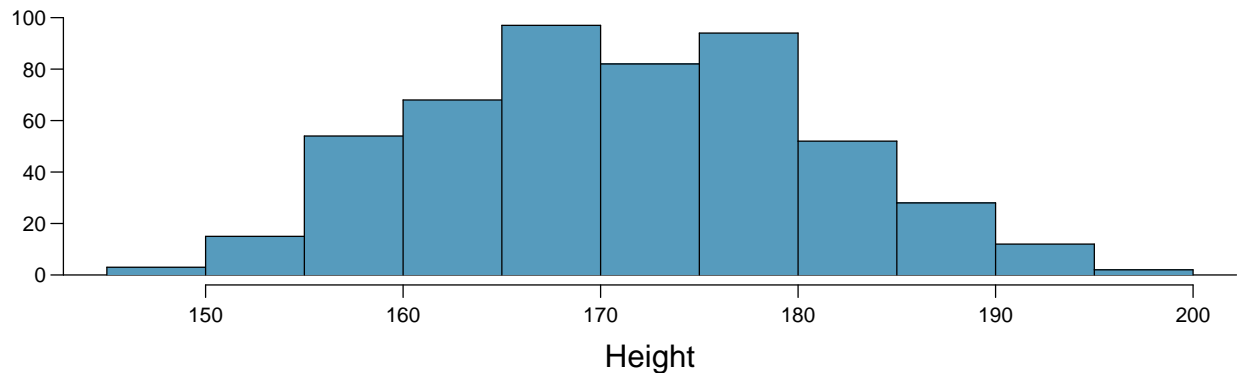


Chapter 5 - Foundations for Inference

Cameron Smith

Heights of adults. (7.7, p. 260) 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.



- What is the point estimate for the average height of active individuals? What about the median?
- What is the point estimate for the standard deviation of the heights of active individuals? What about the IQR?
- 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.
- 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 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 $SD_x = \frac{\sigma}{\sqrt{n}}$)? Compute this quantity using the data from the original sample under the condition that the data are a simple random sample.

Answers

- The estimate for avg height (i.e. mean) is 171.1, and the estimated is 170.3.
- The SD is 9.4, and the IQR is 14.
- Given that the mean is approximately 171, and the SD is 9.4, someone who is 180 would be about 1 SD more than “normal” people, i.e. in the top 32%.
- I would expect it to be similar (given that it is a normal distribution), but likely not the same due to random chance. That being said though, if the sample size were large enough then it would probably be the same.
- The measure we use is standard error (SE), and in this case it is .418.

```
# Calculate summary stats
bdims$ht %>% summary()
```

```
##      Min. 1st Qu.  Median      Mean 3rd Qu.      Max.
##    147.2   163.8   170.3   171.1   177.8   198.1
```

```
bdims$hgt %>% sd()
```

```
## [1] 9.407205
```

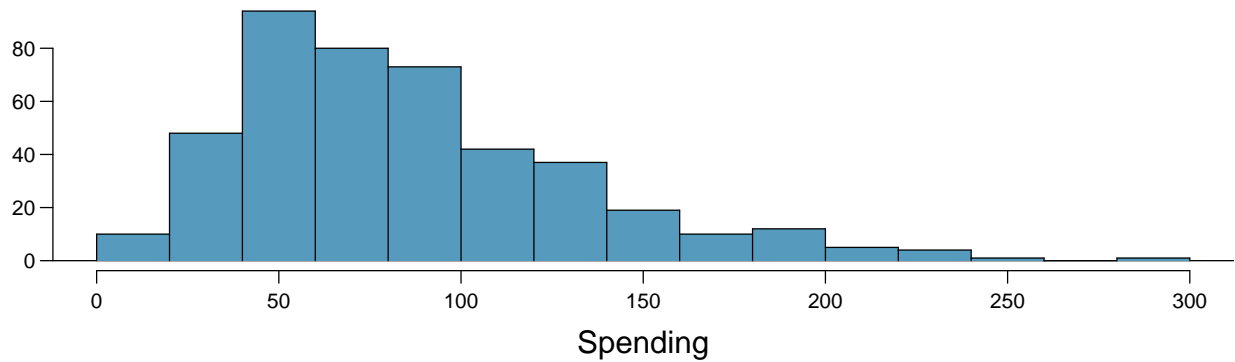
```
bdims$hgt %>% IQR()
```

```
## [1] 14
```

```
# Calculation of standard error
sd(bdims$hgt) / sqrt(length(bdims$hgt))
```

```
## [1] 0.4177887
```

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.



- We are 95% confident that the average spending of these 436 American adults is between \$80.31 and \$89.11.
- This confidence interval is not valid since the distribution of spending in the sample is right skewed.
- 95% of random samples have a sample mean between \$80.31 and \$89.11.
- We are 95% confident that the average spending of all American adults is between \$80.31 and \$89.11.
- A 90% confidence interval would be narrower than the 95% confidence interval since we don't need to be as sure about our estimate.
- 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.
- The margin of error is 4.4.

Answers

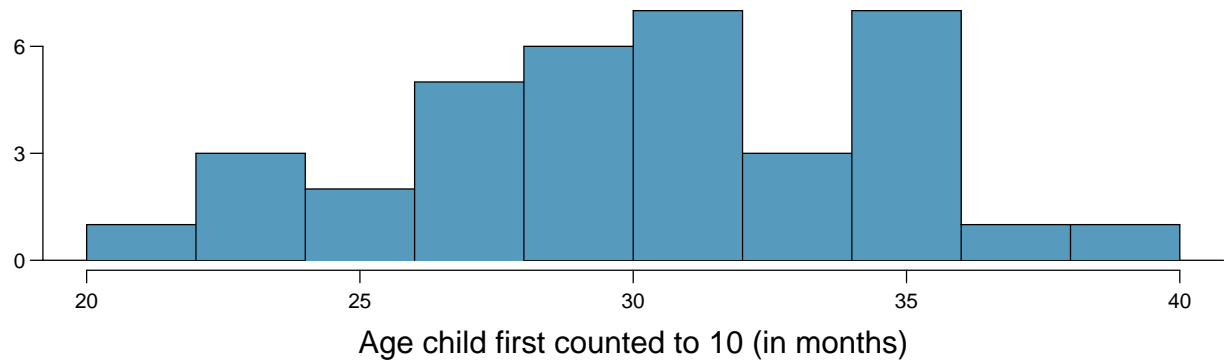
- True, as that reflects the confidence interval
- False, as it still resembles a normal distribution
- False, as random chance can affect the specific figures for individuals selected in the sample, but this gets 'more true' as the number of samples increases.
- True, the intention of this exercise is to enable inference on a larger scale
- True, the interval gets narrower as the confidence interval decreases.
- False, a larger number if needed.
- True, it is 4.4 based on the calculation below.

```
# Calculate margin of error
```

```
1.96 * (sd(thanksgiving_spend$spending) / sqrt(436))
```

```
## [1] 4.405038
```

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

- Are conditions for inference satisfied?
- 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 first count to 10 successfully is less than the general average of 32 months. Use a significance level of 0.10.
- Interpret the p-value in context of the hypothesis test and the data.
- Calculate a 90% confidence interval for the average age at which gifted children first count to 10 successfully.
- Do your results from the hypothesis test and the confidence interval agree? Explain.

Answers

- Yes. It's a seemingly independent sample with a sample size greater than 30, and the data appears to follow a normal distribution (slightly skewed to the right). Of note though is that it does fail the success-failure condition since $1(n-p)$ is not greater than 10.
- Per the calculations below, we have a p value of .06, which is less than the significance level of .1 and therefore we must reject the null hypothesis.
- The value is essentially saying that there's an ~93% chance that the findings in the online article is not true.
- The CI is 29.51 to 31.87. Since 32 falls outside of this range we can reject the null hypothesis.
- Yes, the CI confirms that the true mean lies within that range.

```
# Calculations for part a
```

```
36 * 30.69
```

```
## [1] 1104.84
```

```
sqrt(36 * (1 - 30.69) / 36)
```

```
## Warning in sqrt(36 * (1 - 30.69)/36): NaNs produced
```

```
## [1] NaN
```

```
# Hypothesis test to evaluate null hypothesis of  $p = 32$ .
stderror <- sd(gifted$count) / sqrt(36)
# Zscore, null value is used for mu (note to self: see page 195 for more info)
zscore <- (30.69-32) / stderror

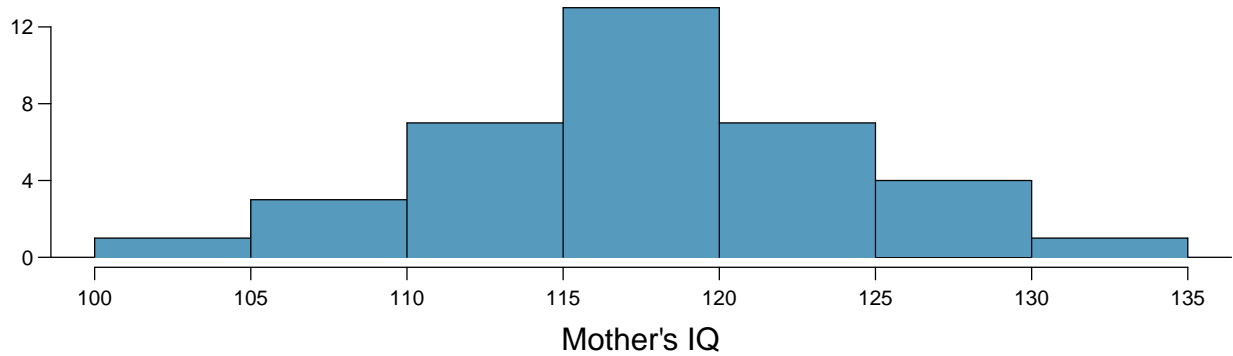
# Calculate pvalue, multiplying by 2 to account for both sides of the distribution (both tails)
pvalue <- pnorm(zscore) * 2
pvalue
```

```
## [1] 0.06851567
```

```
# Calculate 90% confidence interval
q <- qnorm(.05)
(exercisec.ci <- c(mean(gifted$count) - q * stderror, mean(gifted$count) + q * stderror))
```

```
## [1] 31.87734 29.51155
```

Gifted children, Part II. Exercise above 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

- 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.
- Calculate a 90% confidence interval for the average IQ of mothers of gifted children.
- Do your results from the hypothesis test and the confidence interval agree? Explain.

ANSWERS

- The null hypothesis must be rejected as the pvalue is less than 10%.
- The CI (calculated below) is from 119.38 to 117.02.
- Yes, as the observed mean of 118.2 falls within the CI and thus supports the hypothesis that it's not 100 (which falls outside of the CI).

```
n <- 36
min <- 101
x <- 118.2
null_mean <- 100
sd <- 6.5
max <- 131
se <- sd / sqrt(36)
se
```

```
## [1] 1.083333
```

```
# Hypothesis test for A
zscore <- (x-null_mean) / se
zscore
```

```
## [1] 16.8
```

```
# P value
pvalue <- pnorm(zscore, lower.tail = FALSE ) * 2
round(pvalue, 2)
```

```
## [1] 0
```

```
# Confidence interval for B
q <- qnorm(.05)
(exercisec.ci <- c(mean(x) - q * stderror, x + q * stderror))
```

```
## [1] 119.3829 117.0171
```

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.

Answer

The sampling distribution is a distribution of sample proportions (for example in simulations). Per the book’s text, the shape, center and spread are as follows:

Shape is bell curved / normal, assuming conditions have been met. Center is the mean of the distributions. Spread is the standard deviation of the distribution.

As the sample size increases the shape gets more normal, the center reflects the actual mean of the population, and the spread gets smaller.

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?

ANSWER

- a) There is roughly a 7% probability.
- b) Uncertain with that sample size (due to random chance), though once past 30 it would likely be normal based on the central limit theorem.
- c) Nearly zero
- d) See charts below
- e) Not without more info on the skewness, but if we had a bigger sample size we could based on the central limit theorem (which would give us a normal distribution of the sample distribution means).

```
# Hypothesis test for A
null_mean <- 9000
sd <- 1000
x <- 10500
zscore <- (x - null_mean) / sd
pvalue <- pnorm(zscore)
# Subtract from 1 for probability that it happens
1 - round(pvalue, 2)
```

```
## [1] 0.07
```

```
# Central limit theorem conditions test for B
15*null_mean
```

```
## [1] 135000
```

```
15*(1-15)
```

```
## [1] -210
```

```
# P value for C
se <- sd / sqrt(15)
se
```

```
## [1] 258.1989
```

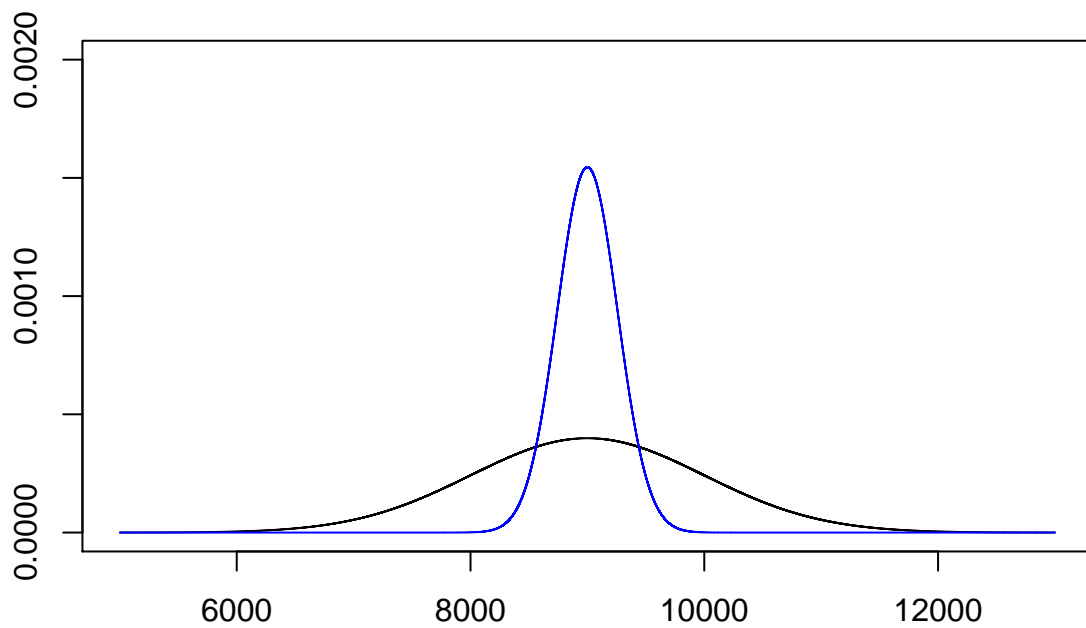
```
zscore_new <- (x - null_mean) / se
pvalue <- pnorm(zscore_new, lower.tail=FALSE)
round(pvalue, 2)
```

```
## [1] 0
```

```

# Charts for D.
# Note: These charts use a modified version of code originally from a previous
# student's work accessible at the following link:
# https://rpubs.com/srahnuma/431237
s <- seq(5000,13000,0.01)
plot(s, dnorm(s,9000, 1000), type="l", ylim = c(0,0.002), ylab = "", xlab = "Lifespan of light bulbs: P",
lines(s, dnorm(s,9000, 258.1989), col="blue")

```



Lifespan of light bulbs: Population and Sampling distributions

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.

ANSWER

The p value would likely decrease based on the fact that the larger sample size would enable a reduced standard error.