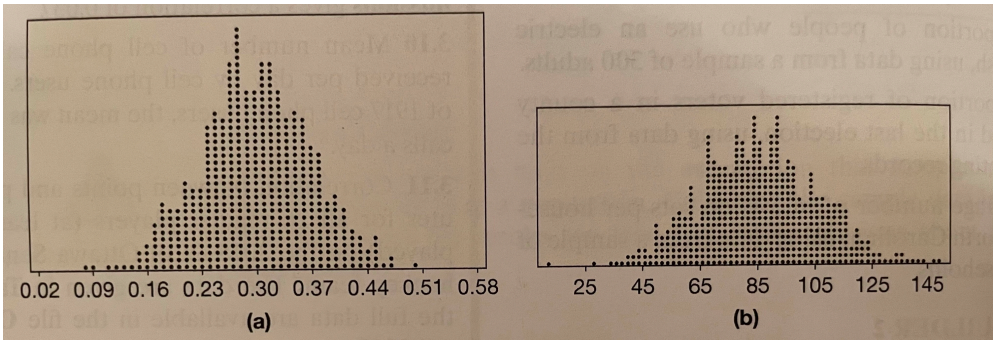
Answers below Questions

Math 341 Spring 2021, Homework 5

1a. Graph (a) below is a sampling distribution of sample proportions from samples of size $n = 40$. Use the graph to give estimates for the value of the population parameter and the standard error for the sample statistic. Use the correct notation.

b. Graph (b) below shows sample means from samples of size $n = 30$ from a population. Use the graph to give estimates for the value of the population parameter and the standard error for the sample statistic. Use the correct notation.



2. Use the figures above to answer the following questions.

a. For the sampling distribution in part (a), how likely are these sample proportions? For each, decide whether it is (i) reasonably likely to occur from a sample of this size, (ii) unusual but might occur occasionally, or (iii) extremely unlikely to ever occur.

$\hat{p} = 0.35$

$\hat{p} = 0.1$

$\hat{p} = 0.6$

b. For the sampling distribution in part (a), how likely are these sample means? For each, decide whether it is (i) reasonably likely to occur from a sample of this size, (ii) unusual but might occur occasionally, or (iii) extremely unlikely to ever occur.

$\overline{x} = 70$

$\overline{x} = 100$

$\overline{x} = 140$

3. For each of the following, construct an interval giving a range of plausible values for the given parameter using the given sample statistic and margin of error.

a. For μ, using $\overline{x} = 20$ with margin of error 4.

b. For p, using $\hat{p} = 0.35$ with margin of error 0.01.

4. Is each of the following true or false?

a. A population parameter varies.

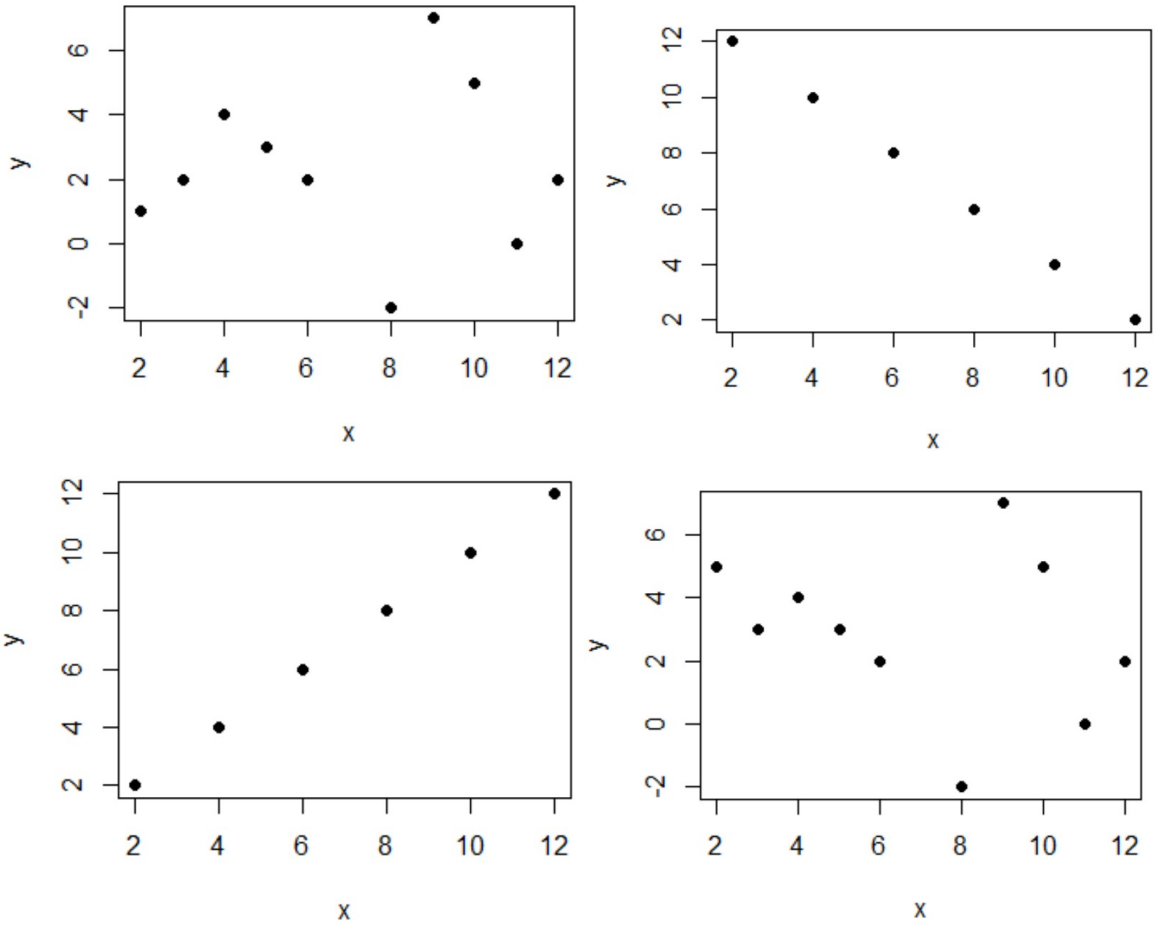
b. A sample statistic varies from sample to sample.

c. If samples are randomly chosen from the population, then the center of the sampling distribution should be very close to the value of the population parameter.

d. The standard error (SE) is the standard deviation of the sampling distribution.

e. The standard error (SE) is the standard deviation of one sample.

5. Match the scatterplots with the correlations: $r = -1$, $r = -0.24$, $r = 0.05$, $r = 1$. Write the each correlation on the corresponding scatterplot.



6. Two quantitative variables are given in each part below. Do you expect a positive or negative association between the two variables?

a. Size of a house and cost to heat the house.

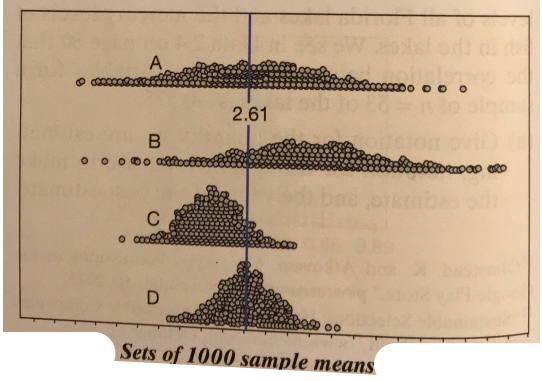
b. Distance driven since the last fill-up of the gas tank and amount of gas left in the tank.

c. Amount of time spent studying and grade on the exam.

7. US Census data from 2010 listed the average household size as 2.61. The Figure below shows possible distributions of means for 1000 samples of household sizes.

(a) Assume that two of the distributions show results from 1000 random samples, while two others show distributions from a sampling method that is biased. Which two dotplots appear to show samples produced using a biased sampling method? Explain your reasoning. Pick one of the distributions that you listed as biased and describe a sampling method that might produce this bias.

(b) For the two distributions that appear to show results from random samples, suppose that one comes from 1000 samples of size $n = 100$ and one comes from 1000 samples of size $n = 500$. Which distribution goes with which sample size? Explain.



8.

Proportion of US Residents Less Than 25 Years Old The US Census indicates that 35% of US residents are less than 25 years old. Figure 3.8 shows possible sampling distributions for the proportion of a sample less than 25 years old, for samples of size $n = 20$, $n = 100$, and $n = 500$.

(a) Which distribution goes with which sample size?

(b) If we use a proportion $\hat{p}$, based on a sample of size $n = 20$, to estimate the population parameter $p = 0.35$, would it be very surprising to get an estimate that is off by more than 0.10 (that is, the sample proportion is less than 0.25 or greater than 0.45)? How about with a sample of size $n = 100$? How about with a sample of size $n = 500$?

(c) Repeat part (b) if we ask about the sample proportion being off by just 0.05 or more.

(d) Using parts (b) and (c), comment on the effect that sample size has on the precision of an estimate.

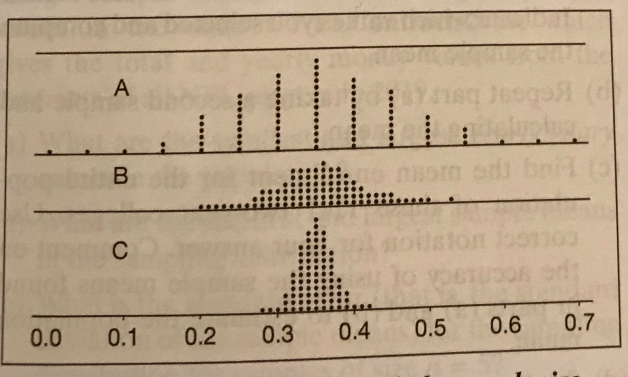


Figure 3.8 *Match the dotplots with the sample size*

9. ***Is it getting harder to win a hot dog eating contest?*** Every Fourth of July, Nathan's Famous in New York City holds a hot dog eating contest, in which contestants try to eat as many hot dogs as possible in 10 minutes. In 2019, thousands of people watched the event live on Coney Island, and it was broadcast live to many more on ESPN. HotDogs2019 gives the winning number of hot dogs eaten for each year through 2019. The R code is in RcodeHotDogs.R.

a. Find the mean and standard deviation of the number of hot dogs eaten.

b. How many of the 18 values are above the mean?

c. How many are above the mean in the years 2002-2010?

d. How many are above the mean in the years 2011-2019?

f. Create a scatterplot of the data, using the year as the explanatory variable.

g. Is the trend in the data mostly positive or negative?

h. Is the residual (actual value - predicted value (aka vertical deviation)) larger in 2007 or 2008? Is the residual positive or negative in 2014?

i. Find the correlation coefficient.

j. Find the equation of the regression line. What does its slope represent? Interpret it in terms of hot dogs.

k. Use the regression line to predict the winning number of hot dogs in 2020.

l. Did Nathan's Famous Hot Dog Eating Contest occur in 2020? If so, what was the winning number of hot dogs? How does that compare to the prediction you gave in part (f)? That is, compute the 2020 residual.

m. Why would it not be appropriate to use this regression line to predict the winning number of hot dogs in 2030?

10. A little theory.

In class we showed $\hat{y} = b + mx$ is the line of regression for data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ where:

$$m = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \quad \text{and} \quad b = \bar{y} - m\bar{x}$$

We also discussed: $r = \frac{1}{n-1} \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y}$

r ↑ correlation coefficient

The covariance of x and y measures the joint variability of x and y and is defined by:

$$\text{cov}(x, y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

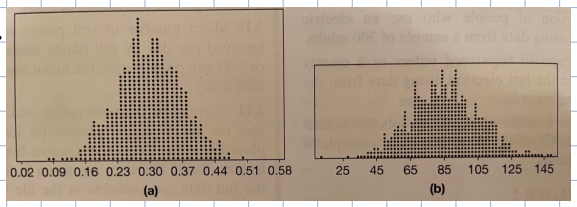
a. How is $\text{cov}(x, x)$ related to s_x, the sample standard deviation of x?

b. Write the correlation coefficient, r, in terms of $\text{cov}(x, y)$

c. Show $\text{cov}(x, y) = \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right)$.

d. Write the slope of the regression line, m, in terms of $\text{cov}(x, y)$.

1.



a) Sample: $n = 40$

$\bar{x} = .27$

$SE = .05$

b) Sample: $n = 30$

$\bar{x} = 85$

$SE = 10$

2. a) i 0.35
ii 0.1
iii 0.6

b) i 76
ii 190
iii 140

3. a) $\bar{x} = 20$ error of 4

$\bar{x} \pm$ error

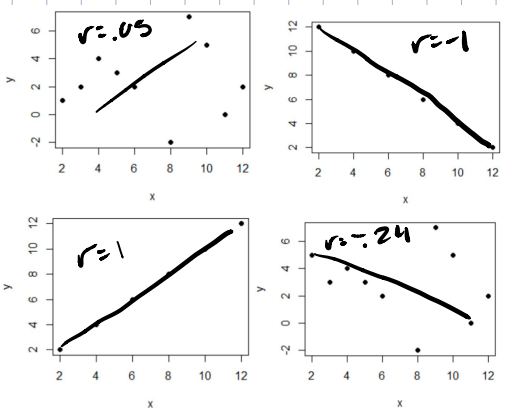
$20 \pm 4 \Rightarrow$ interval $(16, 24)$

b) $\hat{p} = .35$ error: .01

$.35 \pm .01 \Rightarrow (.34, .36)$

4. a) False
b) True
c) True
d) True
e) False

5.

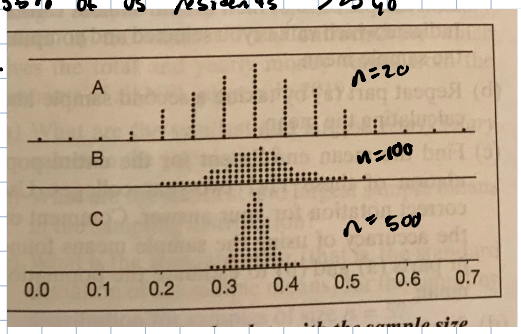


6. a) Positive
   b) Negative
   c) Positive

7. a) Graphs c and d were most likely produced using biased sampling because they are very concentrated at the center of the graph. There are not many outliers for the data. A sampling method that might produce this bias could be cluster sampling, which only selected groups with similar views
   b) Most likely graph A came from the sample of 500 because the mean is more centered and lined up with the other graphs. Where graph B's center is more to the right than the other graphs.

8. 35% of US residents >25 yo

   a.



   b. More than .10, No
   Sample size 100, No
   Sample size 50, Yes

   c. Being off by .05 more, Yes, No, No

   d. The smaller the n, the smaller margin of error

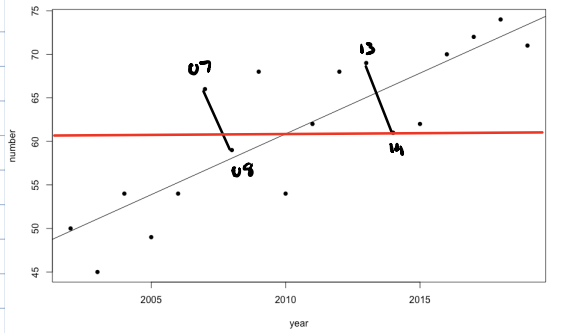
9. a. Mean: 61.556
SD: 8.833

b. 10

c. 2

d. 8

f.



g. Trend line is positive.

h. Larger in 2007
Negative in 2014

i. r = .8432

j. $y = 1.4x - 2743.6$

The slope represents about how many hot dogs eaten increase every year approximately.

k. Based on the regression line, 72.4 hot dogs will be eaten in 2020

l. Yes, the winning number of hot dogs was 75. Based on my graph, it was off by a little over 2 hot dogs. The residual is $75 - 72.4 = 2.6$

m. It would not be appropriate to use the regression line for 2030 because it likely will not go up 2.6 hot dogs a year and in 2030 be approximately 101 hot dogs. Also, the same person has won 12 of the last 13 contests and will likely start dropping his numbers.

```r
#Code for Nathan's Famous Hot Dog Eating Contest
hotdogs<-read.csv(paste("https://docs.google.com/spreadsheets/d/e/2PACX-1vRVF",
                  "F8zXttXwLKISNuv_0RoW4AfwsWLNxjINK5jV7XoRdqamzxZwY_XfzR3ZSzS",
                  "vwtRElDQtrfP2QZw/pub?output=csv",sep=""),header=T)

year<-hotdogs$Year
number<-hotdogs$HotDogs

plot(year,number,pch=16)
abline(lm(number~year), col ="red")


mean <- mean(number)
sd <- sd(number)

count <- 0
for (val in number) {
  if(val > mean)
    count = count= count + 1
}
print(count)

cor(number, year)

coef(lm(number~year))

paste('y =', coef(lm(number~year))[[2]], '* x', '+', coef(lm(number~year))[[1]])
```

```
> mean <- mean(number)
> sd <- sd(number)
>
> count <- 0
> for (val in number) {
+   if(val > mean)
+     count = count= count + 1
+ }
> print(count)
[1] 10
>
> cor(number, year)
[1] 0.8432479
>
> coef(lm(number~year))
 (Intercept)         year
-2743.600275     1.395253
>
> paste('y =', coef(lm(number~year))[[2]], '* x', '+', coef(lm(number~year))[[1]])
[1] "y = 1.39525283797733 * x + -2743.60027519787"
>
```

10. a. $Cov(x,x)$ and sample standard deviation are related because when you divide $cov(x,x)$ by the sample sd, your result is r.

b. $r = \frac{cov(x,x)}{S_x \cdot S_y}$

c. $r = \frac{1}{(n-1)S_x S_y} \sum_{i=1}^{n} (x_i y_i - n\bar{x}\bar{y})$

d. $\frac{cov(x,y)}{S_x \cdot S_y} = \frac{1}{(n-1)S_x S_y} \sum_{i=1}^{n} (x_i y_i - n\bar{x}\bar{y})$

$cov(x,y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i y_i - n\bar{x}\bar{y})$

e. $m = \frac{\sum_{i=1}^{n} (x_i y_i - n\bar{x}\bar{y})}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{cov(x,y)(n-1)}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$