first few scores:
head(mydata$Grades)
[1] 55 56 56 57 60 60

# Use 'hist' to make a histogram.
# The simplest way is like this:
hist(mydata$Grades)
# By default, R chooses the class interval and axis labels.
#
# Let's chose our own class intervals or 'breaks' using
# R's 'seq' function.  'seq' returns a sequence of numbers
# beginning with the first value, ending with the second
# value, and stepping with the third.  To generate our
# class interval boundaries, we can define a new variable
# 'class.interval' like this:
class.interval <- seq(54,80,2)
# Note, we could have called this variable whatever we want.

# You can your histogram by defining parameters like:
# 'main' for the title
# 'xlab' for the xlabel
# 'col' for the color
# 'xlim' for the x axis limits and
# 'breaks' for the class intervals:
hist(mydata$Grades,
     main="Histogram of Grades",
     xlab="Score",
     col="blue",
     xlim=c(54,80),
     breaks =class.interval
)
# I don't like R's choice for the X-axis and y-axis
# ticks.  For one thing, frequencies are whole
# numbers, so there's no reason to have 1/2 increments.
# in the y-axis.
#
# You can customize the x and y axes by first using
# 'xaxt' = n and 'yaxt' = n in 'hist' to turn off the
# x and y axis labels:
hist(mydata$Grades,
     main="Histogram of Grades",
     xlab="Score",
     col="blue",
     xlim=c(54,80),
     xaxt='n',
     yaxt = 'n',
     breaks =class.interval
)
```

```
# and then adding your own axes with the 'axis' function
# Axis 1 is 'x' and 2 is 'y':
axis(1, at=class.interval)
axis(2, at=seq(0,4),las = 1)
# In the tutorial we made a cumulative percentage curve. We can do
# this in R too.
#
# First, we'll find out how many scores fall into each class
# interval.  We aready plotted this with 'hist'.  'hist' will
# return these values if we ask it to.  Here we'll have 'hist'
# send the information into the variable 'freq', and suppress
# the plotting by using 'plot = FALSE':
freq <- hist(mydata$Grades,
     breaks =class.interval,
     plot = FALSE)
# The field 'counts' in freq holds the frequencies for the class intervals:
print(freq$counts)
 [1] 3 1 2 3 1 0 0 0 2 0 3 3 2

# Next we'll accumulate these frequencies like we did in the tutorial
# using R's 'cumsum' function.  We'll also scale it by 100 and divide
# by the total number of scores, which can be found with the 'length'
# function:
y <- 100*cumsum(freq$counts)/length(mydata$Grades)
# We'll concatinate a zero to the beginning of the list:
y = c(0,y)
# And plot:
plot(class.interval,y,
     xlab ='Score',
     ylab = 'Cumulative Frequency (%)',
     xaxt = 'n',
     yaxt ='n')
# That just made symbols.  To add lines we use:
lines(class.interval,y)
# And set our x and y axes ticks like we did with 'hist':
axis(1, at=class.interval)
axis(2, at=seq(0,100,10),las = 1)
# This should look like the cumulative frequency percentage
# curve in the tutorial
```

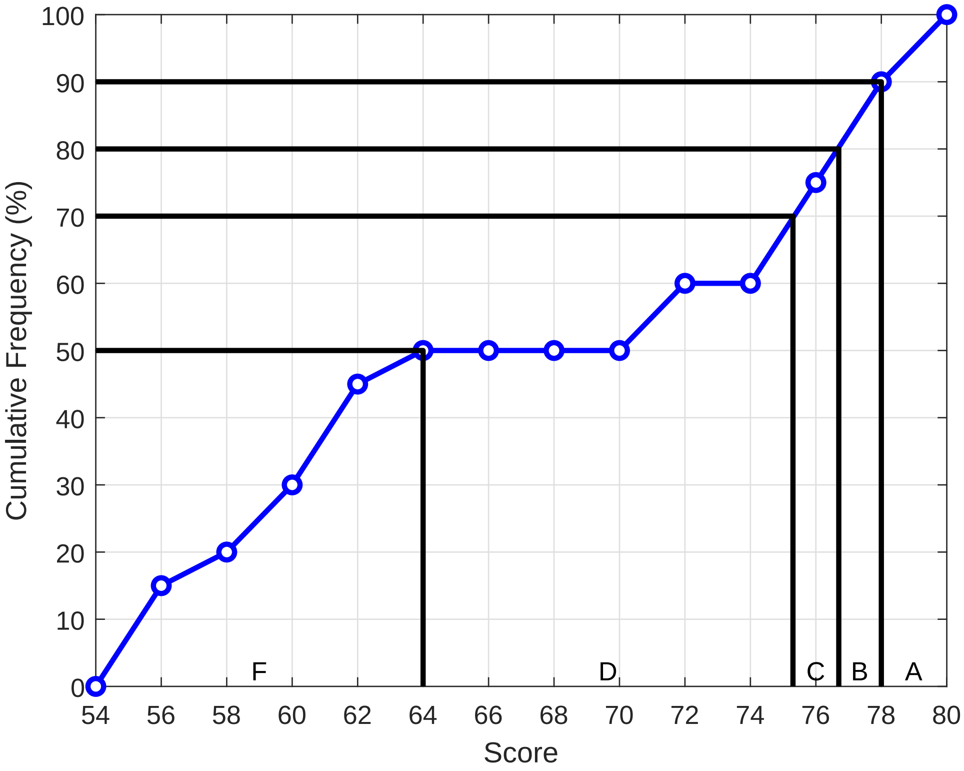
Using the Cumulative Frequency Graph to Estimate Percentile Points

We can use this graph to eyeball how to assign scores to grades. For example, remember that we wanted to assign a grade of A to the top 90% of scores. Looking at the cumulative frequency graph, take a value of 90% on the Y-axis and move rightward until you hit the cumulative frequency curve and drop down to the X-axis. This X-axis value is the corresponding percentile point, which is about 78.

A grade of B goes to scores between the 80 and the 90 percentile ranks. Looking again at the graph, this corresponds to scores roughly between 76.7 and 78.

And so on... A grade of C goes to scores between the 70 and the 80 percentile ranks. This corresponds to scores roughly between 75.3 and 76.7.

We can connect the percentile ranks and percentile points for all grades with lines on the cumulative frequency graph:



Percentile Ranks to Percentile Points, the proper way

This method using the cumulative frequency graph should be considered only as a way of estimating a way for converting percentile ranks to points. That's because the values you get depend on your choice of class intervals. The real way to do it is to use all of the scores in the distribution.

We'll go through this now. Note, **this is not covered in the book**. Also, you should know that there is not a consensus for how to do this across different computer programs. MATLAB, Excel, R, and SPSS all give slightly different answers when it comes to repeated values in the list. But the numbers are similar and for large samples they're similar enough.

The procedure we'll do here is what MATLAB uses which is the simplest, and some consider the most rational.

The first step is to make a table of raw scores, ranked from lowest to highest. We then add subsequent columns to the right. The next column counts from 1 to the total number of scores (20 for our example). We'll call these values 'C' for 'count'.

The next column is simply C-.5.

The final column is the conversion of C-.5 to percentile ranks, R, which is $\frac{(C-.5)}{n}$, or for our example, $\frac{(C-.5)}{20}$

Here's the table for our scores:

Score (P)	Rank (C)	C-.5	R = $100\frac{(C-.5)}{20}$
55	1	0.5	2.5
56	2	1.5	7.5
56	3	2.5	12.5
57	4	3.5	17.5
60	5	4.5	22.5
60	6	5.5	27.5
61	7	6.5	32.5
61	8	7.5	37.5
62	9	8.5	42.5
64	10	9.5	47.5
72	11	10.5	52.5
72	12	11.5	57.5
76	13	12.5	62.5
76	14	13.5	67.5
76	15	14.5	72.5
77	16	15.5	77.5
77	17	16.5	82.5
77	18	17.5	87.5
79	19	18.5	92.5
79	20	19.5	97.5

This table tells us the exact percentile rank (R) for every score (percentile point, P). For example, a score (or percentile point) of 64 has a percentile rank of 47.5 (or, $P_{47.5} = 64$).

Things are a just a little more complicated when we have repeated scores. For example, there are 2 scores of 79. To compute the percentile rank for 79 we take the mean of the ranks corresponding to the repeated scores: $\frac{92.5+97.5}{2} = 95$. So, therefore $P_{95} = 79$.

What about percentile ranks that are not on the list? For example, the cutoff for a grade of A is at the percentile rank of 90 which is not on the list. So, how do we find P_{90}?

Looking at the table, you can see that the percentile point for a rank of 90 must fall between the scores of 77 and 79. The exact percentile point is found using **linear interpolation**.

First we find where our percentile rank sits in the range of ranks on in the table. Since 90 is 2.5 percentile ranks above the lower range of 87.5, and the entire range is 92.5 - 87.5 = 5 percentile ranks, the percentile rank of 90 is $\frac{2.5}{5} = 0.5$ of the interval above the lower bound of 87.5.

The corresponding percentile point will therefore be 0.5 of the interval above the lower percentile point of 77. The length of the interval containing our percentile is 79-77 = 2 percentile points, so our percentile point sits (0.5)(2) = 1 percentile points above the lower percentile point of 77.

So, $P_{90} = 77 + 1 = 78$.

That was kind of ugly. It's probably easier to show how to do this with a formula. If R is the known percentile rank, then the corresponding percentile point can be calculated by:

$$P = PL + (PH - PL)\frac{(R{-}RL)}{(RH{-}RL)}$$

Where R is the known percentile rank, RH and RL are the lower and higher percentile ranks in the table that bracket R, and PH and PL are the corresponding percentile points for RH and RL. For this example:

$$P_{90} = 77 + (79 - 77)\frac{(90{-}87.5)}{(92.5{-}87.5)} = 78$$

Now it's your turn. Use the formula to find the score that is the cutoff for a grade of B. This should correspond to P_{80}.

Here's the answer:

$$P_{80} = 77 + (77 - 77)\frac{(80{-}77.5)}{(82.5{-}77.5)} = 77$$

Do these percentile points match the values we estimated above using the cumulative frequency graph? Why or why not?

You're now ready to curve the scores for the class. Verify using the formula that these are the corresponding grade ranges, assuming that grades that fall on the boundary are rounded up to the higher grade:

Grade	Scores
A	greater than or equal to 78
B	less than 78 and greater than or equal to 77
C	less than 77 and greater than or equal to 76
D	less than 76 and greater than or equal to 68
F	less than 68

Percentile Points to Percentile Ranks, the proper way

Linear interpolation is also used to go the other way - from percentile points to percentile ranks. Let's find the percentile rank for a score of 78, which is not in our list of scores.

We'll use the same logic and find the scores in our list that bracket our desired score. The formula looks a lot like the one we used to convert from percentile ranks to percentile points (in fact, you can derive it by solving that equation for R):

$$R = RL + (RH - RL)\frac{(P-PL)}{(PH-PL)}$$

Our score of 78 falls between the existing scores of 77 and 79, which correspond to percentile ranks of 87.5 and 92.5 respectively. So:

$PL = 77$, $PH = 79$, $RL = 87.5$, and $RH = 92.5$

With our percentile point of P = 78, plugging these values into the formula gives:

$$R = 87.5 + (92.5 - 87.5)\frac{(78-77)}{(79-77)} = 90$$

So for a percentile point of 78, the percentile rank is 90, or $P_{90} = 78$.

Percentile Points and Percentile Ranks in R

R has commands for computing percentile points and ranks.

The R commands shown below can be found here: PercentilePointExample.R

```
# Clear the workspace:
rm(list = ls())
# Load in the grades from the .csv file on the course website
mydata <-read.csv("http://www.courses.washington.edu/psy315/datasets/ExampleGrades.csv")
# R's function 'quantile' give you percentile points from percentile ranks.  For
# Example, here's how get P90, the percentile point for a rank of 90%
quantile(mydata$Grades,.9,type = 5)
90%
 78

# Note the option 'type=5'.  R allows for 9 different ways for computing percentile
# points!  They're all very similar.  Type 5 is the method described in the
# tutorial and is the simplest and most commonly used.

# If you want to calculate more than one percentile rank at a time, you can
# add a list of ranks using the 'c' command.  Remember, 'c' allows you to
# concatenate a list of numbers together.
#
# Let's generate the cutoff percentile points for the grades of A, B, C, D and F.
# These correspond to ranks of 90, 80, 70 and 50%.
```

```
quantile(mydata$Grades,c(.9,.8,.7,.5),type = 1)
90% 80% 70% 50%
 77  77  76  64

# Going the other way, from percentile points to ranks isn't as straightforward
# in R. The most recommended way is with the 'ecdf' function ('Emperical Cumulative
# Distribution Function').  Here's how to calculate the percentile rank for a point
# of 68:
ecdf(mydata$Grades)(68)
[1] 0.5

# You'll notice that 'ecdf' doesn't give you the exact same answers as the method
# in the tutorial. That's because it's using a different method for interpolation.
#
# For large data sets, 'ecdf' will give a number very similar to the method in the
# tutorial.
```

Your turn: Study the Weather

Let's look at the average temperatures for the month of March in Seattle over the years between 1950 and 2015.

You can download the csv file containing these temperatures here: SeattleMarchTemps.csv

What is the temperature corresponding to a percentile rank of 95?

To do:

1) Sort the temperatures from low to high

2) Create columns like those in the example for Grades above

3) Use the formula to calculate the percentile point.

Here's the answer:

$$P_{95} = 49.2 + (49.3 - 49.2)\frac{(95 - 94.2623)}{(95.9016 - 94.2623)} = 49.245$$

$P_{95} = 49.245$ degrees

By the way, the average temperature in March in 2015 was 50.5 degrees Farenheit. What can you say about the percentile rank for this temperature?

Or, if you have a computer, here's how to calculate P_{95} in R:

```
# Load in the data"
mydata <-read.csv("http://www.courses.washington.edu/psy315/datasets/SeattleMarchTemps.csv")
# Temperatures are in the field 'Temp':
head(mydata$Temp)
[1] 42.0 39.3 42.3 43.0 40.7 39.1
# The 95th percentile point is:
quantile(mydata$Temp,.95,type=5)
   95%
49.245
```