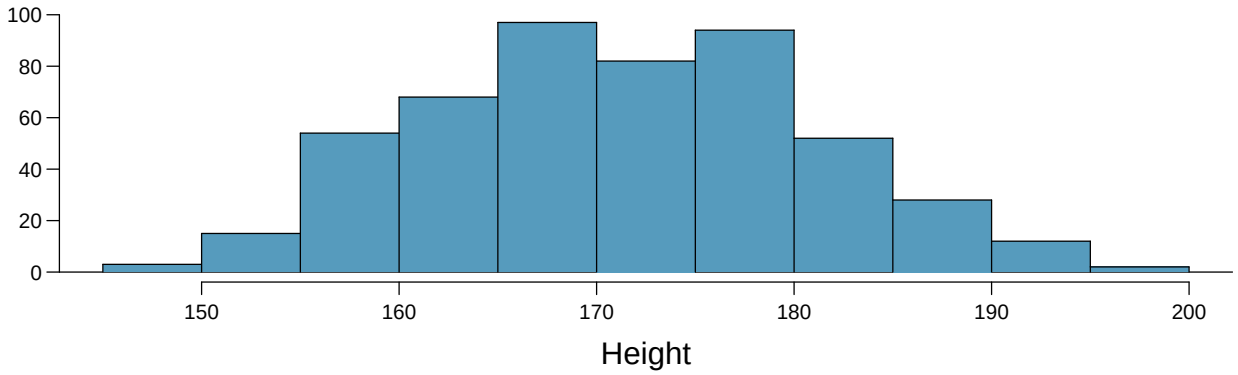
Chapter 5 - Foundations for Inference

functions

```
confidence_interval95<-function(p,se){
  b<-p-1.96*se
  a<-p+1.96*se
  return(cat("95% confidence interval is" ,b,"-",a))
}

confidence_interval99<-function(p,se){
  b<-p-2.58*se
  a<-p+2.58*se
  return(cat("95% confidence interval is" ,b,"-",a))
}
success_failure<-function(n,p){
  n<-n
  p<-p
  a<-(n*p>10)
  b<-n*(1-p)>10
  a
  b
  if(a==TRUE & b==TRUE){
    return("reasonable to construct confidence interval")
  }else{
    return("CANNOT use a confidence interval")
  }


}

standard_error<-function(p,n){
  p<-p
  n<-n
  return(sqrt((p*(1-p)/n)))
}

z_score<-function(point_estimate,null_mean,se){
  x<-point_estimate
  p<-null_mean
  se=se
  z<-((x-p)/se)
  return(z)
}
```

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?

```
pt_est<-round(mean(bdims$hgt),2)
med<-round(median(bdims$hgt),2)

cat(c("the point estimate, or mean is ",pt_est," and the median is ",med))
```

```
## the point estimate, or mean is  171.14  and 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?

```
sd<-round(sd(bdims$hgt),2)
IQR<-round(IQR(bdims$hgt,2))

cat(c("the standard deviation is ",sd," and the IQR is ",IQR))
```

```
## the standard deviation is  9.41  and the IQR is  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.

```
z<-z_score(180,pt_est,sd)
percentile_180<-round(pnorm(180,pt_est,sd),2)
z2<-z_score(155,pt_est,sd)

z2
```

```
## [1] -1.715197
```

```
percentile_155<-round(pnorm(155,pt_est,sd),2)

cat(c("A 180cm person is in the ",percentile_180," percentile and a 155cm person is in the ",percentile
```

```
## A 180cm person is in the  0.83  percentile and a 155cm person is in the  0.04  percentile.
```

(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 guess there is some corrolation between height and attractiveness, so I would bet it would have a higher mean. I would imagine being very tall would decrease attractiveness as well, so maybe the standard deviation would be smaller.

(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.

Use standard error.

```
n<-507
SE<-sd/sqrt(n)
cat(c("the standard error is",round(SE,2)))
```

```
## the standard error is 0.42
```

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.

Since this refers to the sample only. it is 100% certain that the point estimate is in this range.

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

FALSE.

The sample is large enough that the skew does not play a role.

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

```
var<-vector()
for(i in 1:1000){
samp<-mean(sample(thanksgiving_spend$spending,10,replace = TRUE))
var[i]<-samp
}
mu<-mean(var)
sd<-sd(var)

confidence_interval95(mu,sd)
```

```
## 95% confidence interval is 54.52785 - 114.5563
```

It looks like if you sample the data set the confidence interval is much wider.

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

Yes, because we deemed that this SRS is representative.

(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

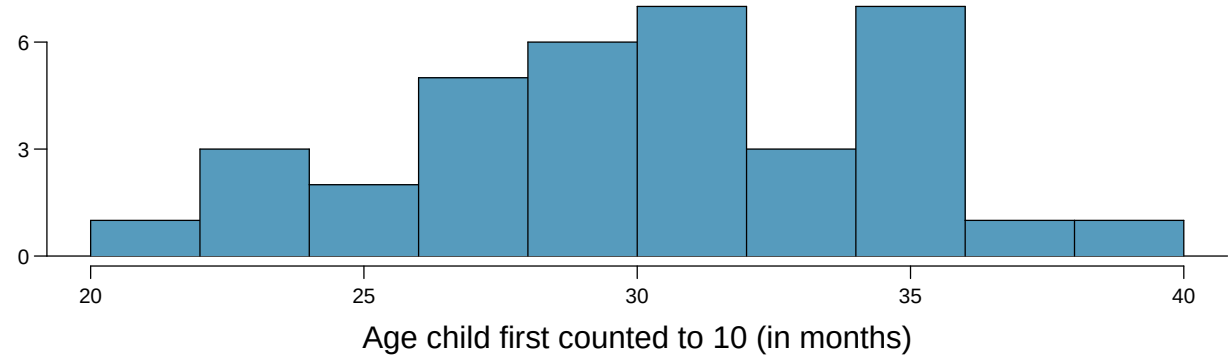
I don't like the wording of "need to be as sure about" because the estimate is for a different percentage of the population. But yes, the range would be tighter.

(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.

Because the sample size (n) is inside a square root in the denominator, and we want it to be three times larger. We would have to square the increase. It would need to be nine times larger. (g) The margin of error is 4.4.

TRUE

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?

YES

n>30 and it was a SRS. It meets the qualifications for independence (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.

HYPOTHESIS TEST H_0=32 H_A!=32 alpha=.1

```
SE<-4.31/(sqrt(36))

#my z score function
z<-z_score(30.69,32,SE)
#two tailed score because if it were the null hypothesis, the data could be in the top tail as well
p<-round(pnorm(z)*2,2)
```

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

```
cat(c("since the p-value ", p," is less than the alpha, .1 we can reject the null hypothesis. "))
```

```
## since the p-value  0.07  is less than the alpha, .1 we can reject the null hypothesis.
```

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

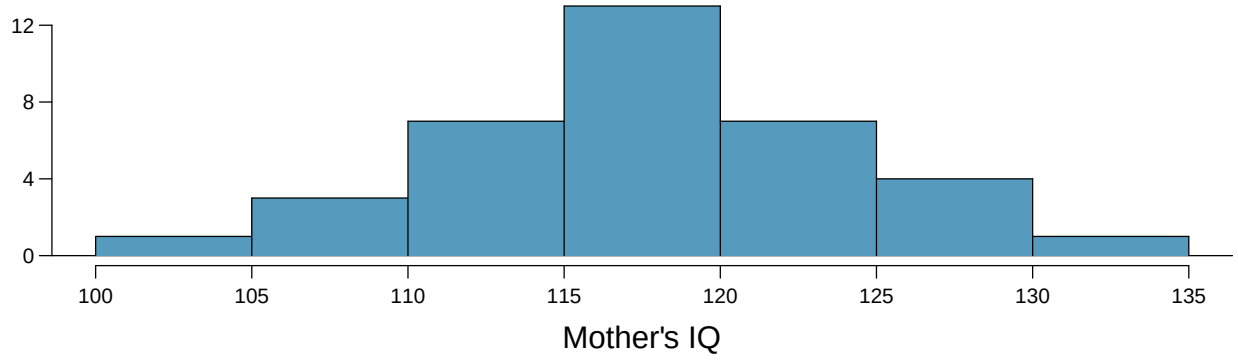
```r
p<-32
z<-qnorm(.9)
se<-SE



  b<-p-z*se
  a<-p+z*se
  cat("90% confidence interval is" ,b,"-",a)
```

```
## 90% confidence interval is 31.07942 - 32.92058
```

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

No, because we are 90% confident we captured the 30.69 value in our 90% confidence interval. If the alpha was at a higher standard, it would not have passed our hypothesis test. This might be the reason for the discrepancy.

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.

```
p<-100
x<-118.2
sd<-6.5
n<-36

SE<-6.5/6
z<-z_score(x,p,SE)


p<-2*(1-pnorm(z))

cat("the p value ",p," is lower than .1 so we reject the null hypothesis.")
```

```
## the p value  0  is lower than .1 so we reject the null hypothesis.
```

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

```
p<-100
z<-qnorm(.9)
se<-6.5/6



  b<-p-z*se
  a<-p+z*se
  cat("90% confidence interval is" ,b,"-",a)
```

```
## 90% confidence interval is 98.61165 - 101.3883
```

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

Yes they do agree, since the mean of the high IQ moms is outside of the 90% confidence interval.

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 distribution of sampling proportions is called the sampling distribution. (sample the dataset and take the mean a bunch of times.)

the shape of a sampling distribtion is always normal, no matter the shape of the dataset (I think there is one special case where this isn't true, but it can be ignored.) The center of the sampling distribution is the mean of the dataset. As n increases the spread decreases. This makes sense as it hones in on the mean of the dataset when fed more information.

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?

```
p<-9000
sd<-1000
x<-10500

p<-round(1-pnorm(x,p,sd),2)

cat("the probability of it lasting this long is ",p)
```

```
## the probability of it lasting this long is  0.07
```

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

```
#sim a dataset
bulb<-rnorm(15,9000,1000)
summary(bulb)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6672    8284    8918    8841    9374   10925
```

```
sd(bulb)
```

```
## [1] 1037.545
```

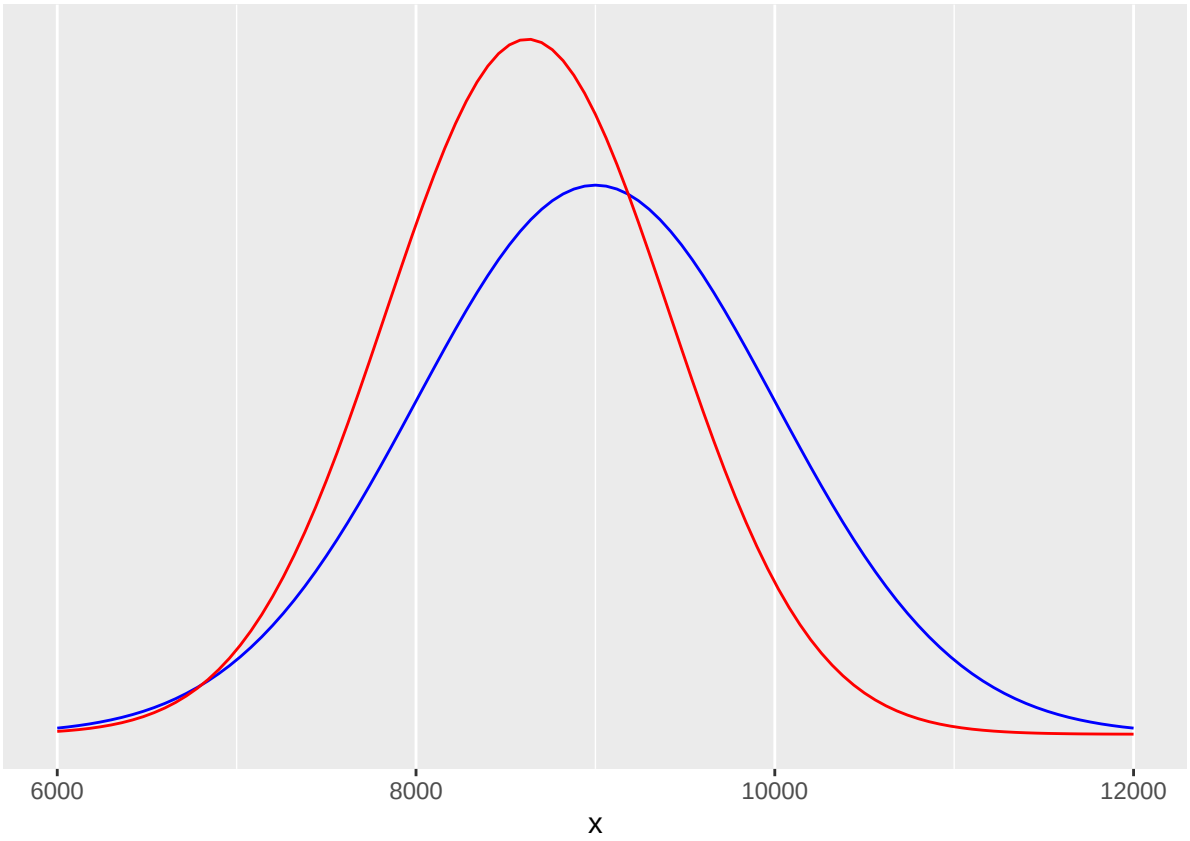
(c) What is the probability that the mean lifespan of 15 randomly chosen light bulbs is more than 10,500 hours?

```
null<-9000
p<-10500
se<-1000/sqrt(15)
z<-z_score(p,null,se)
p<-1-pnorm(z)

cat("The probability of the life span being this long is ",p)
```

```
## The probability of the life span being this long is  3.133452e-09
```

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

```
library(ggplot2)
p1 <- ggplot(data = data.frame(x = c(6000, 12000)), aes(x)) +
  stat_function(fun = dnorm, n = 101, args = list(mean = 9000, sd = 1000), col="blue") + ylab("") +
  scale_y_continuous(breaks = NULL)+
  stat_function(fun=dnorm, n=101, args=list(mean = 8622, sd = 790), col="red")


p1
```



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

You could estimate the probability from (a) because n>30, but not c. Skew doesn't come into play when n is large.

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.