Bootstrap

Stats 102A

Miles Chen

Department of Statistics

Week 9 Wednesday

UCLA

Section 1

Sampling Distributions

Sampling distributions

A key idea behind statistical inference is the sampling distribution.

The **Sampling Distribution** of a statistic is the distribution of all possible values of a sample statistic from all possible random samples drawn from the population.

Key uses of the sampling distribution:

- By studying the sampling distribution, we get a sense of how much a sample statistic can vary.
- If we have data, we can compare the statistic from our observed sample with the sampling distribution under the null hypothesis. We can find the probability that the random sampling process can produce a sample statistic like the one observed (the p-value).

Central Limit Theorem

The **Central Limit Theorem** says that the sampling distribution of the *sample mean* will follow a normal distribution if the sample is randomly drawn from the population and the sample size is large enough.

$$\bar{x} \sim N\left(\text{mean} = \mu, \text{sd} = \frac{\sigma}{\sqrt{n}}\right)$$

Sampling distributions exist for **any** statistic that can be calculated from a sample. The sampling distribution of other statistics, like the standard deviation, are not governed by the central limit theorem, and will not follow a normal distribution.

Simulating Sampling distributions

For statistics other than the mean, we may need to rely on our mathematical knowledge of probability distributions. But even if we don't remember which distributions to use, we can always use simulation to approximate the sampling distribution.

These simulations require that we **make assumptions about the population**.

We might need to make assumptions about:

- the population's shape
- the population's mean or standard deviation
- what values may appear in the population

Once you have established your assumptions about the population, you can simulate the sampling distribution by repeatedly drawing samples of a given size and recording the sample statistic for each sample.

Section 2

Parametric Bootstrap

Parametric Bootstrap

The procedure of simulating the sampling distribution by making *parametric* assumptions about the population is known as **parametric bootstrap**.

- The term **bootstrap** comes from the phrase to “pull yourself up by your bootstraps” which describes accomplishing a task without additional help (or in this case, without gathering additional data).
- Origins of the phrase are disputed, but one theory is that it came from a folk tale where the hero was stuck in a swamp. To get out, he pulled himself out by his bootstraps, an absurd thing that only folk heroes could achieve.
- **With parametric bootstrap, we assume the population follows a specified parametric distribution**.
  - We often use statistics from our observed sample to help us specify the distribution (but not always).
- We then simulate drawing many random samples from the assumed population.
- For each random sample drawn, we calculate some statistic. We record every value of this statistic to create our sampling distribution.
- With our sampling distribution, we can perform statistical inference.

Example: Simulating the sampling distribution of the variance

SAT verbal scores (officially: Evidenced-Based Reading and Writing - EBRW) are designed to be approximately normally distributed with a mean of 500 and a standard deviation of 100 for the general population of high school seniors. Scores have a maximum of 800 and a minimum of 200 and have increments of 10. (e.g. 640 is a valid score, but 638 is not)

(In reality, scores are not normally distributed because only high school seniors who intend to go to college take the exam, but we won't worry about that for this example. We'll assume scores are normally distributed with mean 500 and sd 100.)

Let's say you take random samples of 15 students.

What is the sampling distribution of the sample variance s^2?

Pop quiz: What probability distribution can be used to model the sample variance of a sample of random normal values?

Example: Simulating the sampling distribution of the variance

We can simulate drawing a random sample from the assumed population using `rnorm` along with `round`

```r
set.seed(1)
samp <- rnorm(15, mean = 500, sd = 100)
samp <- round(samp, -1) # round to nearest 10
samp <- ifelse(samp > 800, 800, samp) # max is 800
samp <- ifelse(samp < 200, 200, samp) # min is 200
samp
```

```
##  [1] 440 520 420 660 530 420 550 570 560 470 650 540 440 280 610
```

```r
var(samp)
```

```
## [1] 10120.95
```

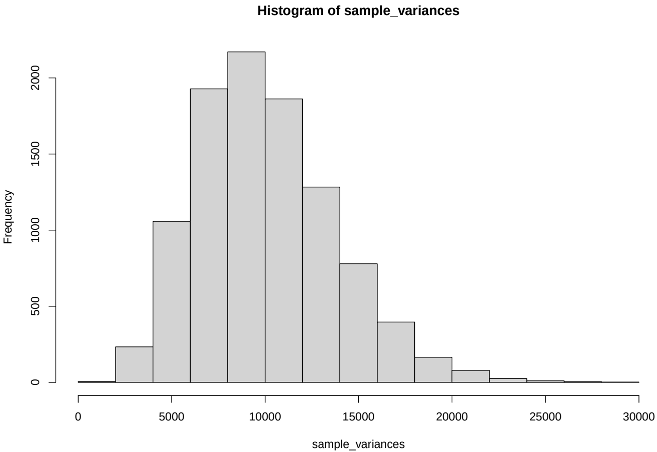
Example: Simulating the sampling distribution of the variance

```
set.seed(1)
sample_variances <- rep(NA, 10^4)
for(i in seq_along(sample_variances)){
  samp <- rnorm(15, mean = 500, sd = 100) # generate 15 scores from norm dist
  samp <- round(samp, -1) # round all scores to nearest 10
  samp <- ifelse(samp > 800, 800, samp) # force the max score to be 800
  samp <- ifelse(samp < 200, 200, samp) # force the min score to be 200
  sample_variances[i] <- var(samp) # record the variance of the sample
}
summary(sample_variances)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1540    7307    9626   10045   12327   29838
```

Example: Simulating the sampling distribution of the variance

```r
hist(sample_variances)
```



Answer to the pop quiz

What probability distribution can be used to model the sample variance of a sample of random normal values?

Answer: The Chi-squared distribution

Suppose:

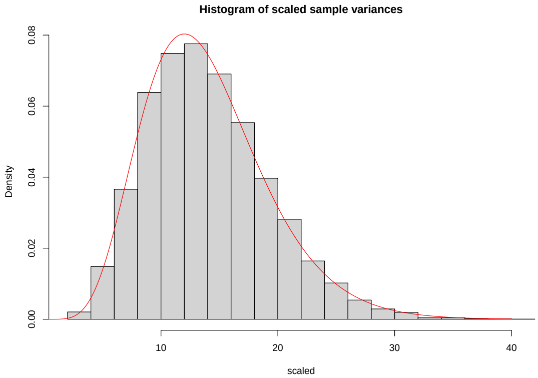
- $X_1, X_2, \ldots, X_n$ are i.i.d. from a Normal distribution with mean μ and variance σ^2
- s^2 is the sample variance

Then the distribution of the sample variance (after multiplied by a constant) follows the chi-squared distribution with n-1 degrees of freedom:

$$\frac{(n-1)}{\sigma^2} s^2 \sim \chi^2(n-1)$$

Example: Simulating the sampling distribution of the variance

```r
x <- seq(0, 40, by = .01)
d <- dchisq(x, 14)
scaled <- (15 - 1) / (100 ^ 2) * sample_variances
hist(scaled, freq = FALSE, main = "Histogram of scaled sample variances")
lines(x, d, col = 'red')
```



Benefits of Bootstrap

We might not have remembered the sampling distribution of the variance follows a chi-squared distribution. The parametric bootstrap gets around this by simulating the sampling distribution directly.

Sample Problem: Let's say you have a sample of 15 SAT scores where the sample variance is 20,000. Is it reasonable to believe that this sample was not randomly drawn from a population with mean 500 and sd $=$ 100? $(H_0 : \sigma = 100,\ H_A : \sigma > 100)$

We can get an empirical p-value estimate by comparing our sample variance of 20,000 to the sampling distribution of the variance that we simulated via parametric bootstrap.

```r
mean(sample_variances >= 20000)
```

```
## [1] 0.012
```

According to our simulated sampling distribution, a population with a sd 100 could produce a random sample where the variance is 20,000 or more, but only with a probability of 0.012. This empirical p-value is low enough for me to reject the null hypothesis. I conclude that my sample of scores is not a random sample from a population with mean 500 and sd 100.

Another Example: Highest score in the class

Let's keep working with SAT Verbal scores. Normal distribution. Mean 500, SD 100.

Let's pretend a classroom has 25 students and that these students are a random sample (probably a bad assumption). What is the sampling distribution of the highest score in the class? What is the probability that the highest score is 750 or higher?

There is a theoretic distribution (the Gumbel distribution) that can be used to describe the extreme value (like the max value) in a sample. However, I doubt that many people know this or are familiar enough with it to use it.

So instead, let's use the parametric bootstrap.

Again, we make the big assumption that we know what the population looks like. We then use the computer's ability to generate random samples and simulate the sampling distribution.

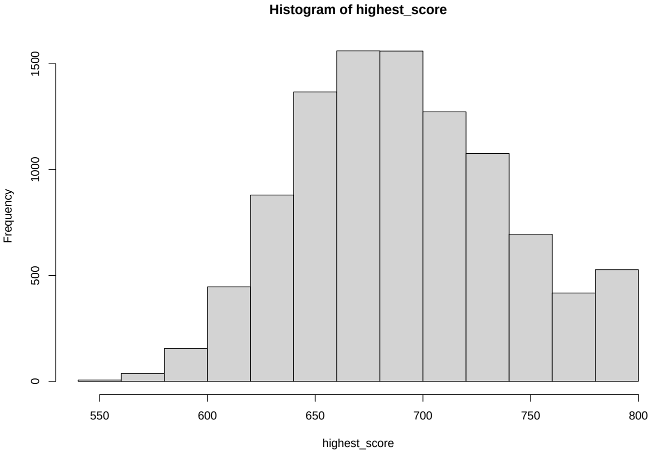
Example: Highest score in the class

```r
set.seed(1)
highest_score <- rep(NA, 10^4)
for(i in seq_along(highest_score)){
  samp <- rnorm(25, mean = 500, sd = 100) # generate 25 scores from norm dist
  samp <- round(samp, -1) # round all scores to nearest 10
  samp <- ifelse(samp > 800, 800, samp) # force the max score to be 800
  samp <- ifelse(samp < 200, 200, samp) # force the min score to be 200
  highest_score[i] <- max(samp) # record the max value in the sample
}
summary(highest_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   540.0   660.0   690.0   695.4   730.0   800.0
```

Example: Distribution of the highest score in the class

```r
hist(highest_score)
```



Example: Highest score in the class

What is the probability that the highest score is 750 or higher?

```r
mean(highest_score >= 750)
```

```
## [1] 0.1639
```

With a sample of 25 random students, there is about a 0.16 probability that the highest score is 750 or above.

What if we had a sample of 50 random students instead of 25?

Example: Highest score in the class

Easy: change the size of the random sample to 50.

```r
set.seed(1)
highest_score <- rep(NA, 10^4)
for(i in seq_along(highest_score)){
  samp <- rnorm(50, mean = 500, sd = 100) # generate 50 scores from norm dist
  samp <- round(samp, -1) # round all scores to nearest 10
  samp <- ifelse(samp > 800, 800, samp) # force the max score to be 800
  samp <- ifelse(samp < 200, 200, samp) # force the min score to be 200
  highest_score[i] <- max(samp) # record the max value in the sample
}
mean(highest_score >= 750)
```

```
## [1] 0.2965
```

If the sample size is 50 instead of 25, the probability of having the highest score be 750 or higher goes up to 0.2965.

Example: Graduating class claims

A high school has 200 students in its senior graduating class.

The school administration is very proud of itself because 10 of its students scored 700 or higher on the SAT verbal.

“We’re doing a great job preparing our students for college! Look at these SAT scores!”

Does this sample provide evidence that the student body is something other than a random sample drawn from the general population with a mean of 500 and sd of 100?

Simulating the Sampling distribution

We want to know the sampling distribution of the 10th highest score in a random sample of 200 scores.

```r
set.seed(1)
tenth_highest <- rep(NA, 10^4)
for(i in seq_along(tenth_highest)){
  samp <- rnorm(200, mean = 500, sd = 100) # generate 200 scores from norm dist
  samp <- round(samp, -1) # round all scores to nearest 10
  samp <- ifelse(samp > 800, 800, samp) # force the max score to be 800
  samp <- ifelse(samp < 200, 200, samp) # force the min score to be 200
  tenth_highest[i] <- sort(samp, decreasing = TRUE)[10] # record the 10th highest value
}
```

P-value and Conclusion

```r
mean(tenth_highest >= 700)
```

```
## [1] 0.035
```

Empirical p-value is 0.035. The simulation shows that if 200 scores were randomly drawn from a normal population, it is unlikely for the top 10 students to all have scores of 700 or higher. It is reasonable to conclude that the sample of students is not randomly drawn from a population with mean 500 and sd 100.

The results, however, cannot speak to whether the scores are the result of the school's efforts or if there are other circumstances surrounding the high scores.

Summary: Parametric Bootstrap

Summary:

If you are comfortable with stating the population follows a particular parametric distribution, you can use the parametric bootstrap to simulate the sampling distribution of any sample statistic.

That sampling distribution can then be used to perform hypothesis tests and find p-values.

Section 3

Non-parametric bootstrap

Non-parametric bootstrap

If someone talks about ‘bootstrap’, they are probably refering to non-parametric bootstrap

- **Non-parametric bootstrap does not make any parametric assumptions on distribution of the population.**
- **The assumption in non-parametric bootstrap is that our sample has been drawn randomly from the population, and therefore, can be expected to be representative of the population.**
- As such, we can duplicate of our current sample many many times as a stand in for the actual population. (Think of using the clone tool in photoshop to clone a patch of grass to make a field)
- Or rather than actually duplicating the data, we draw samples **with replacement** from the very data that we have.
- This creates a sampling distribution which we can use to make conclusions.

Bootstrap resampling vs Randomization test

- If the data you have has been obtained via *random sampling* from the population, you can use *non-parametric bootstrap*.
- If the data you have has been obtained via *random sampling* from the population, and you are willing to assume the population has a *parametric distribution*, you can use *parametric bootstrap*.
- If the data was obtained via an *experiment where random assignment* was used, we can use a *randomization test*.
- In both cases, we are creating a sampling distribution by simulating many instances of where the sole source of variation in our statistics is random chance.

Example problem and dataset

This problem comes from Zieffler chapters 6 (pg 118) and used in chapter 7.

> *The Center for Immigration Studies at the United States Census Bureau has reported that despite shifts in the ethnic makeup of the immigrant population, Latin America-and Mexico specifically-remains this country's greatest source of immigrants. Although the average immigrant is approximately 40 years old, large numbers are children who enroll in U.S. schools upon arrival. Their subsequent educational achievement affects not only their own economic prospects but also those of their families, communities, and the nation as a whole. Stamps and Bohon (2006) studied the educational achievement of Latino immigrants by examining a random sample of the 2000 decennial Census data, a subset of which is provided in LatinoEd.csv. One interesting research question that has emerged from their research is whether there is a link between where the immigrants originated and their subsequent educational achievement. Specifically, the question is if there is a difference in the educational achievement of immigrants from Mexico and that of immigrants from other Latin American countries.*

The data

Year 2000 Census data. Comes from Year 2000 Census data. The data consist of 150 Latino immigrants living in Los Angeles.

- ID: ID Number
- Achieve: Measure of educational achievement level. This is a scale of educational achievement, ranging from 1 to 100, in which higher values indicate higher levels of educational achievement.
- ImmYear: Year in which the immigrant arrived in the US (This indicates the year – to the nearest tenth – in which the immigrant arrived. To get the full year, 1900 must be added to each value. For example, 81.5 is half way through 1981.)
- ImmAge: Immigrant's age at time of immigration. It has been rounded to the nearest tenth of a year.
- English: Is the individual fluent in English? ($0 = $ No; $1 = $ Yes)
- Mex: Did the individual immigrate from Mexico? ($0 = $ No; $1 = $ Yes)

The data

```
latino <- read.csv("LatinoEd.csv")
latino
```

```
##      ID Achieve ImmYear ImmAge English Mex
## 1     1    59.2    77.7    9.6       1   1
## 2     2    63.7    65.8    1.1       1   1
## 3     3    62.4    63.6    6.1       0   1
## 4     4    46.8    55.3    2.1       1   1
## 5     5    67.6    73.1    2.3       1   1
## 6     6    63.1    75.7    8.4       1   0
## 7     7    63.7    72.0    4.9       0   1
## 8     8    63.1    69.9    5.2       1   1
## 9     9    67.6    63.9   12.7       1   1
## 10   10    30.6    77.5    9.2       0   1
## 11   11    39.0    72.2    9.4       1   1
## 12   12    46.2    81.6    8.7       1   0
## 13   13    39.0    59.0   10.5       0   1
## 14   14    57.9    72.3    1.4       0   1
## 15   15    37.1    80.2   10.5       1   1
## 16   16    66.3    72.9    5.1       1   0
```

Exploratory analysis

```
library(dplyr)
results <- latino %>% group_by(Mex) %>% summarise(mean = mean(Achieve), sd =
results
```

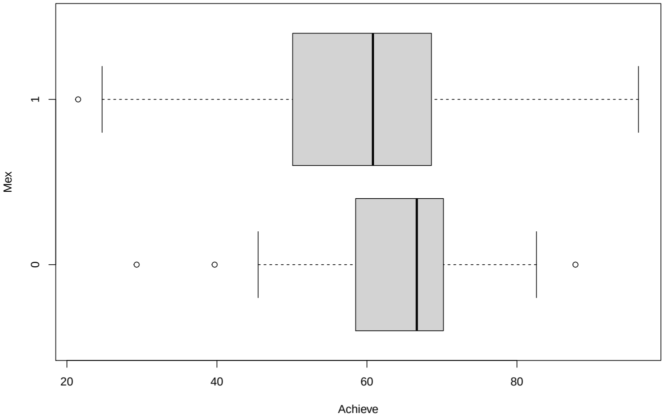
```
## # A tibble: 2 x 4
##       Mex  mean    sd     n
##     <int> <dbl> <dbl> <int>
## 1       0  64.5  13.0    34
## 2       1  58.6  15.6   116
```

```
observed_dif <- as.numeric(results[2,2]) - as.numeric(results[1,2])
observed_dif # (mex = 1) - (mex = 0)
```

```
## [1] -5.921602
```

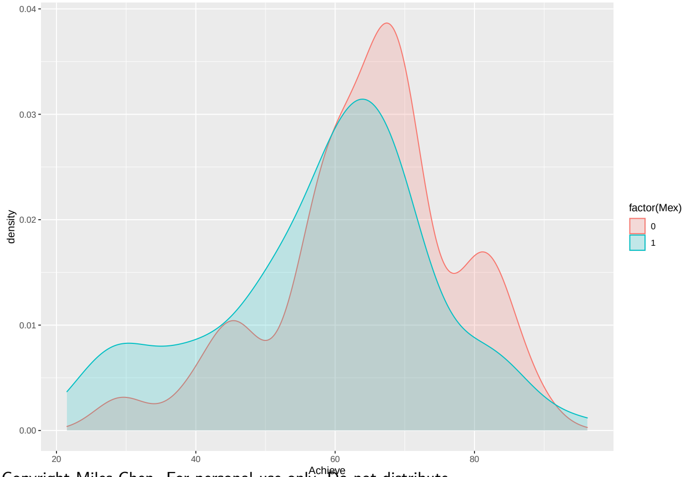
Exploratory analysis

```
boxplot(Achieve ~ Mex, horizontal = TRUE, data = latino)
```



Exploratory analysis

```r
library(ggplot2)
ggplot(latino, aes(Achieve, colour = factor(Mex), fill = factor(Mex))) + geom_density(alpha = 0.2)
```



Traditional analysis

Traditional t-test for two independent samples

$H_0 : \mu_{Mex} - \mu_{other} = 0$

$H_A : \mu_{Mex} - \mu_{other} \neq 0$

```r
t.test(Achieve ~ Mex, data = latino, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  Achieve by Mex
## t = 2.0126, df = 148, p-value = 0.04597
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##   0.1073928 11.7358121
## sample estimates:
## mean in group 0 mean in group 1
##        64.51471        58.59310
```

Non-parametric bootstrap

For non-parametric bootstrap, we will resample our original data with replacement.

We assume that our sample is representative of the population.

Sampling with replacement is equivalent to making an infinite number of copies of our sample and having that stand-in for our population.

Non-parametric bootstrap

We will build our sampling distribution of the quantity $(\bar{x}_{Mex} - \bar{x}_{other})$ if the null hypothesis were true.

$H_0 : \mu_{Mex} - \mu_{other} = 0$

The null hypothesis states that the means of education achievement level are equal for both groups.

If we assume variances are equal, then the null hypothesis effectively says that values for both groups come from the same population

Non-parametric bootstrap

- If the null hypothesis were true, the achievement scores for Mexican immigrants come from the same population as the achievement scores for other Latin American immigrants.
- With parametric bootstrap, we assume our sample of 150 observed achievement scores is representative of the entire population.
- We will draw two random samples (one of size 116, and the other size 34) by resampling the 150 scores with replacement.
- We calculate the difference between group means.
- We repeat this many times.
- We estimate the p-value by seeing how often our random samples of data produced results that were as extreme as the actual data we observed.

Non-parametric bootstrap

```r
set.seed(1)
differences <- rep(NA, 10000)
for(i in seq_along(differences)){
  # both samples are drawn from the same population: all scores in latino$Ach
  # note the use of replace = TRUE
  groupA <- sample(latino$Achieve, 116, replace = TRUE)
  groupB <- sample(latino$Achieve, 34, replace = TRUE)
  differences[i] <- mean(groupA) - mean(groupB)
}
```

Empirical p-value

To estimate our p-value, we want to know how often our random differences were as large as or greater than the difference we observed in our sample's data. Note that because the alternative is two-sided, we use the absolute value of the bootstrapped differences. However, the `observed_dif` is negative, so 100% of the absolute values of `differences` will be greater than the negative `observed_dif`. So we will compare the absolute value of differences to the absolute value of the `observed_dif`

```r
mean(abs(differences) >= observed_dif) # not correct
```

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

```r
mean(abs(differences) >= abs(observed_dif))
```

```
## [1] 0.0465
```

We get an estimated p-value of 0.0465. This tells us that even when we draw random samples from the same population, sometimes we get mean differences of 5.9 points from random chance alone. This happens a little bit less than 5% of the time. Our conclusion would then depend on the significance level we are using.

Parametric Bootstrap

We can also use Parametric bootstrap for the same problem.

We will build our sampling distribution of the quantity $(\bar{x}_{Mex} - \bar{x}_{other})$ if the null hypothesis were true.

$H_0 : \mu_{Mex} - \mu_{other} = 0$

The null hypothesis states that the means of education achievement level are equal for both groups.

If we assume variances are equal, then the null hypothesis effectively says that values for **both groups come from the same population**.

p-value

- We are ultimately trying to calculate a p-value:
  - "what is the probability of observing our data or something more unusual, if the null hypothesis were true?"
- For us, that means "how often do two randomly drawn samples have a difference of 5.9 points or greater, if both samples are drawn from the same population?"

simulate the sampling distribution

- To answer this, we will perform parametric bootstrap by repeatedly drawing two random samples from the same population.
- We can simulate drawing random values from a defined population using R's built-in random functions.
- Here we make a **major assumption** about the data: the population from which our data is drawn is normally distributed.
- We will use the dataset's **overall mean** and **overall standard deviation** as estimates of the mean and standard deviation of our parametric distribution.

Parametric Bootstrap

We perform parametric bootstrap 10,000 times to build our sampling distribution. We will draw samples using `rnorm`. We use the mean and standard deviation of all 150 achievement scores to define our assumed normal population.

```r
m <- latino$Achieve %>% mean()
s <- latino$Achieve %>% sd()

set.seed(1)
differences <- rep(NA, 10000)
for(i in seq_along(differences)){
  # both samples are drawn from the same population
  groupA <- rnorm(116, m, s)
  groupB <- rnorm( 34, m, s)
  differences[i] <- mean(groupA) - mean(groupB)
}
```

empirical p-value

To estimate our p-value, we want to know how often our random differecnces were as large as or greater than the difference we observed in our dataset. (Again we will compare the absolute value of differences to the absolute value of the `observed_dif`)

```
mean(abs(differences) >= abs(observed_dif))
```

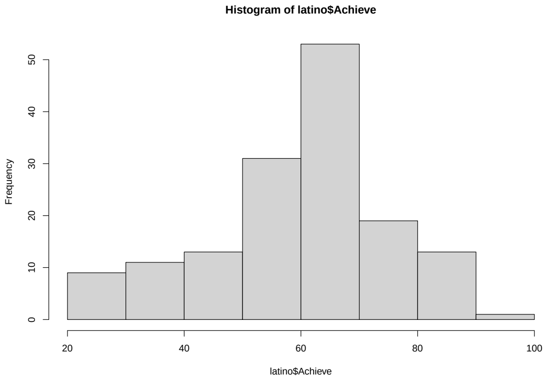
```
## [1] 0.0466
```

We get an estimated p-value of 0.0466. This tells us that even when we draw random samples from the same population, sometimes we get mean differences of 5.9 points from random chance alone. This happens a little bit less than 5% of the time. Our conclusion would then depend on the significance level we are using. The estimated p-value we got from the non-parametric bootstrap was 0.0465, which agrees very much with what we have here.

Checking our assumptions

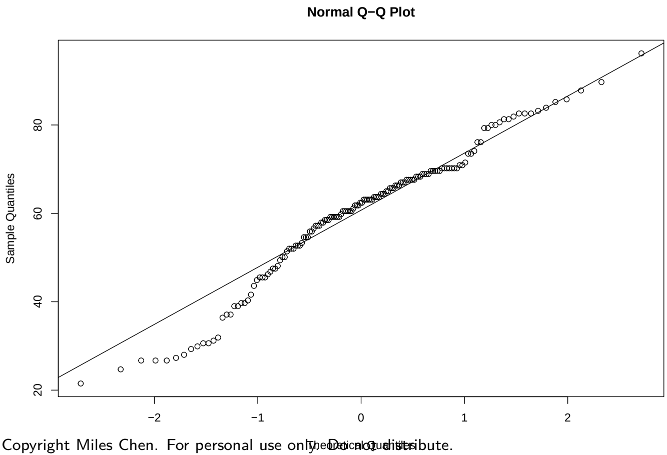
Was our assumption about the population following the normal distribution okay? We can look at the histogram and quantile-quantile plot of our Achievement data.

```r
hist(latino$Achieve)
```



Checking our assumptions

```r
qqnorm(latino$Achieve)
qqline(latino$Achieve)
```


Normal Q–Q Plot
Sample Quantiles
Theoretical Quantiles