数理统计数据分析报告

吴璐欢

12/17/2017

1. Data Preprocessing

1.1 Read Data

```r
library(readxl)
boys <- read_xlsx("boy_data.xlsx", sheet = 1)
```

```
## Warning in strptime(x, format, tz = tz): unknown timezone 'zone/tz/2017c.
## 1.0/zoneinfo/Asia/Shanghai'
```

```r
girls <- read_xlsx("girl_data.xlsx", sheet = 1)
boys <- data.frame(boys)
girls <- data.frame(girls)
girls$foot <- girls$foot / 10 # foot length unit: m -> cm
nrow(boys)
```

```
## [1] 84
```

```r
nrow(girls)
```

```
## [1] 29
```

1.2 Data description

1. We obtain 84 samples for boys, 29 samples for girls. Each sample include 5 vlues physical measurements:height, armspan, weight, foot-length ("foot" for short), leg-length ("leg" for short), and weight.
2. The unit for height / armspan / foot-length / leg-length is cm. The unit for weight is kg. We will ignore the unit in the folllowing context.

We store the information of boy-samples and girl-samples in two seperate dataframes.

```r
head(boys)
```

```
##   height armspan foot leg weight
## 1    186     181 25.0 115     72
```

```
## 2    175      175 27.0 100      65
## 3    175      176 26.0  97      67
## 4    176      198 27.0 104      72
## 5    175       17 26.0  98      80
## 6    180      178 24.5 101      54
```

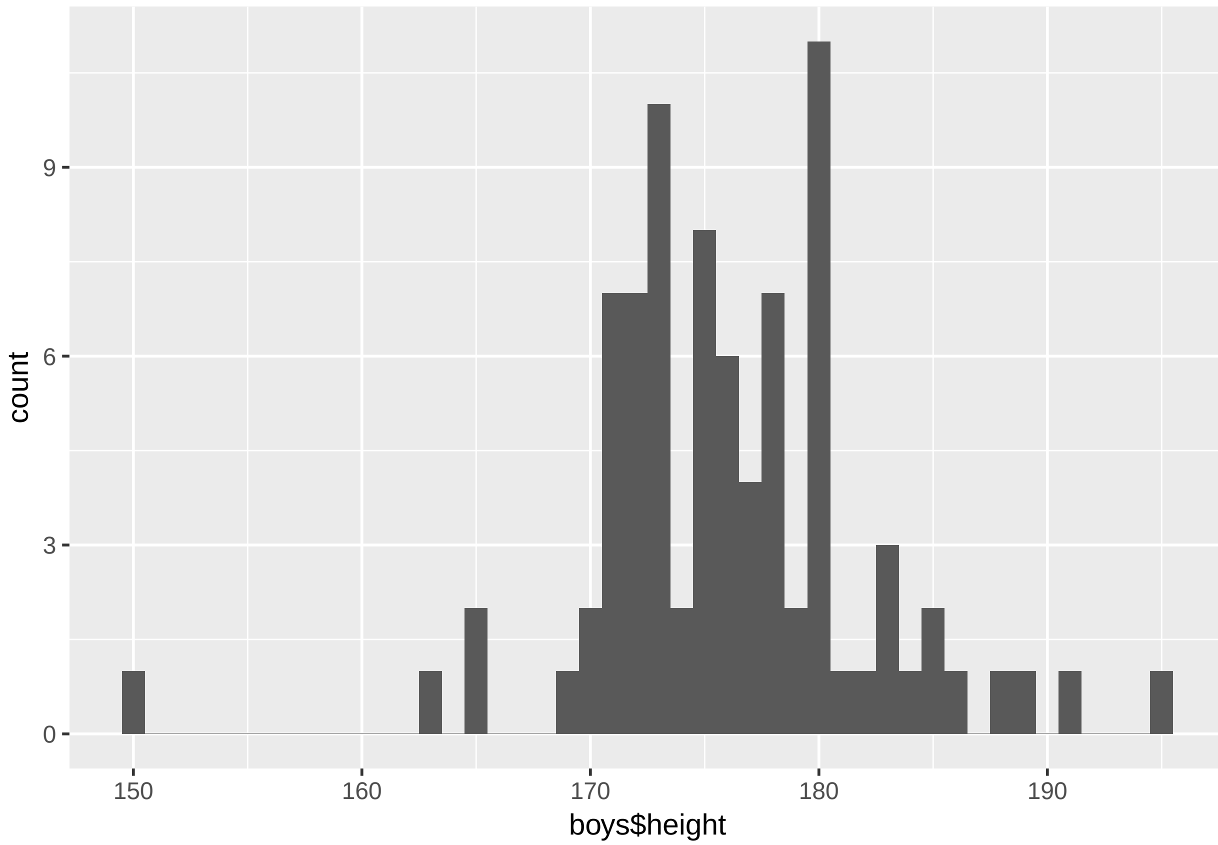
```r
head(girls)
```

```
##   height armspan foot leg weight
## 1    170   169.0 24.5 104      60
## 2    166   163.0 24.0  96      55
## 3    167   165.0 23.0  88      53
## 4    164   153.0 23.3 100      50
## 5    160   156.0 23.0  91      45
## 6    165   157.5 23.8 101      58
```

1.3 Outlier Detection and Treatment

In this part, we use histogram to detect and delete extreme values. Boy height:

```r
library(ggplot2)
qplot(boys$height,binwidth=1)
```

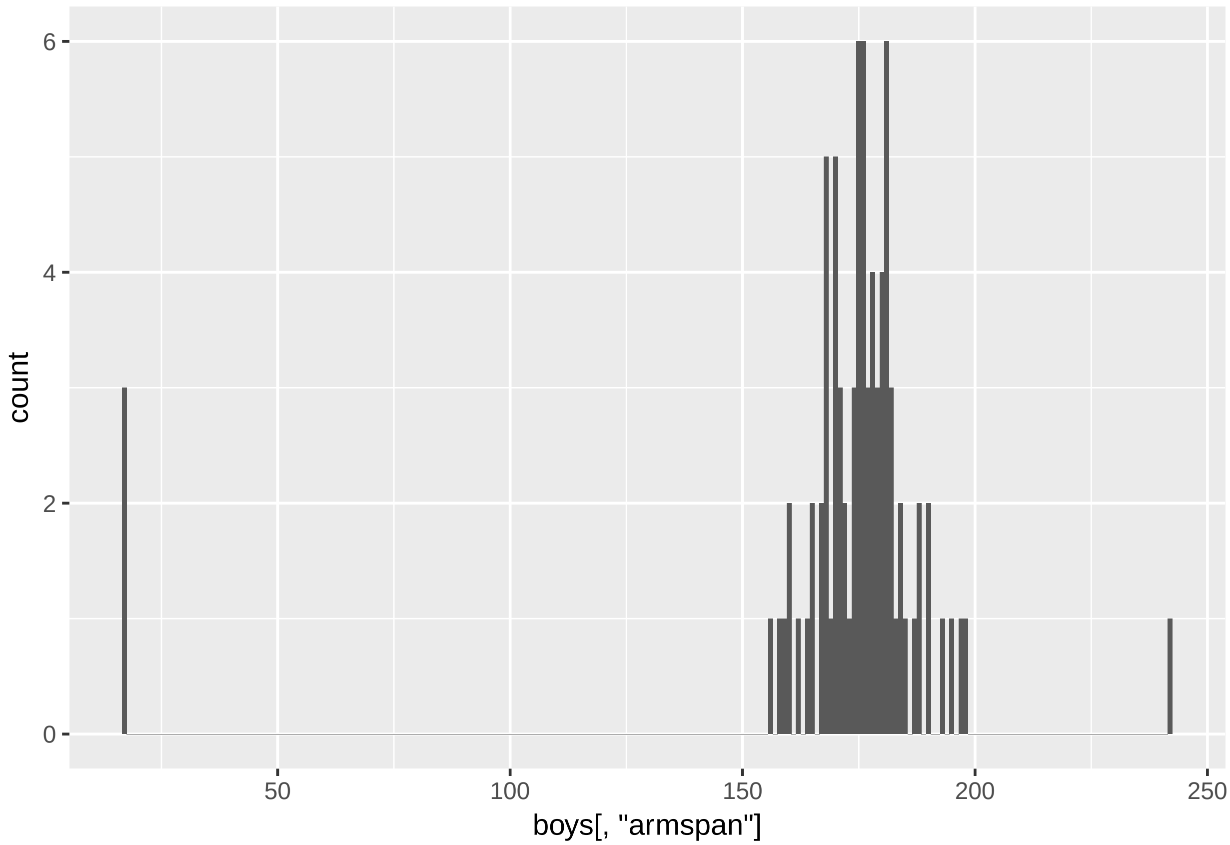


We should drop boy student whose height = 150

```
boys <- boys[boys$height!=150,]
```

Boy armspan:

```
qplot(boys[,"armspan"],binwidth=1)
```




```
min(boys[,"armspan"])
```

```
## [1] 17
```

```
max(boys[,"armspan"])
```

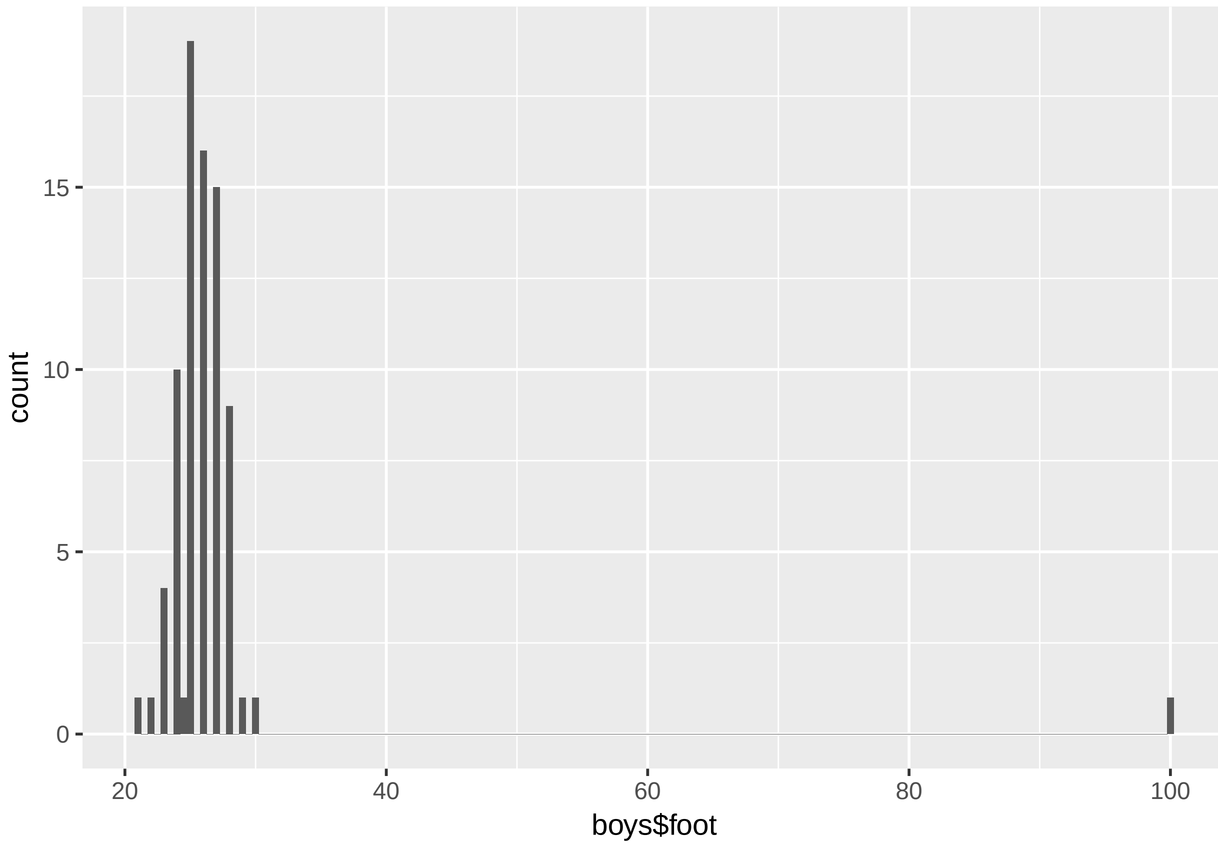
```
## [1] 242
```

We should drop whose armspan = 17(min), 242(max).

```
boys <- boys[boys$armspan!=17,]
boys <- boys[boys$armspan!=242,]
```

Boy foot-length:

```
qplot(boys$foot,binwidth=0.5)
```

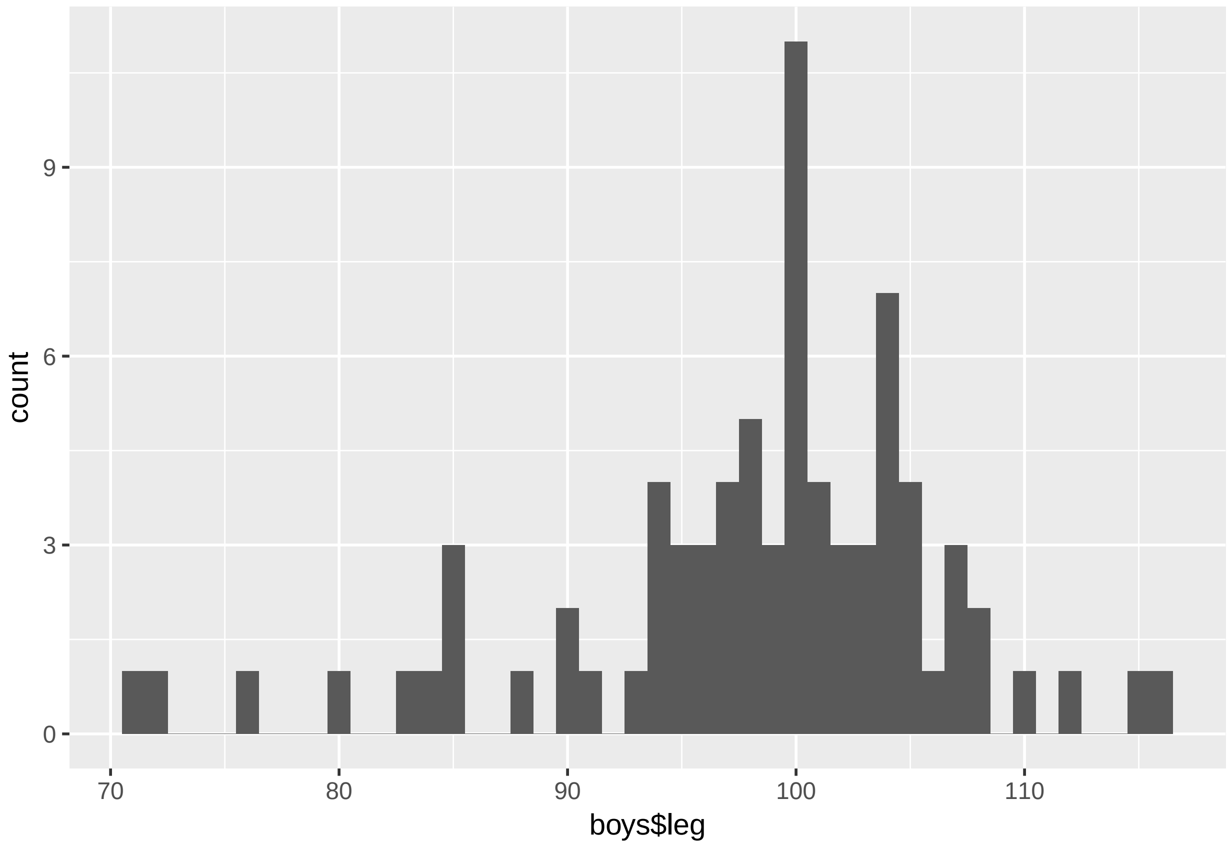


We should drop boy student whose foot length = 100.

```
boys <- boys[boys$foot!=100,]
```

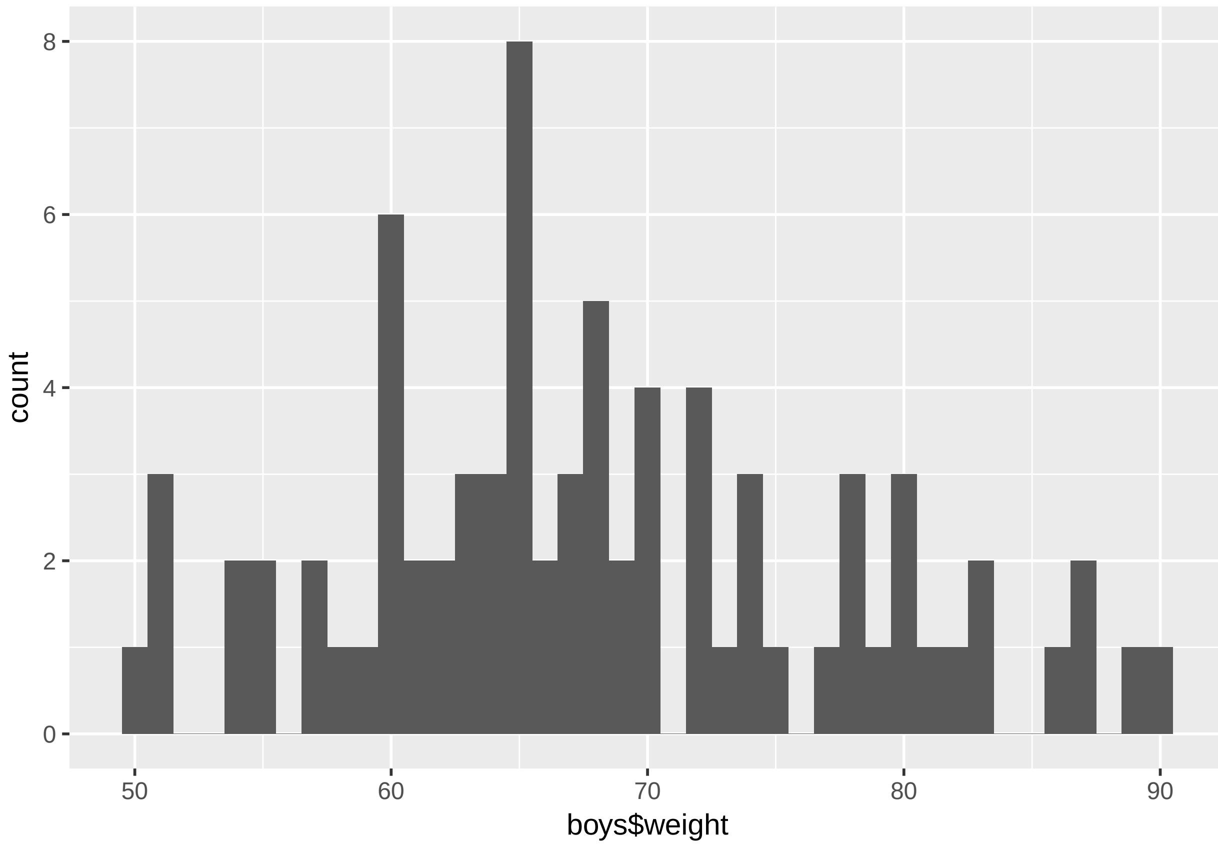
Boy leg-length:

```
qplot(boys$leg,binwidth=1)
```



Boy weight:

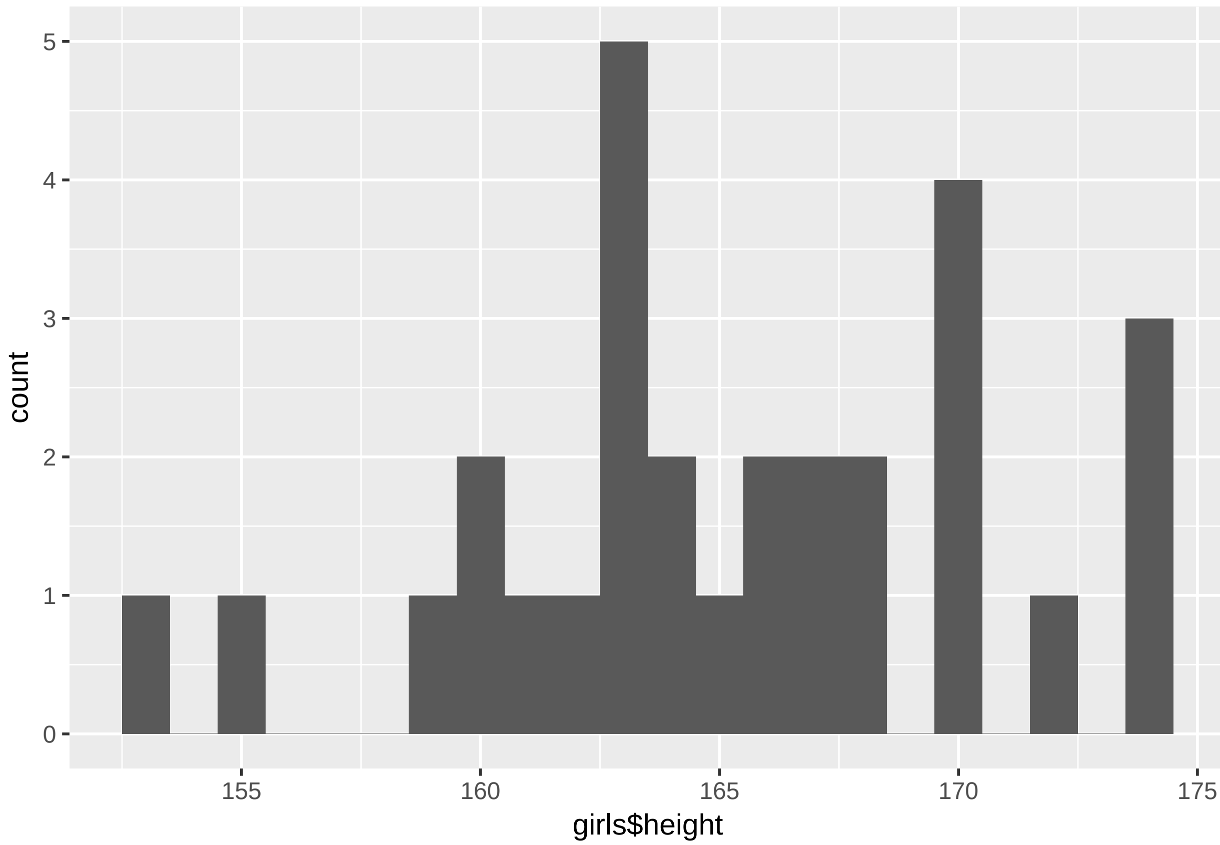
```
qplot(boys$weight,binwidth=1)
```



Not outliers detected for boy leg-length and weight.

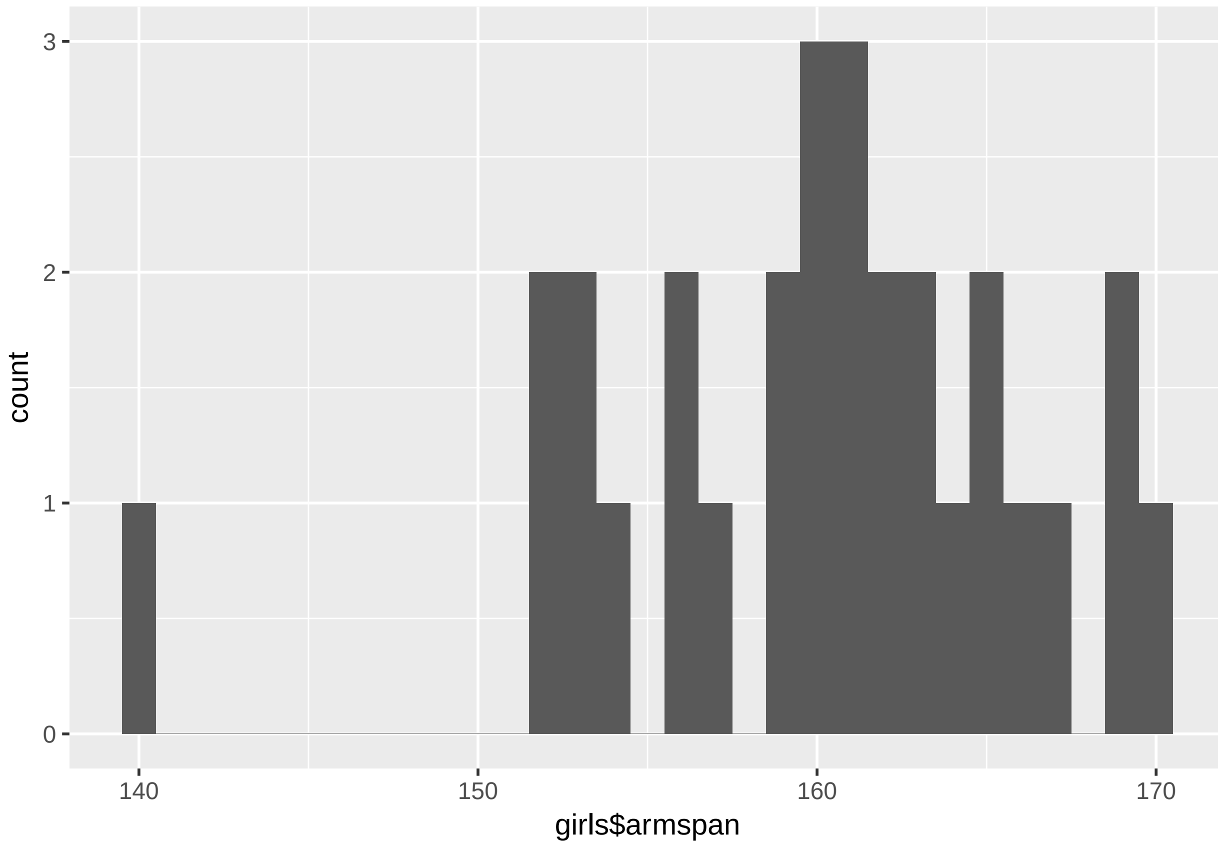
Similarly, we detect and clean outliers for girls data. Girl height:

```
qplot(girls$height,binwidth=1)
```



No significant outliers for girl height. Gril armspan:

```
qplot(girls$armspan,binwidth=1)
```

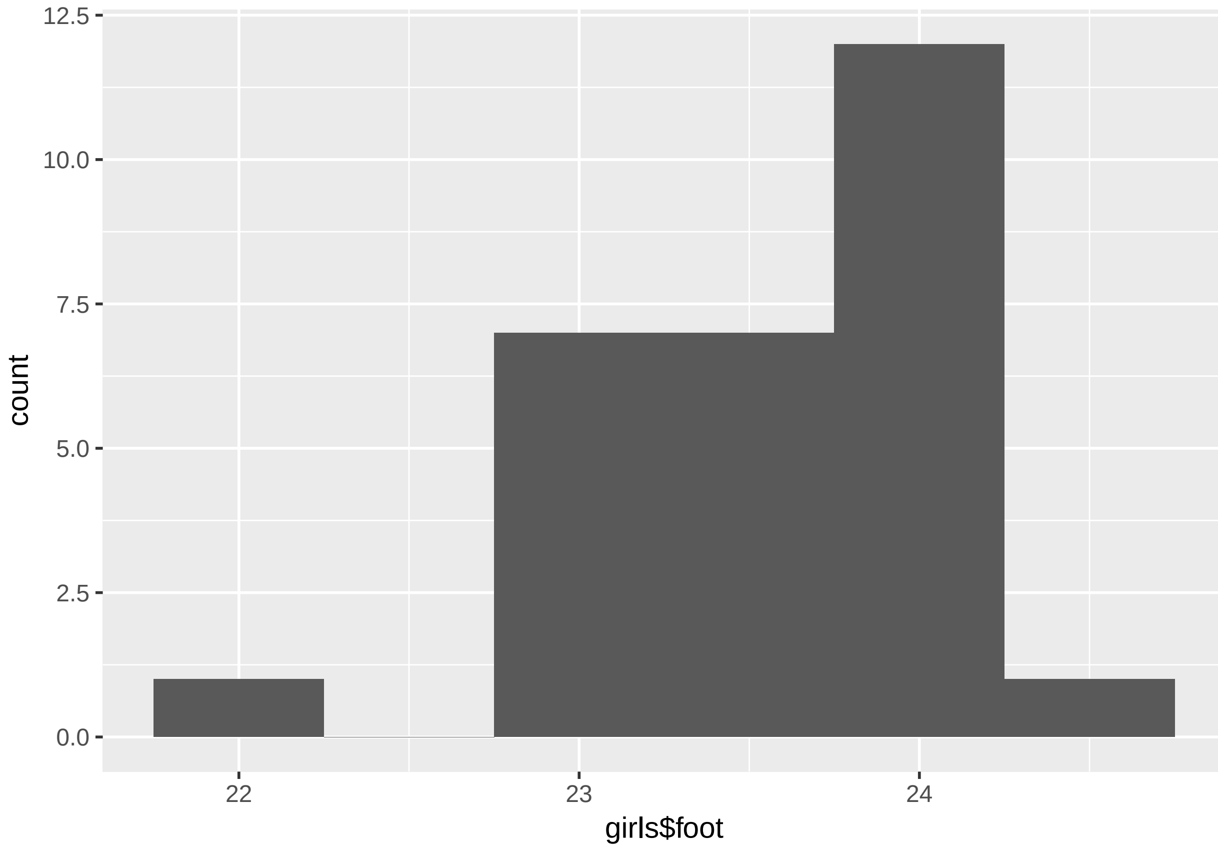


Delete the min value.

```
girls <- girls[girls$armspan!=140,]
```

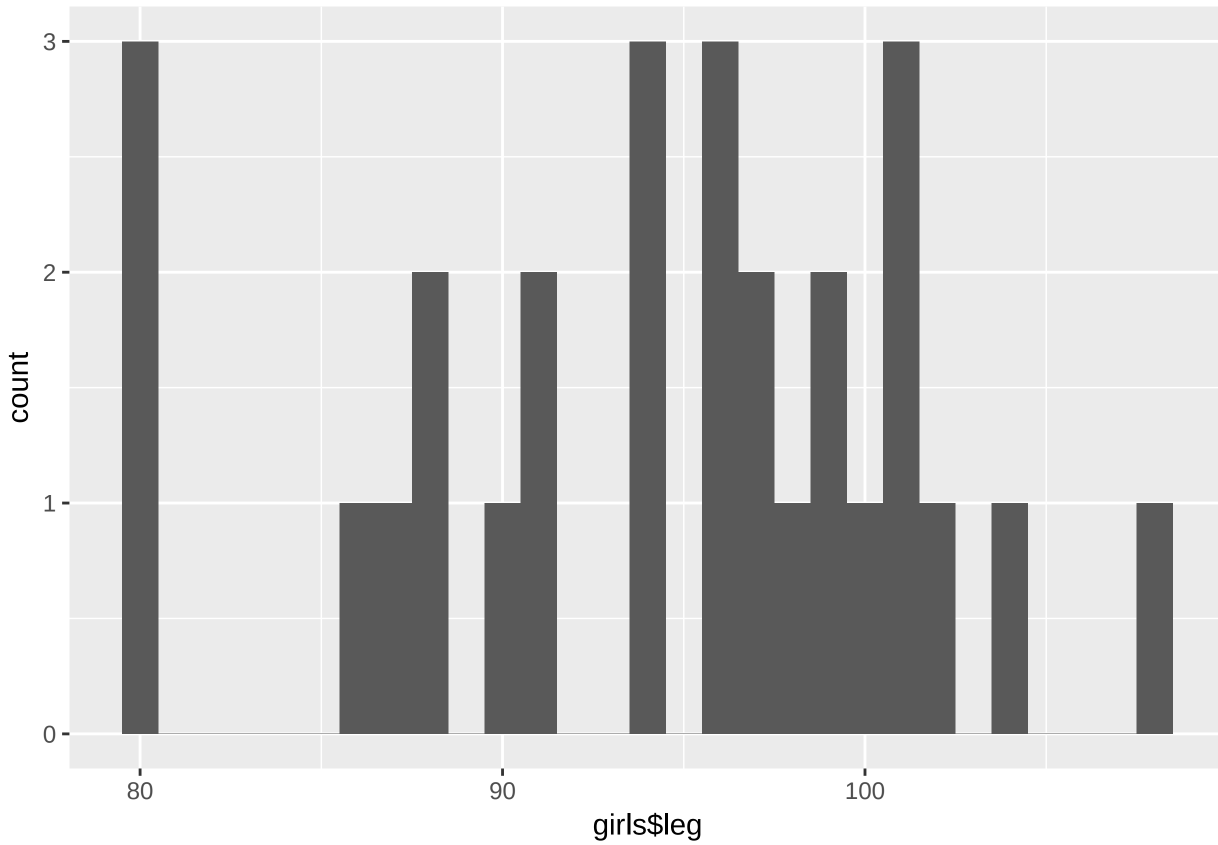
Girl Foot-length

```
qplot(girls$foot, binwidth=0.5)
```



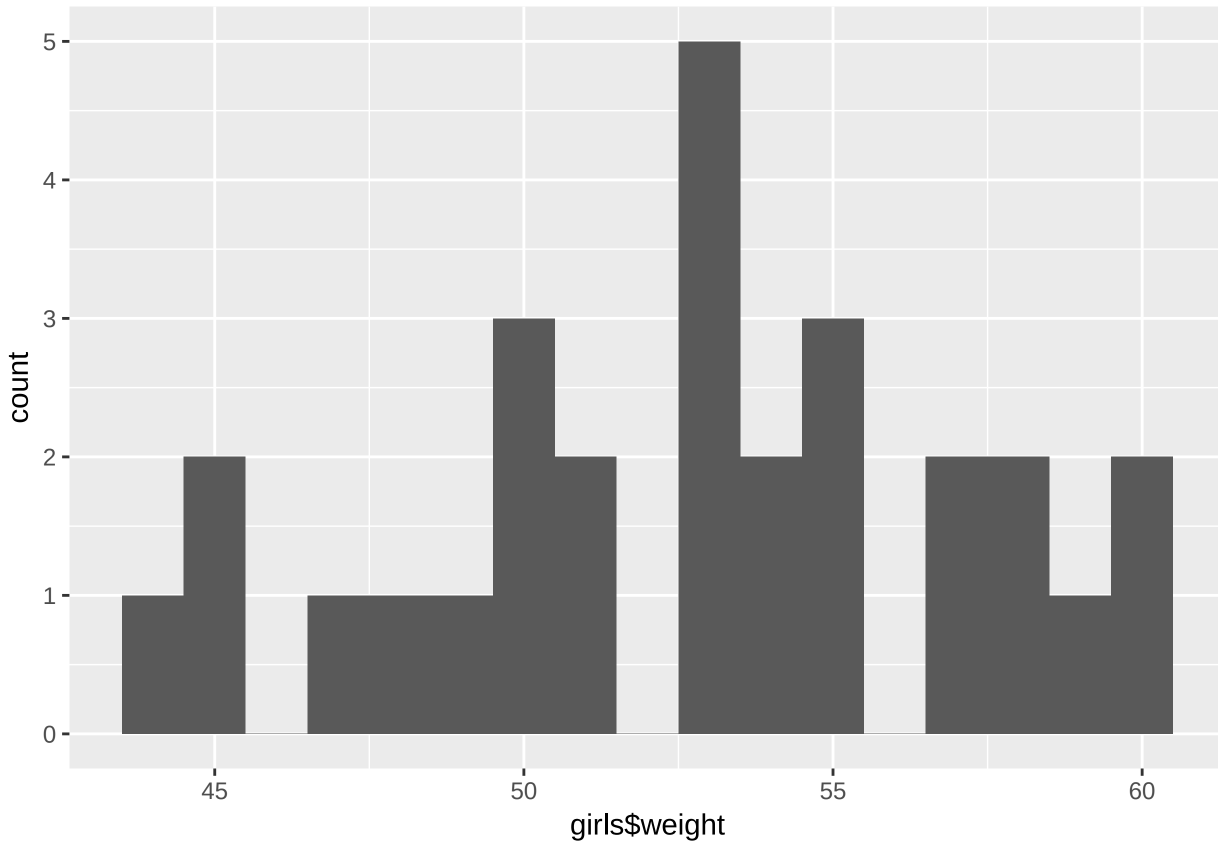
Girl leg-length:

```
qplot(girls$leg, binwidth=1)
```



Girl weight:

```
qplot(girls$weight, binwidth=1)
```



No significant outliers for girl foot-length, leg-legnth,weight.

```
total <- rbind(boys, girls)
nrow(total)
```

```
## [1] 106
```

```
nrow(boys)
```

```
## [1] 78
```

```
nrow(girls)
```

```
## [1] 28
```

After data-cleaning, we have 78 samples for boys, 28 samples for girls and 106 for total students.

2. Descriptive Statistics

2.1 Measures of Statistical Posistion

We could use function *summary* to check the mean, min value, max value, first quantitle, Median and third quantitle for each statistics.

```r
summary(boys)
```

```
##      height         armspan            foot             leg        
##  Min.   :163.0   Min.   :156.0   Min.   :21.00   Min.   : 71.00  
##  1st Qu.:172.2   1st Qu.:170.0   1st Qu.:25.00   1st Qu.: 95.00  
##  Median :175.5   Median :176.0   Median :26.00   Median :100.00  
##  Mean   :176.3   Mean   :175.8   Mean   :25.72   Mean   : 98.21  
##  3rd Qu.:180.0   3rd Qu.:181.0   3rd Qu.:27.00   3rd Qu.:104.00  
##  Max.   :195.0   Max.   :198.0   Max.   :30.00   Max.   :116.00  
##      weight     
##  Min.   :50.00  
##  1st Qu.:61.25  
##  Median :67.00  
##  Mean   :68.03  
##  3rd Qu.:74.00  
##  Max.   :90.00
```

```r
summary(girls)
```

```
##      height         armspan            foot             leg        
##  Min.   :155.0   Min.   :152.0   Min.   :22.00   Min.   : 80.00  
##  1st Qu.:163.0   1st Qu.:157.1   1st Qu.:23.00   1st Qu.: 89.50  
##  Median :165.5   Median :161.0   Median :23.50   Median : 96.00  
##  Mean   :165.8   Mean   :160.8   Mean   :23.56   Mean   : 94.21  
##  3rd Qu.:170.0   3rd Qu.:164.2   3rd Qu.:24.00   3rd Qu.: 99.25  
##  Max.   :174.0   Max.   :170.0   Max.   :24.50   Max.   :108.00  
##      weight     
##  Min.   :44.00  
##  1st Qu.:50.00  
##  Median :53.00  
##  Mean   :52.79  
##  3rd Qu.:55.50  
##  Max.   :60.00
```

```r
summary(total)
```

```
##      height         armspan            foot             leg        
##  Min.   :155.0   Min.   :152.0   Min.   :21.00   Min.   : 71.00  
##  1st Qu.:170.0   1st Qu.:164.0   1st Qu.:24.00   1st Qu.: 94.00  
##  Median :173.0   Median :171.5   Median :25.00   Median : 99.00  
##  Mean   :173.5   Mean   :171.8   Mean   :25.15   Mean   : 97.15  
##  3rd Qu.:178.0   3rd Qu.:179.0   3rd Qu.:26.00   3rd Qu.:102.00
```

```
##  Max.   :195.0   Max.   :198.0   Max.   :30.00   Max.   :116.00
##      weight
##  Min.   :44.0
##  1st Qu.:55.0
##  Median :63.5
##  Mean   :64.0
##  3rd Qu.:70.0
##  Max.   :90.0
```

2.2 Measures of Dispersion

2.2.1 Sample Variance – unbiased estimator of population variance

$$\hat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

In R, we use function *var* to the sample variance.

```
#Boys
apply(boys, 2, var)
```

```
##    height   armspan      foot       leg    weight
## 32.397602 76.962537  2.705503 73.515818 92.051282
```

```
#Girls
apply(girls, 2, var)
```

```
##     height    armspan       foot        leg     weight
## 23.5459656 27.4537037  0.2884656 53.5079365 20.3042328
```

```
#Total
apply(total, 2, var)
```

```
##     height    armspan       foot        leg     weight
##  51.611343 107.989937   2.979854  70.796047 118.300000
```

From the output above, we could see that for the 5 physical measurements, the sample values of boys are greater than thaose of girls, which is consistent wiht our common sense. ###2.2.2 Biased Sample Variance – biased estimator of population variance

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Since r does not have buil-in function for biased sample variance, we define the function as below.

```r
var2 <- function(h) {
  h.l <- length(h)
  h.v2 <- (h.l - 1) / h.l*var(h)
  h.v2
}
```

```r
#Boys
apply(boys, 2, var2)
```

```
##    height   armspan      foot       leg    weight
## 31.982249 75.975838  2.670817 72.573307 90.871137
```

```r
#Girls
apply(girls, 2, var2)
```

```
##     height    armspan       foot        leg     weight
## 22.7050383 26.4732143  0.2781633 51.5969388 19.5790816
```

```r
#Total
apply(total, 2, var2)
```

```
##     height    armspan       foot        leg     weight
##  51.124444 106.971164   2.951742  70.128159 117.183962
```

We could see that, for sample variance and biased sample variance, values in boys are greater than girls, while values in total is the largest.

2.2.3 Range

$$R = x_{(n)} - x_{(1)} = max(x) - min(x)$$

In R, we define function *getRange* to get the minimum and maximum of data.

```r
getRange <- function(h) {
  max(h) - min(h)
}
```

```r
#Boys
apply(boys, 2, getRange)
```

```
##  height armspan    foot     leg  weight
##      32      42       9      45      40
```

```r
#Girls
apply(girls, 2, getRange)
```

```
##   height armspan     foot      leg   weight
##     19.0     18.0      2.5     28.0     16.0
```

```r
#Total
apply(total, 2, getRange)
```

```
##   height armspan     foot      leg   weight
##       40       46        9       45       46
```

2.3 Measures of Distribution

2.3.1 Skewness

$$g_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1} n(x_i - \bar{x})^3 = \frac{n^2 \mu_3^3}{(n-1)(n-2)s^3}$$

where s is the sample variance and μ is the third central moment. In R, we use function *skewness* in libary *moments* for computation.

```r
library(moments)
#Boys
apply(boys, 2, skewness)
```

```
##     height    armspan       foot        leg     weight
##  0.6510605  0.1750956 -0.1394418 -1.0207482  0.3318014
```

```r
#Girls
apply(girls, 2, skewness)
```

```
##       height      armspan         foot          leg       weight
##  0.008821704 -0.059486525 -0.778134258 -0.450301862 -0.214181371
```

```r
#Total
apply(total, 2, skewness)
```

```
##     height    armspan       foot        leg     weight
##  0.1000425  0.1851263  0.3179549 -0.7926093  0.4064730
```

2.3.2 Kurtosis

In R, we use function *kurtosis* in library *moments* for computation.

```r
#Boys
apply(boys, 2, kurtosis)
```

```
##   height  armspan     foot      leg   weight
## 4.041638 3.153141 3.122427 4.569236 2.565326
```

```r
#Girls
apply(girls, 2, kurtosis)
```

```
##   height  armspan     foot      leg   weight
## 2.452677 2.170296 3.588708 2.563484 2.263160
```

```r
#Total
apply(total, 2, kurtosis)
```

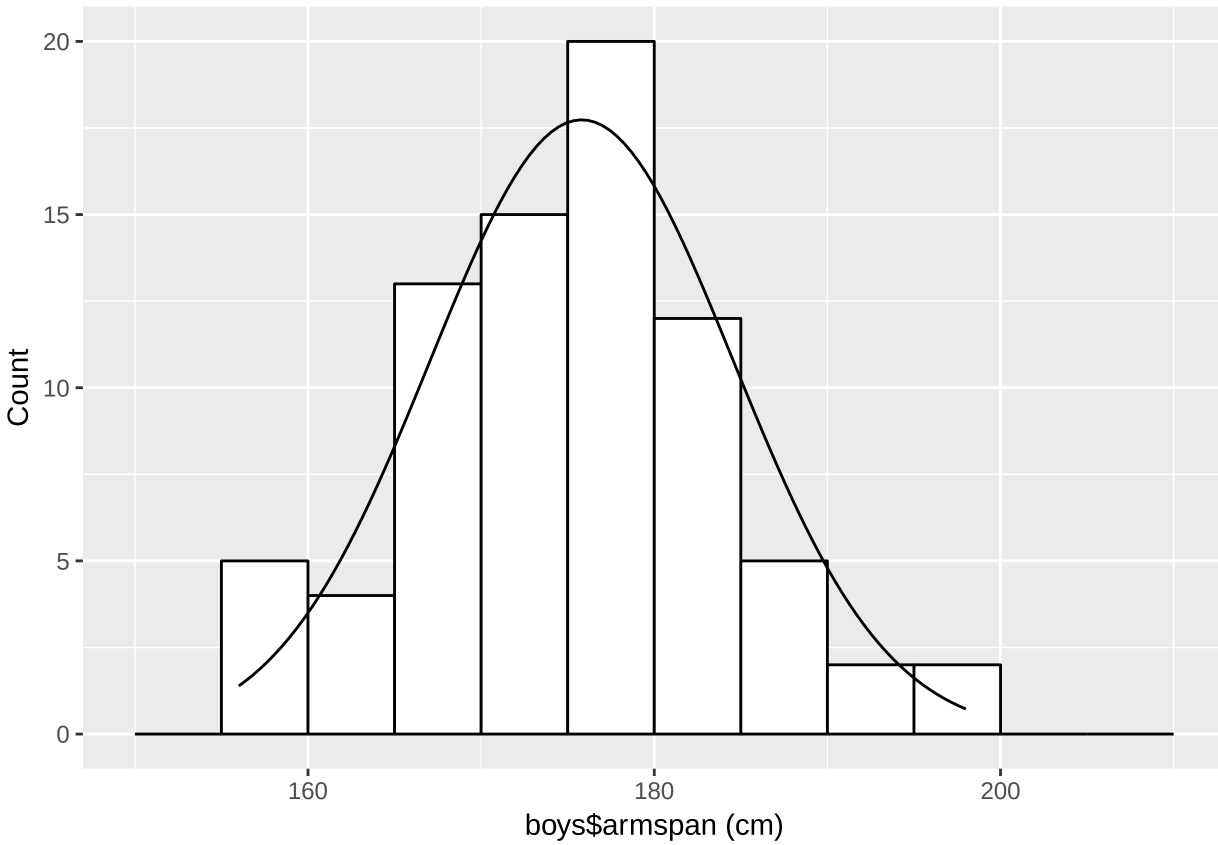
```
##   height  armspan     foot      leg   weight
## 3.287369 2.610451 2.571282 3.894495 2.491807
```

3. Data Distribution

3.1 Histogram.

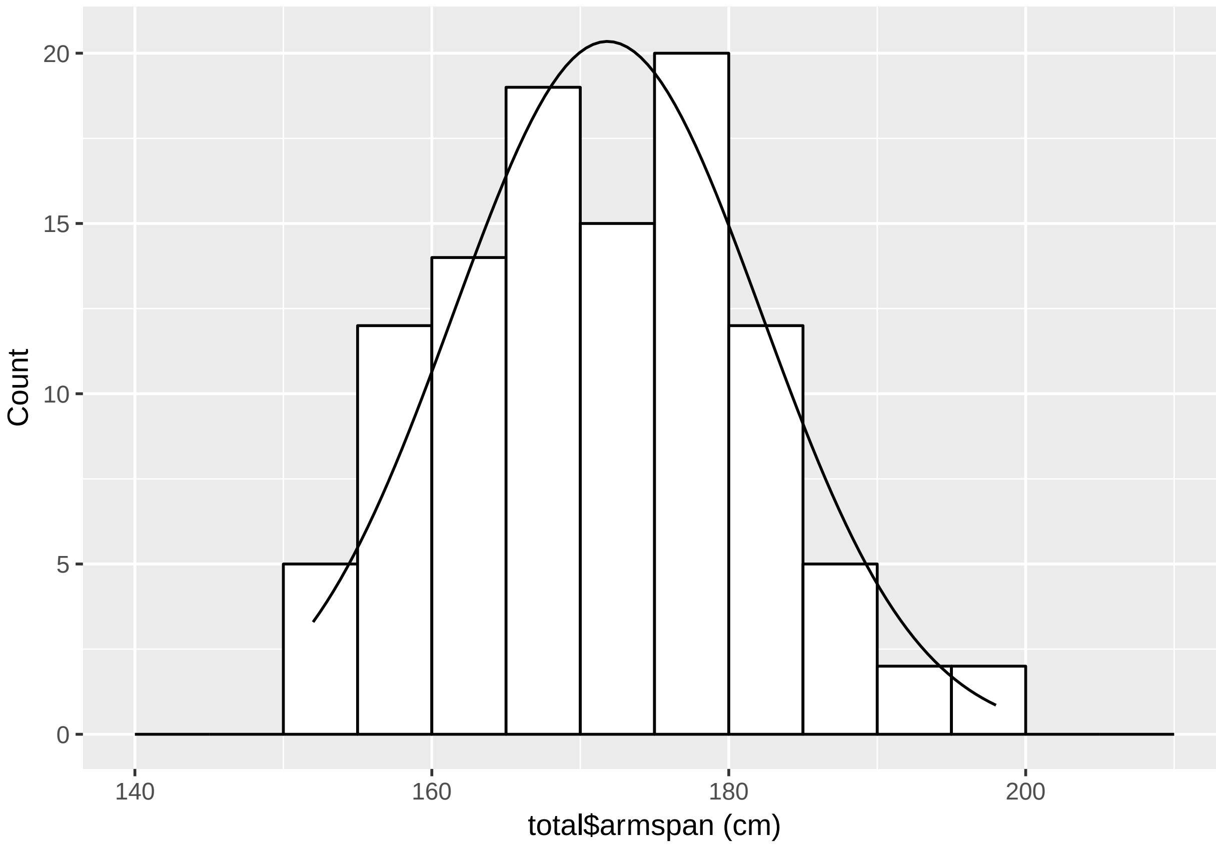
A histogram is an accurate representation of the distribution of numerical data. It is an estimate of the probability distribution of a continuous variable (quantitative variable). In R, we could use function *qplot* in package *ggplot*2 to plot the histrogram. For example, for boys height, we have

```r
n <- nrow(boys)
mean <- mean(boys$armspan)
sd <- sd(boys$armspan)
binwidth <- 5
#Try to fit a normal curve
qplot(boys$armspan, geom = "histogram", breaks = seq(150, 210, binwidth),
      colour = I("black"), fill = I("white"),
      xlab = "boys$armspan (cm)", ylab = "Count") +
  # Create normal curve, adjusting for number of observations and binwidth
  stat_function(
    fun = function(x, mean, sd, n, bw){
      dnorm(x = x, mean = mean, sd = sd) * n * bw
    },
    args = c(mean = mean, sd = sd, n = n, bw = binwidth))
```



After suppossing a normal curve that has mean equal to sample mean of Boys armspan and stand deviation equal to sample stanfard deviation, we could hypothesize that Boys armspan is normal distributed. Simarily, for Total armspan, we make the normal distribution hypothesis.

```
n <- nrow(total)
mean <- mean(total$armspan)
sd <- sd(total$armspan)
binwidth <- 5
#Try to fit a normal curve
qplot(total$armspan, geom = "histogram", breaks = seq(140, 210, binwidth),
      colour = I("black"), fill = I("white"),
      xlab = "total$armspan (cm)", ylab = "Count") +
  # Create normal curve, adjusting for number of observations and binwidth
  stat_function(
    fun = function(x, mean, sd, n, bw){
      dnorm(x = x, mean = mean, sd = sd) * n * bw
    },
    args = c(mean = mean, sd = sd, n = n, bw = binwidth))
```

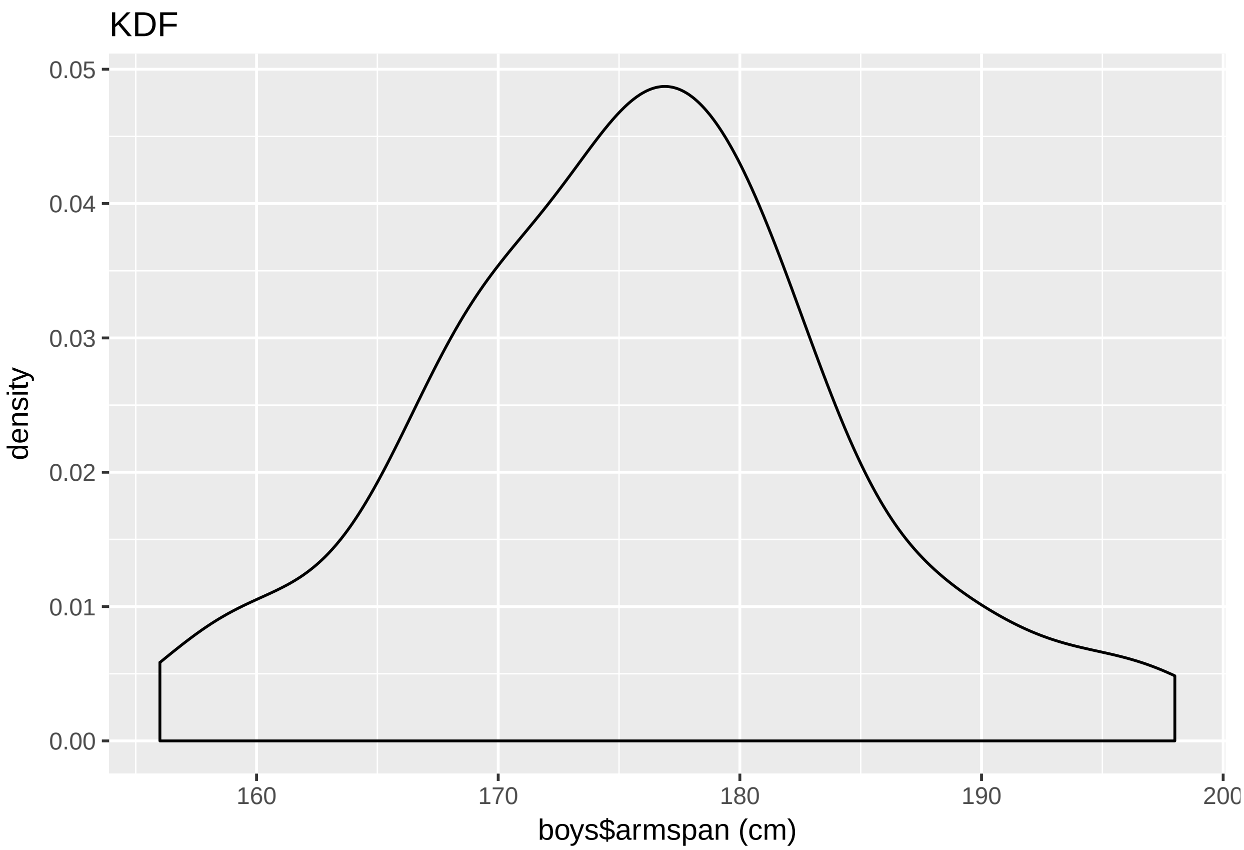


The following discussion is mainly focused on validation of the normal distribution hypothesis of Boys armspan and Total armspan.

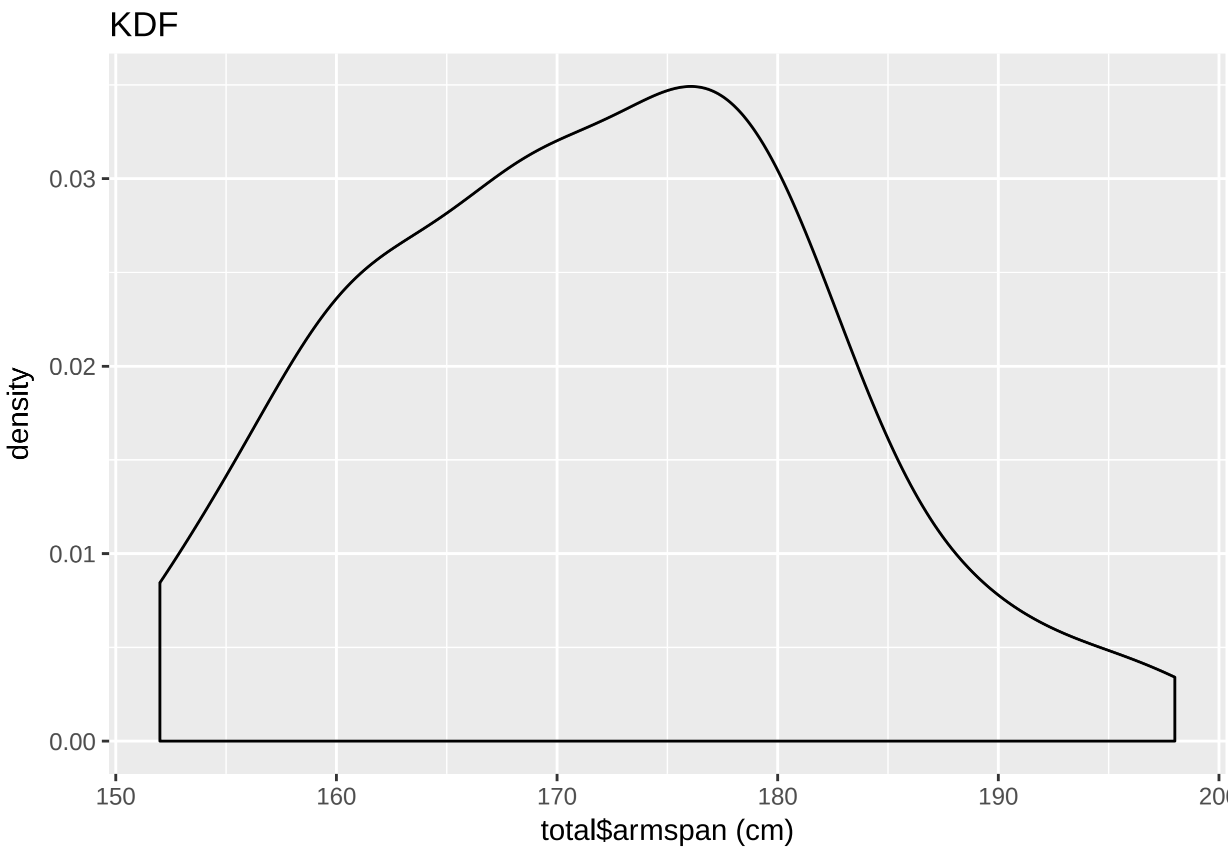
3.2 Kernel Density Estimation

Kernel density estimation (KDE) is a non-parametric way to estimate the probability density function of a random variable. Kernel density estimation is a fundamental data smoothing problem where inferences about the population are made, based on a finite data sample. In R, we use function in $ggplot2 :: geom_density$ to plot the density curve of Boys armspan and total armspan.

```
par(mar=c(1,1,1,1))
ggplot(boys, aes(x=armspan)) + geom_density() + labs(x = "boys$armspan (cm)") + labs(title = "KDF"
```

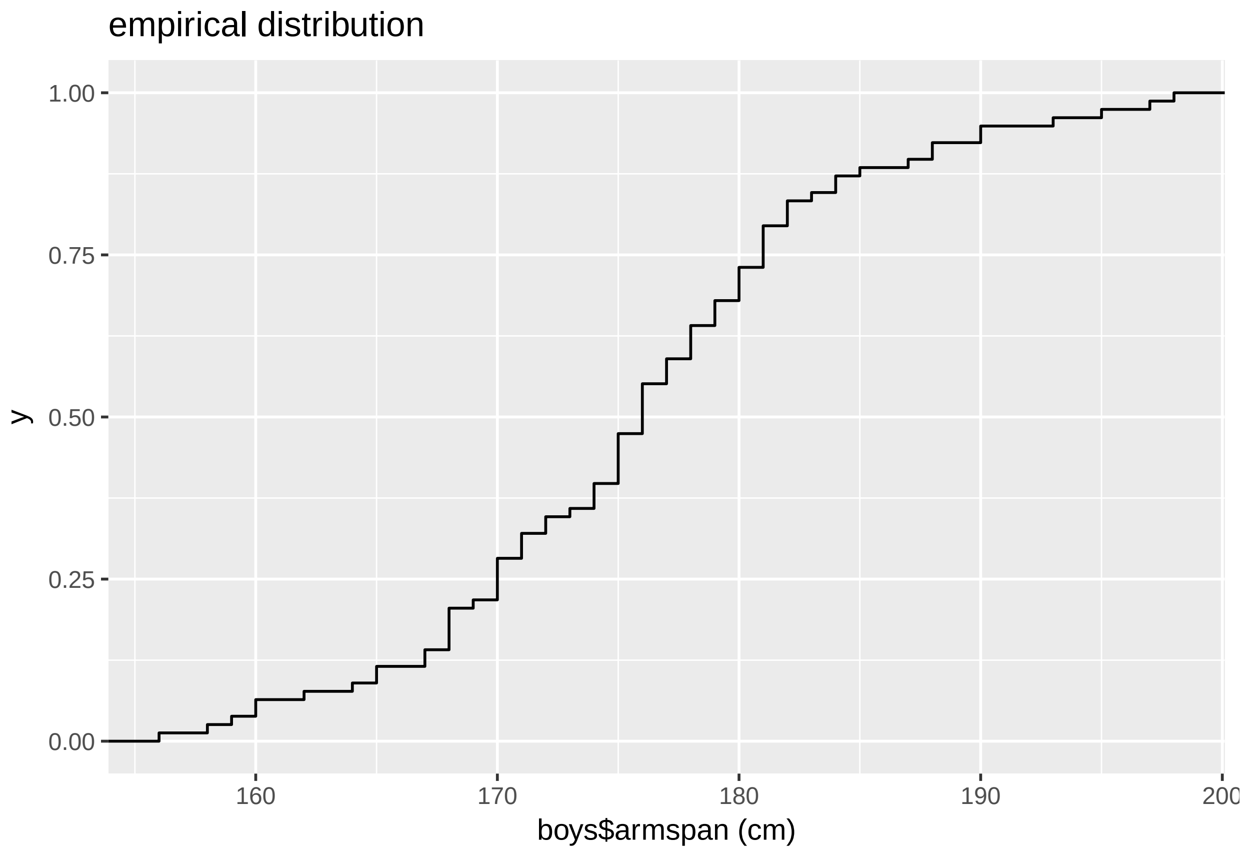



```
ggplot(total, aes(armspan)) + geom_density() + labs(x = "total$armspan (cm)") + labs(title = "KDF"
```

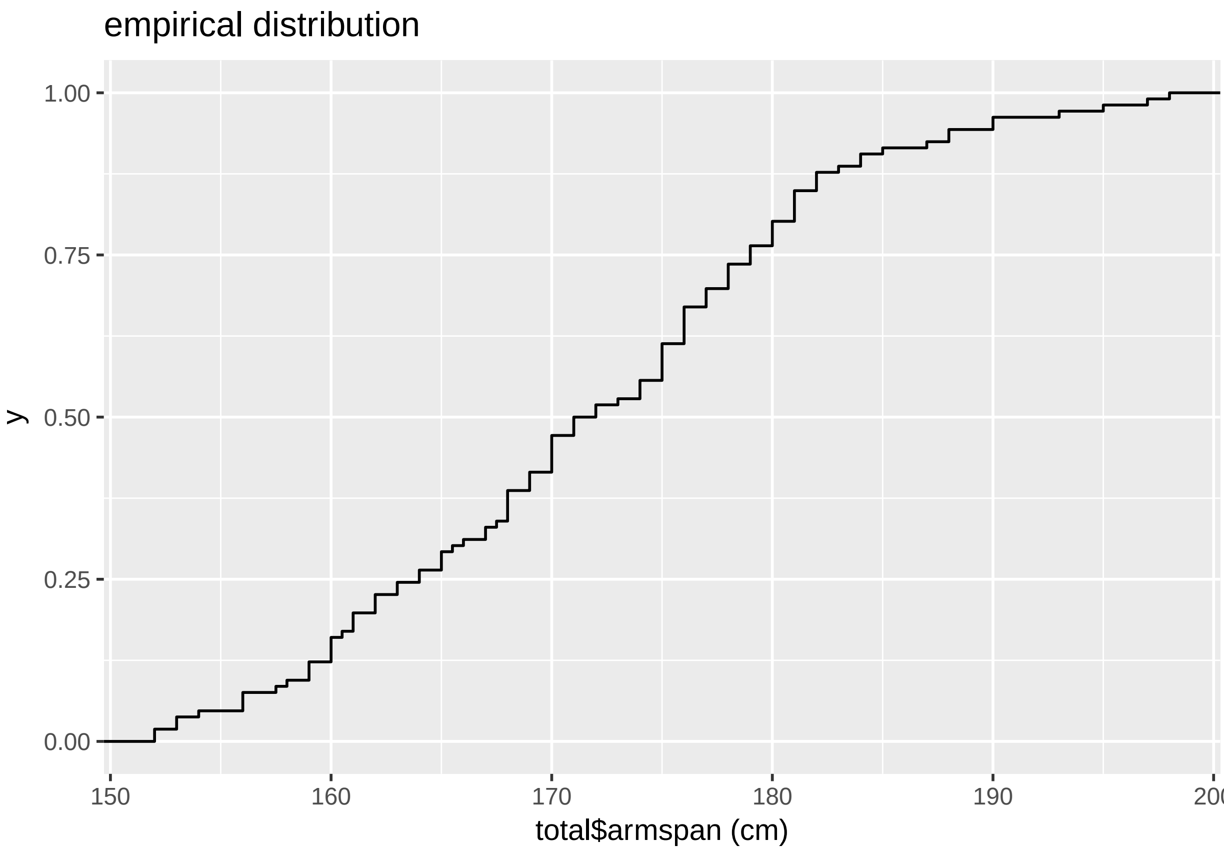


3.2.1 Empidrical distribution

```
par(mar=c(1,1,1,1))
#Boys
ggplot(boys, aes(x=armspan)) + stat_ecdf(geom = "step")+ labs(x = "boys$armspan (cm)") + labs(titl
```



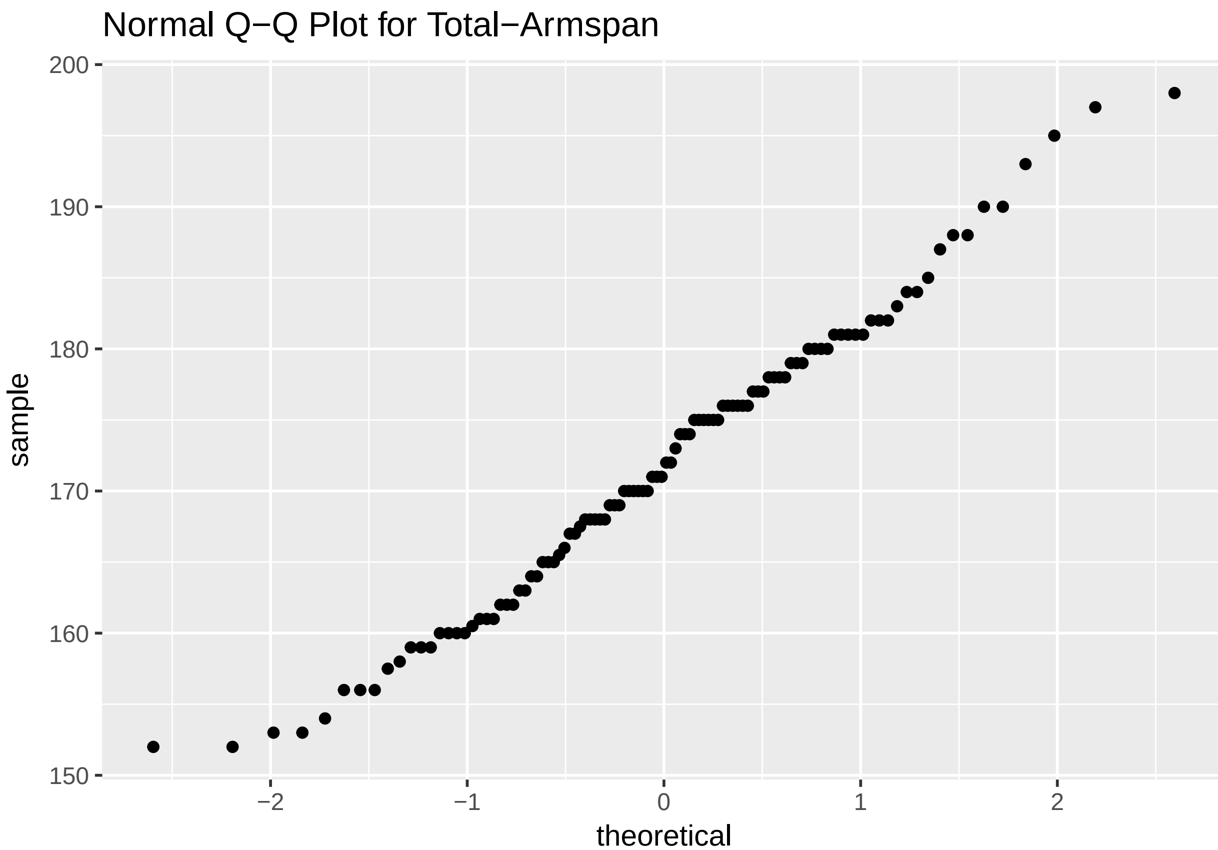

```r
#Total
ggplot(total, aes(armspan)) + stat_ecdf(geom = "step") + labs(x = "total$armspan (cm)") + labs(tit
```



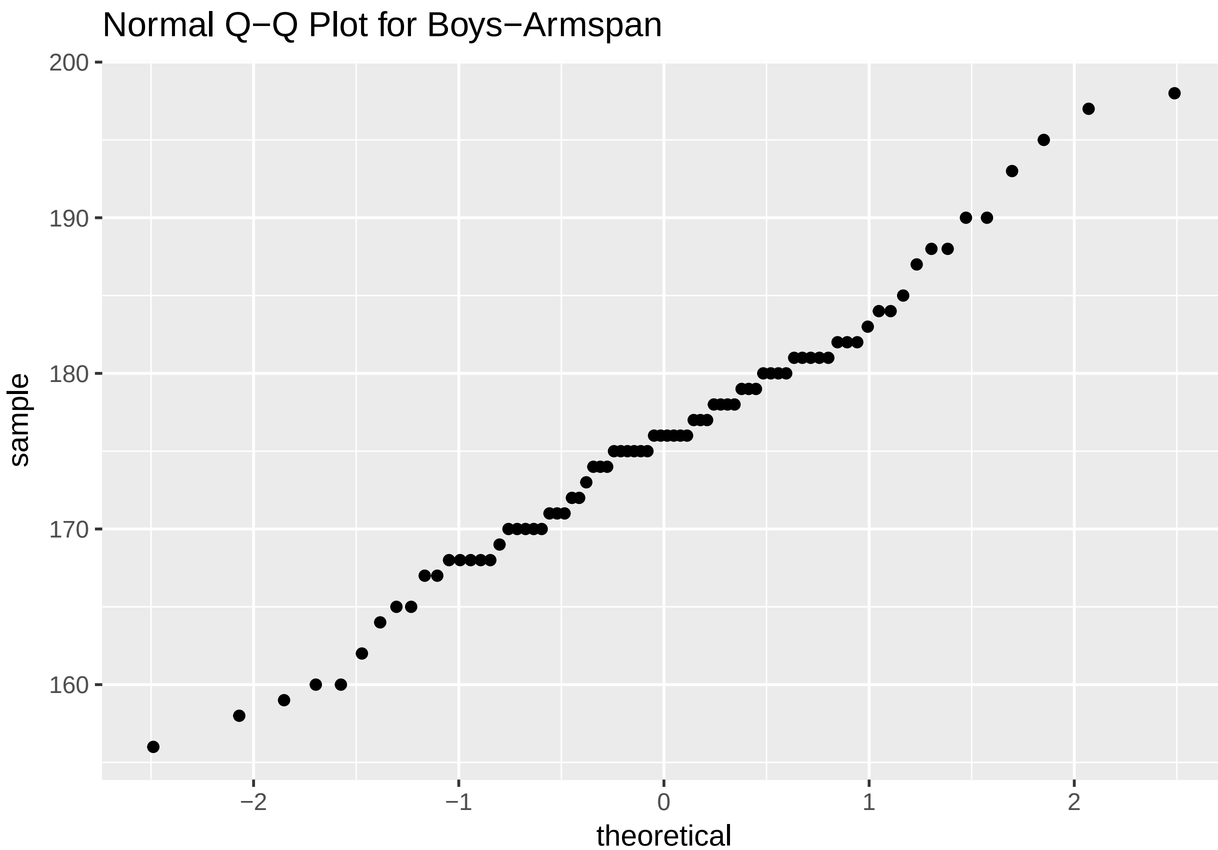
3.2.2 Quantile-Quantile Plot

The quantile-quantile (q-q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. Suppose the population is subject to normal distribution $N(\mu, \sigma^2)$, for sample $x_1, x_2, ..., x_n$, its order statistics are $x_{(1)}, x_{(2)}, ..., x_{(n)}$. Let $\Phi(x)$ be the CDF of standard normal distribution $N(0,1)$, $\Phi^{-1}(x)$ be the inverse function, then the Q-Q plot for normal distribution is a scatter plot consisting of points $(\Phi^{-1}(\frac{i-0.375}{n+0.25}), x_{(i)}), i = 1, 2, ..., n$. If the sample data is subject to normal distribution approximately, then Q-Q plot would be near the straight line $y = \delta x + \mu$. In R, we could use `ggplot2:stat_qq()`.

```
ggplot(total, aes(sample=armspan))+stat_qq()+labs(title = "Normal Q-Q Plot for Total-Armspan")
```




```
ggplot(boys, aes(sample=armspan))+stat_qq()+labs(title = "Normal Q-Q Plot for Boys-Armspan")
```

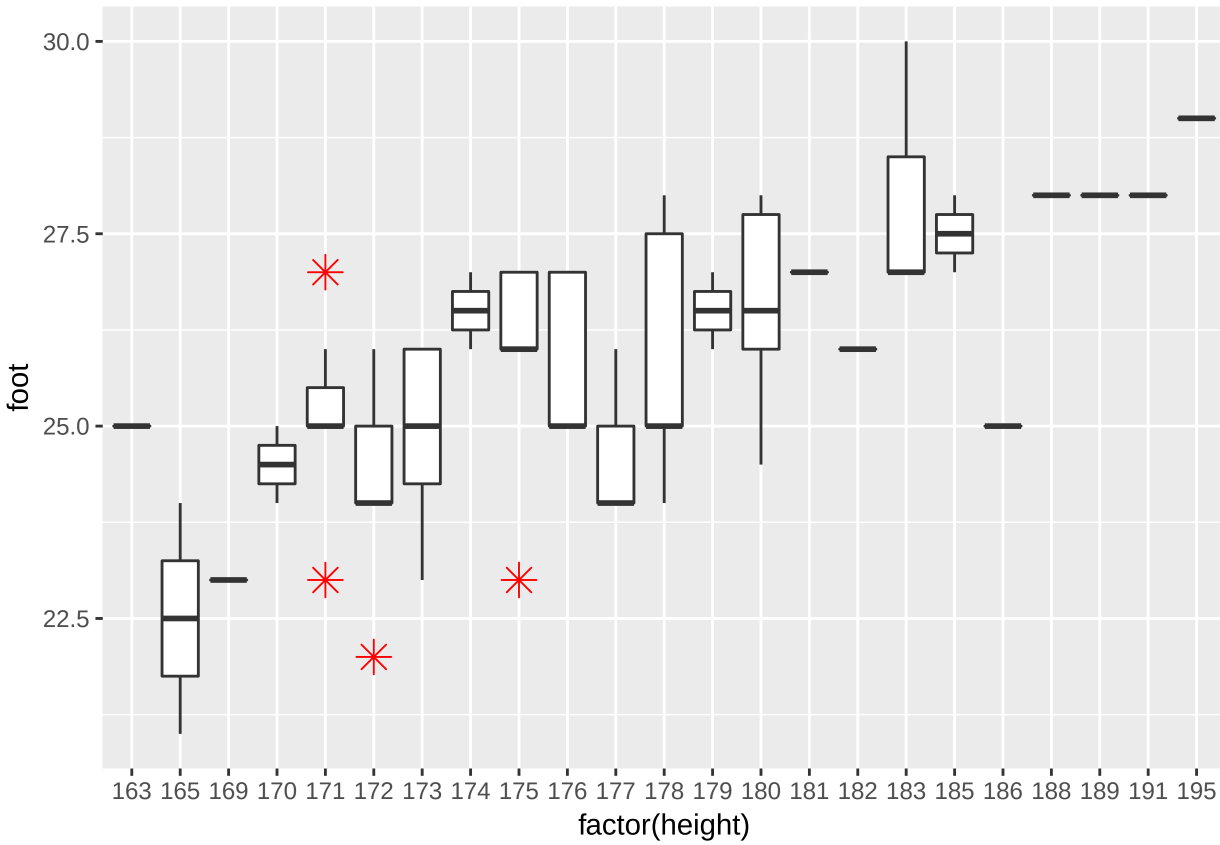


we could infer from above that Boys-Armspan and Total-Armspan are both approximately subject to normal distribution.

3.2.3 Box Plot

A box plot is a method for graphically depicting groups of numerical data through their quartiles. In R, we could use `ggplot2::geom_boxplot()`. For example, we plot the Box Plot of Boys-Height and Boys-Foot-Length, and use red start to denote outliers.

```
ggplot(boys, aes(x=factor(height), y=foot)) + geom_boxplot(outlier.colour="red", outlier.shape=8,
                outlier.size=4)
```



From the above plot, we could see there are 4 outliers, which means the corresponding boy student either has "too long" or "too short" foot length for his height.

4. Hypothesis Test

First, we would perform non-parametric hypothesis test on Total-Height, Total-Armspan, Total-Foot-Length, to infer their distributions. Since we do not have historical data and experience, we could not choose a value to perform parametric hypothesis test. If we use sample mean for the hypothesis test for mean, then the hypothesis would not be rejected, which is make the statisitcal inference meaningless. Therefore, we focus on hypothesis test for the mean, variance in the situation of two normal poplutaions.

4.1 Skewness / Kurtosis Test

First, define Skewness: $g_1 = \frac{B_3}{B_2^{3/2}}$ Kurtosis: $g_2 = \frac{B_4}{B_2^2}$ where $B_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$

Note: the definitions above are different from those in Chapter 2 of the textbook. We make null hypothesis $H_0 : F(x) = F_0(x)$, alternative hypothesis$H_1 : F(x) \neq F_0(x)$, where $F_0(x)$ is the CDF of normal distribution. It could be proven that when $n >> 1$, we have

$$g_1 \sim N(0, \frac{6(n-2)}{(n+1)(n+3)}$$

$$g_2 \sim N(3 - \frac{6}{n+1}, \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$$

Note: It is generally required that $n \geq 100$, so here we only test on Total Students sample.

Denote:

$$\sigma_1^2 = \frac{6(n-2)}{(n+1)(n+3)}$$

$$\sigma_2^2 = \frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}$$

then we have the rejeciton region

$$X_0 = |\frac{B_3}{\sigma_1 B_2^{3/2}}| > u_{1-\alpha/4}$$

or

$$|\frac{B_4/B_2^2 - 3(n-1)/(n+1)}{\sigma_2}| > u_{1-\alpha/4}$$

In R, we define function:

```r
test1 <- function(H) {
   alpha <- 0.05
   H.l <- length(H)
   H.m <- mean(H)
   x<-sum((H-H.m)^3)/H.l;
   y<-sum((H-H.m)^2)/H.l;
   z<-sum((H-H.m)^4)/H.l;
   a1<-6*(H.l-2)/(H.l+1)/(H.l+3);
   a2<-24*H.l*(H.l-2)*(H.l-3)/(H.l+1)^2/(H.l+3)/(H.l+5);
   b<-3*(H.l-1)/(H.l+1);
  r1<-abs(x/(sqrt(a1)*y^(3/2)));
  r2<-abs((z/y^2-b)/sqrt(a2));
  r3<-qnorm(1-alpha/4,0,1);
  r<-(r1>r3)||(r2>r3);
}
```

If return FALSE, then reject; else accept. For example

```r
apply(boys, 2, test1)
```

```
##  height armspan    foot     leg  weight
##    TRUE   FALSE   FALSE    TRUE   FALSE
```

4.2 Sign Test

The **sign test** is a statistical method to test for consistent differences between pairs of observations. In R, we could use *binom.test()* to perform sign test. We make the null hypothesis: there is not significante diference in heights of boys and girls, and the alternative hypothesis: there is significante diference in heights of boys and girls.

```
ngirls<-nrow(girls)
binom.test(sum(sample(boys$height, ngirls)<girls$height), ngirls)
```

```
##
##  Exact binomial test
##
## data:  sum(sample(boys$height, ngirls) < girls$height) and ngirls
## number of successes = 2, number of trials = 28, p-value =
## 3.032e-06
## alternative hypothesis: true probability of success is not equal to 0.5
## 95 percent confidence interval:
##  0.008770497 0.235034773
## sample estimates:
## probability of success
##             0.07142857
```

We couls see from the test result, that the probability of success is very small, which means we should reject the null hypothesis.

4.3 Rank Correlation Test

4.3.1 Spearman's Rank Correlation Test

In R, we use function *cor.test()* to perform the test. For example, we test wether there is a correlation between Girl-Height and Girl-Weight.

```
cor.test(girls$height, girls$weight)
```

```
##
##  Pearson's product-moment correlation
##
## data:  girls$height and girls$weight
## t = 5.253, df = 26, p-value = 1.726e-05
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
```

```
##  0.4703886 0.8603148
## sample estimates:
##       cor
## 0.7175431
```

The Spearman's coefficent is 0.7175431, which suggests a rather high positive correlation, that is, Girl-Height and Girl-Weight is highly positively correlated.

4.3.2 Kendall's Rank Correlation Test

In R, we use *cor.test(,method="kendall")* to perform the test. For example, we test whether there is a correlation between Boys-Foot-Length and Boys-Leg-Length.

```r
cor.test(boys$foot, boys$leg, method="kendall")
```

```
##
##  Kendall's rank correlation tau
##
## data:  boys$foot and boys$leg
## z = 2.9035, p-value = 0.00369
## alternative hypothesis: true tau is not equal to 0
## sample estimates:
##       tau
## 0.2450441
```

We could see that the p-value is very samll, which means a high correlation.

4.3.3 Wilcoxon Coefficient Test

4.3.3.1 Paired Sample Test

In R, we could use *wilcox.test()* to perform the test. In this part, we use rank test to infer whether there is a significant difference in Boys-Armspan and Girls-Armspan. Since the sample number of boys and girls are not equal, we bootstrap the Boys sample, and then perform the pairing.

```r
wilcox.test(sample(boys$armspan, ngirls),girls$armspan, alternative="g", paired=TRUE)
```

```
## Warning in wilcox.test.default(sample(boys$armspan, ngirls), girls
## $armspan, : cannot compute exact p-value with ties

## Warning in wilcox.test.default(sample(boys$armspan, ngirls), girls
## $armspan, : cannot compute exact p-value with zeroes

##
##  Wilcoxon signed rank test with continuity correction
```

```
##
## data:  sample(boys$armspan, ngirls) and girls$armspan
## V = 366, p-value = 1.1e-05
## alternative hypothesis: true location shift is greater than 0
```

Since $p = 3.276e - 06 < 0.05$, we could reject the null hypothesis, which means Boys-Armspan is generally larger than Girls-Armspan.

4.3.3.2 Unpaired Sample Test

In R, we still use _wilcox.test(). In this case, we do not need to bootstrap Boys sample.

```r
wilcox.test(sample(boys$armspan, ngirls),girls$armspan, alternative="g", paired=FALSE)
```

```
## Warning in wilcox.test.default(sample(boys$armspan, ngirls), girls
## $armspan, : cannot compute exact p-value with ties

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  sample(boys$armspan, ngirls) and girls$armspan
## W = 714, p-value = 6.771e-08
## alternative hypothesis: true location shift is greater than 0
```

Since $p = 1.772e - 07 < 0.05$, we could reject the null hypothesis, which means Boys-Armspan is generally larger than Girls-Armspan.

5. Estimation of Parameter

Due to the limit of sample size, we focus on Boys sample in this section.

5.1 Point Estimation

5.1.1 Method of Moments

We could use sample mean and sample variance for population mean and population variance. We have computed the values in previous chapter.

5.1.2 Maximum Likelihood Estimation

In R, we define the function:

```r
mle <- function(H) {
  n <- length(H)
  H.m <- mean(H)
  H.v <- sum((H-H.m)^2)/n
  c(H.m, H.v)
}
```

Then, we perform the MLE for Boys, Girls and Total. Note that in the output, the first row is the estimated mean, and the second row is the estimated variance.

```r
#Boys
apply(boys, 2, mle)
```

```
##         height   armspan      foot      leg   weight
## [1,] 176.30769 175.80769 25.724359 98.20513 68.02564
## [2,]  31.98225  75.97584  2.670817 72.57331 90.87114
```

```r
#Girls
apply(girls, 2, mle)
```

```
##         height   armspan       foot      leg   weight
## [1,] 165.76786 160.75000 23.5571429 94.21429 52.78571
## [2,]  22.70504  26.47321  0.2781633 51.59694 19.57908
```

```r
#Total
apply(total, 2, mle)
```

```
##         height  armspan      foot      leg  weight
## [1,] 173.52358 171.8302 25.151887 97.15094  64.000
## [2,]  51.12444 106.9712  2.951742 70.12816 117.184
```

5.2 Interval Estimation

5.2.1 One Normal Population

1. Mean In R, we could use *t.test()*. For example,

```r
apply(boys, 2, t.test)
```

```
## $height
##
##  One Sample t-test
##
## data:  newX[, i]
```

```
## t = 273.57, df = 77, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  175.0244 177.5910
## sample estimates:
## mean of x
##  176.3077
##
##
## $armspan
##
##  One Sample t-test
##
## data:  newX[, i]
## t = 176.99, df = 77, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  173.8297 177.7857
## sample estimates:
## mean of x
##  175.8077
##
##
## $foot
##
##  One Sample t-test
##
## data:  newX[, i]
## t = 138.12, df = 77, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  25.35350 26.09521
## sample estimates:
## mean of x
##  25.72436
##
##
## $leg
##
```

```
##  One Sample t-test
##
## data:  newX[, i]
## t = 101.16, df = 77, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##   96.27196 100.13830
## sample estimates:
## mean of x
##  98.20513
##
##
## $weight
##
##  One Sample t-test
##
## data:  newX[, i]
## t = 62.619, df = 77, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  65.86245 70.18883
## sample estimates:
## mean of x
##  68.02564
```

From above, we could get the mean estimation and the corresponding confidence intervals for Boys sample. 2. Variance In R, we define the function:

```
interval_var1<-function(x,mu=Inf,alpha=0.05){
   n<-length(x);
   if(mu<Inf){
   S2<-sum((x-mu)^2)/n;
   df<-n;
   }
   else {
   S2<-var(x);
   df<-n-1;
   }
  a<-df*S2/qchisq(1-alpha/2,df)
  b<-df*S2/qchisq(alpha/2,df)
  data.frame(var=S2,df=df,a=a,b=b)
```

```
}
```

For example, on Boys sample:

```
apply(boys,2,interval_var1)
```

```
## $height
##       var df        a        b
## 1 32.3976 77 24.18245 45.66939
##
## $armspan
##        var df        a        b
## 1 76.96254 77 57.44692 108.4905
##
## $foot
##        var df       a        b
## 1 2.705503 77 2.01946 3.813821
##
## $leg
##        var df        a        b
## 1 73.51582 77 54.87419 103.6318
##
## $weight
##        var df        a        b
## 1 92.05128 77 68.70956 129.7604
```

For the output above, [a,b] stands for the 95% CI.

5.2.2 Two Normal Populations

Interval estimation of $\mu_1 - \mu 2$

```
interval_estimate2<-function(x,y,sigma=c(-1,1),var.equal=FALSE,alpha=0.05){
  n1<-length(x); n2<-length(y);
    xb<-mean(x); yb<-mean(y)
 if(all(sigma>=0)){
  tmp<-qnorm(1-alpha/2)*sqrt(sigma[1]^2/n1+sigma[2]^2/n2);
 df<-n1+n2;}
 else {
 if (var.equal ==  TRUE){
  Sw<-((n1-1)*var(x)+(n2-1)*var(y))/(n1+n2-2)
  tmp<-sqrt(Sw*(1/n1+1/n2))*qt(1-alpha/2,n1+n2-2)
```

```
df<-n1+n2=2;}
else {
S1<-var(x);S2<-var(y);
 nu<-(S1/n1+S2/n2)^2/(S1^2/n1^2/(n1-1)+S2^2/n2^2/(n2-1))
tmp<-qt(1-alpha/2, nu)*sqrt(S1/n1+S2/n2)
df<-nu
}
}
 data.frame(mean=xb-yb, df=df, a=xb-yb-tmp, b=xb-yb+tmp)
}
```

For example, we perform the interval estimation for Boys-Height and Girls-Height.

```
interval_estimate2(boys$height, girls$height)
```

```
##       mean       df        a        b
## 1 10.53984 55.51021 8.294087 12.78558
```

From the output, we could see the estimated $\mu_1 - \mu_2 = 10.53994$, and the CI $= [8.294087, 12.78558]$.

6. Regression Analysis

6.1 Covariance of multivariate data

We could use *cor()* to get the covaraince matrix of Boys, Girls, and Total.

```
cov(boys)
```

```
##            height   armspan     foot       leg    weight
## height  32.397602 40.280719 6.020979 25.611389 23.511489
## armspan 40.280719 76.962537 9.530719 27.286713 26.953047
## foot     6.020979  9.530719 2.705503  5.154679  4.929237
## leg     25.611389 27.286713 5.154679 73.515818 29.371295
## weight  23.511489 26.953047 4.929237 29.371295 92.051282
```

```
cov(girls)
```

```
##            height    armspan      foot        leg    weight
## height  23.545966 15.8379630 0.9341270  5.1441799 15.689153
## armspan 15.837963 27.4537037 0.8592593  7.4814815 15.129630
## foot     0.934127  0.8592593 0.2884656  0.2613757  1.146032
## leg      5.144180  7.4814815 0.2613757 53.5079365  4.029101
## weight  15.689153 15.1296296 1.1460317  4.0291005 20.304233
```

```r
cov(total)
```

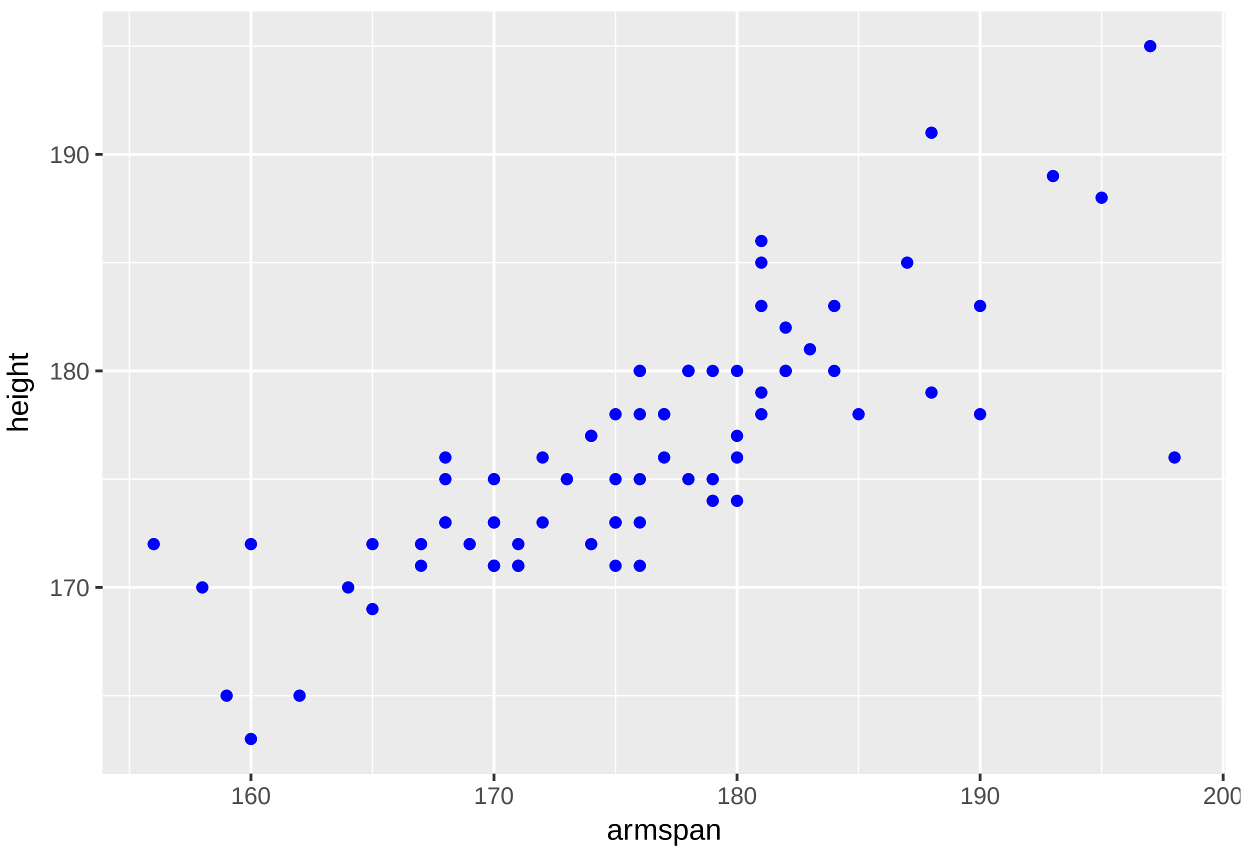
```
##            height   armspan      foot       leg    weight
## height  51.611343  64.75404  9.137812 28.358311  52.79524
## armspan 64.754043 107.98994 13.613657 33.725876  68.68571
## foot     9.137812  13.61366  2.979854  5.544474  10.39048
## leg     28.358311  33.72588  5.544474 70.796047  34.50952
## weight  52.795238  68.68571 10.390476 34.509524 118.30000
```

6.2 Linear Regression Analysis

In R, we could use *lm()* to fit the linear model In this part, we test whether there is a linear relationship between Height and Armspan, Foot-Length and Leg-Lenghth. Due to the limit of sample size, we perform the test on Boys sample.

1. Boys-Height & Boys-Weight

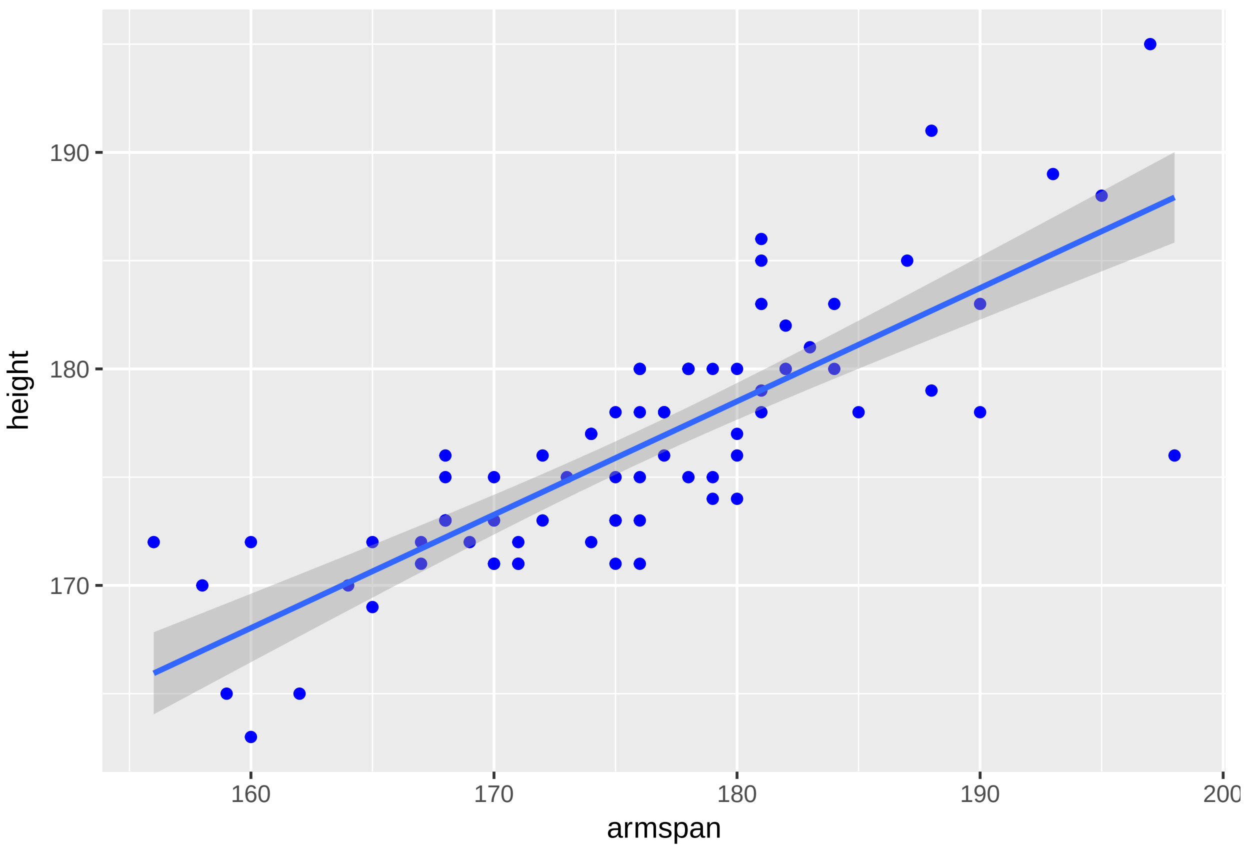
```r
#Plot the data
bhba.plot <- ggplot(boys, aes(x=armspan, y=height)) + geom_point(color='blue')
bhba.plot
```



```r
#Fit the linear model
bhba.lm <- lm(boys$height ~ 1 + boys$armspan)
bhba.lm
```

```
##
## Call:
## lm(formula = boys$height ~ 1 + boys$armspan)
##
## Coefficients:
##  (Intercept)  boys$armspan
##      84.2933        0.5234
```

```r
#Add to the plot
bhba.plot + geom_smooth(method='lm')
```



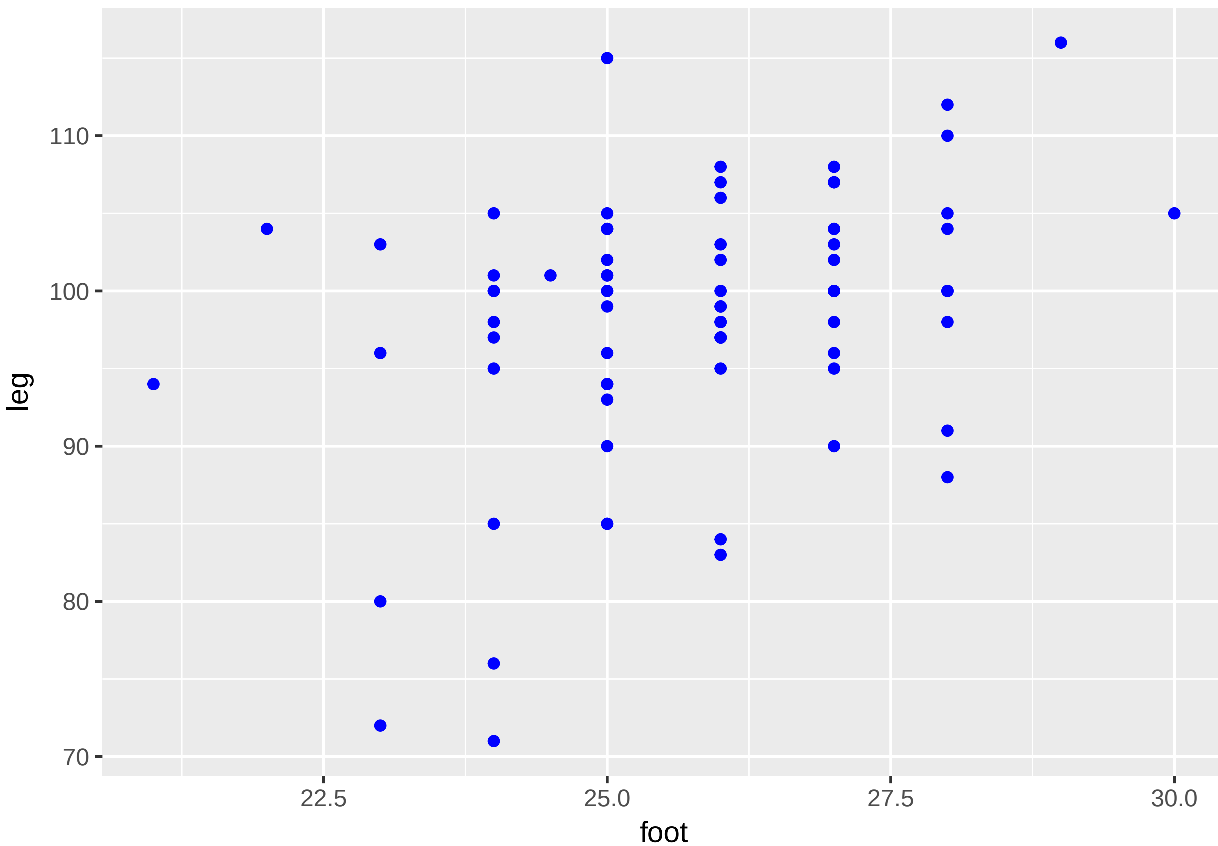
From the output of the *lm()*, we could obtain the regression function:

$$Boys.Height = 84.2933 + 0.5234 * Boys.Armspan$$

2. Boys-Foot-Length & Boys-Leg-Length

```r
#Plot the data
bfbl.plot <- ggplot(boys, aes(x=foot, y=leg)) + geom_point(color='blue')
```

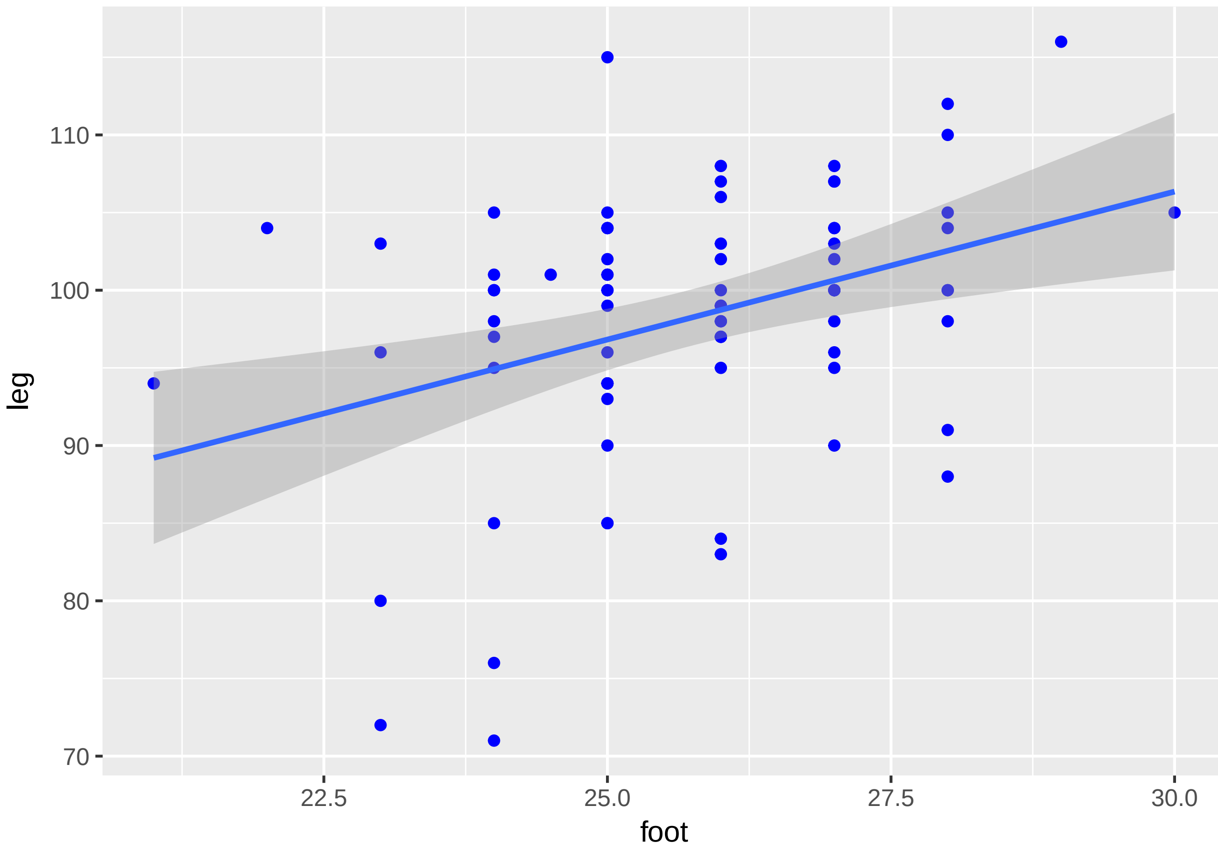
```r
bfbl.plot
```




```r
#Fit the linear model
bfbl.lm <- lm(boys$leg ~ 1 + boys$foot)
bfbl.lm
```

```
##
## Call:
## lm(formula = boys$leg ~ 1 + boys$foot)
##
## Coefficients:
## (Intercept)    boys$foot
##      49.194        1.905
```

```r
#Add to the plot
bfbl.plot + geom_smooth(method='lm')
```



From the last plot, we could see the linear relaitonship merely exists.