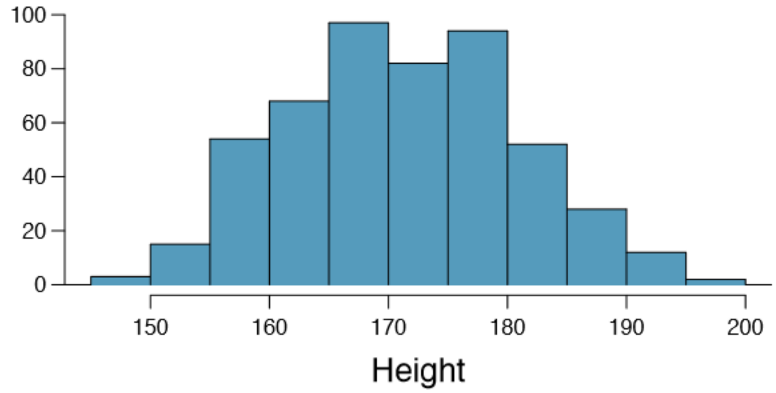
Leland Randles DATA606

<u>Homework, Chapter 4</u>

4.4 Heights of adults. Researchers studying anthropometry collected body girth measurements and skeletal diameter measurements, as well as age, weight, height and gender, for 507 physically active individuals. The histogram below shows the sample distribution of heights in centimeters.[38]



Min	147.2
Q1	163.8
Median	170.3
Mean	171.1
SD	9.4
Q3	177.8
Max	198.1

(a) What is the point estimate for the average height of active individuals? What about the median? **(See the next page for parts (b)-(e).)**

(b) What is the point estimate for the standard deviation of the heights of active individuals? What about the IQR?

(c) Is a person who is 1m 80cm (180 cm) tall considered unusually tall? And is a person who is 1m 55cm (155cm) considered unusually short? Explain your reasoning.

(d) The researchers take another random sample of physically active individuals. Would you expect the mean and the standard deviation of this new sample to be the ones given above? Explain your reasoning.

(e) The sample means obtained are point estimates for the mean height of all active individuals, if the sample of individuals is equivalent to a simple random sample. What measure do we use to quantify the variability of such an estimate (Hint: recall that $SD_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$)? Compute this quantity using the data from the original sample under the condition that the data are a simple random sample.

(a) I loaded the dataset in R as 'bdims'. The point estimate is given by mean(bdims$hgt), which returns 171.1438, which matches the mean given in the chart above. The median is provided in the chart as well (170.3), or in R - median(bdims$hgt) also returns 170.3.

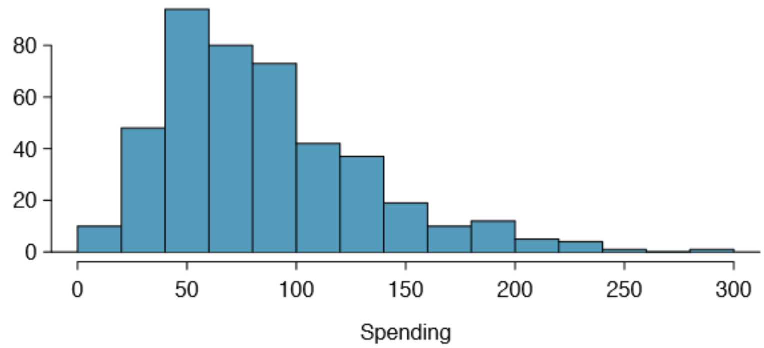
(b) Is provided in the chart above (9.4) or can be computed in R - sd(bdims$hgt) returns 9.407205.

(c) The interquartile range is shown in the chart Q1-Q3, which is 163.8 to 177.8. Alternatively, if we run summary(bdims$hgt) in R, it returns the same range for 1st Qu. and 3rd Qu.

(d) No, I would not expect a different random sample to have the same mean and standard iation point estimates, because there will be variability between samples.

(e) The measure used to quantify the variability of such an estimate is the "standard error of the sample mean". We can calculate in R - sd(bdims$hgt) / sqrt(507) returns 0.4177887.

4.14 Thanksgiving spending, Part I. The 2009 holiday retail season, which kicked off on November 27, 2009 (the day after Thanksgiving), had been marked by somewhat lower self-reported consumer spending than was seen during the comparable period in 2008. To get an estimate of consumer spending, 436 randomly sampled American adults were surveyed. Daily consumer spending for the six-day period after Thanksgiving, spanning the Black Friday weekend and Cyber Monday, averaged \$84.71. A 95% confidence interval based on this sample is (\$80.31, \$89.11). Determine whether the following statements are true or false, and explain your reasoning.



(a) We are 95% confident that the average spending of these 436 American adults is between \$80.31 and \$89.11.

(b) This confidence interval is not valid since the distribution of spending in the sample is right skewed.

(c) 95% of random samples have a sample mean between \$80.31 and \$89.11.

(d) We are 95% confident that the average spending of all American adults is between \$80.31 and \$89.11.

(e) A 90% confidence interval would be narrower than the 95% confidence interval since we don't need to be as sure about our estimate.

(f) In order to decrease the margin of error of a 95% confidence interval to a third of what it is now, we would need to use a sample 3 times larger.

(g) The margin of error is 4.4.

(a) False. We are about 95% confident that the average spending for all American adults (the full population) are within this range.

(b) False. Though the sample is skewed, and there are some high-side outliers, the sample size is large enough to accept the skew.

(c) True – approximately. If we took an infinite number of samples, the percentage which had a sample mean within the range would approach 95%.

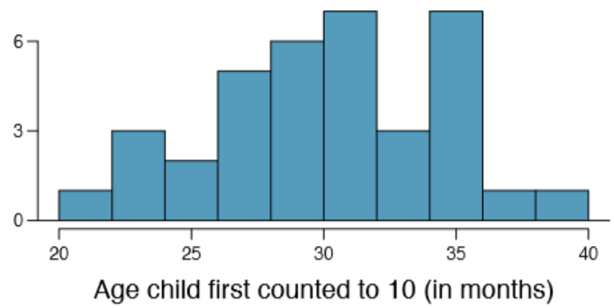
(d) True, though I would say "about 95% confident".

(e) True, a 90% confidence interval would be narrower.

(f) False. We don't need to change the sample size to chanage the margin of error, we just increase the number of standard deviations used.

(g) True. Z times the SE is the margin of error, which can be deterined by the ± from the mean, which in this case is 4.4.

4.24 Gifted children, Part I. Researchers investigating characteristics of gifted children collected data from schools in a large city on a random sample of thirty-six children who were identified as gifted children soon after they reached the age of four. The following histogram shows the distribution of the ages (in months) at which these children first counted to 10 successfully. Also provided are some sample statistics.[43]



n	36
min	21
mean	30.69
sd	4.31
max	39

(a) Are conditions for inference satisfied?

(b) Suppose you read online that children first count to 10 successfully when they are 32 months old, on average. Perform a hypothesis test to evaluate if these data provide convincing evidence that the average age at which gifted children fist count to 10 successfully is less than the general average of 32 months. Use a significance level of 0.10.

(c) Interpret the p-value in context of the hypothesis test and the data.

(d) Calculate a 90% confidence interval for the average age at which gifted children first count to 10 successfully.

(e) Do your results from the hypothesis test and the confidence interval agree? Explain.

(a) Yes. Based on the description provided, the sample observations are independent. Secondly, the sample size is greater than or equal to 30. Lastly, the population is not strongly skewed.

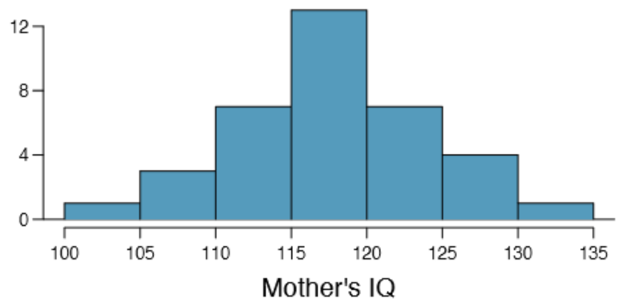
(b) z <- (32 - mean(gifted$count)) / ((sd(gifted$count) / sqrt(36)))
pnorm(-abs(z))
0.03472969 (p-value)

(c) The p-value states that probably of observing our sample mean given a true null hypothesis is only 0.03472969. Thus we would reject the null hypothesis.

(d) H_0 – The average age at which chidren first count to 10 successfully is the same for gifted children as children in general.
H_A – The average age at which children first count to 10 successfully is different for gifted children than children in general.
In R:
> gifted <- read.csv("C:/Users/Lelan/Desktop/gifted.csv", header=TRUE, sep=",")
> m <- mean(gifted$count)
> sd <- sd(gifted$count)
> se <- sd / sqrt(36)
c(m - (qnorm(0.95)*se),m + (qnorm(0.95)*se))
[1] 29.51155 31.87734
The mean for the general population doesn't fall within the confidence interval, so we would reject the null hypothesis.

(e) Yes, in both cases we are rejecting the null hypothesis.

4.26 Gifted children, Part II. Exercise 4.24 describes a study on gifted children. In this study, along with variables on the children, the researchers also collected data on the mother's and father's IQ of the 36 randomly sampled gifted children. The histogram below shows the distribution of mother's IQ. Also provided are some sample statistics.



n	36
min	101
mean	118.2
sd	6.5
max	131

(a) Perform a hypothesis test to evaluate if these data provide convincing evidence that the average IQ of mothers of gifted children is different than the average IQ for the population at large, which is 100. Use a significance level of 0.10.

(b) Calculate a 90% confidence interval for the average IQ of mothers of gifted children.

(c) Do your results from the hypothesis test and the confidence interval agree? Explain.

(a) Compute p-value in R:

```
> m <- mean(gifted$motheriq)
> sd <- sd(gifted$motheriq)
> se <- sd /sqrt(36)
> z <- (m - 100) / se
> 2*pnorm(-abs(z))
[1] 5.077477e-63
```

The p-value is easily within the signicance level, so we would reject the null hypothesis.

(b)
```
c(m - (qnorm(0.95)*se), m + (qnorm(0.95)*se))
[1] 116.3834 119.9499
```

(c) Yes. The mean for the regular population does not fall within the confidence interval, so we would reject the null hypothesis.

4.34 CLT. Define the term "sampling distribution" of the mean, and describe how the shape, center, and spread of the sampling distribution of the mean change as sample size increases.

The sampling distribution of the mean is a term to describe the distribution of sample means from multiple samples taken from the population. As the sample size increases, the shape of the distribution becomes more normal, more sharp in the middle, and the spread becomes smaller.

4.40 CFLBs. A manufacturer of compact fluorescent light bulbs advertises that the distribution of the lifespans of these light bulbs is nearly normal with a mean of 9,000 hours and a standard deviation of 1,000 hours.

(a) What is the probability that a randomly chosen light bulb lasts more than 10,500 hours?

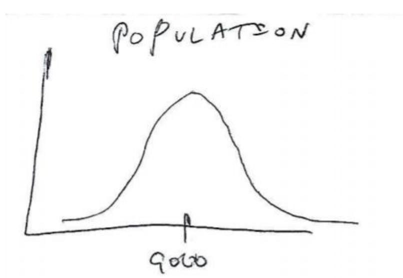
(b) Describe the distribution of the mean lifespan of 15 light bulbs.

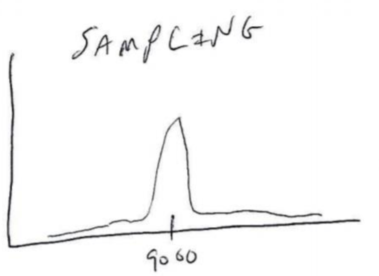
(c) What is the probability that the mean lifespan of 15 randomly chosen light bulbs is more than 10,500 hours?

(d) Sketch the two distributions (population and sampling) on the same scale.

(e) Could you estimate the probabilities from parts (a) and (c) if the lifespans of light bulbs had a skewed distribution?

(a) > (10500 - 9000) / 1000
    [1] 1.5
    > pnorm(1.5)
    [1] 0.9331928
(b) The sample mean is estimated as 9,000.
    > 1000 / sqrt(15)
    [1] 258.1989 # Standard Error of the Mean
(c) Nearly 100%
    > 1000 / sqrt(15)
    [1] 258.1989
    > (10500 - 9000) / 258.1989
    [1] 5.809475
    > pnorm(5.809475)
    [1] 1
(d) See sketch below
(e) No, to use the normal distribution the distribution cannot be skewed





4.48 Same observation, different sample size. Suppose you conduct a hypothesis test based on a sample where the sample size is $n = 50$, and arrive at a p-value of 0.08. You then refer back to your notes and discover that you made a careless mistake, the sample size should have been $n = 500$. Will your p-value increase, decrease, or stay the same? Explain.

Your p-value will increase because the standard error will increase, which means the number of standard errors from the mean will decrease, which means the p-value will increase.