ISCAM 3: CHAPTER 0 EXERCISES

1. Weights of Olympic Rowers

Below is the roster of the 2016 U.S. Men's Olympic Rowing Team, with weight listed in pounds.

Name	Event	Wt	Name	Event	Wt
Andrew Campbell, Jr.	Lwt double sculls	155	Matt Miller	Men's four	217
Charlie Cole	Men's four	200	Rob Munn	Men's eight	215
Mike DiSanto	Men's eight	200	Tyler Nase	Lwt men's four	157
Sam Dommer	Men's eight	202	Glenn Ochal	Men's eight	210
Anthony Fahden	Lwt men's four	158	Sam Ojserkis	Men's eight	122
Nareg Guregian	Men's pair	215	Robin Prendes	Lwt men's four	160
Austin Hack	Men's eight	220	Henrik Rummel	Men's four	215
Alex Karwoski	Men's eight	200	Hans Struzyna	Men's eight	200
Steve Kasprzyk	Men's eight	229	Seth Weil	Men's four	215
Edward King	Lwt men's four	170	Anders Weiss	Men's pair	205
Joshua Konieczny	Lwt double sculls	165	Matt Miller	Men's four	217

(a) Create a dotplot of this distribution, remember to clearly label the horizontal axis.

(b) Summarize the behavior of this distribution as if to someone who could not view the graph directly. Remember to put your comments in context.

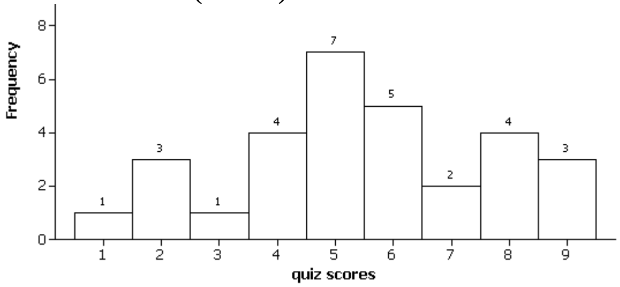
(c) Why do you think there is one weight that is much smaller than the rest?

(d) Why do you think there is a cluster of weights between 155 and 175?

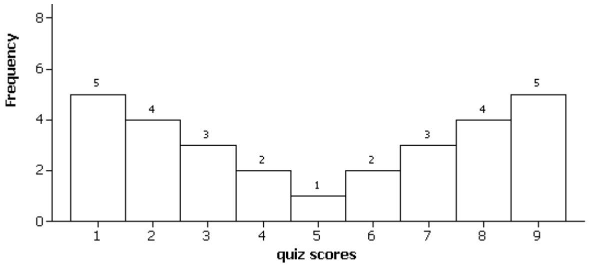
2. Hypothetical Quiz Scores

Consider the following four histograms of (hypothetical) quiz scores in four classes (scores are integers ranging from 1 to 9):

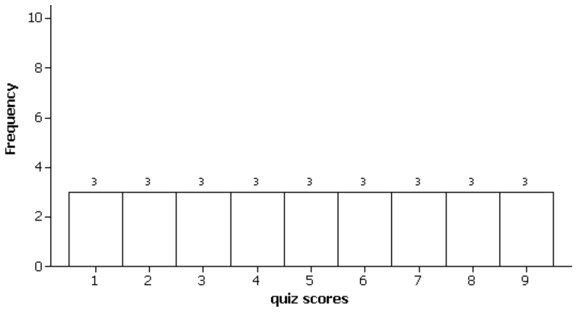
Class A (*n* = 30)



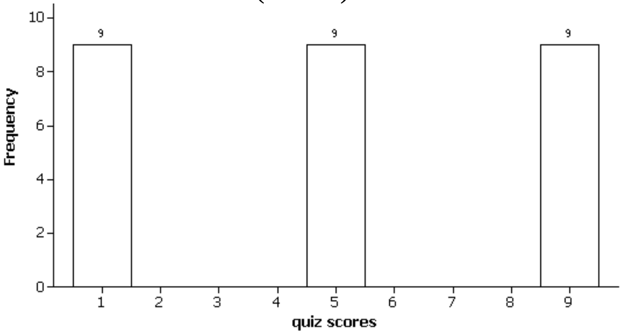
Class B (*n* = 29)



Class C (*n* = 27)



Class D (*n* = 27)

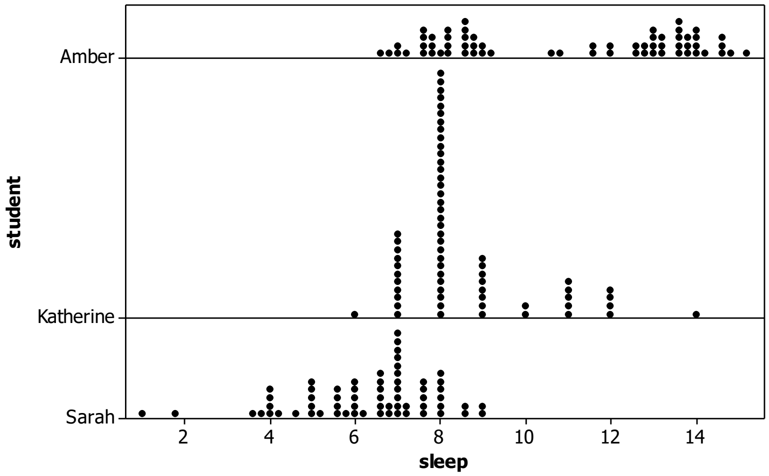


(a) Between class A and class B, which has more variability in quiz scores? [*Hint*: Think which will have the larger standard deviation and/or larger interquartile range. Explain your reasoning.]

(b) Between classes C and D, which has more variabilty in quiz scores? Explain your reasoning.

3. Student Sleep Times

The following dotplots display the distribution of sleeping times (per day, in hours) of three college students (Amber, Katherine, Sarah) for a nine-week period in the fall of 2004.



(a) One of these students developed mononucleosis during the term and so was told to get as much rest as possible for several weeks. Which student do you think this is? Explain your reasoning.
(b) One of these students is the mother of two small children. Which student do you think this is? Explain your reasoning.
(c) Which student recorded her sleeping times only to the nearest hour? Explain.
(d) Which student generally got the most sleep? Which generally got the least?

4. Nenana Ice Break

Nenana is a small, interior Alaskan town that holds a famous competition to predict the exact moment that "spring arrives" every year. The arrival of spring is defined to be the moment when the ice on Tanana River breaks, which is measured by a tripod erected on the ice with a trigger to an official clock. A contest is now held everyone to see who can predict the timing of this break. The minute at which the ice breaks has been recorded in every year since 1917. For example, the dates and times for the years 2000-2004 were:

2000	2001	2002	2003	2004
May 1, 10:47am	May 8, 1:00pm	May 7, 9:27pm	April 29, 6:22pm	April 24, 2:16pm

The data file `NenanaIceBreak2020.txt` contains all of the data from 1917 to 2020, Nenana Ice Classic. Edited by W. N. Meier and C. F. Dewes. 2020. *Nenana Ice Classic: Tanana River Ice Annual Breakup Dates, Version 2*. Boulder, Colorado USA. NASA National Snow and Ice Data Center Distributed Active Archive Center. doi: https://doi.org/10.5067/CAQ58H42LQY2. [1/19/2021]

(a) Examine and comment on graphical displays of the "day of the year" variable. [*Hint*: Look at both a time plot and a dotplot or histogram. For the latter, remember to comment on shape, center, and spread, and relate your comments to the context. You may want to use a chart like this to help track the days of the year.]
(b) Predict the date for the ice break in 2021. Which graph is more useful to you?
(c) Repeat (a)–(b) for one of the *time of day* variable (0-24 hours).

5. Age of Diabetes Diagnosis

The National Health and Nutrition Examination Survey (NHANES) is a large-scale study conducted annually by the National Center for Health Statistics. It involves over 10,000 Americans, randomly selected according to multistage sampling plans. All sampled subjects are asked to complete a survey and take a physical examination. One of the questions asked in the 2003-2004 NHANES survey pertained only to subjects who had been diagnosed with diabetes. Subjects were asked to indicate the age at which they were first diagnosed with diabetes by a health professional (diabetes.txt). The *age* column contains these data for the 548 subjects who were diagnosed with diabetes.

(a) Use the Descriptive Statistics applet to create a dotplot and/or a histogram of these data.
(b) Describe the shape of this distribution, in context, commenting on any unusual features.
(c) What would you say is a "typical" age at diagnosis based on these data?
(d) One consideration in constructing a histogram is how large to make the "bins." Change the value in the "bin width" box from 20 to 10 and press Enter. Explain why this graphical display may not be the most effective.
(e) Do you think the standard deviation would be an effective measure of the variability in this distribution? Justify your answer.

6. Old Faithful Geyser

Millions of people from around the world flock to Yellowstone Park in order to watch eruptions of Old Faithful geyser. How long does a person usually have to wait between eruptions, and has the timing changed over the years? In particular, scientists have investigated whether a 1998 earthquake lengthened the time between eruptions at Old Faithful. The data in OldFaithful.txt are the inter-eruption times (in minutes) for all 108 eruptions occurring between 6am and midnight on August 1–8 in 1978 (from Weisberg, 1985) and for 95 eruptions for the same week in 2003 (http://www.geyserstudy.org/geyser.aspx?pGeyserNo=OLDFAITHFUL).

(a) Use the Descriptive Statistics applet to create a parallel dotplots and/or histograms of these data. (Check the **Stacked** box and then paste into the Sample data box. The press the **Use Data** button.)
(b) Compare the two distributions in terms of shape, center, and variability (include the values of the means and standard deviations). What are the most striking differences between these distributions and speculate as to what you think could have caused these differences?
(c) Paste in only the 1978 data, produce and examine the timeplot. Repeat for the 2003 data and compare the graphs.
(d) Using these data, in which year (1978 or 2003) would you rather have been a visitor to see Old Faithful? Justify your answer.

7. Fan Cost Index

Every year the Team Marketing Report publishes data on the cost to attend a home game for the four major profession leagues (NFL, MLB, NBA, NHL), FanCostIndex®. The 2015 data for Major League Baseball is in MLB_FCI_15.txt. The "fan cost index" calculates the price of four adult tickets, two small draft beers, four small soft drinks, four regular-size hot dogs, parking for one car, and two (cheapest) adult-size adjustable caps.

(a) Produce a graph of the *FCIPctChange* column (don't include the "average data) and summarize what this reveals, being sure to relate your comments to the context (e.g., what does the percentage change measure here). Identify, by name, any unusual observations. Can you offer an explanation for the largest value? (If you aren't familiar with sports, do a bit of research on the 2014 MLB season.)
(b) Determine the average *FCIPctChange*. How far is the team identified in (a) from this average?
(c) Determine the *price per ounce* for a small soft drink for each team (e.g., if using the applet, use Excel to manipulate the data before pasting in?). Produce a histogram or dotplot of this variable and describe the distribution. Which team offers the best deal on a soda?
(d) Examine a dotplot of the *cap prices*. You should see some *granularity* in the data – cluster of data at specific numerical values. What are the four most common values and why does that make sense in this context? There is one observation that does not have an integer or a .99 value. Which team is that and why does it's unusual behavior make sense in this context?

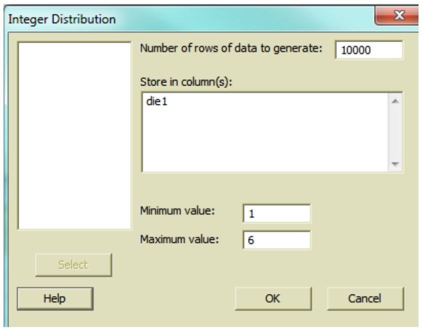
8. Random Ice Cream Prices

Suppose that an ice cream shop offers a special deal one day: The price of a small ice cream cone will be determined by rolling a pair of ordinary, six-sided dice. The price (in cents) will be the larger number followed by the smaller number. So, rolling a 2 and 5 results in a price of 52 cents, rolling a 4 and 3 produces a price of 43 cents and so on. Use R or Minitab to conduct a simulation analysis of this random process, and then perform an exact analysis of this random process.

(a) Use R or Minitab to simulate 10,000 repetitions of rolling a pair of fair, six-sided dice:

R:
```
> die1 = sample(1:6,10000, replace = TRUE)
> die2 = sample (1:6,10000, replace = TRUE)
```

Minitab: Select Calc > Random data > Integer
Number of rows to generate: 10000; Store in column(s) die1 die2;
Minimum value 1; Maximum value: 6 > OK.



Integer Distribution

Number of rows of data to generate: 10000

Store in column(s): die1

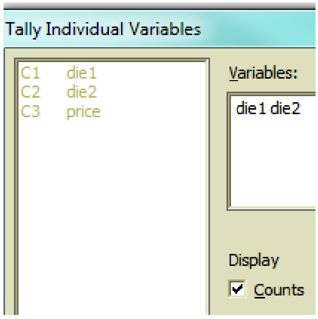
Minimum value: 1

Maximum value: 6

Tally the results of the 10,000 simulated rolls for each die.

R:
```
> table (die1); table(die2)
```

Minitab: Select Stat > Tables > Tally Individual Variables
Variables: die1 die2 > Check the Counts box under Display. Click OK.



Tally Individual Variables

C1 die1
C2 die2
C3 price

Variables: die1 die2

Display
☑ Counts

Report the tallies and comment on whether they appear to be consistent with 1–6 being equally likely results of a die roll.

(b) Use software to calculate the 10,000 ice cream prices:

R:
```
> price = 10*pmax(die1,die2)  + pmin(die1,die2)
```

Minitab: Select Calc > Calculator > Store result in variable price;
Expression: `10*RMAX(die1,die2) + RMIN(die1,die2)` then click OK.

Explain what the command does and why it calculates the price correctly.

(c) Produce a histogram of the 10,000 prices:

R: `> hist(price)`

Minitab: Select Graph > Histogram > Simple > OK. Graph variables: price, then click OK.

Submit this graph, and comment on what it reveals about the distribution of these prices.

(d) Calculate and report the average of these 10,000 prices:

R: `> mean(price)`

Minitab: `MTB> desc 'price'`

Now suppose that you walk into the ice cream shop having only two quarters and no other money.

(e) Approximate the probability that the price is 50 cents or less by determining the proportion of these 10,000 prices that are 50 cents of less:

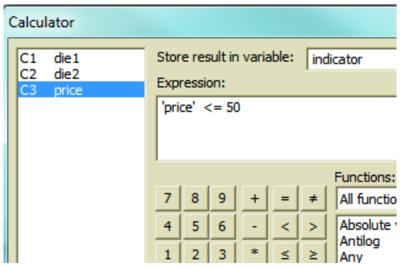
R: `> table(price)`

Minitab: Select Stat > Tables > Tally Individual Variables
Variables: price
Check the Cumulative counts box under Display. Click OK.

Or, create an indicator variable that equals 1 if the condition is satisfied, 0 if not, as follows:

R: `> indi50 = (price <= 50); table(indi50)`

Minitab: Select Calc > Calculator > Store result in variable indicator;
Expression: `'price' < = 50` then click OK.



Calculator

C1 die1
C2 die2
C3 price

Store result in variable: indicator

Expression:
'price' <= 50

Functions:
All functio
Absolute
Antilog
Any

Then tally the values in the indicator column.

Exact Analysis

(f) To determine the exact probabilities, start by listing all 36 possible and equally likely outcomes (ordered pairs) of rolling two six-sided dice.

(g) Determine the (exact) probability that the price is 50 cents or less by counting how many outcomes produce such a price and dividing by the total number of possible outcomes. Is the approximate probability from your simulation close to this exact probability?

(h) Suppose that I offer to pay for your ice cream cone if the price turns out to be an odd number. Determine the (exact) probability that I pay for your ice cream cone.

Different Dice

You will now investigate how much difference it would make to use a pair of fair, *ten*-sided dice that contain the numbers 0–9 on the ten sides. The rule for determining the price is the same: larger number followed by smaller number (in cents).

(i) Use software to simulate 10,000 repetitions of rolling a pair of fair, *ten*-sided dice containing the numbers 0–9 on the ten sides. (*Hint*: Use essentially the same instructions as before, with one or two small adjustments.) Produce and submit a histogram of the resulting prices. Also determine and report the average price. Comment on how this average price has changed from using six-sided dice. Is the change what you would have expected? Explain briefly.

(j) Use your simulation results to approximate the probability that the price is 50 cents or less. Then determine the exact, theoretical probability that the price is 50 cents or less. (*Hint*: You do not have to submit a complete list of all possible outcomes, but you do need to consider all possible outcomes. Be sure to explain/show how you determine these probabilities.) Comment on how this probability has changed from using six-sided dice.

9. Consecutive Lottery Numbers

The California Lottery's SuperLottoPlus game involves picking five numbers between 1 and 47. Many people who play the lottery refuse to choose consecutive integers among their five number choices, believing that it's very unlikely for consecutive numbers to appear? But how unlikely is it? We will use simulation to approximate this probability.

(a) First, make a *guess* for the probability that consecutive numbers appear.

Simulation Analysis:

(b) Use R or Minitab to conduct a simulation of the California game by randomly selecting 5 numbers (without replacing) from the numbers 1 – 47, repeating this process 25 times.

R:

```
winner <- sample(1:47,5)
```

Minitab: To sample *without replacement* in Minitab, you need to create a column of the values you want to sample from:

```
set c1
1:47
end
```

Then you can randomly sample from that column, saving the results into another:

```
sample 5 c1 c2
```

Repeat this process a total of 25 times, (you can simply copy-and-paste the commands above over and over), keeping a running tally of whether or not the (sorted) set includes consecutive integers.
Report the proportion of these 25 drawings in which the numbers include at least one set of consecutive integers.

(c) Now we want to automatic this process to run it a very large number of times.

R:

```
> winner = 0; y =0; diff=0
> for (j in 1:1000){
    winner <- sample(1:47,5)
    sortwin = sort(winner)
    for (i in 1:4) diff[i] = sortwin[i+1]-sortwin[i]
    y[j] = (min(diff)==1)
  }
  table(y)
```

Minitab: Save the worksheet you had been using (with C1) into a specific directory. Save the `lottery.mac` file from the textbook page into the same directory. (You can view this file using a text editor.) The main commands are:

```
Do k1=1:1000
   sample 5 c1 c2
   sort c2 c3
   diff c3 c4
   let c5(k1)=(min(c4)=1)
endDo
```

Then in Minitab, you should be able to execute the macro:

```
%lottery.mac
tally c5
```

Write a description of what each commands does, and explain how the computer achieves the goal of determining whether a drawing produces at least one set of consecutive integers.

(d) From your results in (c), report the estimate of the probability that this lottery game produces at least one set of consecutive integers. How does this compare to your prediction in (a)?

Exact Analysis:

(e) How many different drawings are possible? In other words, how many ways are there to choose five numbers from 47 numbers?

(f) It can be shown that when choosing *k* numbers between 1 and *n*, the number of ways to get consecutive integers is $\binom{n}{k} - \binom{n-k+1}{k}$. Use this result to determine the number of different drawings that include consecutive integers with the California game.

(g) Use your answers to (e) and (f) to determine the theoretical probability that the five winning numbers in the California game will include consecutive integers.

(h) How does the theoretical probability compare to your simulation results in (d)?

(i) Even though picking consecutive numbers won't increase your chances of winning in any one drawing, explain another advantage to doing so considering other players' tendencies.

10. Roulette

An American roulette wheel has 38 slots: 18 contain black numbers, 18 red numbers, and 2 green numbers. The wheel is spun and the ball falls at random into one of the 38 slots. If you bet $1 on a color (either red or black) and win (the ball lands on the color you picked), you receive $2 for a net gain of $1. If you lose, your "net gain" is –$1. If you bet $1 on a number (1–36) and win, you receive $36 for a net gain of $35. Let the random variable *X* denote your *net* gain from one bet on a color, and let *Y* be your *net* gain from one bet on a number.

(a) Specify the probability distribution of *X* (net gain in a color bet). [*Hint*: List the (two) possible values and their probabilities (to four decimal places), where the 38 slots are equally likely to occur.]

(b) Perform a simulation of this probability distribution. First, set up the probability distribution (outcomes and probabilities) and then sample with replacement from that distribution.

R:

```
> x= c(-1,1); prob=c(.5263, .4737)
> outcomes=sample(x, 1000, replace = TRUE, prob)
```

Minitab: Enter the values –1 and 1 into C1 and their respective probabilities (.5263, .4737) in the corresponding rows in C2.

```
random 1000 c3;
discrete c1 c2.
```

Simulates 1000 spins of roulette wheel using the probability distribution specified by c1 and c2

Then look at the total "gain" for these 1000 bets and the average net gain among these 1000 bets.

R:

```
>  sum(outcomes); mean(outcomes)
```

Minitab:

```
sum c3
mean c3
```

Report these values.

(c) Example the behavior of the average gain as you increase the number of spins:

R:

```
> spinnum=1:1000
> avgcolor = cumsum(outcomes)/spinnum
> plot(spinnum, avgcolor, type="l")
```

Minitab:

```
set c4
1:1000
end
let c5=parsum(c3)/c4
name c4 'spin number' c5 'cum avg'
plot c5*c4;
connect.
```

Include and describe the behavior of this graph. Explain what these commands have done (*Hint*: You may want to look at the cumsum or parsum output directly to see what that command does.)

(d) Now calculate the theoretical expected value of *X*. How does this result compare to your simulation results? [Show your work and use the proper notation, e.g., *E*(*X*).]

(e) Write an interpretation of the expected value calculated in (d). (*Hint*: Use the results from (c).)

(f) In a similar manner, specify the probability distribution of *Y*, produce a simulation of 1000 (or more) spins, and examining the behavior of the average gain across the numbers of spins, and calculate the theoretical expected value of *Y*.

(g) How do the two expected values for the two bets (color vs. number) compare? Does this mean that the two bets should be considered equivalent? Explain.

(h) Now determine and compare the theoretical *variance* of the gain for the two bets. How do they compare? (*Hint*: Create a graph as in (c) and consider how the variability in the two sets of results compares.)

11. Early Die Game

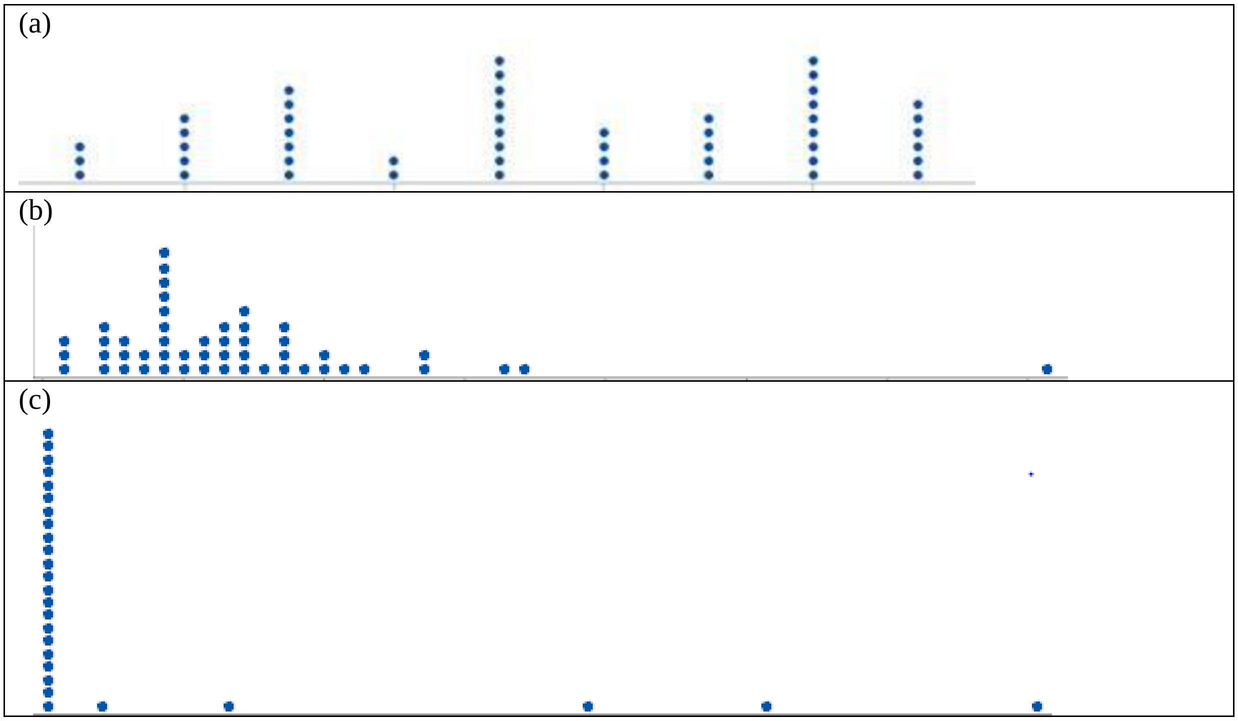
In the 1600s, French nobleman Chevalier de Mere and mathematician Blaine Pascal developed formal study of probability. One of the first questions they looked at was “which is more likely, rolling at least one six in four rolls of a die, or rolling at least one pair of double sixes in 24 rolls of a die.” Use simulation and/or probability rules to answer this question.

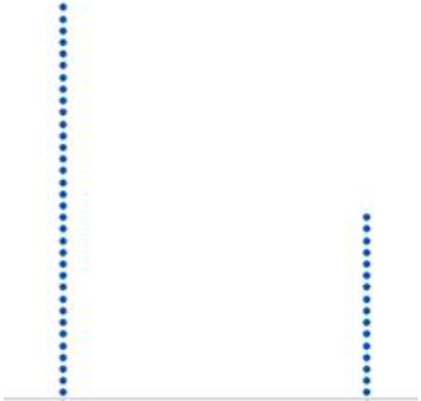
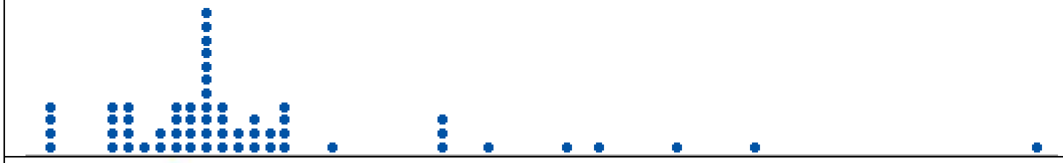
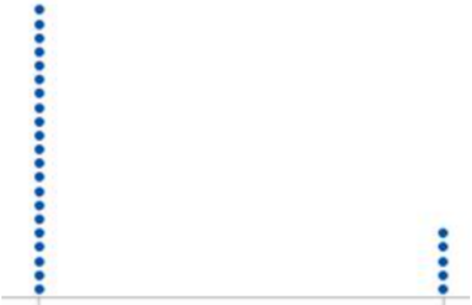
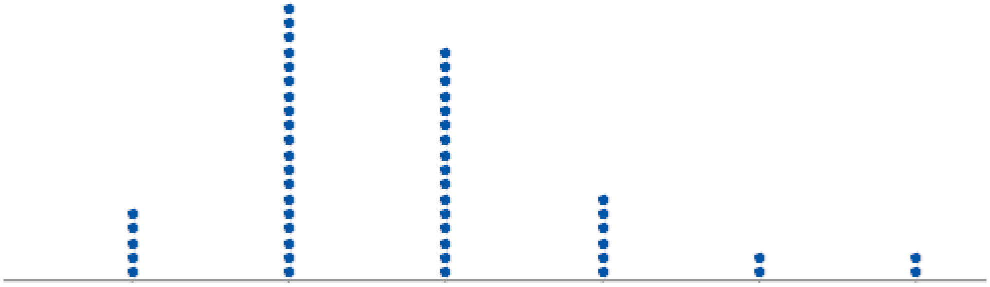
New problems

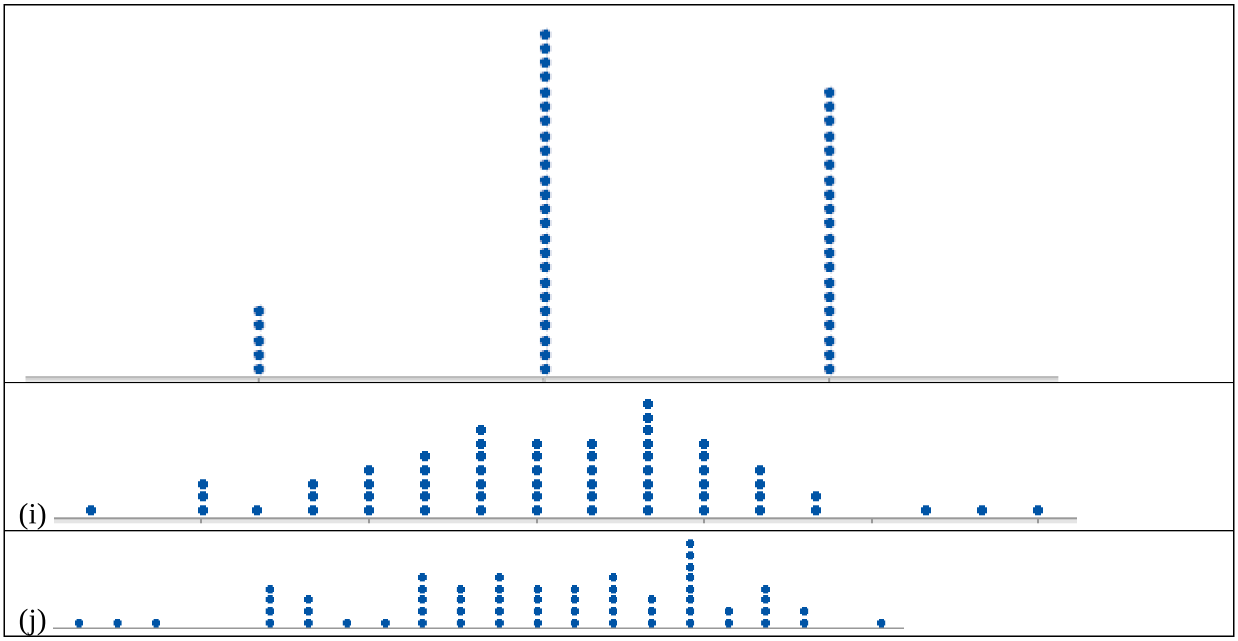
12. Match the variable to the graph

Below are some variables and some graphs. Match the graph with the variable. Write a paragraph explaining how you decided which graph belonged with which variable. (You can cite “process of elimination” for at most one graph but should give justifications for the others, clearly state any assumptions you make along the way. For example, you might consider whether reasonable numerical values can be placed along the horizontal axis as well as what shape you expect the distribution to have.) You will be graded more on your justification than the correctness of your matches.

- Heights of students in class
- Do you use a Mac or a PC?
- Number of siblings
- Number of states visited
- Political inclination (conservative, moderate, or liberal)
- Amount of change in pockets (dollar amount)
- Number of heads reported in 50 tosses of a coin
- Cost of last haircut
- Ratings of the value of statistics on a scale of 1 – 9.
- Do you prefer Coke or Pepsi?



(d) 
(e) 
(f) 
(g) 
(h)



13. Interpreting probability

Reconsider what you learned about the definition of *probability* from Investigation B. Use the same long-run relative frequency interpretation of probability to *interpret* what we mean by "probability" in each of the following statements. (*Hint*: Don't use the words probability, chance, or likelihood in your interpretation.)

(a) The probability of getting a red M&M candy is 0.2. [*Hint*: For 20% of . . . what happens? . . .]
(b) The probability of a four-of-a-kind in a five-card poker hand is 0.0240%.
(c) There is a 30% probability of rain tomorrow.
(d) I heard that Clemson has a 53% probability of winning the game against LSU.