Distribution Fitting

Tonnar Castellano

Introduction

In the following blog post we are going to look at different ways of fitting a model. It is important to realize that none of these models are perfect. However, we will see how different methods do different things for fitting.

```r
library(dplyr)
library(tidyverse)
library(bbmle)
```

```r
Hmisc::getHdata(nhgh)
d1 <- nhgh %>%
  filter(sex == "female") %>%
  filter(age >= 18) %>%
  select(gh, ht) %>%
  filter(1:n()<=1000)
```

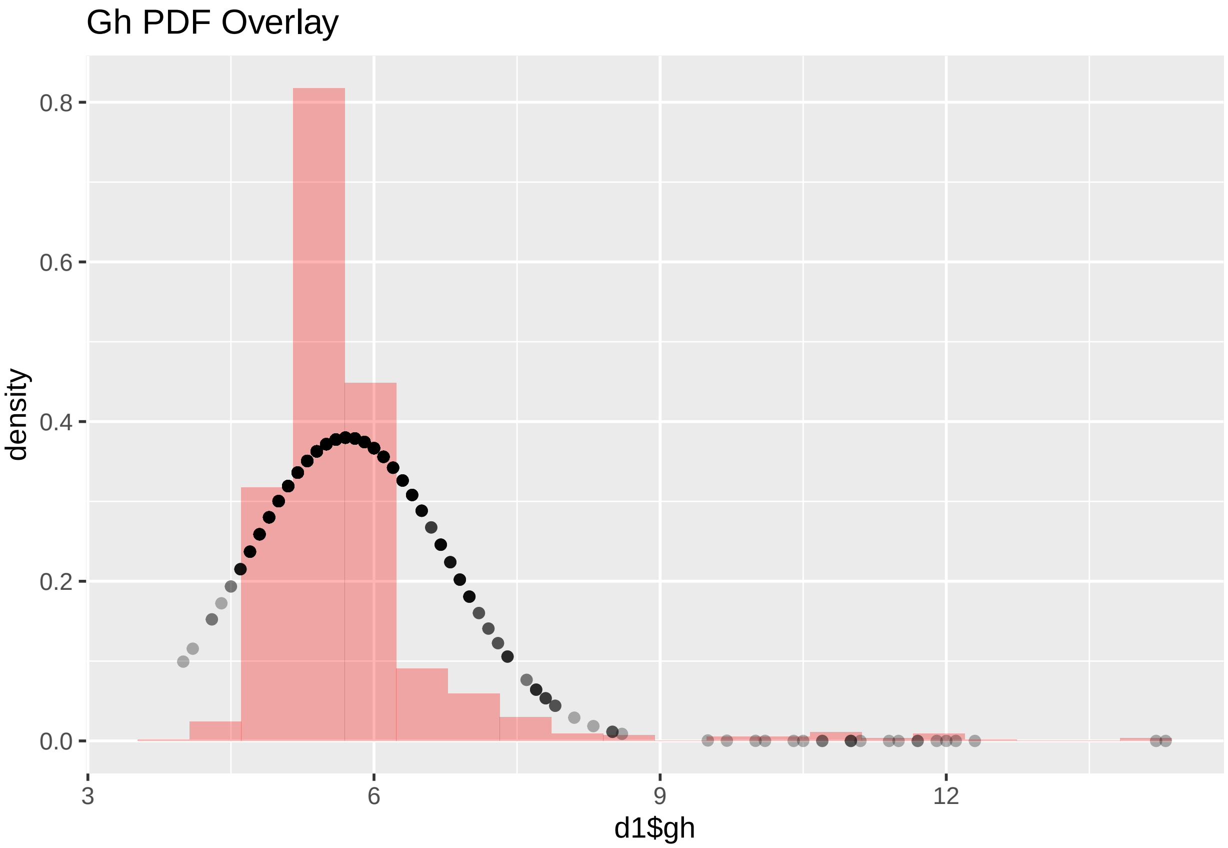
Normal

Normal MLE

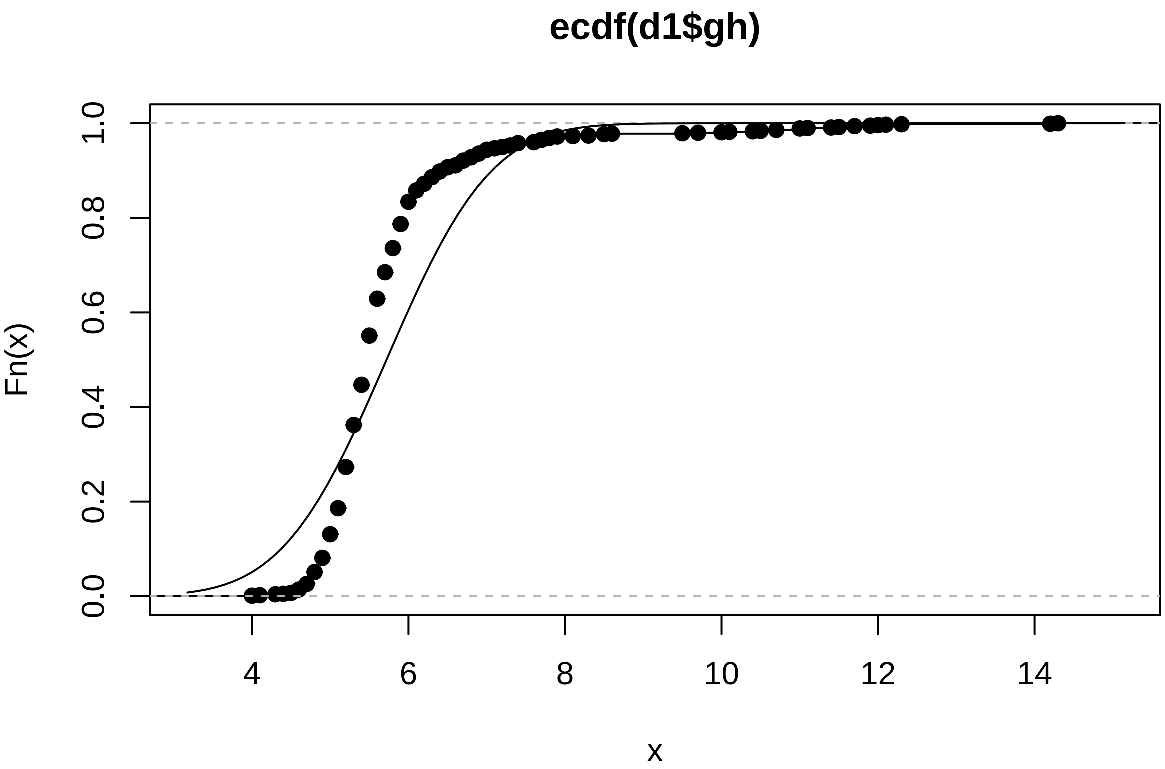
```r
neg_log_lik_gaussian <- function(mu=0,sigma=0.01) {
  -sum(dnorm(d1$gh, mean=mu, sd=sigma, log=TRUE))
}

mle_norm_gh <- mle(neg_log_lik_gaussian)
pdf_norm_gh <- dnorm(d1$gh,mean = 5.72,sd = 1.05)

df_gh <- data.frame(d1$gh,pdf_norm_gh)

neg_log_lik_gaussian <- function(mu=0,sigma=0.01) {
  -sum(dnorm(d1$ht, mean=mu, sd=sigma, log=TRUE))
}

mle_norm_ht <- mle(neg_log_lik_gaussian)
pdf_norm_ht <- dnorm(d1$ht,mean= 160.74,sd = 7.32)

df_ht <- data.frame(d1$ht,pdf_norm_ht)
```

Normal MLE Graphs

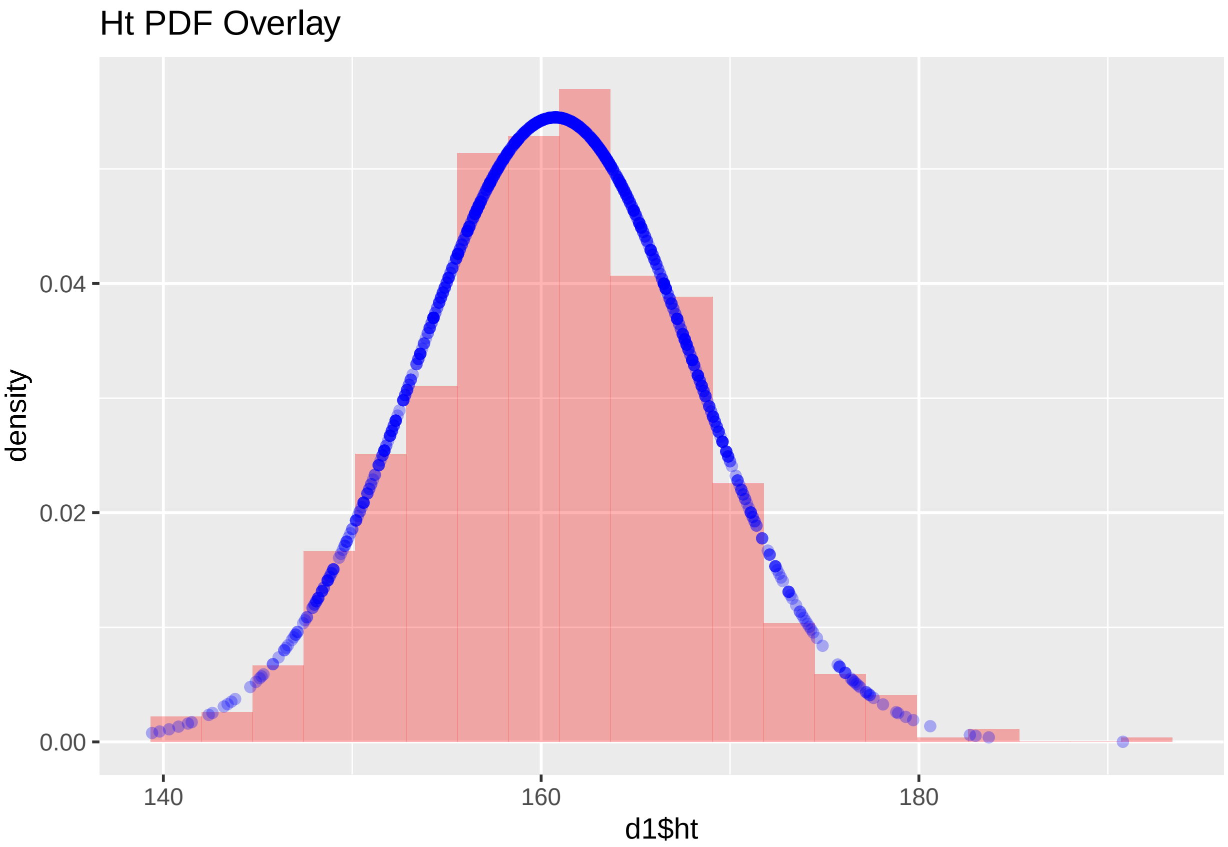
```r
ggplot(df_gh) +
  geom_histogram(aes(x = d1$gh, y =..density..), fill = 'red', position = 'identity', alpha =.3, bins =
  geom_point(aes(y = pdf_norm_gh, x = d1$gh),fill = 'blue', position = 'identity', alpha =.3)+
  labs(
    title = 'Gh PDF Overlay'
  )
```



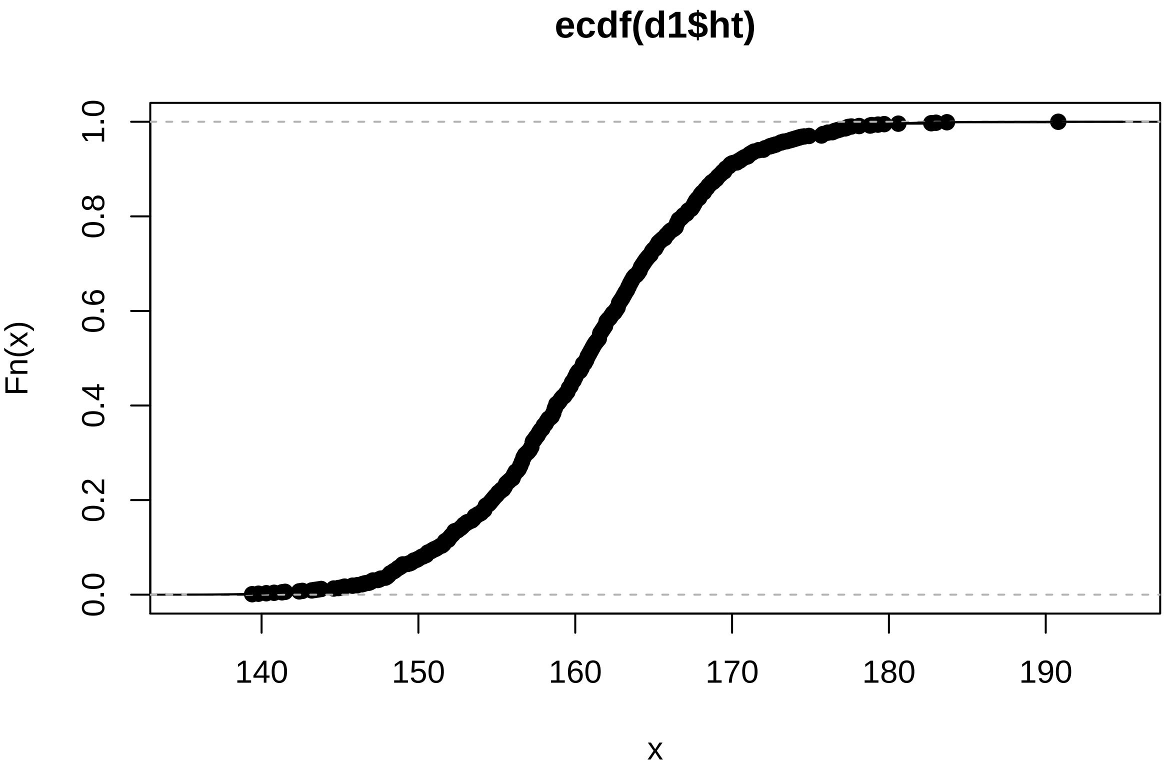
```
plot(ecdf(d1$gh))
curve(pnorm(x,mean = 5.72,sd = 1.05), add = TRUE)
```



```
ggplot(df_ht) +
  geom_histogram(aes(x = d1$ht, y= ..density..), fill = 'red', position = 'identity', alpha =.3, bins =
  geom_point(aes(y = pdf_norm_ht, x = d1$ht), color = 'blue', position = 'identity', alpha =.3)+
  labs(
    title = 'Ht PDF Overlay'
  )
```



```r
plot(ecdf(d1$ht))
curve(pnorm(x,mean = 160.74,sd = 7.32),add = TRUE)
```



Normal MLE Median

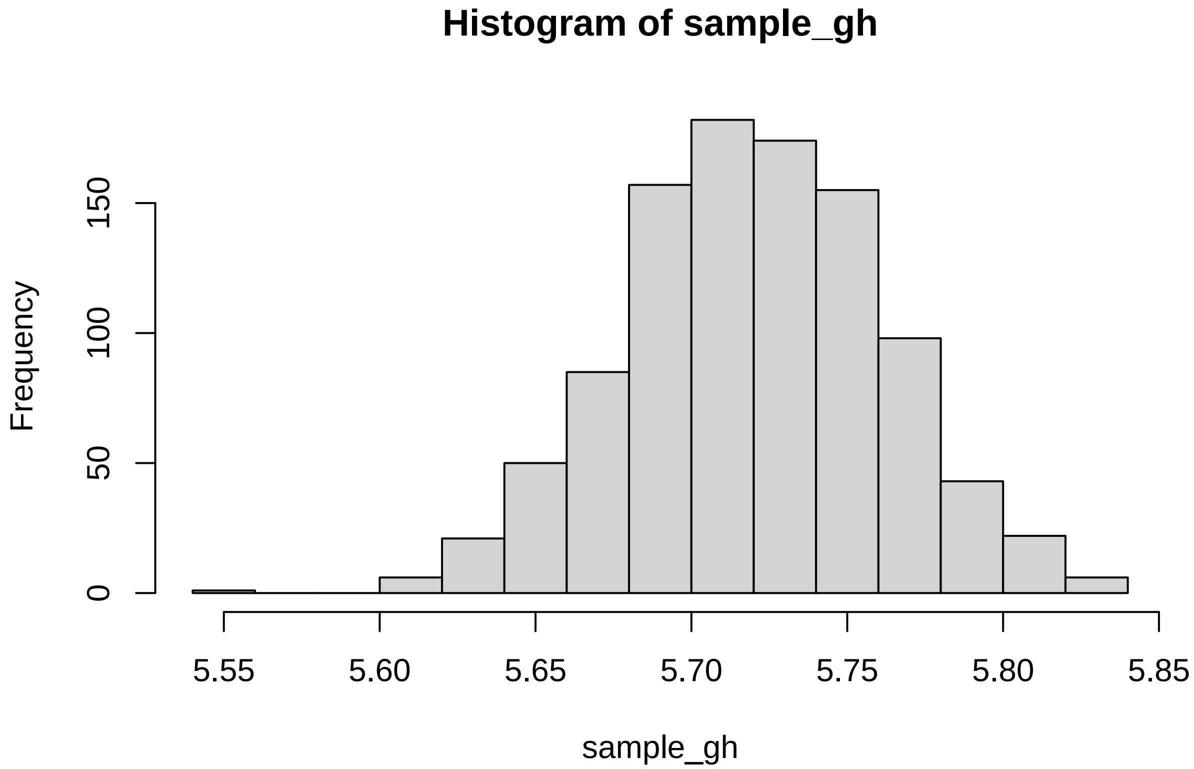
```
est_gh_med <- qnorm(.5,mean = 5.72,sd = 1.05)
est_ht_med <- qnorm(.5,mean = 160.74,sd = 7.32)

print(est_gh_med)
```

```
## [1] 5.72
```

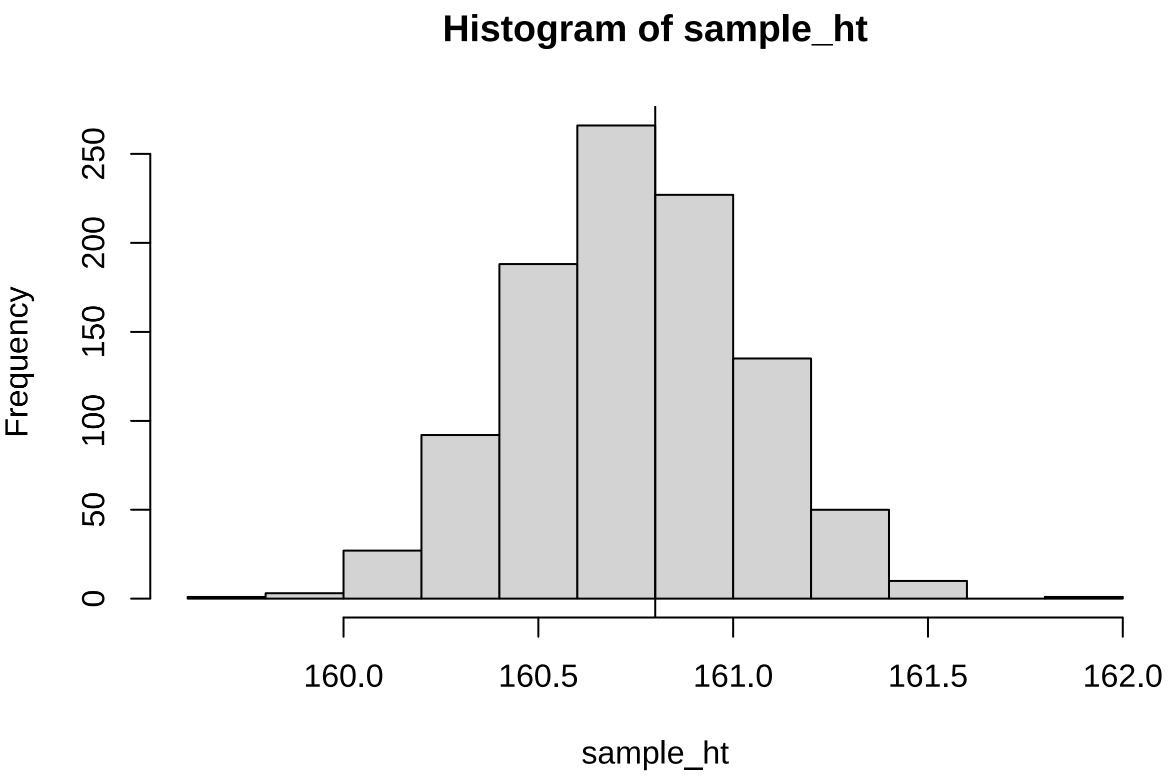
```
print(est_ht_med)
```

```
## [1] 160.74
```

```
sample_gh <- rep(NA,1000)
for(i in c(1:1000)){
sample_gh[i] <- median(rnorm(1000,mean = 5.72,sd = 1.05))
}

hist(sample_gh)
abline(v=median(d1$gh))
```



```
sample_ht <- rep(NA,1000)
for(i in c(1:1000)){
sample_ht[i] <- median(rnorm(1000,160.74,sd = 7.32))
}

hist(sample_ht)
abline(v=median(d1$ht))
```



Normal MLE Range

```
norm_gh_range <- quantile(probs = c(.025,.975),sample_gh)
norm_ht_range <- quantile(probs = c(.025,.975),sample_ht)

print(norm_gh_range)
```

```
##     2.5%    97.5%
## 5.638256 5.800253
```

```
print(norm_ht_range)
```

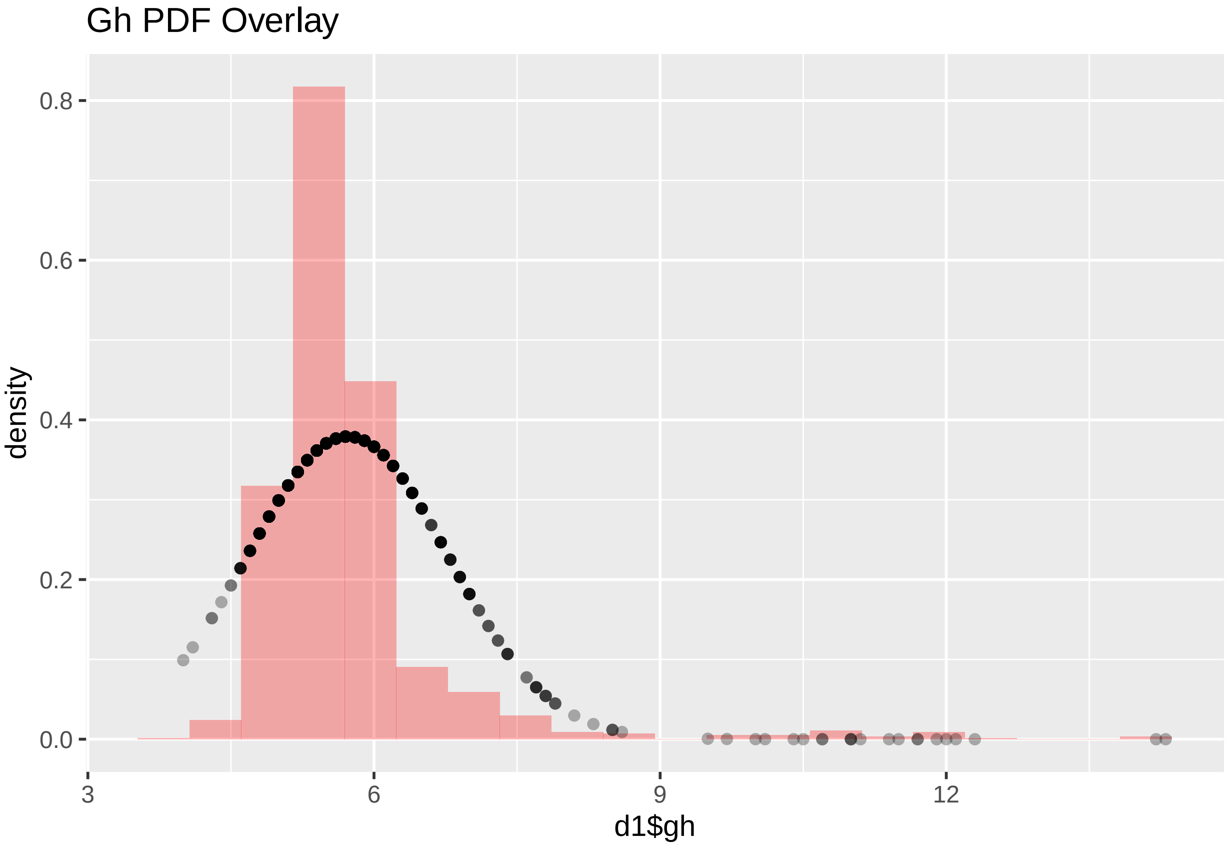
```
##     2.5%    97.5%
## 160.1772 161.3216
```

Normal MM

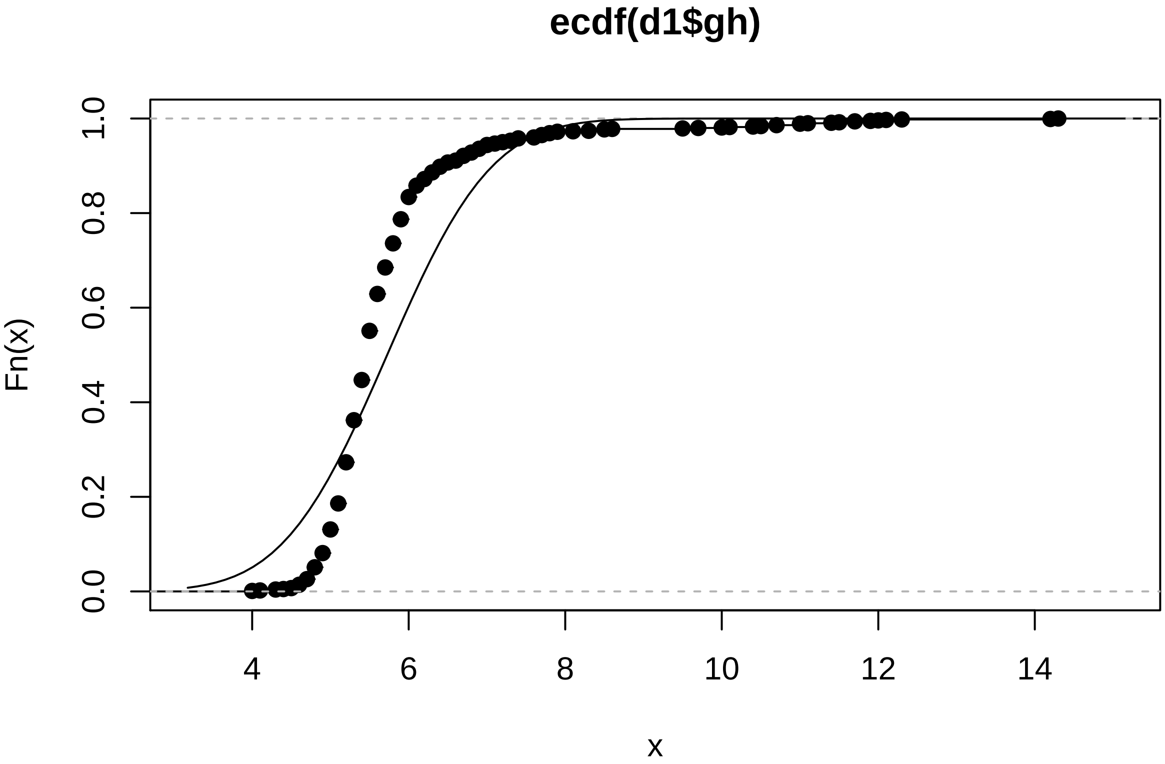
```
gh_norm_mean <- mean(d1$gh)
gh_norm_sd <- sqrt(var(d1$gh))
pdf_norm_gh <- dnorm(d1$gh,mean = gh_norm_mean,sd = gh_norm_sd)

df_gh <- data.frame(d1$gh,pdf_norm_gh)

ht_norm_mean <- mean(d1$ht)
ht_norm_sd <- sqrt(var(d1$ht))
pdf_norm_ht <- dnorm(d1$ht,mean = ht_norm_mean,sd = ht_norm_sd)

df_ht <- data.frame(d1$ht,pdf_norm_ht)
```

Normal MM Graphs

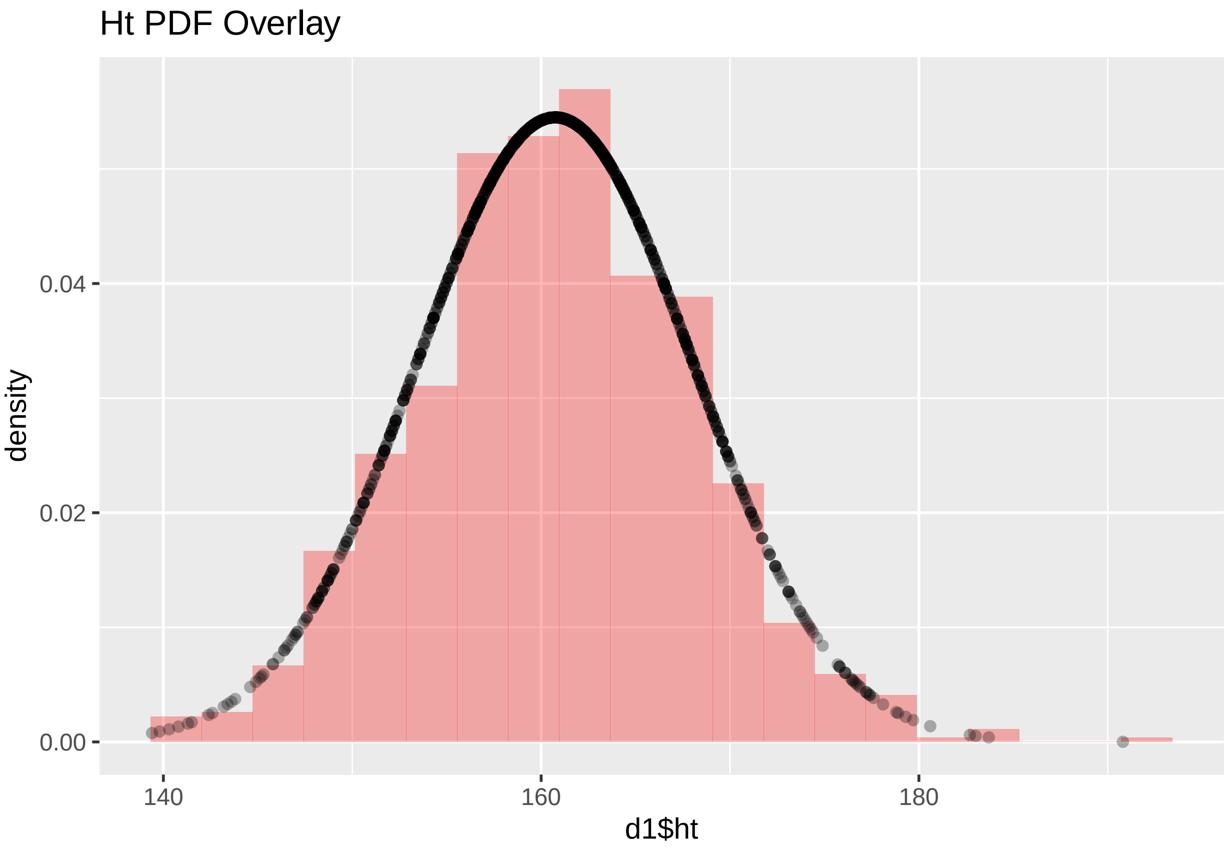
```r
ggplot(df_gh) +
  geom_histogram(aes(x = d1$gh, y =..density..), fill = 'red', position = 'identity', alpha =.3, bins =
  geom_point(aes(y = pdf_norm_gh, x = d1$gh),fill = 'blue', position = 'identity', alpha =.3)+
  labs(
    title = 'Gh PDF Overlay'
  )
```



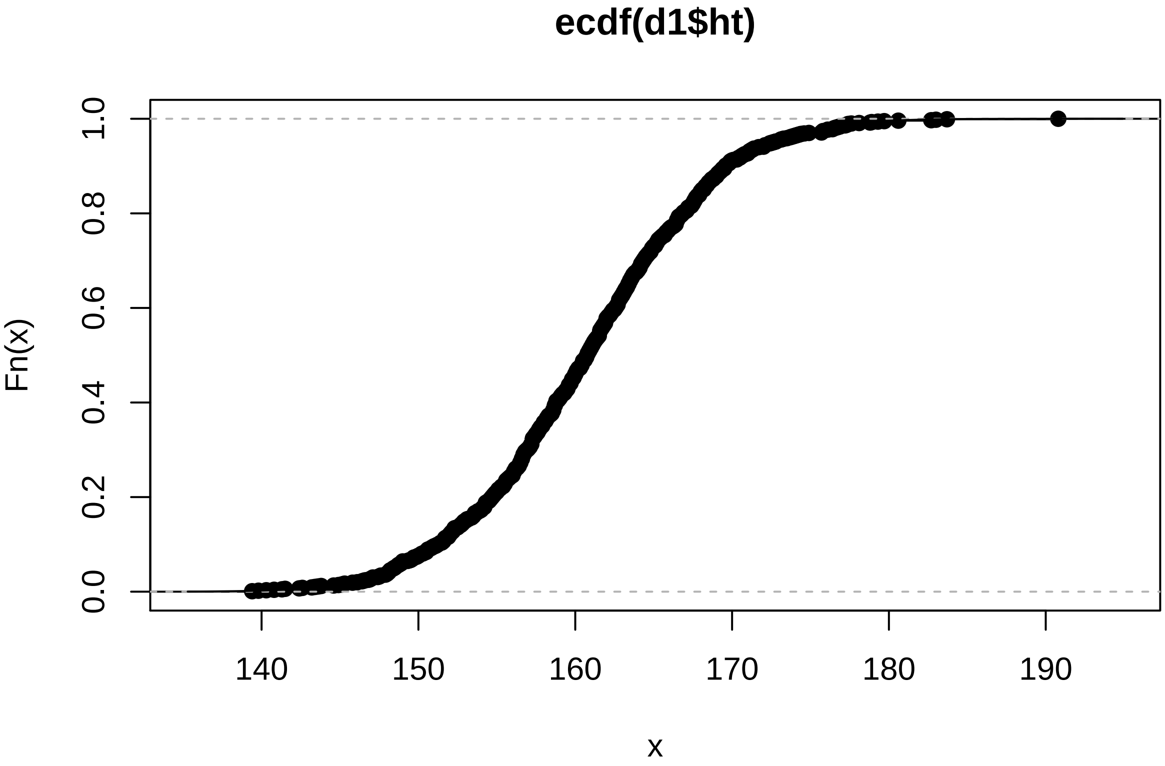
```r
plot(ecdf(d1$gh))
curve(pnorm(x,mean = gh_norm_mean, sd = gh_norm_sd), add = TRUE)
```



```
ggplot(df_ht) +
  geom_histogram(aes(x = d1$ht, y =..density..), fill = 'red', position = 'identity', alpha =.3, bins =
  geom_point(aes(y = pdf_norm_ht, x = d1$ht),fill = 'blue', position = 'identity', alpha =.3)+
  labs(
    title = 'Ht PDF Overlay'
  )
```



```
plot(ecdf(d1$ht))
curve(pnorm(x,mean = ht_norm_mean, sd = ht_norm_sd), add = TRUE)
```



Normal MM Median

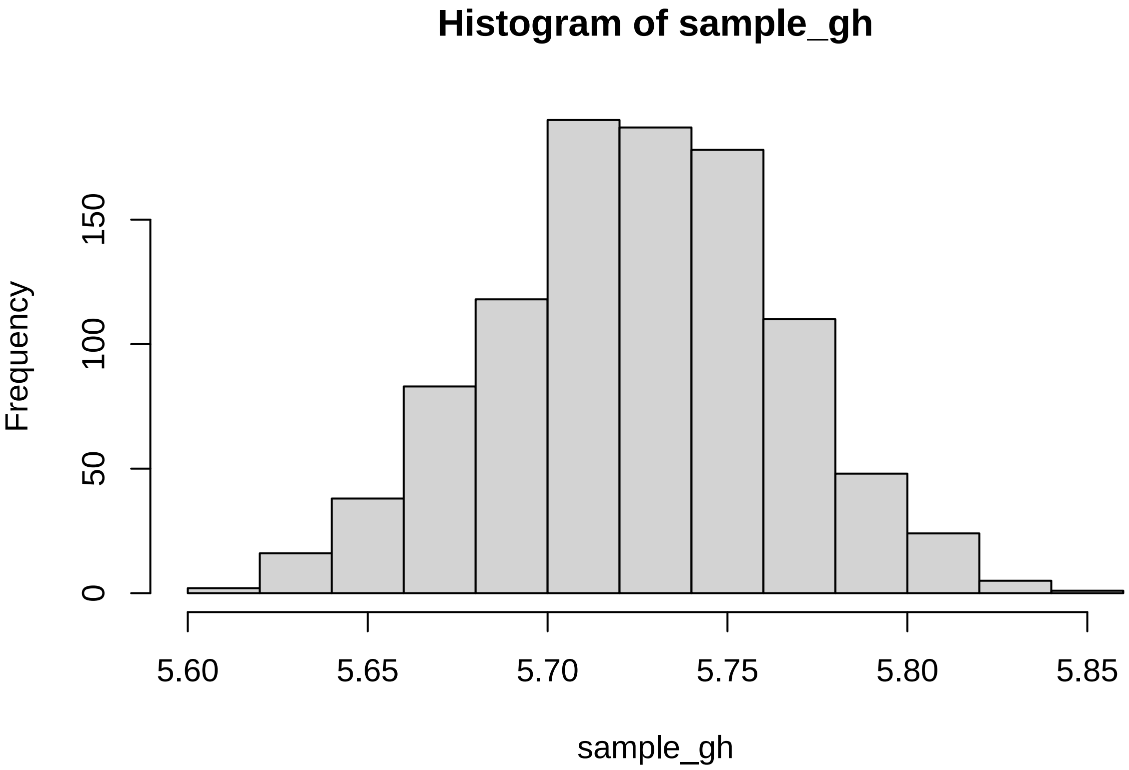
```
est_gh_med <- qnorm(.5,gh_norm_mean, gh_norm_sd)
est_ht_med <- qnorm(.5,ht_norm_mean, ht_norm_sd)
print(est_gh_med)
```

```
## [1] 5.7246
```

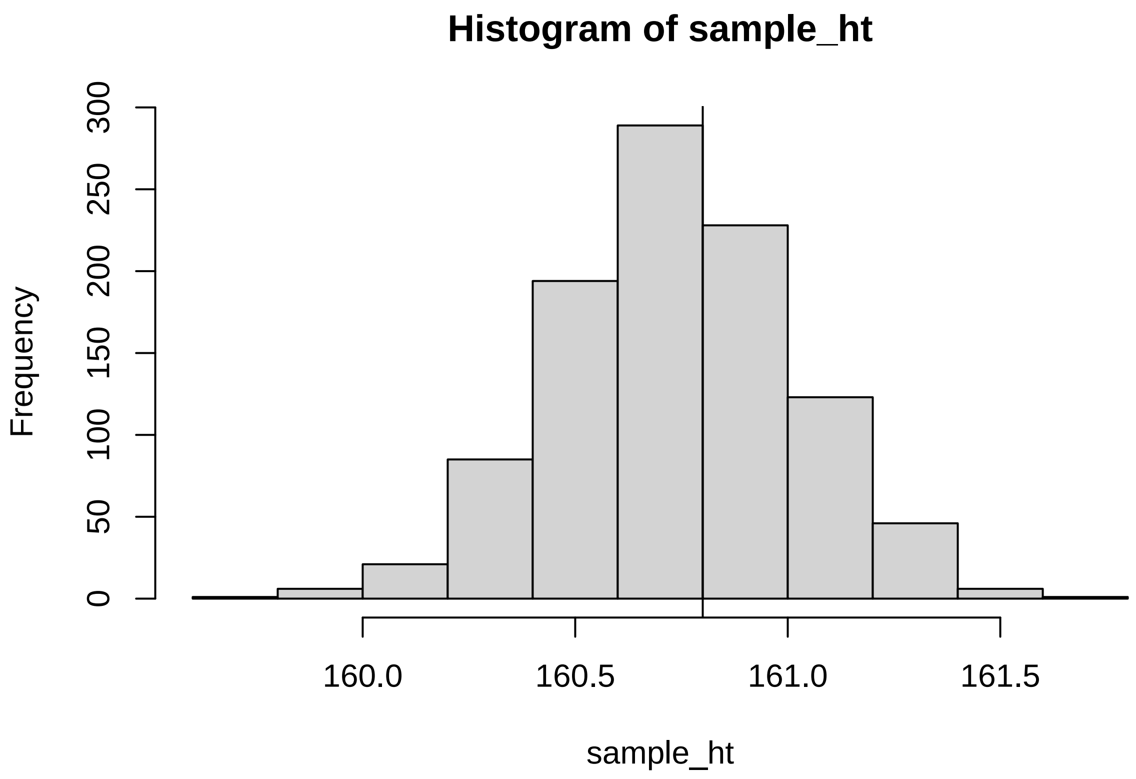
```
print(est_ht_med)
```

```
## [1] 160.7419
```

```
sample_gh <- rep(NA,1000)
for(i in c(1:1000)){
sample_gh[i] <- median(rnorm(1000,mean = gh_norm_mean, sd = gh_norm_sd))
}

hist(sample_gh)
abline(v=median(d1$gh))
```



```
sample_ht <- rep(NA,1000)
for(i in c(1:1000)){
sample_ht[i] <- median(rnorm(1000,mean = ht_norm_mean, sd = ht_norm_sd))
}

hist(sample_ht)
abline(v=median(d1$ht))
```



Normal MM Range

```
norm_gh_range <- quantile(probs = c(.025,.975),sample_gh)
norm_ht_range <- quantile(probs = c(.025,.975),sample_ht)
print(norm_gh_range)
```

```
##     2.5%    97.5%
## 5.642927 5.802874
```

```
print(norm_ht_range)
```

```
##     2.5%    97.5%
## 160.1835 161.2768
```

Gamma

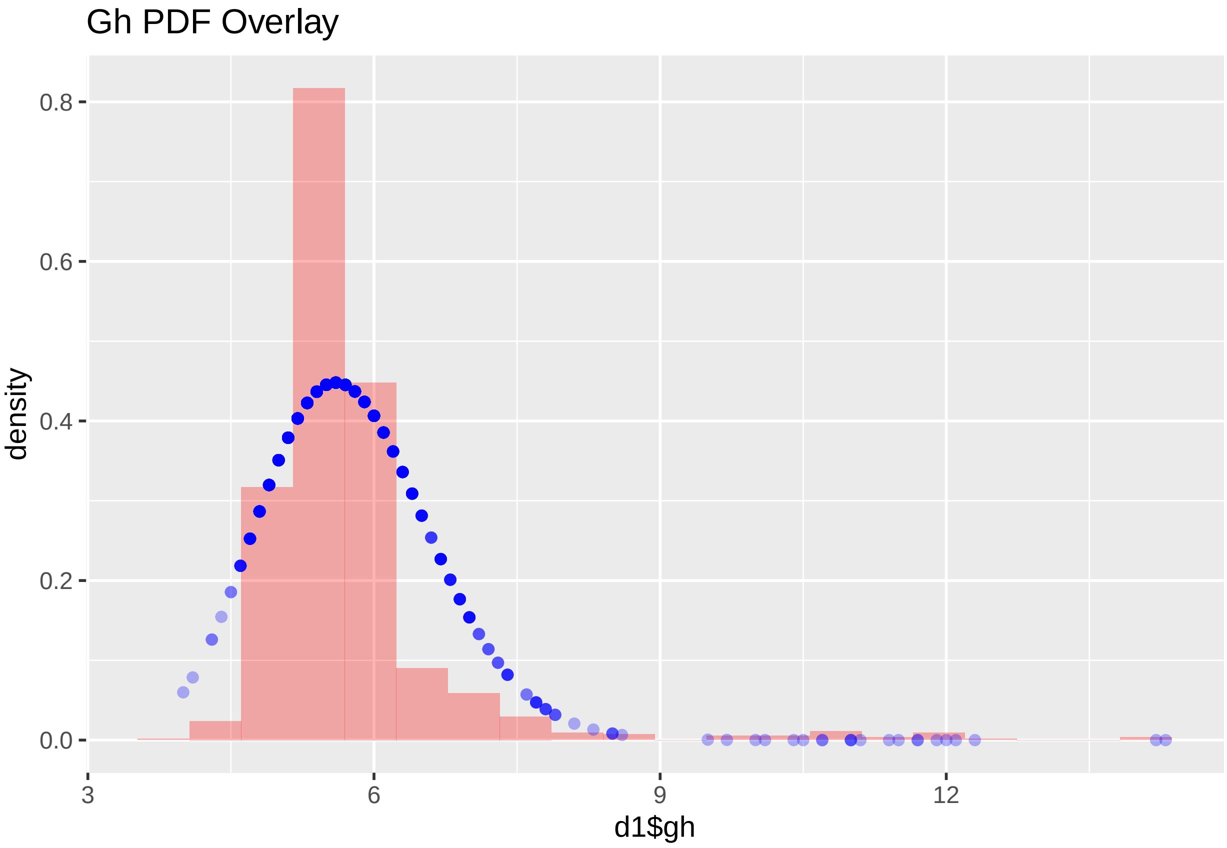
Gamma MLE

```
neg_log_lik_gamma <- function(alpha,beta) {
  -sum(dgamma(d1$gh, shape=alpha, scale = beta, log=TRUE))
}

mle_gamma_gh <- mle(neg_log_lik_gamma,
          start=list(alpha=0.01, beta=0.01))

pdf_gamma_gh <- dgamma(d1$gh,shape = 40.706,scale = 0.141)

df_gh <- data.frame(d1$gh,pdf_gamma_gh)
```

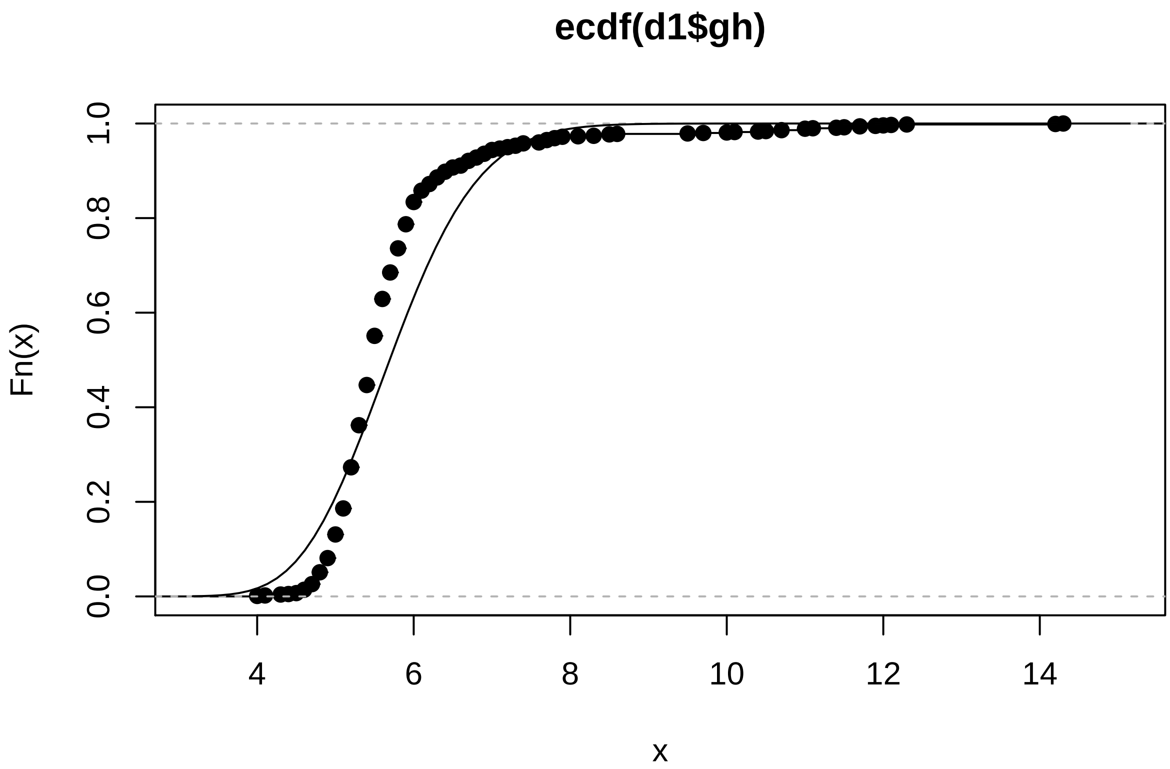
```
neg_log_lik_gamma <- function(alpha,beta) {
  -sum(dgamma(d1$ht, shape=alpha, scale = beta, log=TRUE))
}

mle_gamma_ht <- mle(neg_log_lik_gamma,
          start=list(alpha=0.01, beta=0.01))

pdf_gamma_ht <- dgamma(d1$ht,shape = 400.1318856,scale = 0.401)

df_ht <- data.frame(d1$ht,pdf_gamma_ht)
```

Gamma MLE Graphs

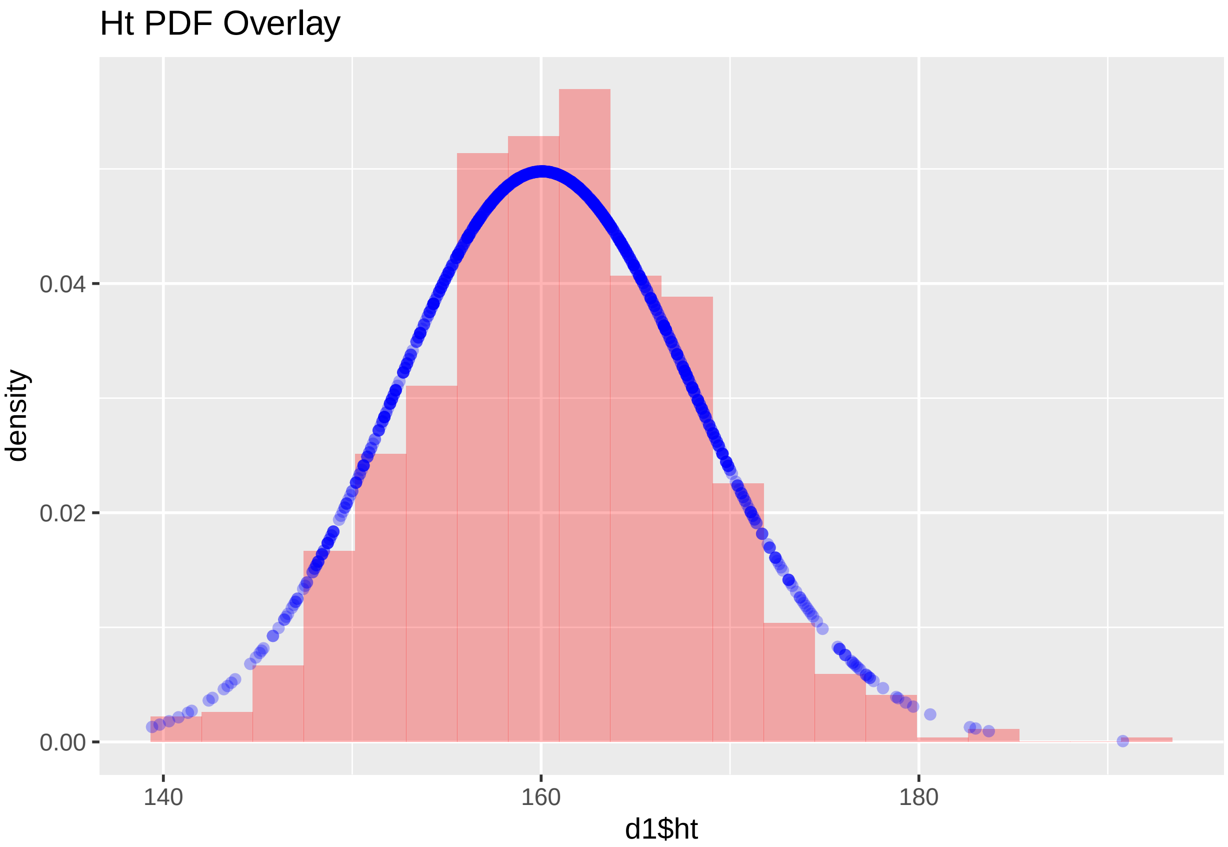
```
ggplot(df_gh) +
  geom_histogram(aes(x = d1$gh, y= ..density..), fill = 'red', position = 'identity', alpha =.3, bins =
  geom_point(aes(y = pdf_gamma_gh, x = d1$gh), color = 'blue', position = 'identity', alpha =.3)+
  labs(
    title = 'Gh PDF Overlay'
  )
```



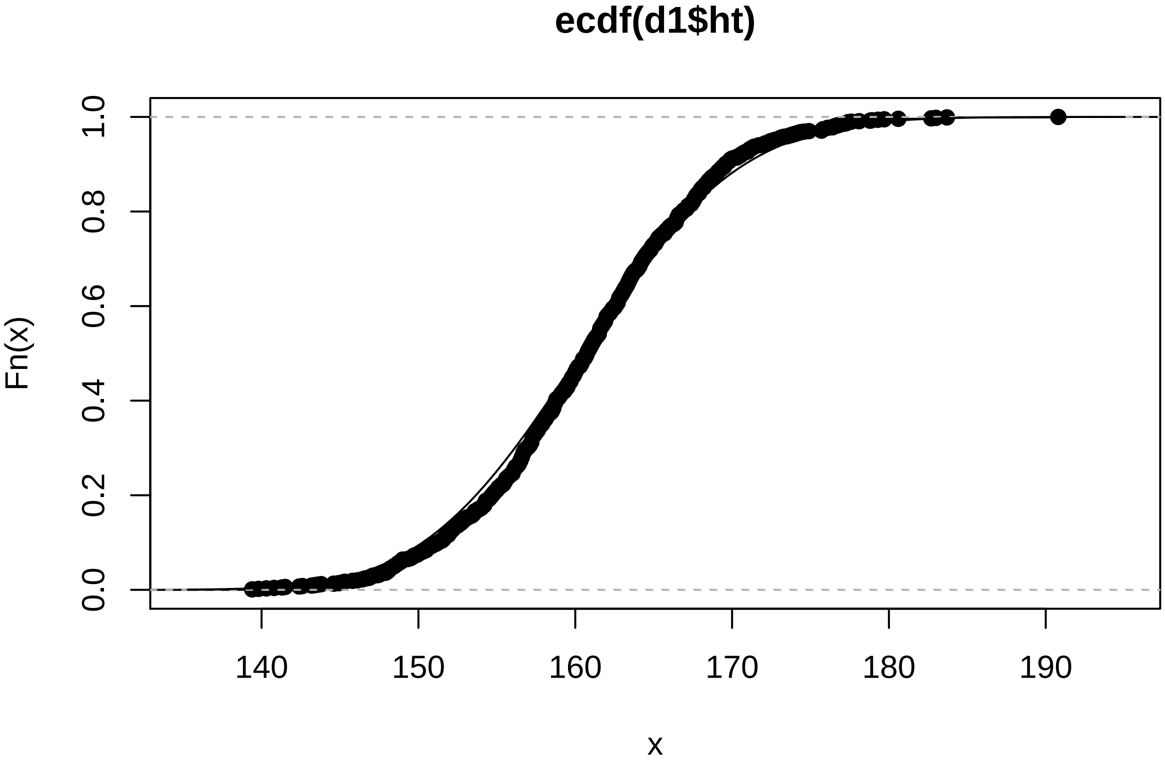
```
plot(ecdf(d1$gh))
curve(pgamma(x,shape = 40.706,scale = 0.141), add = TRUE)
```



```
ggplot(df_ht) +
  geom_histogram(aes(x = d1$ht, y= ..density..), fill = 'red', position = 'identity', alpha =.3, bins =
  geom_point(aes(y = pdf_gamma_ht, x = d1$ht), color = 'blue', position = 'identity', alpha =.3)+
  labs(
    title = 'Ht PDF Overlay'
  )
```



```
plot(ecdf(d1$ht))
curve(pgamma(x,shape = 400.1318856,scale = 0.401), add = TRUE)
```



Gamma MLE Median

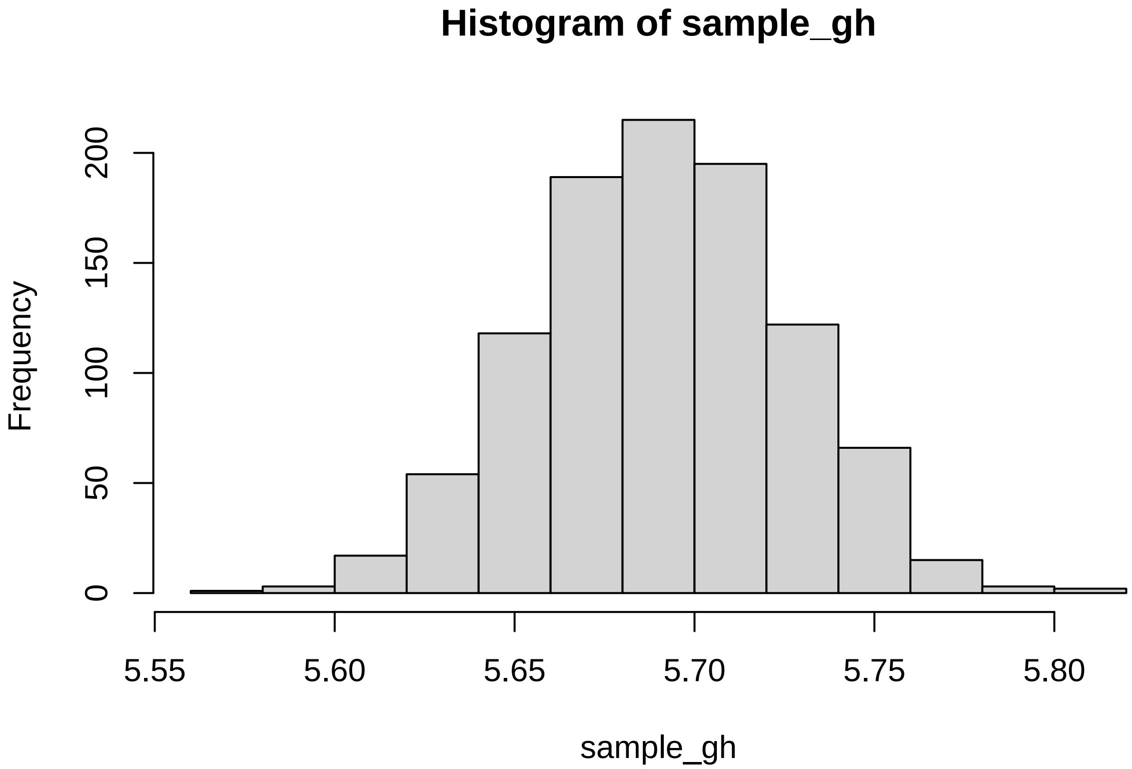
```r
est_gh_med <- qgamma(.5,shape = 40.706,scale = 0.141)
est_ht_med <- qgamma(.5,shape = 400.1318856,scale = 0.401)

print(est_gh_med)
```

```
## [1] 5.692615
```

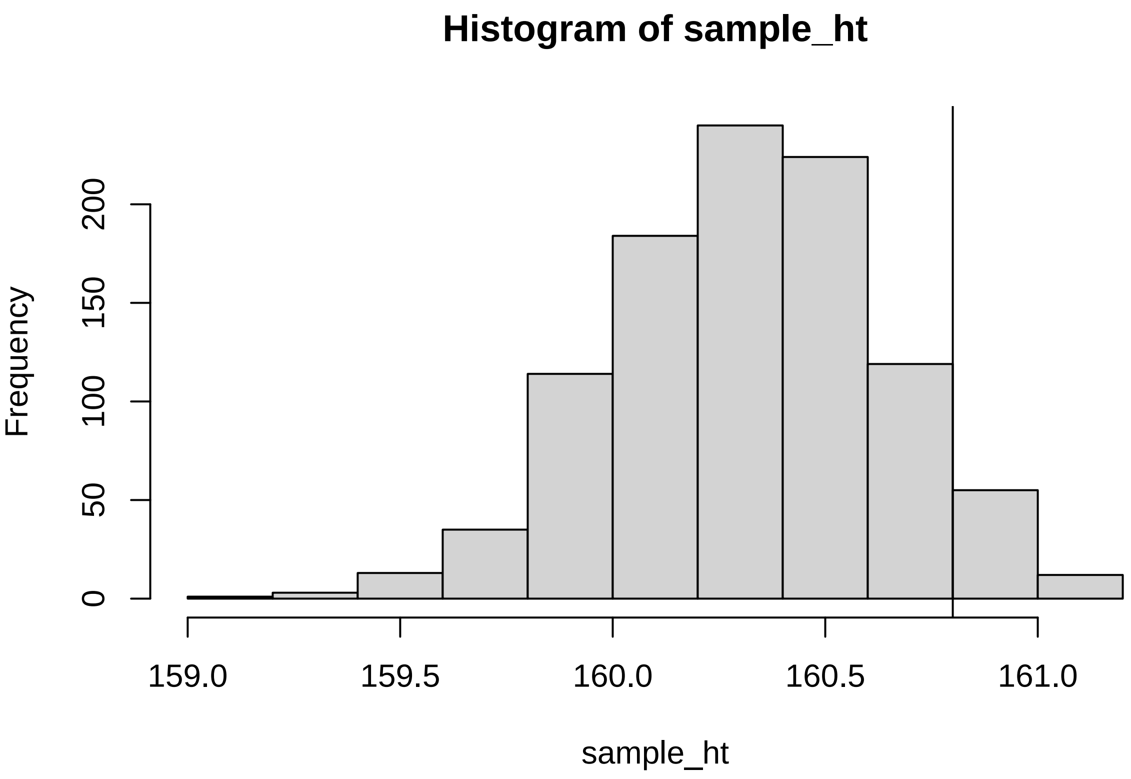
```r
print(est_ht_med)
```

```
## [1] 160.3192
```

```r
sample_gh <- rep(NA,1000)
for(i in c(1:1000)){
sample_gh[i] <- median(rgamma(1000,shape = 40.706,scale = 0.141))
}

hist(sample_gh)
abline(v=median(d1$gh))
```



```r
sample_ht <- rep(NA,1000)
for(i in c(1:1000)){
sample_ht[i] <- median(rgamma(1000,shape = 400.1318856,scale = 0.401))
}

hist(sample_ht)
abline(v=median(d1$ht))
```



Gamma MLE Range

```
gamma_gh_range <- quantile(probs = c(.025,.975),sample_gh)
gamma_ht_range <- quantile(probs = c(.025,.975),sample_ht)
print(gamma_gh_range)
```

```
##     2.5%    97.5%
## 5.621561 5.755421
```

```
print(gamma_ht_range)
```

```
##     2.5%    97.5%
## 159.6757 160.9222
```

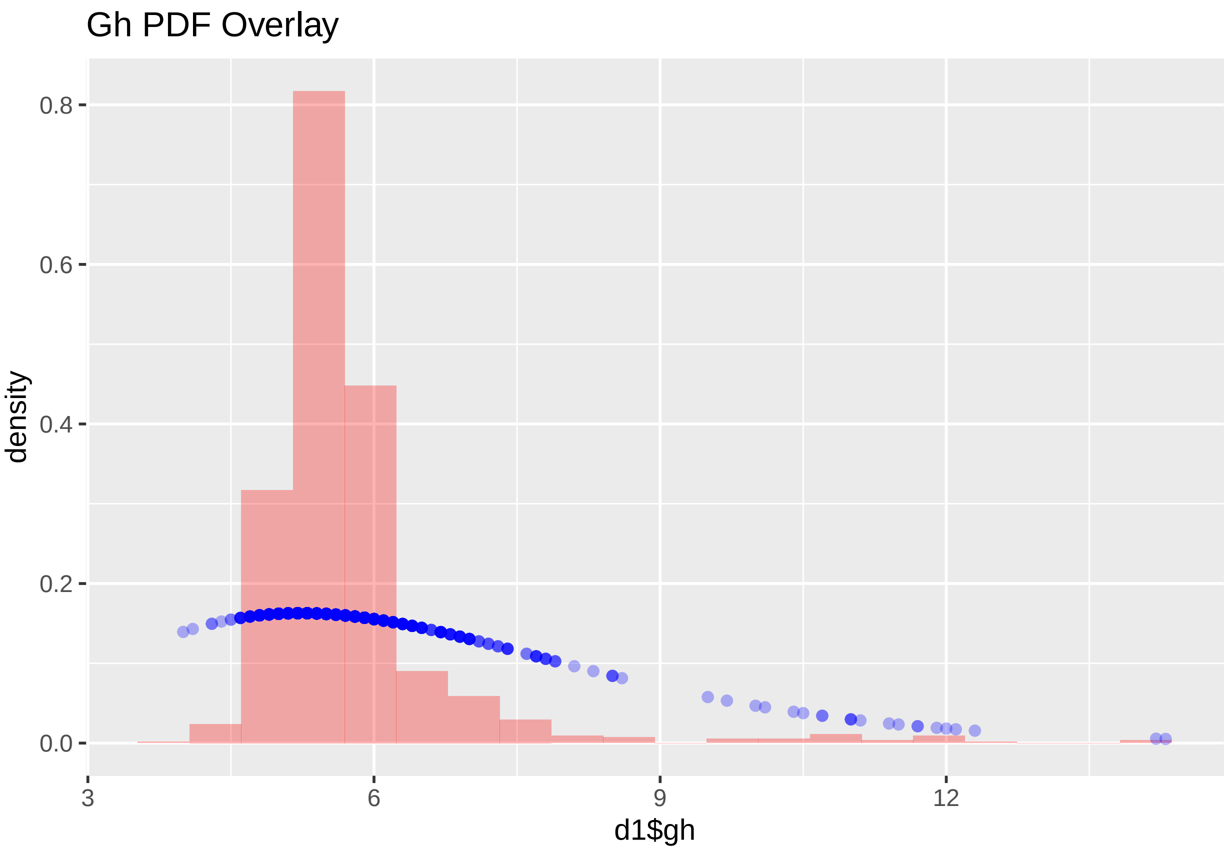
Gamma MM

```
ex <- mean(d1$gh)
vx <- var(d1$gh)

k <- (ex^2)/vx
theta <- vx/ex

gh_gamma_mean <- k * theta
gh_gamma_sd <- k * theta^2

pdf_gamma_gh <- dgamma(d1$gh,shape = gh_gamma_mean,scale = gh_gamma_sd)
df_gh <- data.frame(d1$gh,pdf_norm_gh)

ex <- mean(d1$ht)
```

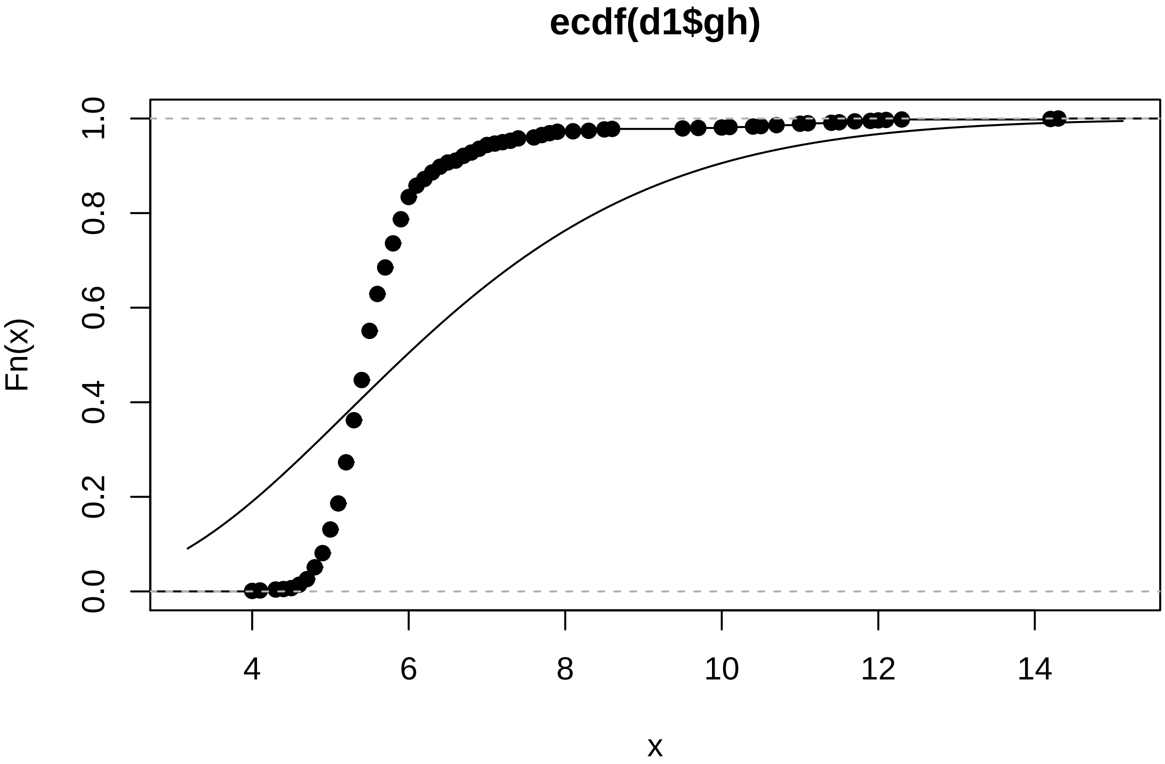
```
vx <- var(d1$ht)

k <- (ex^2)/vx
theta <- vx/ex

ht_gamma_mean <- k * theta
ht_gamma_sd <- k * theta^2

pdf_gamma_ht <- dgamma(d1$ht,shape = ht_gamma_mean,scale = ht_gamma_sd)
df_gh <- data.frame(d1$ht,pdf_norm_ht)
```

Gamma MM Graphs

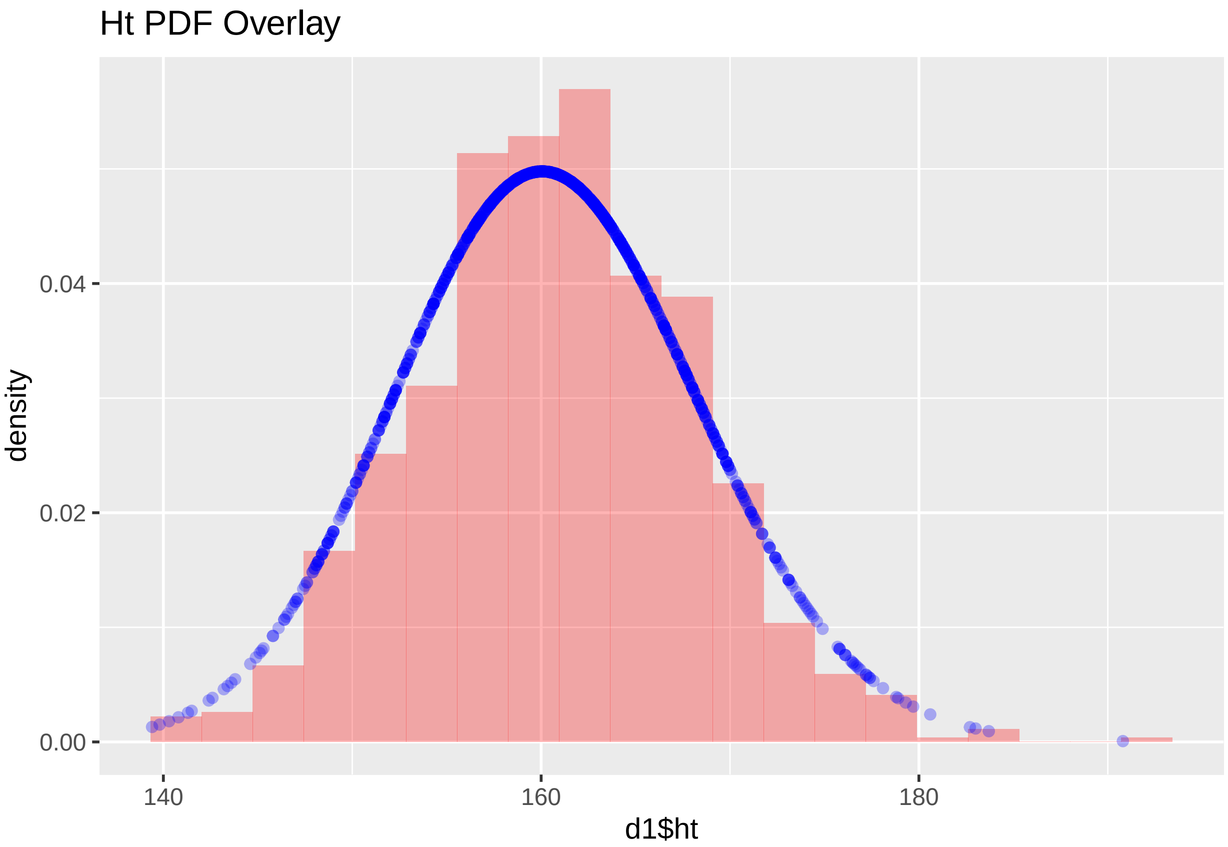
```
ggplot(df_gh) +
  geom_histogram(aes(x = d1$gh, y= ..density..), fill = 'red', position = 'identity', alpha =.3, bins =
  geom_point(aes(y = pdf_gamma_gh, x = d1$gh), color = 'blue', position = 'identity', alpha =.3)+
  labs(
    title = 'Gh PDF Overlay'
  )
```



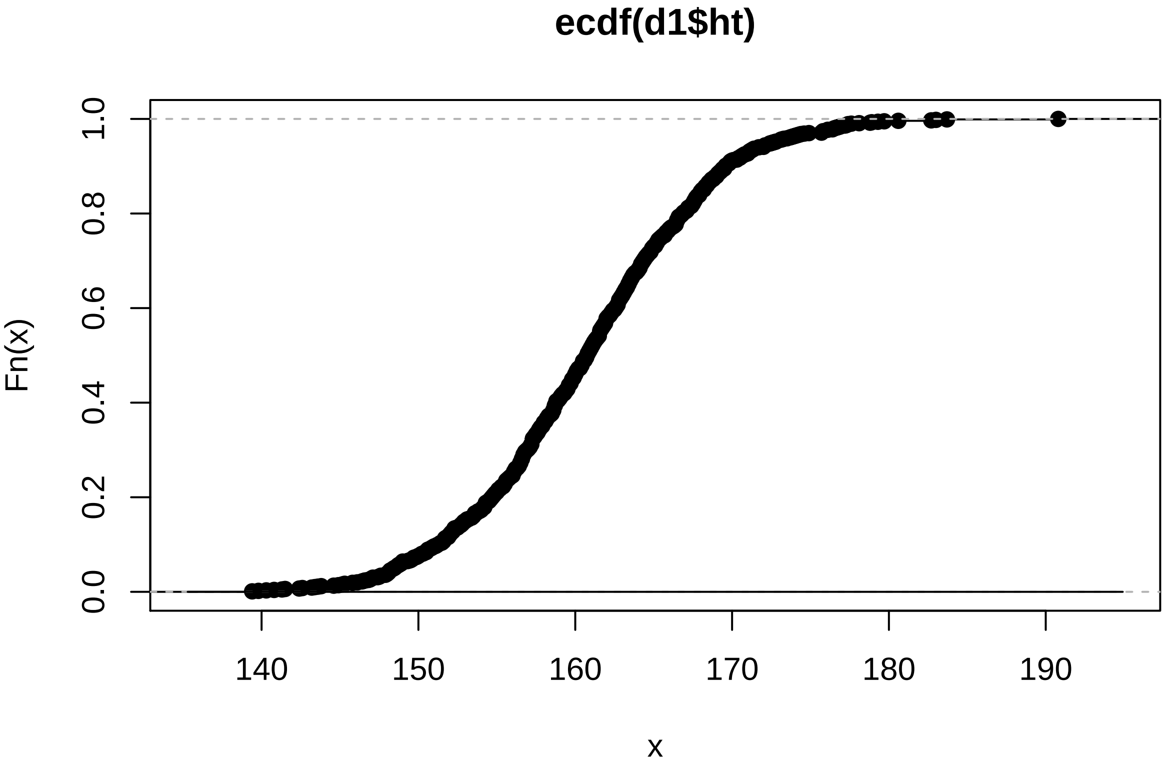
```
plot(ecdf(d1$gh))
curve(pgamma(x,shape = gh_gamma_mean,scale = gh_gamma_sd), add = TRUE)
```



```
ggplot(df_ht) +
  geom_histogram(aes(x = d1$ht, y= ..density..), fill = 'red', position = 'identity', alpha =.3, bins =
  geom_point(aes(y = pdf_gamma_ht, x = d1$ht), color = 'blue', position = 'identity', alpha =.3)+
  labs(
    title = 'Ht PDF Overlay'
  )
```



```r
plot(ecdf(d1$ht))
curve(pgamma(x,shape = ht_gamma_mean,scale = ht_gamma_sd), add = TRUE)
```



Gamma MM Median

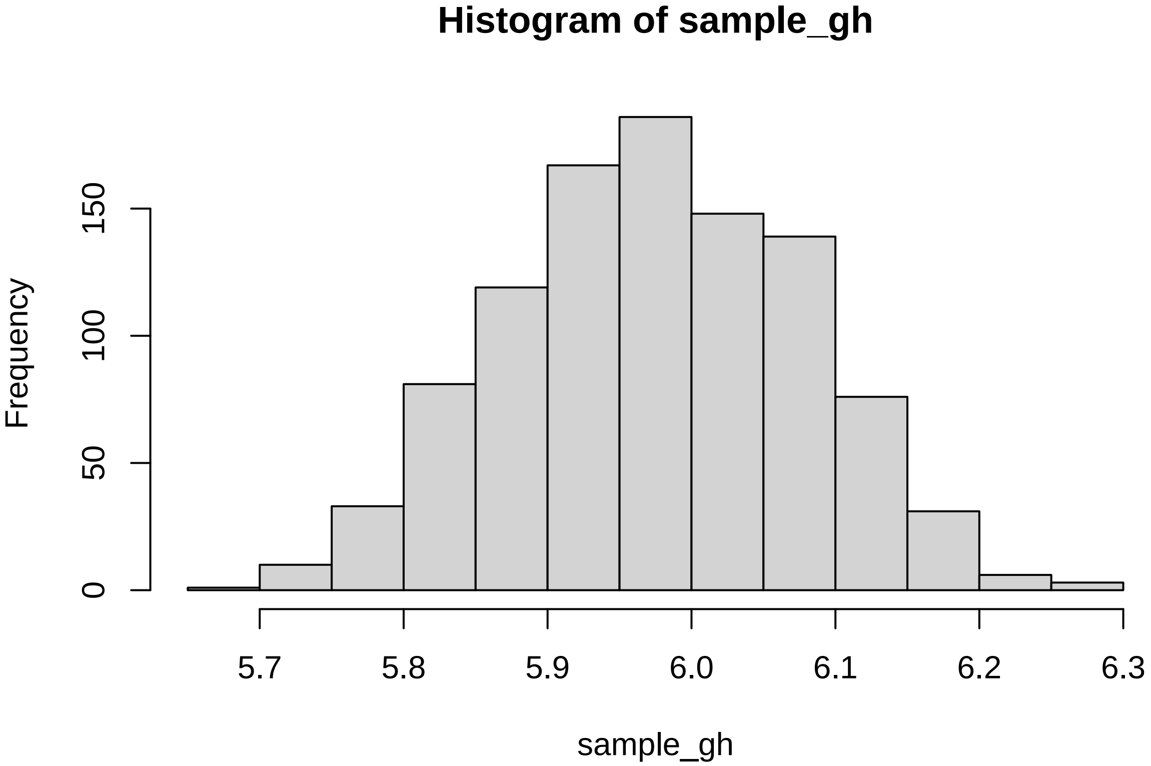
```
est_gh_med <- qgamma(.5,gh_gamma_mean, gh_gamma_sd)
est_ht_med <- qgamma(.5,ht_gamma_mean, ht_gamma_sd)
print(est_gh_med)
```

```
## [1] 4.8725
```

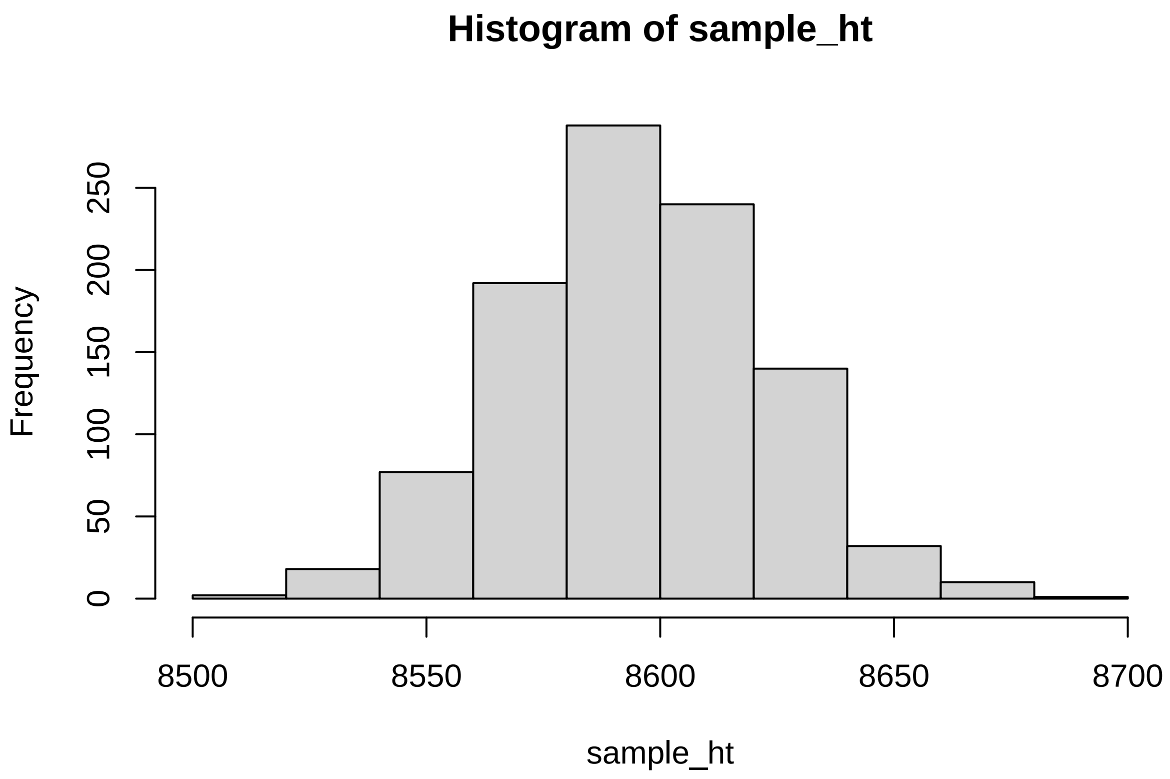
```
print(est_ht_med)
```

```
## [1] 2.993551
```

```
sample_gh <- rep(NA,1000)
for(i in c(1:1000)){
sample_gh[i] <- median(rgamma(1000,shape = gh_gamma_mean,scale = gh_gamma_sd))
}

hist(sample_gh)
abline(v=median(d1$gh))
```



```
sample_ht <- rep(NA,1000)
for(i in c(1:1000)){
sample_ht[i] <- median(rgamma(1000,shape = ht_gamma_mean,scale = ht_gamma_sd))
}

hist(sample_ht)
abline(v=median(d1$ht))
```



Gamma MM Range

```
gamma_gh_range <- quantile(probs = c(.025,.975),sample_gh)
gamma_ht_range <- quantile(probs = c(.025,.975),sample_ht)
print(gamma_gh_range)
```

```
##     2.5%    97.5%
## 5.775455 6.162654
```

```
print(gamma_ht_range)
```

```
##     2.5%    97.5%
## 8543.403 8646.807
```

Weibull

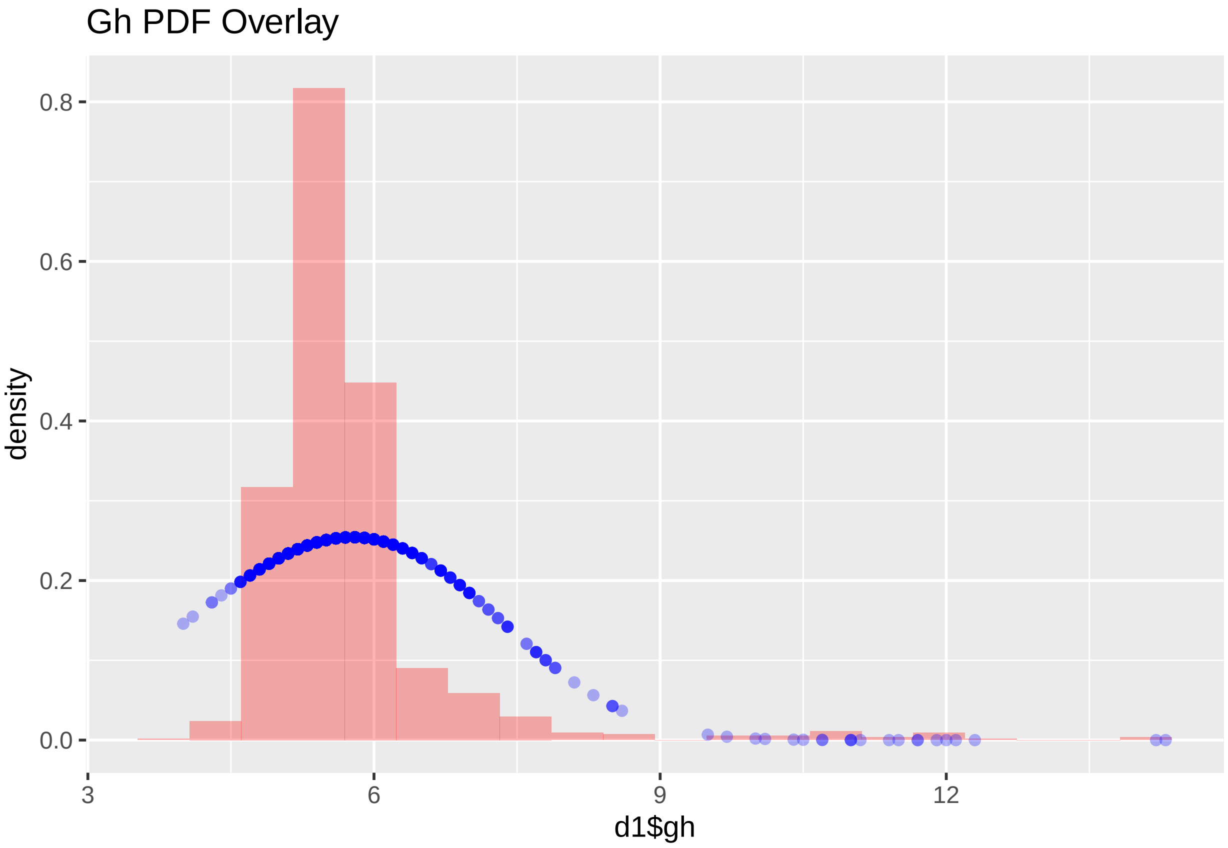
Weibull MLE

```
neg_log_lik_weibull<- function(k,lambda) {
  -sum(dweibull(d1$gh, shape=k, scale = lambda, log=TRUE))
}

mle_weibull_gh <- mle(neg_log_lik_weibull,
          start=list(k=0.01, lambda=0.01))

pdf_weibull_gh <- dweibull(d1$gh,shape = 4.13 ,scale = 6.17)

df_gh <- data.frame(d1$gh,pdf_weibull_gh)
```

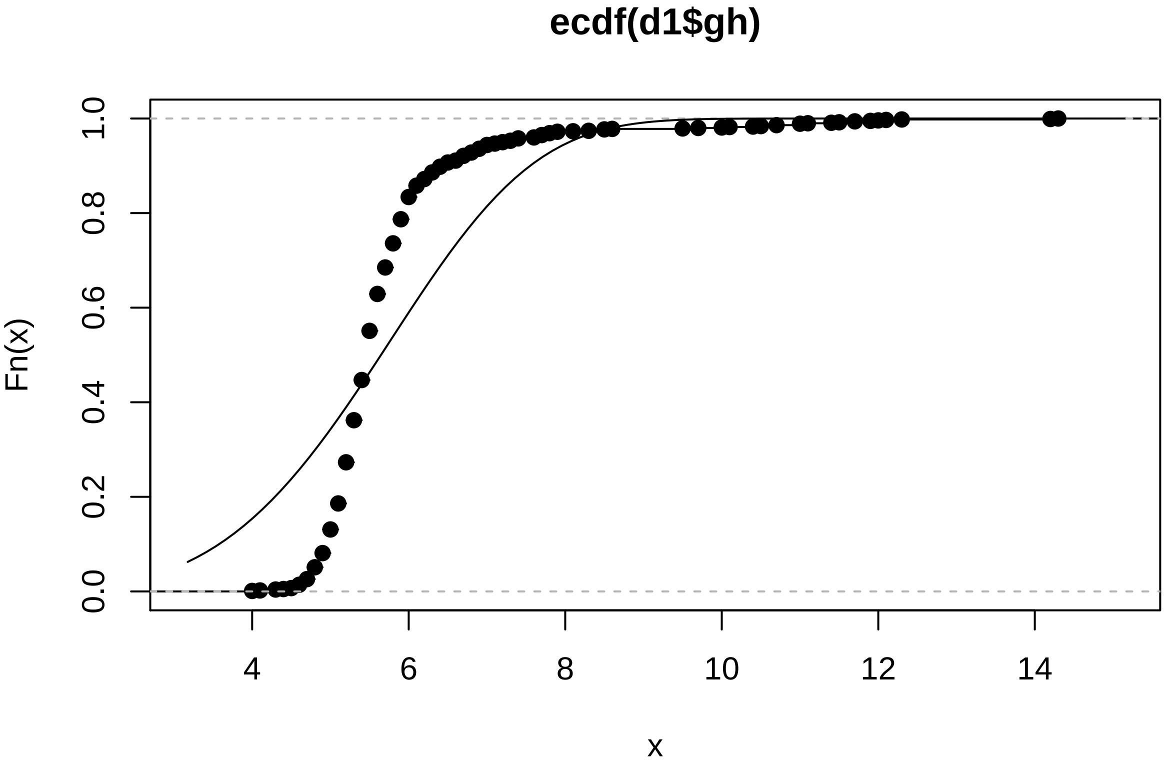
```
neg_log_lik_weibull<- function(k,lambda) {
  -sum(dweibull(d1$ht, shape=k, scale = lambda, log=TRUE))
}

mle_weibull_ht <- mle(neg_log_lik_weibull,
          start=list(k=0.01, lambda=0.01))

pdf_weibull_ht <- dweibull(d1$ht,shape = 29.4,scale = 164.2)

df_ht <- data.frame(d1$ht,pdf_weibull_ht)
```

Weibull MLE Graphs

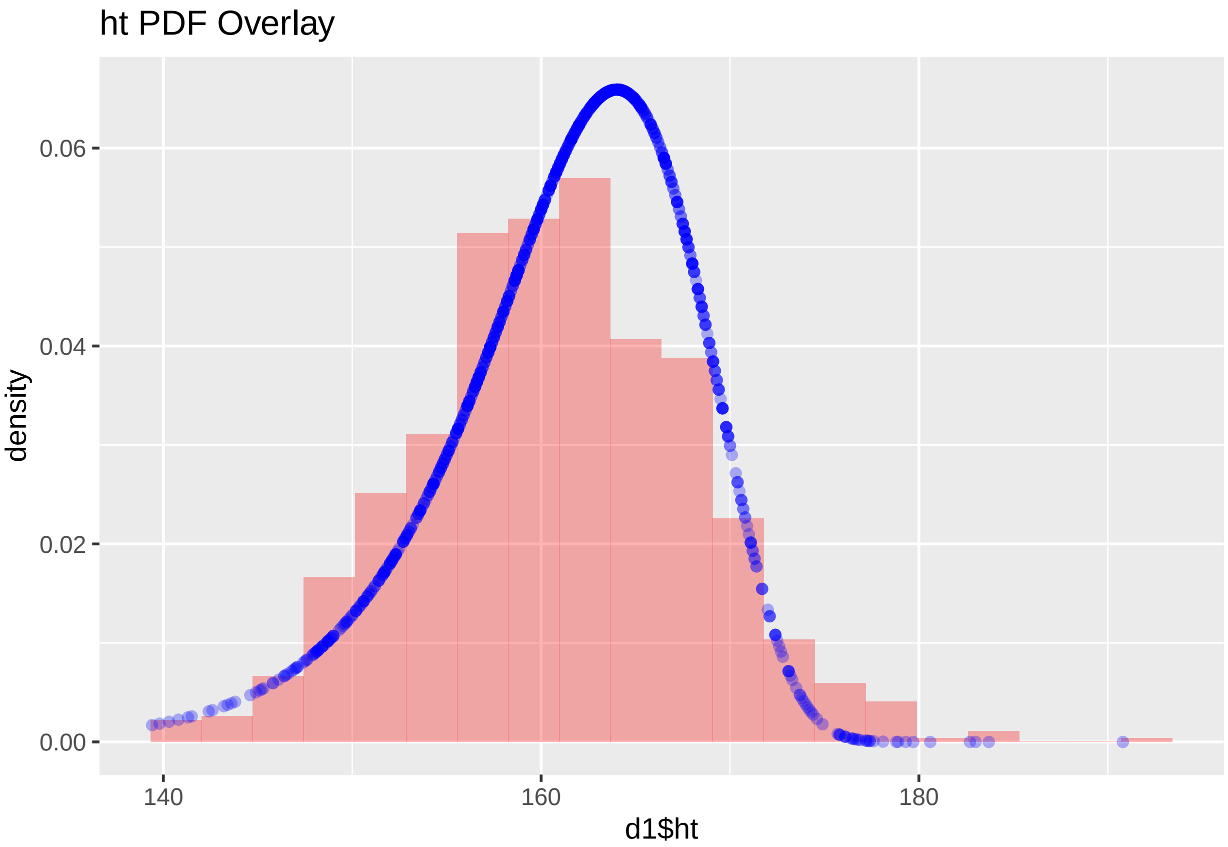
```
ggplot(df_gh) +
  geom_histogram(aes(x = d1$gh, y= ..density..), fill = 'red', position = 'identity', alpha =.3, bins =
  geom_point(aes(y = pdf_weibull_gh, x = d1$gh), color = 'blue', position = 'identity', alpha =.3)+
  labs(
    title = 'Gh PDF Overlay'
  )
```



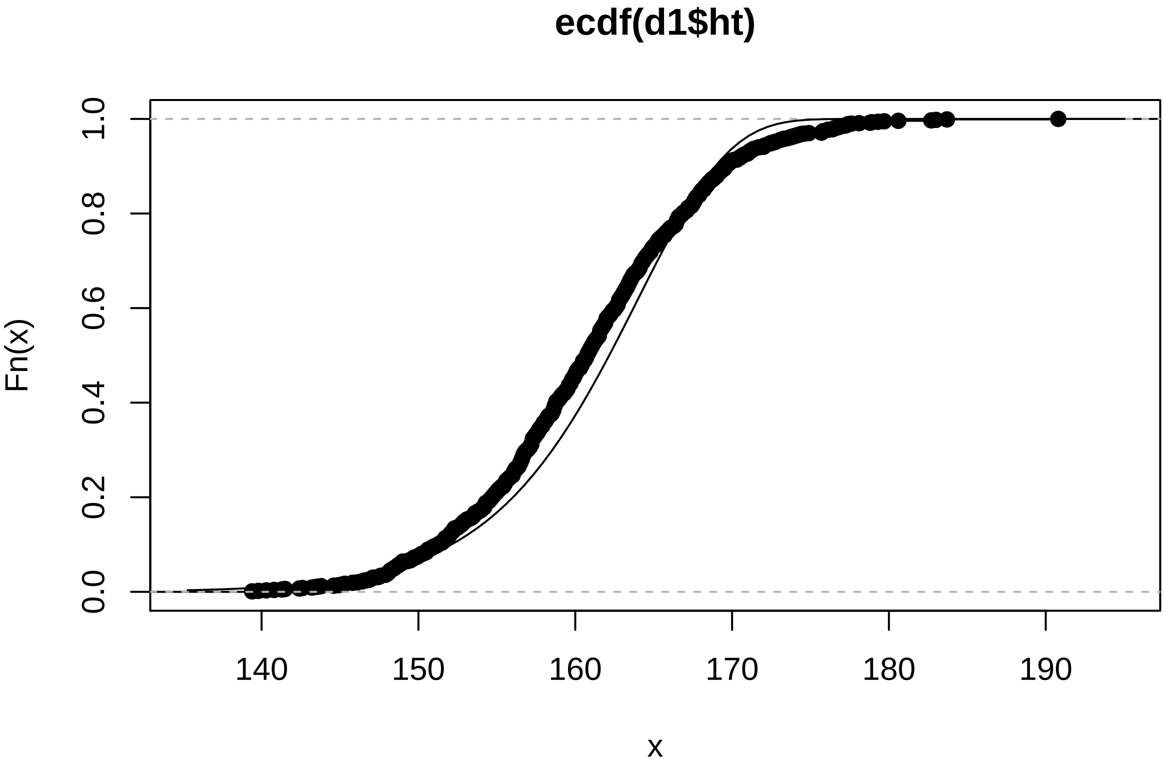
```
plot(ecdf(d1$gh))
curve(pweibull(x,shape = 4.13 ,scale = 6.17), add = TRUE)
```



```
ggplot(df_ht) +
  geom_histogram(aes(x = d1$ht, y= ..density..), fill = 'red', position = 'identity', alpha =.3, bins =
  geom_point(aes(y = pdf_weibull_ht, x = d1$ht), color = 'blue', position = 'identity', alpha =.3)+
  labs(
    title = 'ht PDF Overlay'
  )
```



```
plot(ecdf(d1$ht))
curve(pweibull(x,shape = 29.4,scale = 164.2), add = TRUE)
```



Weibull MLE Median

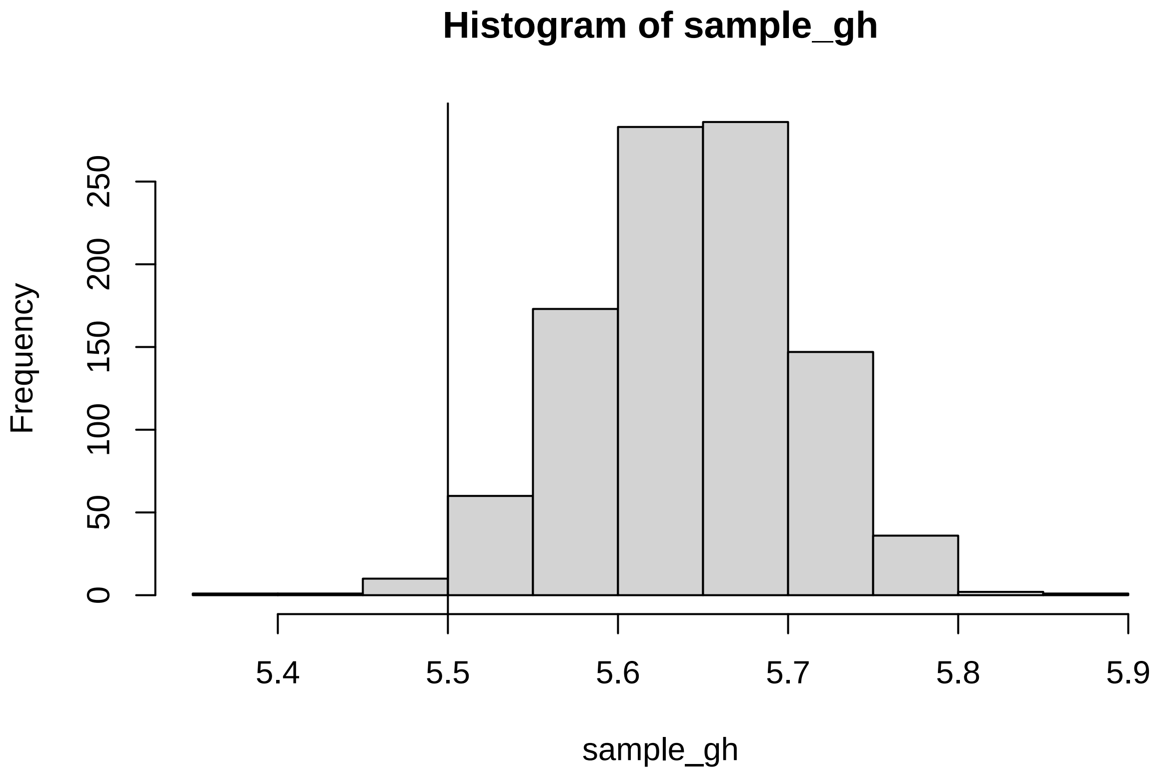
```
gh_median_weibull <- qweibull(.5,shape = 4.13 ,scale = 6.17)
ht_median_weibull <- qweibull(.5,shape = 29.4,scale = 164.2)
print(gh_median_weibull)
```

```
## [1] 5.646042
```

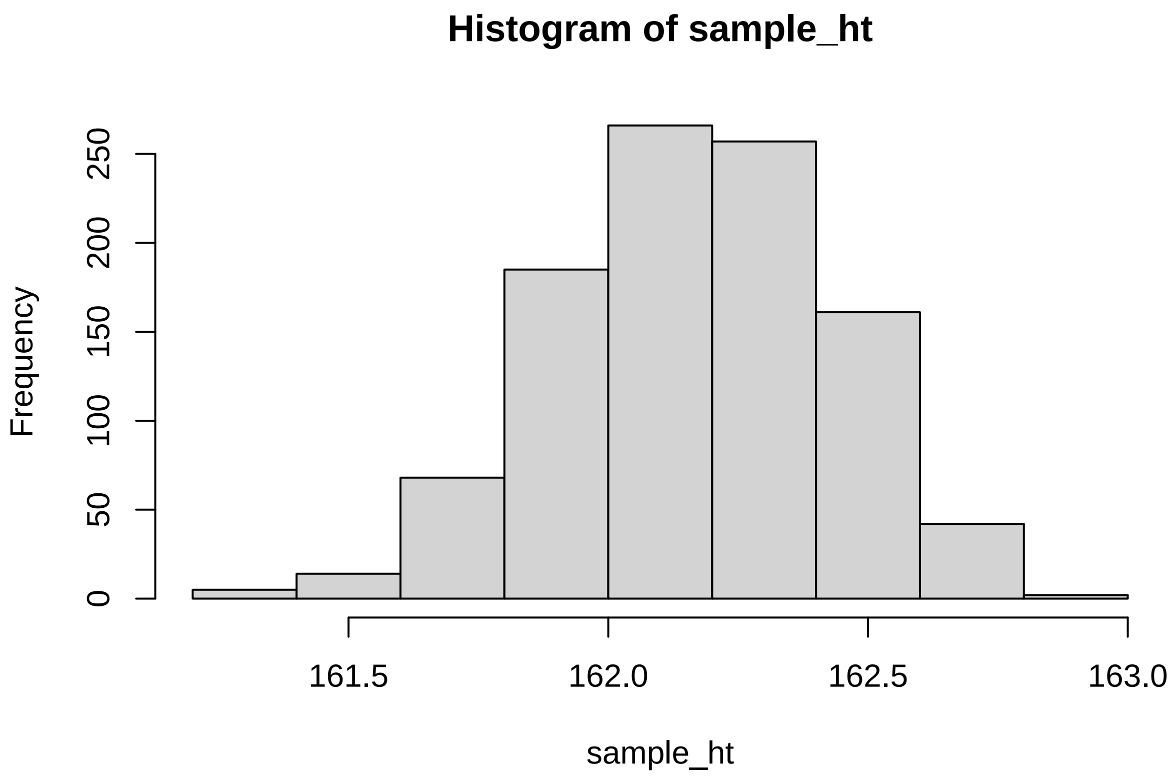
```
print(ht_median_weibull)
```

```
## [1] 162.1657
```

```
sample_gh <- rep(NA,1000)
for(i in c(1:1000)){
sample_gh[i] <- median(rweibull(1000,shape = 4.13 ,scale = 6.17))
}

hist(sample_gh)
abline(v=median(d1$gh))
```



```
sample_ht <- rep(NA,1000)
for(i in c(1:1000)){
sample_ht[i] <- median(rweibull(1000,shape = 29.4,scale = 164.2))
}

hist(sample_ht)
abline(v=median(d1$ht))
```



Weibull MLE Range

```
weibull_gh_range <- quantile(probs = c(.025,.975),sample_gh)
weibull_ht_range <- quantile(probs = c(.025,.975),sample_ht)


print(weibull_gh_range)
```

```
##     2.5%    97.5%
## 5.519614 5.760075
```

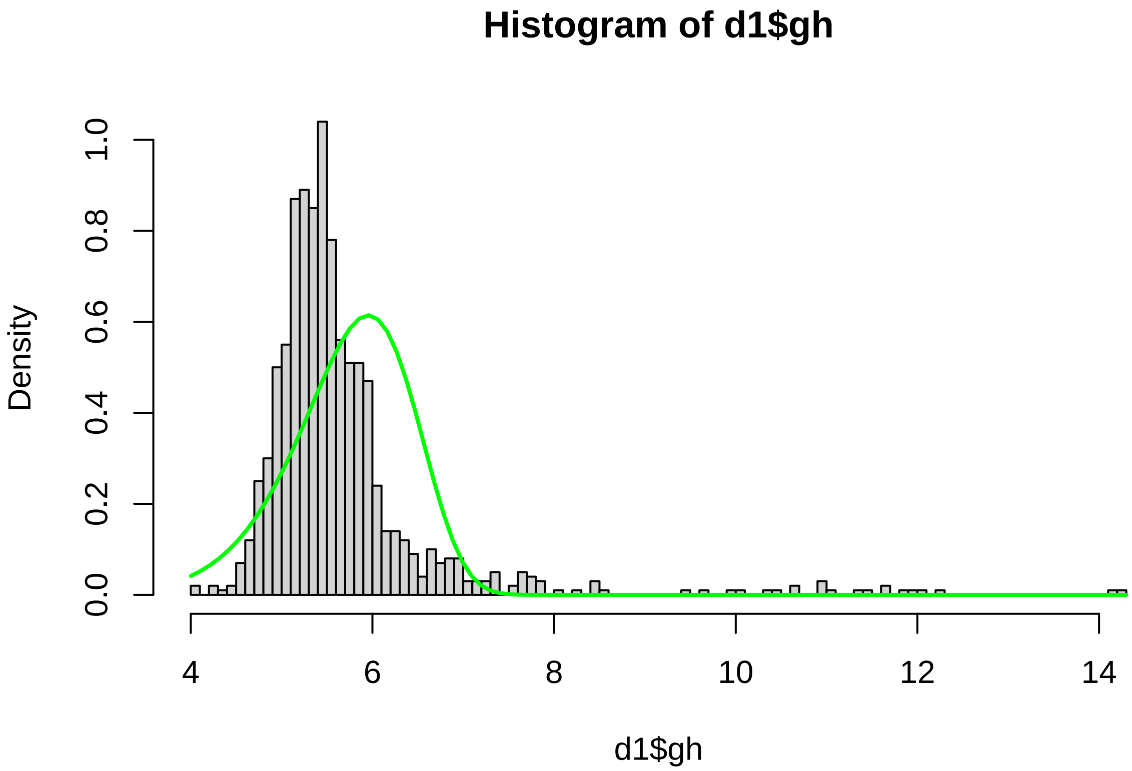
```
print(weibull_ht_range)
```

```
##     2.5%    97.5%
## 161.6372 162.6561
```

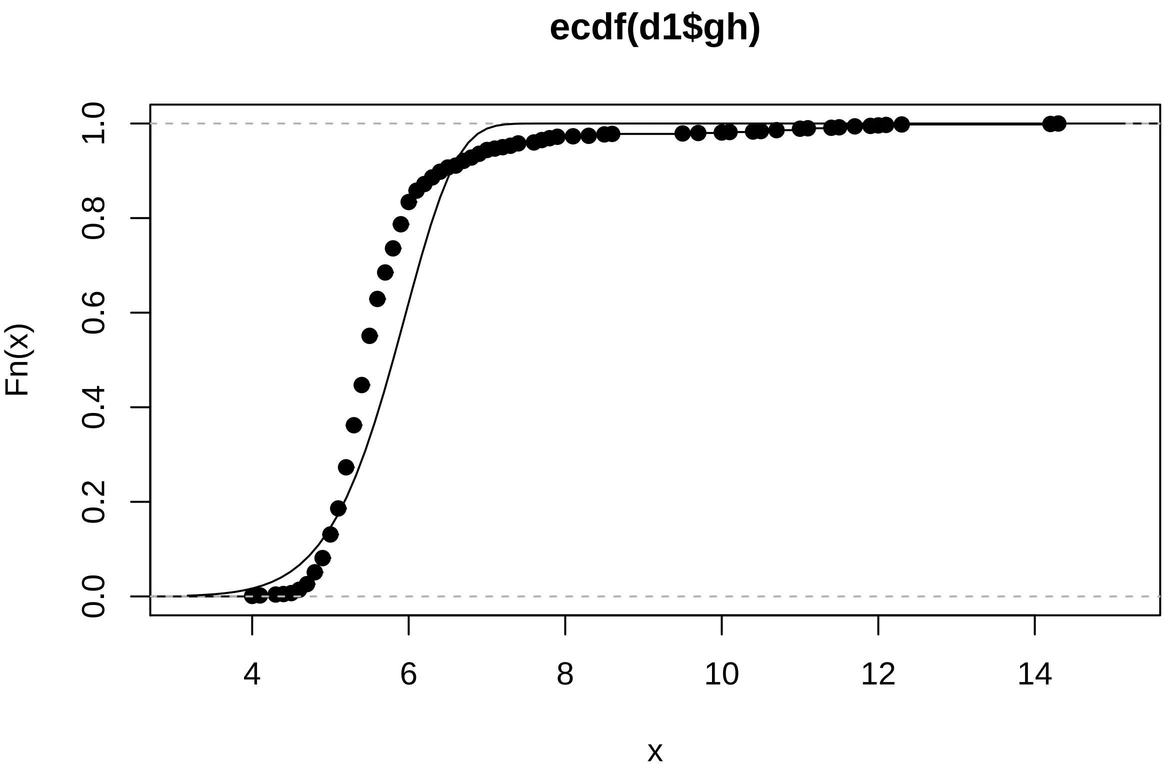
Weibull MM

```
mean.weib = function(lambda, k){
  lambda*gamma(1+1/k)
}

var.weib = function(lambda,k){
  lambda^2*(gamma(1+2/k) - (gamma(1+1/2))^2)
}

lambda = function(samp.mean,k){
  samp.mean/gamma(1+1/k)
}
```

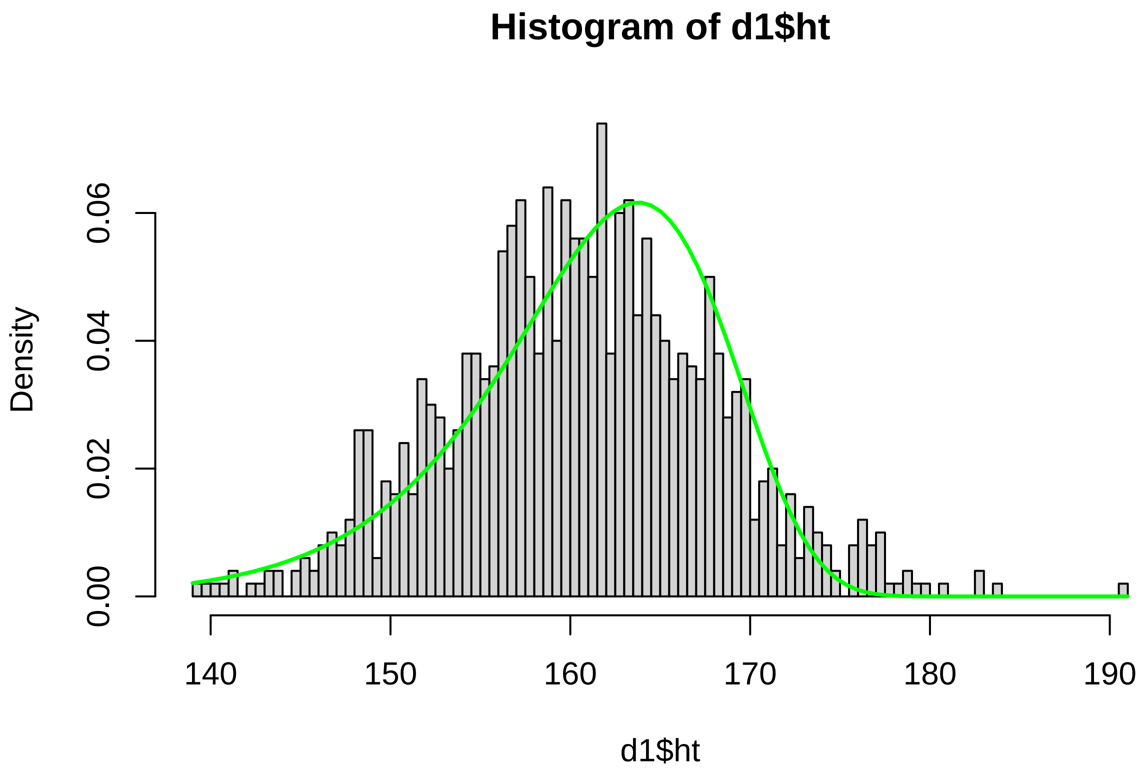
```
var.weib = function(samp.mean,k){
  lambda(samp.mean,k)^2*(gamma(1+2/k)-(gamma(1+1/k))^2)
}

var.weib = function(k,samp.mean,samp.var){
  lambda(samp.mean,k)^2*(gamma(1+2/k)-(gamma(1+1/k))^2) - samp.var
}


mm.opt = optimize(f=function(x){abs(var.weib(k = x,samp.mean=mean(d1$gh),samp.var=var(d1$gh)))},lower=1

mm.weib.k_gh = mm.opt$minimum

mm.weib.lambda_gh = lambda(samp.mean = mean(d1$gh),k=mm.weib.k_gh)

hist(d1$gh, breaks = 100, freq = FALSE)
curve(dweibull(x,shape=mm.weib.k_gh,scale=mm.weib.lambda_gh),add=TRUE,col = 'green',lwd = 2)
```



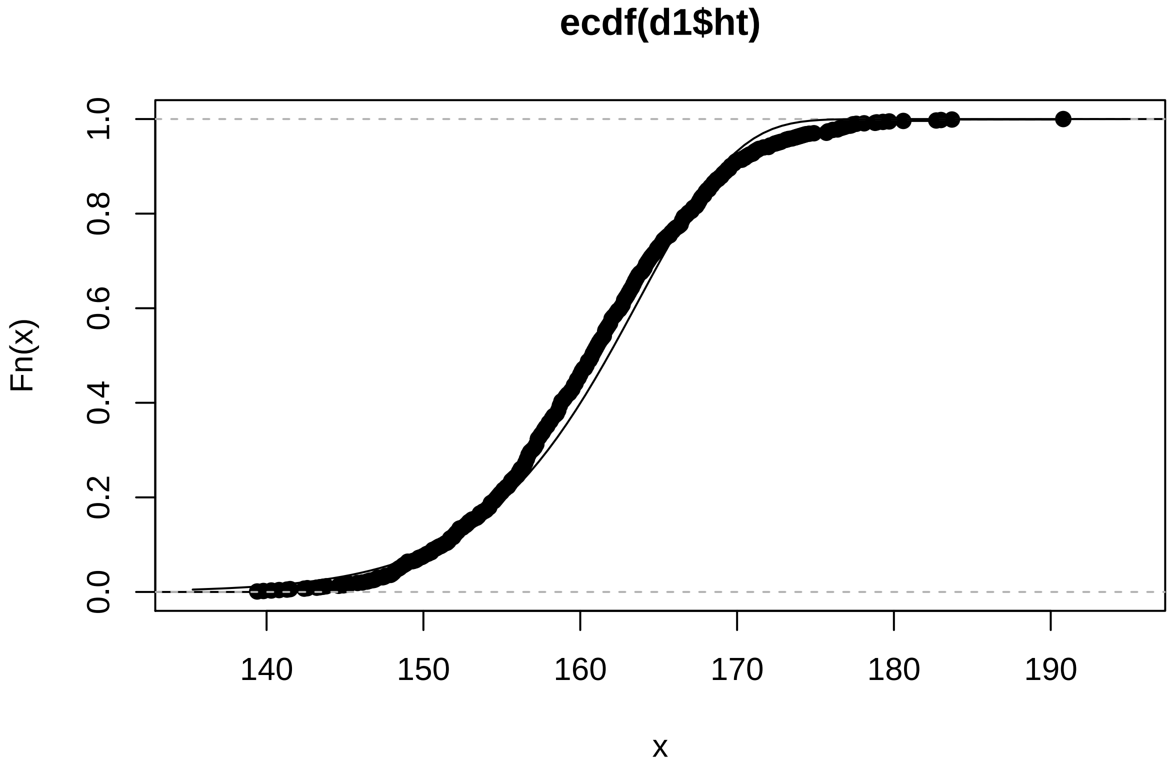
```
plot(ecdf(d1$gh))
curve(pweibull(x,shape = mm.weib.k_gh,scale = mm.weib.lambda_gh), add = TRUE)
```



```
mean.weib = function(lambda, k){
  lambda*gamma(1+1/k)
}

var.weib = function(lambda,k){
  lambda^2*(gamma(1+2/k) - (gamma(1+1/2))^2)
}

lambda = function(samp.mean,k){
  samp.mean/gamma(1+1/k)
}

var.weib = function(samp.mean,k){
  lambda(samp.mean,k)^2*(gamma(1+2/k)-(gamma(1+1/k))^2)
}

var.weib = function(k,samp.mean,samp.var){
  lambda(samp.mean,k)^2*(gamma(1+2/k)-(gamma(1+1/k))^2) - samp.var
}


mm.opt = optimize(f=function(x){abs(var.weib(k = x,samp.mean=mean(d1$ht),samp.var=var(d1$ht)))},lower=1

mm.weib.k_ht = mm.opt$minimum

mm.weib.lambda_ht = lambda(samp.mean = mean(d1$ht),k=mm.weib.k_ht)

hist(d1$ht, breaks = 100, freq = FALSE)
curve(dweibull(x,shape=mm.weib.k_ht,scale=mm.weib.lambda_ht),add=TRUE,col = 'green',lwd = 2)
```



```
plot(ecdf(d1$ht))
curve(pweibull(x,shape = mm.weib.k_ht,scale = mm.weib.lambda_ht), add = TRUE)
```



Weibull MM Median

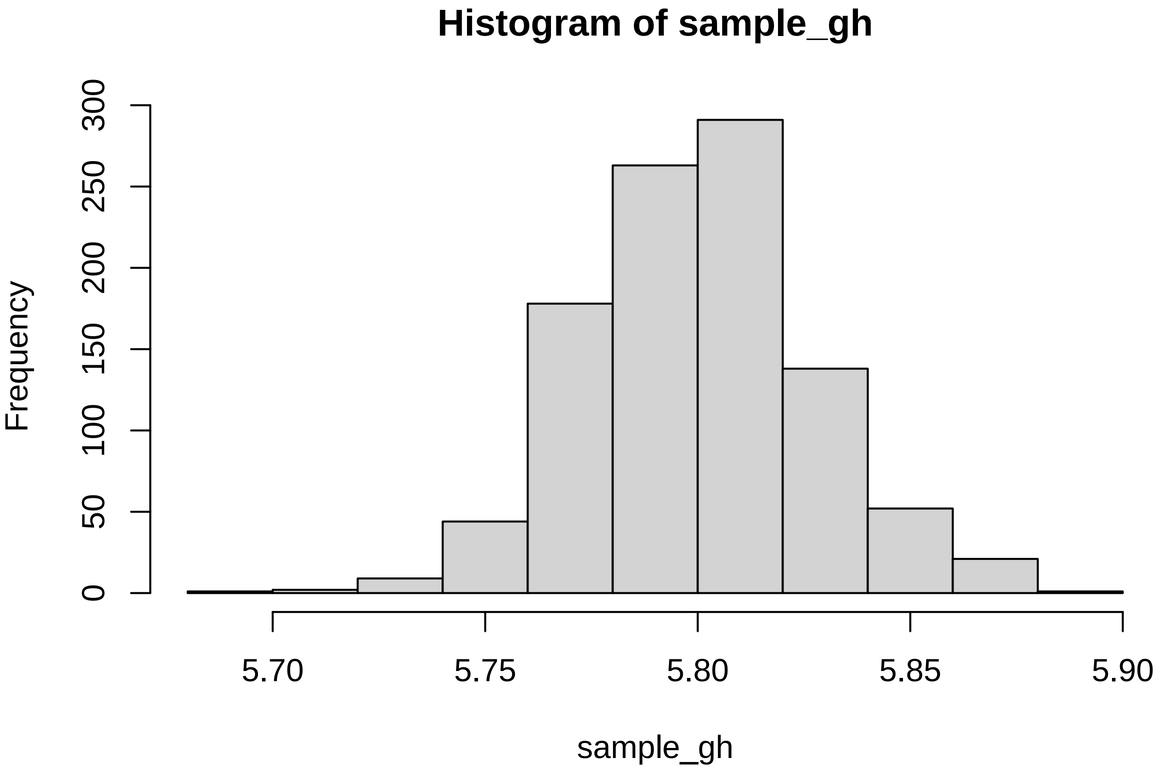
```
gh_median_weibull <- qweibull(.5,shape= mm.weib.k_gh, scale = mm.weib.lambda_gh)

ht_median_weibull <- qweibull(.5,shape= mm.weib.k_ht, scale = mm.weib.lambda_ht)
print(gh_median_weibull)
```

```
## [1] 5.800788
```

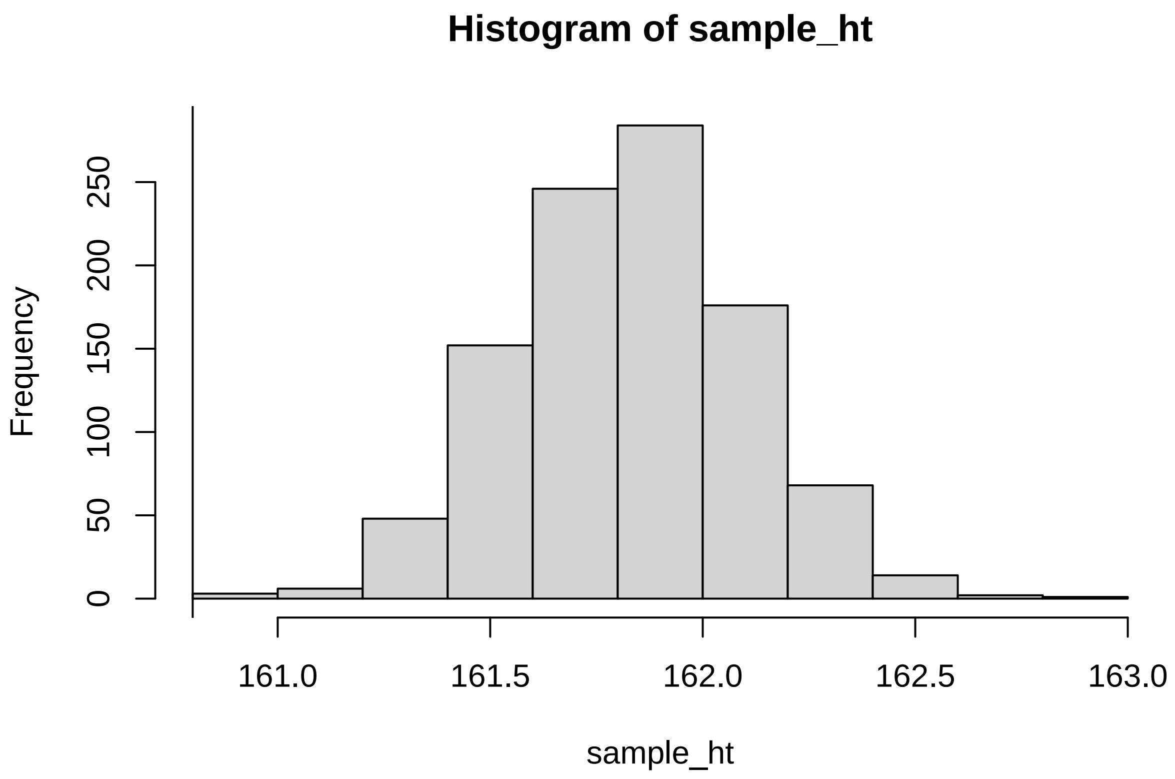
```
print(ht_median_weibull)
```

```
## [1] 161.8065
```

```
sample_gh <- rep(NA,1000)
for(i in c(1:1000)){
sample_gh[i] <- median(rweibull(1000,shape=mm.weib.k_gh,scale=mm.weib.lambda_gh))
}

hist(sample_gh)
abline(v=median(d1$gh))
```



```
sample_ht <- rep(NA,1000)
for(i in c(1:1000)){
sample_ht[i] <- median(rweibull(1000,shape=mm.weib.k_ht,scale=mm.weib.lambda_ht))
}

hist(sample_ht)
abline(v=median(d1$ht))
```



Weibull MM Range

```
weibull_gh_range <- quantile(probs = c(.025,.975),sample_gh)
weibull_ht_range <- quantile(probs = c(.025,.975),sample_ht)


print(weibull_gh_range)
```

```
##     2.5%    97.5%
## 5.750409 5.855164
```

```
print(weibull_ht_range)
```

```
##     2.5%    97.5%
## 161.3169 162.3549
```

Take Aways

The important idea to remember here is that there are multiple ways to estimate a distribution even within the same class of distributions. For example, both the MLE and MM for the normal distribution fit well. The other take home note to remember is to always check agianst multiple distributions to see which is the best.