Sample Final Exam

NAME:

MAT 128/SOC 251, SPRING 2018

Each question is worth 10 points. You are allowed one 8 1/2" x 11" sheet of paper with hand-written notes on both sides.

1. The CSV file `citiesHistPop.csv` contains the historical population for several north-east American cities. Write a **complete** Python program that plots the population of New York over time. Assume the CSV file is in the same directory as your .py file.

   The graph should be a line graph and have a title.

   `citiesHistPop.csv`:

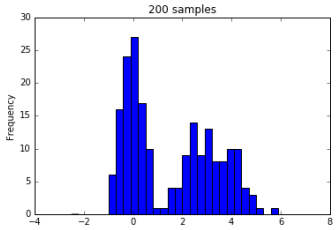
```
The data in this file comes from::,,,
http://www1.nyc.gov,,,
http://www.iboston.org,,,
http://physics.bu.edu/~rednerl,,,
Year,Boston,New York,Philadelphia
1900,560892,3437202,1293697
1910,670585,4766883,1549008
1920,748060,5620048,1823779
1930,781188,6930446,1950961
1940,770816,7454995,1931334
1950,801444,7891957,2071605
1960,697197,7783314,2002512
1970,641071,7894798,1948609
1980,562994,7071639,1688210
1990,574994,7322564,1585577
2000,590433,8008278,1517313
```

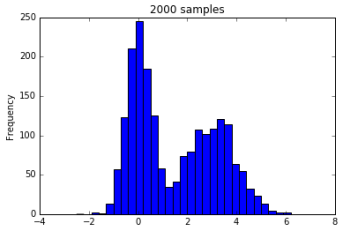
   Answer:

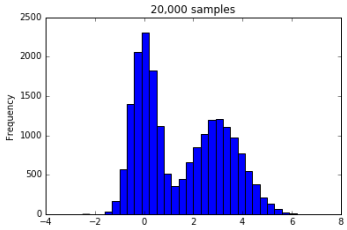
```
import pandas as pd

pop = pd.read_csv("citiesHistPop.csv",skiprows = 3)
pop.plot(x = "Year", y = "New York", title = "Historical population of New York")
```

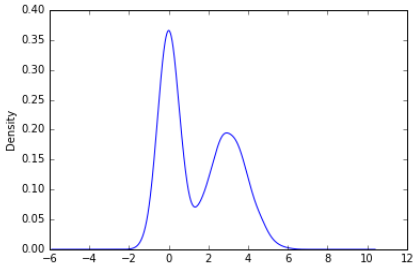
2. (a) Below are 3 histograms. Each histogram is generated by sampling from the same unknown distribution a different number of times. Draw a sketch of the distribution.



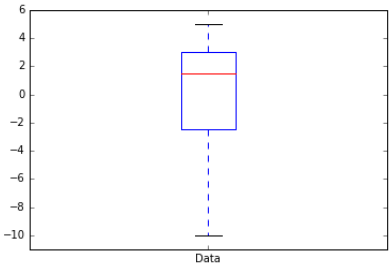




Answer:



(b) Below is a boxplot of some data. Use it to answer the following questions about the data:



(i) What is the minimum data value?
(ii) What is the maximum data value?
(iii) What is the median?
(iv) What is the 25th percentile?
(v) What is the 75th percentile?

Answer: Note: As this question involves reading values off of a graph, your answer just needs to be close to the correct answer to receive full marks.

(i) -10
(ii) 5
(iii) 1.5
(iv) -2.5
(v) 3

3. Assume the NBA dataset (see last page) has been loaded into the dataframe `nba` and that all necessary import statements are already included.

   (a) Write a piece of Python code to calculate and print the variance of the number of free throws made by a player during the season.

   Answer:

```
v = nba['free_throws'].var()
print("The variance in the number of free throws is", v)
```

   (b) Write a piece of Python code to calculate and print the maximum number of points scored by a starting forward (SF) who is 25 years or older.

   Answer:

```
sf_filter = nba['position'] == 'SF'
age_filter = nba['age'] >= 25

filtered_nba= nba [sf_filter & age_filter]
m  = filtered_nba['pts'].max()
print("The maximum number of points is",m)
```

4. Assume the NBA dataset (see last page) has been loaded into the dataframe `nba` and that all necessary import statements are already included.

   (a) Write a piece of Python code that uses the NBA dataset to estimate the probability that a player is a center (position is C) and print this probability.

   Answer:

```
center_filter = nba['position'] == 'C'
num_centers = len(nba[center_filter])
num_players = len(nba)
prob_center = num_centers/num_players
print("The estimated probability that a player is a center is",prob_center)
```

   (b) What is the formula for computing the probability that a center player scores more than 800 points in a season?

   Answer:

$$\text{probability that a center player scores more than 800 points} = \frac{\text{\# of center players who score more than 8000 points}}{\text{\# of center players}}$$

5. Assume the NBA dataset (see last page) has been loaded into the dataframe `nba` and that all necessary import statements are already included.

   The following Python code creates a linear model from the data in `nba`:

   ```
   lm = smf.ols(formula = 'pts ~ games', data = nba).fit()
   ```

   (a) What is this model predicting?

   Answer: The number of points scored by a player in a season.

   (b) What information is this model using to make the prediction?

   Answer: The number of games played by a player during a season.

   (c) Running the code `print(lm.params)` gives the following output:

   ```
   Intercept   -204.083706
   g             13.532706
   dtype: float64
   ```

   What does the number `13.532706` represent?

   Answer: The slope of the best fit line or the number of additional points scored by a player for each game played.

   (d) Write the line of Python code to create and fit a linear model that predicts the number of points scored in a season from the number of games started and the number of free throw points scored in a season.

   Answer:

   ```
   lm2 = smf.ols(formula = 'pts ~ games_started + free_throws', data = nba).fit()
   ```

   (e) Write a piece of Python code to compute and print out the R-Squared value of your model from part (d). **Answer:**

   ```
   print("The R-squared value is",lm2.rsquared)
   ```

6. Assume the NBA dataset (see last page) has been loaded into the dataframe `nba` and that all necessary import statements are already included.

   (a) The correlation matrix for the columns `age`, `games`, `games_started`, `free_throws`, and `pts` in the `nba` dataframe is shown below.

	age	games	games_started	free_throws	pts
age	1.000	-0.012.	0.025	-0.047	-0.012
games	-0.012	1.000	0.611	0.598	0.728
games_started	0.025	0.611	1.000	0.707	0.810
free_throws	-0.047	0.598	0.707	1.000	0.928
pts	-0.012	0.728	0.810	0.928	1.000

(i) Which two columns are the most correlated?

Answer: The columns free_throws and pts are the most correlated.

(ii) Which two columns are the least correlated?

Answer: The columns age and games are the least correlated.

(b) Write a piece of Python code to compute and print out the mean number of games started in during the season, as well as the 90% confidence interval.

Answer:

```
mean_games_started = nba['games_started'].mean()
sd_games_started = nba['games_started'].std()
num_data = len(nba)

ci = st.t.interval(0.9,num_data - 1,loc = mean_games_started, scale = sd_games_s
print("90\% Confidence interval for mean games started is",ci)
```

7. (a) Write a **complete** Python program that computes 500 samples from the normal distribution with a mean of 7 and a standard deviation of 1.5, and plots the samples as a histogram.

The histogram should have 15 bins and a title.

Answer:

```
import random as r

samples_list = []
for i in range(500):
    num = r.gauss(7,1.5)
    samples_list.append(num)
samples = pd.Series(samples_list)
samples.plot.hist(bins = 15, title = "Samples from a normal distribution")
```

(b) Suppose the mean length (in inches) of a New York squirrel's tail is 8 with a standard deviation of 0.5. A park ranger randomly catches a sample of 60 squirrels, measures the lengths of their tails, and computes the mean. What distribution does this sample mean come from? Why?

Answer: By the Central Limit Theorem, the sample mean comes from a normal distribution with a mean of 8 and a standard deviation of $\frac{0.5}{\sqrt{60}}$

8. This question refers to the NBA dataset on the last page.

Our (alternative) hypothesis is that players who start more than 50 games are more likely to score more than 500 points than players who start 50 games or fewer.

(a) What is the null hypothesis?

Answer: Players who start more than 60 games are just as likely to score more than 500 points as players who start 60 games or fewer.

(b) To test our hypothesis, we count the number of players in the data meeting a certain criteria and compare it with a histogram. This histogram is made by counting the number of players meeting the same criteria in samples simulated by assuming the null hypothesis is true.

(i) When we count the number of players in the data meeting a certain criteria, what was that criteria?

Answer: The number of players who have started more than 60 games and scored more than 500 points.

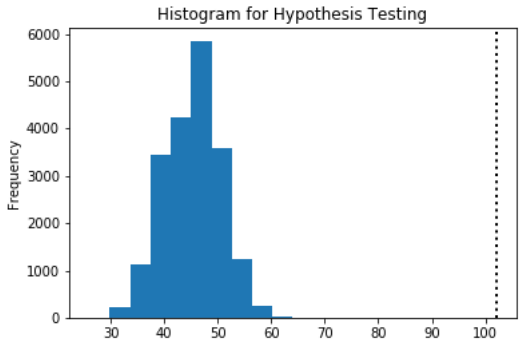
(ii) How many players are simulated with one sample?

Answer: The number of players who started more than 60 games.

(iii) Assume each simulated player is assigned 1 if they scored more than 500 points and 0 if they scored 500 points or fewer. How do you compute the probability that a player is assigned a 1?

Answer: The probability that a player is assigned a 1 is the number of players in the data who scored more than 500 points divided by the total number of players in the data.

(c) The histogram and count from the data (shown as a dashed line on the histogram plot) are below. Would you reject the null hypothesis? Why or why not?



Answer: I would reject the null hypothesis because the data count is much bigger than the possible counts in the histogram that were simulated assuming the null hypothesis is true. Therefore, it seems extremely unlikely that our data count came from this distribution.

9. The CSV file `citiesHistPop.csv` from Question 1 contains the historical population for several north-east American cities. Write a **complete** R program that

computes and prints out the maximum population of Philadelphia.

citiesHistPop.csv:

```
The data in this file comes from::,,,
http://www1.nyc.gov,,,
http://www.iboston.org,,,
http://physics.bu.edu/~rednerl,,,
Year,Boston,New York,Philadelphia
1900,560892,3437202,1293697
1910,670585,4766883,1549008
1920,748060,5620048,1823779
1930,781188,6930446,1950961
1940,770816,7454995,1931334
1950,801444,7891957,2071605
1960,697197,7783314,2002512
1970,641071,7894798,1948609
1980,562994,7071639,1688210
1990,574994,7322564,1585577
2000,590433,8008278,1517313
```

Answer:

```
cities <- read.csv("citiesHistPop.csv",skip = 3)
max_pop <- max(cities$Philadelphia)
print(paste("The maximum population of Philadelphia was",max_pop))
```

10. Using the ggplot2 library, write a piece of R code to plot the number of games started in a season (x-axis) vs. the number of free throws in a season (y-axis) for all players on the New York Knicks (NYK) and the Brooklyn Nets (BRK). Assume the NBA dataset (see last page) has been loaded into the dataframe `nba`. The plot should have:

    - the title "NYC Team Statistics"
    - each data point plotted as a point
    - the points be colored by team

Answer:

```
nyc_teams <- subset(nba,subset = team == 'NYK' | team == 'BRK')

ggplot(nyc_teams,aes(x = games_started, y = free_throws, color = factor(team))) +
  geom_point() + ggtitle("NYC Team Statistics")
```

NBA Dataset

The following dataset is used for questions 3, 4, 5, 6, 8, and 10. It contains statistics on all NBA players for the 2012 -2013 season.
Assume that it has been read in from a csv file and is stored as a Pandas dataframe in the variable `nba`.

player	position	age	team	games	games_started	free_throws	pts
Quincy Acy	SF	23	TOT	63	0	35	171
Steven Adams	C	20	OKC	81	20	79	265
Jeff Adrien	PF	27	TOT	53	12	76	362
Arron Affalo	SG	28	ORL	73	73	274	1330
Alexis Ajinca	C	25	NOP	56	30	56	328
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
Tyler Zeller	C	24	CLE	70	9	87	399

The columns are:

- player = name of NBA player
- position = position of player (eg. C = center, SF = starting forward)
- age = age of player
- team = player's team name, abbreviated
- games = number of games played by the player in the season
- games_started = number of games started in by the player in the season
- free_throws = number of successful free throws during the season
- pts = total number of points scored by player during the season