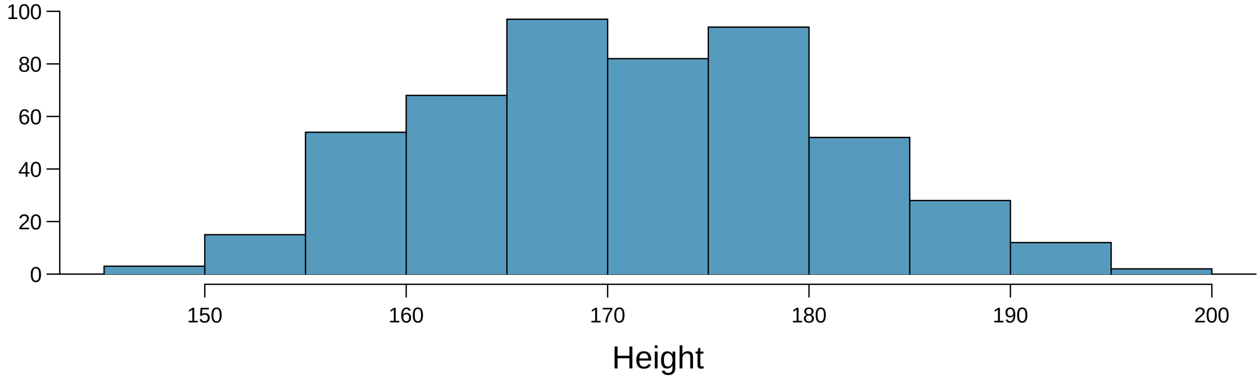
Chapter 5 - Foundations for Inference

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.



(a) What is the point estimate for the average height of active individuals? What about the median?

```
summary(bdims$hgt)
```

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

Point estimate of the population via sample 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?

```
paste("std sample:", sd(bdims$hgt))
```

```
## [1] "std sample: 9.40720520351794"
```

```
paste("IQR sample", IQR(bdims$hgt))
```

```
## [1] "IQR sample 14"
```

(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.

```
tall_z_score<-((180-170.3)/sd(bdims$hgt))
paste("Z score for 180cm", tall_z_score, "Percentile:",pnorm(((180-170.3)/sd(bdims$hgt))))
```

```
## [1] "Z score for 180cm 1.03112452531306 Percentile: 0.848758785927279"
```

One std from the mean height or 84 percentile is tall, but not unusually tall.

```
short_z_score<-((155-170.3)/sd(bdims$hgt))
paste("Z score for 155cm", short_z_score, "Percentile:",pnorm(((155-170.3)/sd(bdims$hgt))))
```

```
## [1] "Z score for 155cm -1.62641291106081 Percentile: 0.0519309230415332"
```

155cm seems to be more unusual - the 5th percentile and nearly 2 std from the mean value.

(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.

I would expect them to be pretty similar. This is a pretty normal distribution and unless the next sample is grossly skewed for some reason, then we should expect similar result. Not identical, but similarly normal.

(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 $SD_x = \frac{\sigma}{\sqrt{n}}$)? Compute this quantity using the data from the original sample under the condition that the data are a simple random sample.

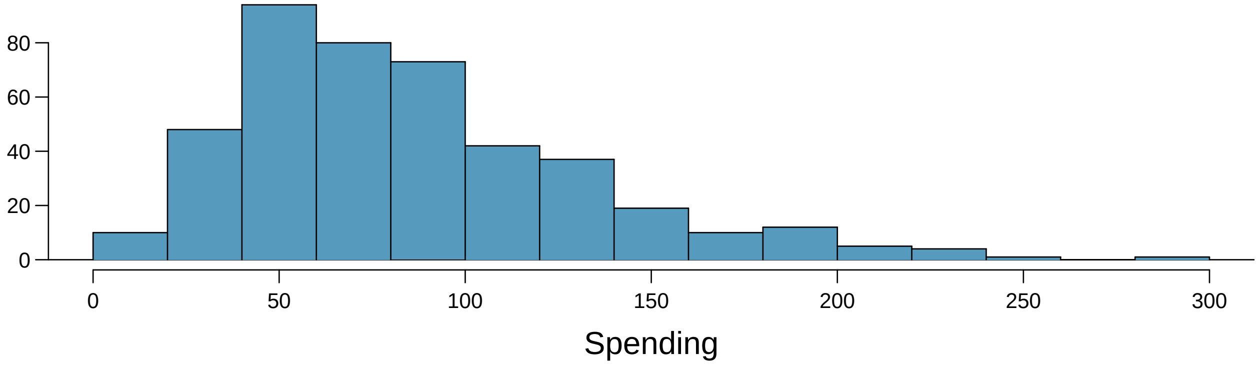
Standard Error can quantify the variability of the point estimates (sample means).

```r
paste("standard error: ",sd(bdims$hgt)/sqrt(dim(bdims)[1]))
```

```
## [1] "standard error:  0.417788650505626"
```

I don't know if this is "good" value or not

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.

False. Confidence interval applies to the population, not the sample.

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

False. Confidence interval is calcultated from the sample distribution using CLT. Our sample is big enough to apply CLT.

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

False. Again the confidence interval applies to the population mean.

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

This is true, we are discussing the population mean in regards to the confidence interval.

(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 data less we need to capture, the narrower the band of z-scsores.

(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. The standard error is proportional to the inverse square of the sample. To reduce the SE you would need a sample 9 times as large.

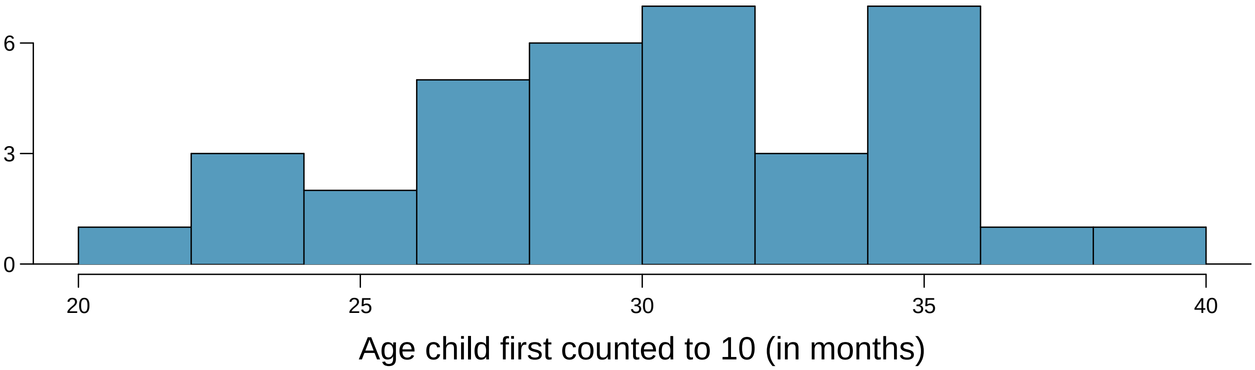
(g) The margin of error is 4.4.

```
(89.11-80.31)/2
```

```
## [1] 4.4
```

True ME is half the width of the 95% CI.

Gifted children, Part I. Researchers investigating characteristics of gifted children col- lected 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 dis- tribution 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?

Sample size is >30, random sample, independent children observations- ok to infer.

(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 first count to 10 successfully is less than the general average of 32 months. Use a significance level of 0.10.

H0: gifted children count to 10 at 32 months HA: gifted children count to 10 BEFORE 32 months.

```r
SE<- 4.31/sqrt(36)
Z<- (30.69-32)/ SE
pnorm(Z)
```

```
## [1] 0.0341013
```

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

Reject H0, p-value less than 0.10 critical value. Gifted children first count to 10 before the age of 32 months.

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

$CI = \mu \mp Z_{ci} * SE$

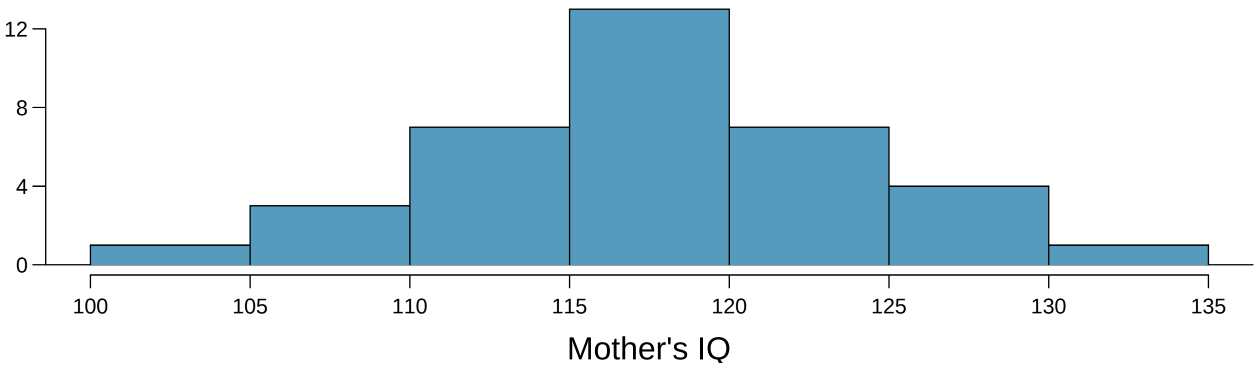
```r
upper<-round(mean(gifted$count) + (1.645*SE),3)
lower<-round(mean(gifted$count) - (1.645*SE),3)
paste("confidence interval: ",lower, upper)
```

```
## [1] "confidence interval:  29.513 31.876"
```

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

Yes. The 90% CI does not contain the population mean (32) for children first counting to 10. This age is where we reject the null hypothesis.

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

(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.

H0: average IQ of moms with gifted children is the same as the poplation average mom IQ. HA: IQ of moms with gifted children is not equal to population mean IQ

```r
SE<- 6.5/sqrt(36)
Z<-(118.2-100)/SE
1-pnorm(Z)
```

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

Our test p-value is less than the critical value for 0.10, we reject the null hypothesis. Moms with gifted children have IQ's that are not 100.

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

```r
upper<-round(mean(gifted$motheriq) + (1.645*SE),3)
lower<-round(mean(gifted$motheriq) - (1.645*SE),3)
paste("confidence interval: ",lower, upper)
```

```
## [1] "confidence interval:  116.385 119.949"
```

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

Yes the population mean of 100 is outside of the 90% CI, which is our Alternative Hypothesis.

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.

Sampling distribution of the mean is the distribution of the means of all the samples. As the number of samples is approaches infinity, the sampling distribution of the mean becomes more to similar to the normal distribution. The spread and mean of the sampling distribution will approximate the true mean and spread of the population as the number of samples approaches infinity.

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?

```r
Z<-(10500-9000)/1000
paste("probability of lasting 10500 hours: ",100*pnorm(Z, lower.tail = F), "%")
```

```
## [1] "probability of lasting 10500 hours:  6.68072012688581 %"
```

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

sample of 15 lightbulbs would have a mean of 9000 and approximate sd of $SE = \frac{\sigma}{\sqrt{15}} = 258.19$.

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

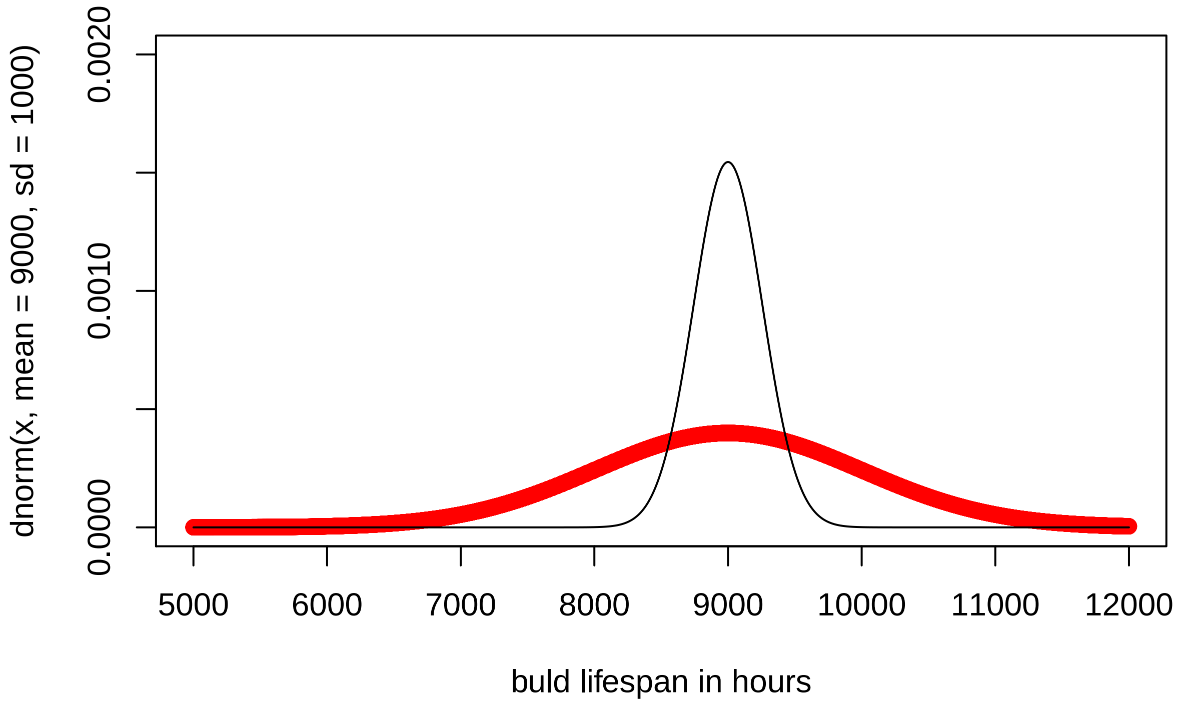
```r
SE<-1000/sqrt(15)
Z<-(10500 - 9000)/SE
paste("probability of sample bulb lasting 10500 hours: ",100*pnorm(Z, lower.tail = F), "%")
```

```
## [1] "probability of sample bulb lasting 10500 hours:  3.13345218907424e-07 %"
```

pretty low.

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

```r
x<-seq(5000,12000, length=10000)
plot(x=x, dnorm(x,mean=9000,sd = 1000),ylim=c(0,0.002),col="red", xlab = "buld lifespan in hours")
lines(x=x, dnorm(x,mean=9000, sd=(1000/sqrt(15))), col='black')
```



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

No. In part (a) we assumed a normal distribution, if the data is clearly not normal (or normal enough) can't compute these statistics. For part (c) CLT allows us to determine probabilities

using the sampling distribution if we have sufficiently large sample size. In this case, 15 is not enough.

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.

The p-value is calculated using the SE, which is inversely proportionaly to the square root of N. In this case, $\sqrt{50}$ is only about 3x less than $\sqrt{500}$, so we will see an decrease in p-value from 0.08 to 0.025. This is only meaningful if our critical value is between these two numbers. A critical value of 0.10, would suggest our change is not meaningful. A critical value of 0.05, would indicate a meaningful change.