Illapani_Homework_HW4

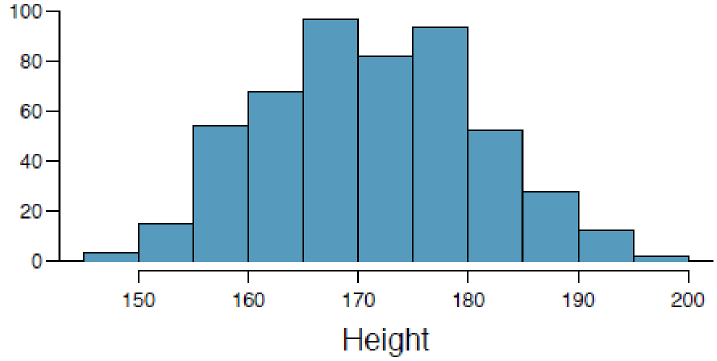
Srini Illapani

October 7, 2015

Foundations for Inference

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.



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?

- From the graph above, The point estimate for the average height of active individuals is 171.1. The median is 170.3

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

- The point estimate for the standard deviation of hieghts of active individuals is 9.4. The IQR is 163.8 - 177.8

(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.

```
ztall <- (180 - 171.1) / 9.4
pnorm(1 - ztall)

## [1] 0.5212103

zshort <- (155 - 171.1) / 9.4
pnorm(1 - zshort)

## [1] 0.9966638
```

- The p-value (pnorm) for both the tall and shorter heights is high (higher than 0.05), it is likely to observe the tall and short heights even if the null hypothesis were true. But these heights are with in the two standard deviations from the mean. So they are not unusually tall or short.

(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.

- The mean and standard deviation of the new sample may or may not be same, although the chances are very slim of them being the same. The random samples of individuals would yield different results but they should not be that far from the numbers we see above.

(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¯x = ! pn)? Compute

this quantity using the data from the original sample under the condition that the data are a

simple random sample.

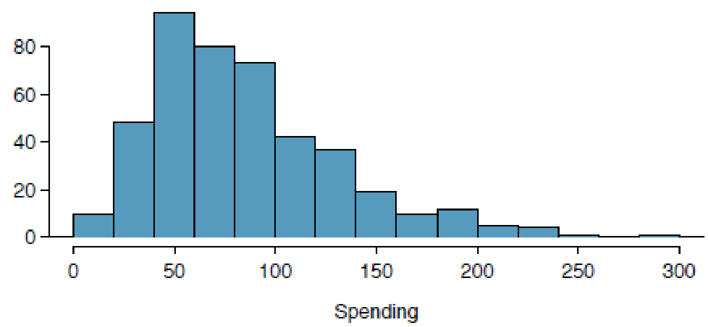
```
sample_error <- 171.1 / sqrt(507)
sample_error

## [1] 7.598818
```

- The variability of the point estimates is measured using the Standard Error (SE), it is the standard deviation of the sampling distribution. The SE in this case is 7.6

4.14 Thanksgiving Spending, Part I

The 2009 holiday retail season, which kicked o??? 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.

- False. We are 95% confident that the average spending of the total Amaerican adults "population", not "sample of 436" has spent between \$80.31 and \$89.11

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

- False. The sample data is slightly right skewed but mostly distributed normal.

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

- False. It is 95% confidence interval not the percentage of the sample.

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

- True. As explained for (d)

(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.

- True. the lower the CI, the narrower the area covered in the plot.

(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.

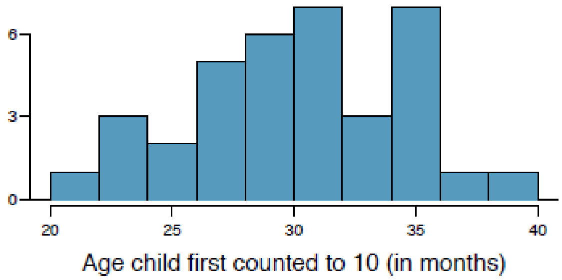
- False. If we use 3 times the sample, the SE = 0.054. Which is still higher than the third of what is now (0.031)

(g) The margin of error is 4.4

- True

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.



n	36
min	21
mean	30.69
sd	4.31
max	39

(a) Are conditions for inference satisfied?

- Yes. The children in the sample are random and independent of each other. The sample size is less than 10% of the population. And the distribution is not extremely skewed.

(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.

Ho: u = 32, Ha: u < 32, sample mean x = 30.69, The sample population s = 36, standard deviation n = 4.31, alpha = 0.1

```
z = (30.69 - 32) / (4.31 /sqrt(36) )
z

## [1] -1.823666

p_val = pnorm(z)
p_val

## [1] 0.0341013
```

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

- The p-value is less than 0.10. The data does not provide convincing evidence that the gifted childrens average age is 32 months. Hence we reject the null hypotheses.

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

```
  z = 1.645
  SE = 4.31 / sqrt(36)
```

```
  ci_lower = 30.69 - (z * SE)
  ci_upper = 30.69 + (z * SE)
  confidence_interval = c(ci_lower, ci_upper)
  confidence_interval

## [1] 29.50834 31.87166
```

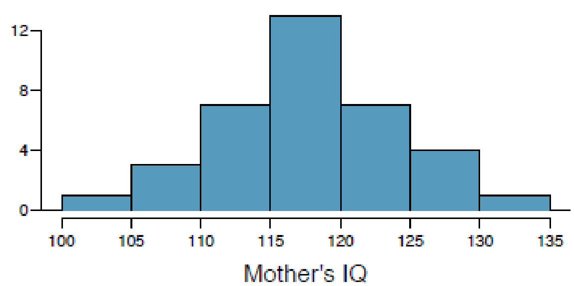
- for a 90% confidence Interval, the range is 29.50834 to 31.87166

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

- Yes, the results agree because the range of the confidence interval is outside the average of 32 months for the gifted children.

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 di???erent than the average IQ for the population at large,

which is 100. Use a significance level of 0.10.

Ho: u = 100, mother's IQ is equal to the average population IQ; Ha: u > 100, mother's IQ is greater than the average population IQ; sample mean x = 118.2, The sample population s = 36, standard deviation n = 6.5, alpha = 0.1

```
z = (118.2 - 100) / (6.5 /sqrt(36) )
z

## [1] 16.8

p_val = 1 - pnorm(z)
p_val

## [1] 0
```

- The p-value is less than 0.10. The data does not provide convincing evidence that the gifted children mothers average IQ is 100. Hence we reject the null hypotheses.

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

```
z = 1.645
SE = 6.5 / sqrt(36)
ci_lower = 118.2 - (z * SE)
ci_upper = 118.2 + (z * SE)
confidence_interval = c(ci_lower, ci_upper)
confidence_interval

## [1] 116.4179 119.9821
```

- for a 90% confidence Interval, the range is 116.4179 to 119.9821

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

- Yes, the results agree because the range of the confidence interval is outside the average IQ of 100.

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 the distribution of mean values from repeated samples from a population. The shape is approximately a normal curve, with a center at the population mean. The shape more closely approximates the normal distribution as more samples are taken and included. This also will move the center closer to the population mean. The sampling distribution of sample proportions is also nearly normal when the sample is sufficiently large.

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?

```
z = (10500 - 9000) / (1000)
z

## [1] 1.5

p_val = 1 - pnorm(z)
p_val

## [1] 0.0668072
```

- The probability is 6.7%

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

- The distribution for a random sample of 15 light bulbs would most likely be a normal one with the average lifespan around 9,000 hrs.

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

Ho: u = 10,500; Ha: u > 10,500; sample mean x = 9000, The sample population s = 15, standard deviation n = 1000, alpha = 0.05

```
z = (9000 - 10500) / (1000 /sqrt(15) )
z

## [1] -5.809475

p_val = pnorm(z)
p_val

## [1] 3.133452e-09
```

- The probability value is very low and hence the chances of picking 15 light bulbs with 10,500 hrs of mean life span is remote.

(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?

- One of the conditions necessary to apply the normal distribution for a given experiment or population is that the skewness should be small or weak. If presented with a skewed distribution, some other techniques could be used.

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.

- Lets compare the SE, z and P-value for both the sample sizes. As we can see from the results, if n increases, SE goes down, z increases and the p-value goes down.

```
SE1 = 1 / sqrt(50)
SE2 = 1 / sqrt(500)

c(SE1, SE2)

## [1] 0.14142136 0.04472136
```

```r
# lets assume the mean and sample mean values of 100 and 101 for both
scenarios

z1 = (101 - 100) / SE1
z2 = (101 - 100) / SE2

c(z1, z2)

## [1]  7.071068 22.360680

p_val1 = 1 - pnorm(z1)
p_val2 = 1 - pnorm(z2)

c(p_val1, p_val2)

## [1] 7.687184e-13 0.000000e+00
```