MLDS 400 Lab 1, Basic Statistics

Adapted from Adam Sandler's Notes

Huiyu Wu

9/25/2023

NORTHWESTERN
UNIVERSITY

Lab Schedule

Labs: Mondays 11:00am-12:00pm CT

Office Hours: TBA

- Week 1: Basic Statistics
- Weeks 2, 3: Markov Chains, MCMC
  - Lab assignment
- Week 4: Cross validation and SVM
  - Lab assignment
- Week 5: Bayesian Statistics
- Week 6: Rare Events and Importance Sampling
  - Lab assignment
- Weeks 7, 8: Data Cleansing, Imputation, and EDA
  - Lab assignment
- Oct 2nd 11am Lab moved to Oct 4th 10:05-11:00am
- There will be an EDA lab session by Veena, Oct 16th (Zoom)

Numerical Summary of Data

- Definitions
  - Data: numerical and/or categorical observations of a phenomenon of interest
  - Population: the complete set of such phenomena
  - Sample: a subset of the population, a portion used for analysis
- Measures
  - Center: mean, median
  - Shape: skew, kurtosis, # of modes/peaks
  - Spread: variance, range, interquartile range

Data Set

- Import a table of the average heights (in m) of men and women of 71 countries

```
height = read.table("height.txt", header=T)
head(height)
```

```
##      Male Female
## 1 1.7348 1.6076
## 2 1.7840 1.6450
## 3 1.7920 1.6760
## 4 1.7180 1.6540
## 5 1.6510 1.5420
## 6 1.7860 1.6810
```

Mean

$X_1, X_2, \cdots, X_n$ form an i.i.d. random sample

- Population mean: $\mu = E[X_1]$
- Sample mean (point estimator of μ): $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

Mean in R

- Calculate the sample mean of male and female

```
m = height$Male  # or height[,1] or height[, "Male"]
f = height$Female
mean(m)  # mean
```

```
## [1] 1.728103
```

```
mean(f)
```

```
## [1] 1.60683
```

Variance

- Population variance: $\sigma^2 = Var(X_1) = E[(X_1 - \mu)^2]$
- Sample variance (point estimator of σ^2):
  $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$

Variance in R

- Calculate the variance of male and female

```
var(m)  # sample variance
```

```
## [1] 0.003693284
```

```
var(f)
```

```
## [1] 0.003174288
```

```
sd(f)  # standard deviation
```

```
## [1] 0.05634082
```

Covariance

- Covariance of two random variables X and Y
  - Measures how the two are linearly correlated
- Population covariance of X and Y:
  $\sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$
  - Normalized version (correlation coefficient): $\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$
- Sample covariance of X and Y:
  $s_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})$

Covariance in R

- Calculate covariance and correlation coefficient of male and female average heights

```
cov(m,f)  # covariance
```

```
## [1] 0.003237705
```

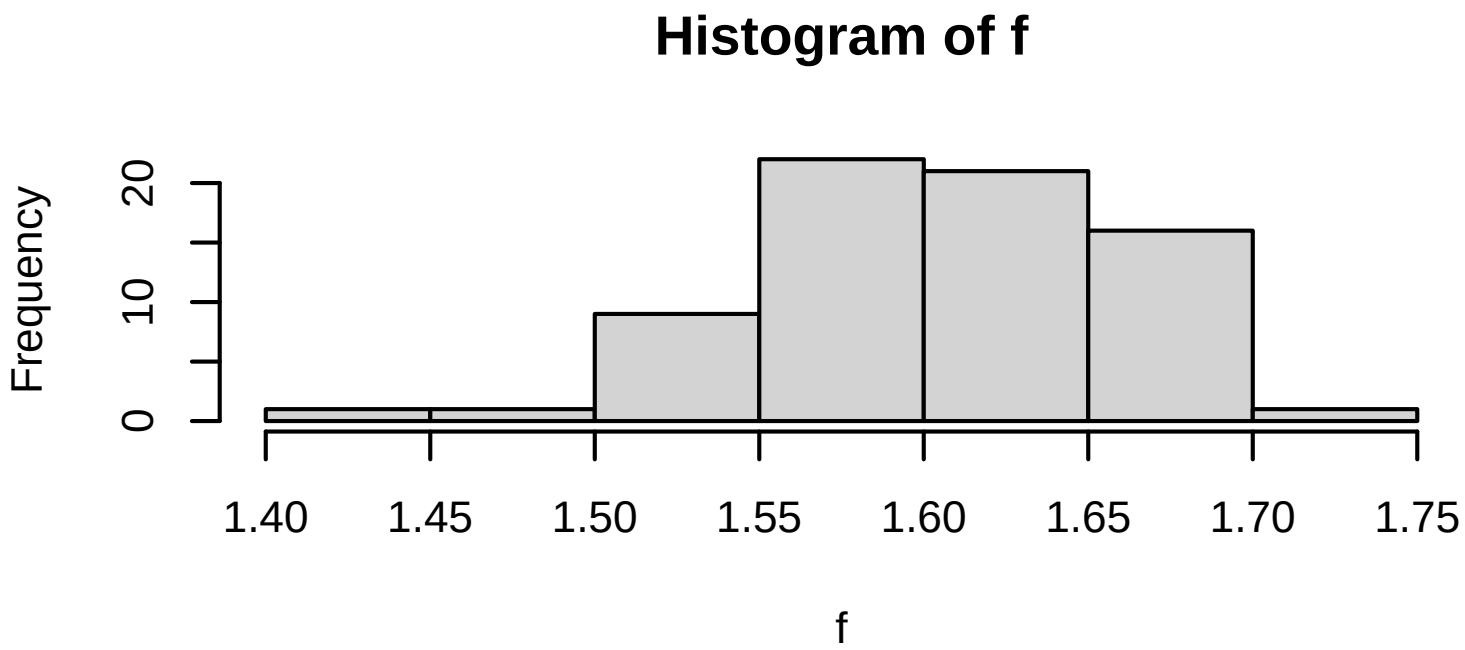
```
cor(m,f)  # correlation
```

```
## [1] 0.9456009
```

Histogram

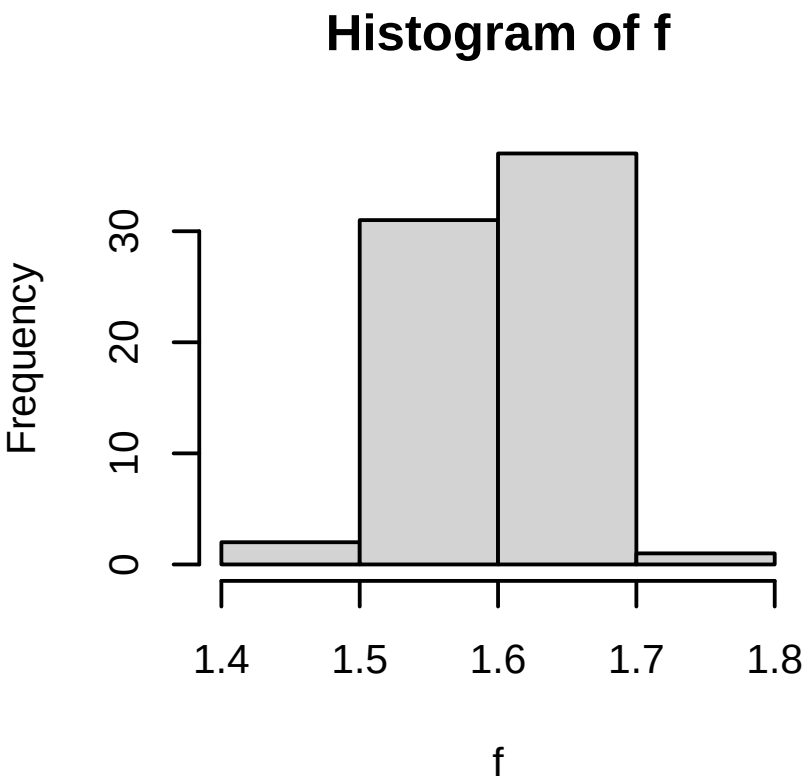
- Histogram (frequency distribution)
  - Represents the counts of observations grouped in pre-specified classes or groups
- R: Generate histograms

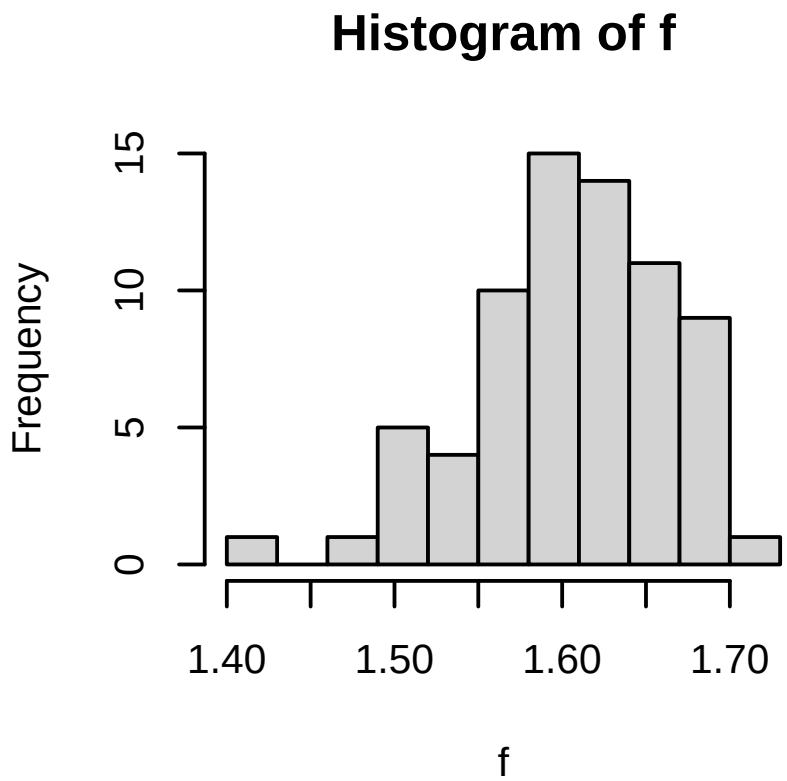
```
par(cex=0.7)  # resize text
hist(f)
```



Histogram in R

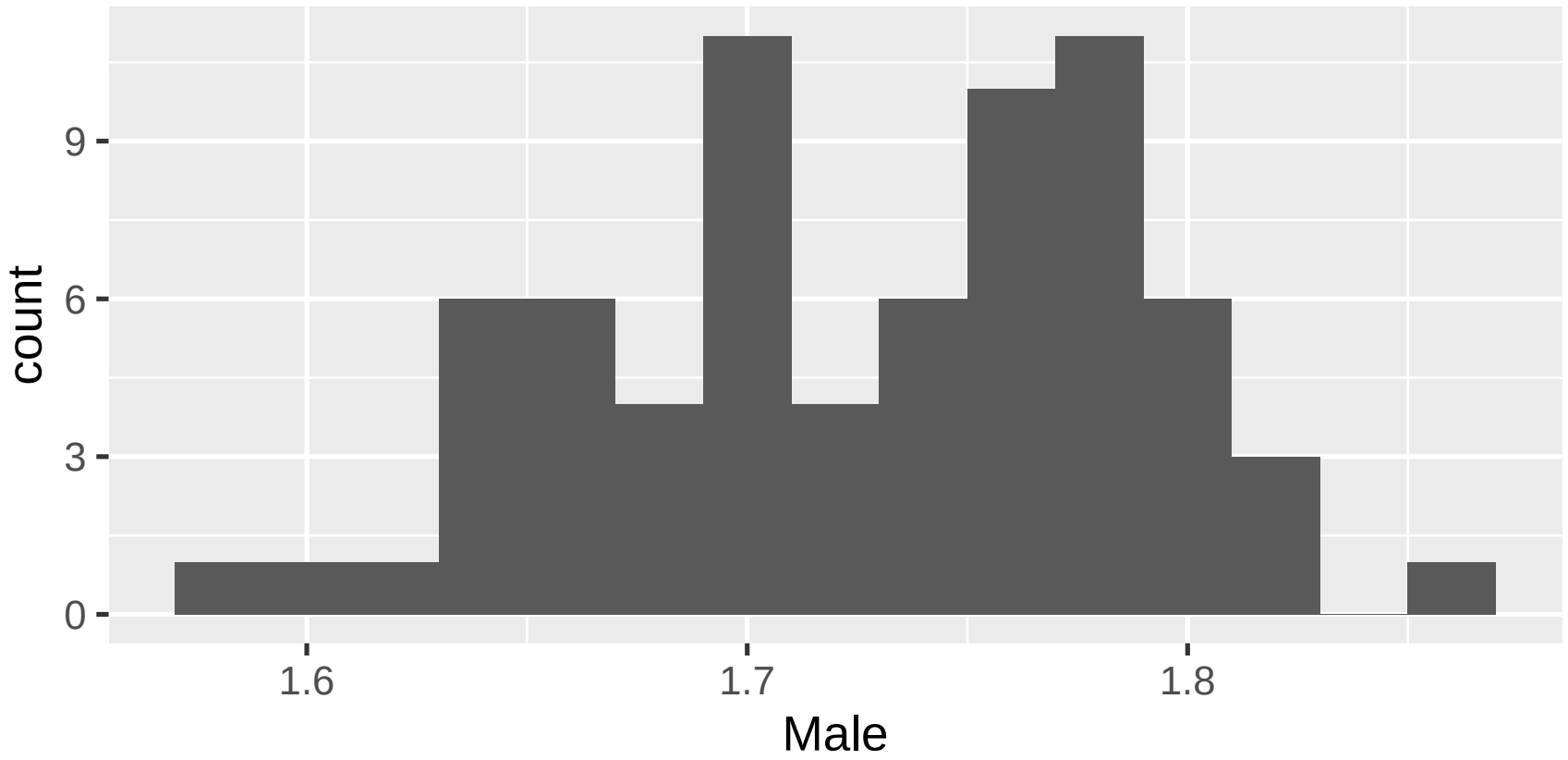
```
par(mfrow=c(1,2), cex=0.7)  # 2 plots side-by-side
hist(f, breaks=4)  # set number of bins
hist(f, breaks=seq(1.40,1.75,by=0.03))  # set bins
```





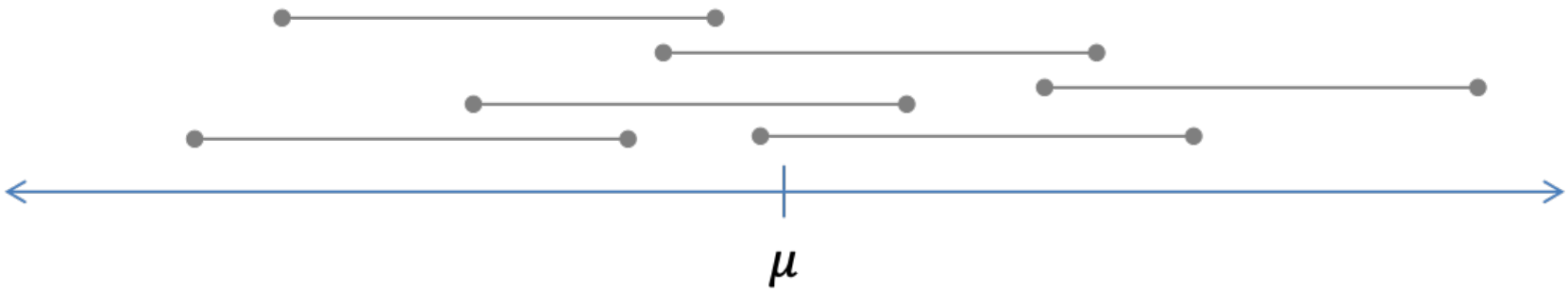
Histogram Using GGPlot2

```
library(ggplot2)  # load package
ggplot(height,aes(Male))+geom_histogram(binwidth=0.02)
```



Confidence Interval

- Confidence interval (CI) on μ with known population variance (σ^2), for given p-value α: $(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}})$
- Interpretation of CI:
  - The confidence interval is a random interval
  - The frequency $1 - \alpha$ of possible confidence intervals that contain the true value of μ
  - For a given interval there is not necessarily a $1 - \alpha$ probability that the true value of μ lies within the interval



Confidence Interval of Population Mean

- Known population variance (σ^2): $\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
- R: Confidence interval of men's height with known $\sigma = 0.0036$ and $\alpha = 0.05$

```
sigma = 0.0036; Mbar = mean(m); n = length(m)
E = qnorm(1-0.05/2)*sigma/sqrt(n)  # margin of error
CI_m = Mbar + c(-E,E)
CI_m
```

```
## [1] 1.727265 1.728940
```

Confidence Interval of Population Mean

- Unkown population variance (S^2): $\bar{X} \pm t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}$
- R: interval estimation of men's heigh with unkown σ and $\alpha = 0.05$

```
S = sd(m)
E = qt(1-0.05/2, df=n-1)*S/sqrt(n)
CI_M = Mbar + c(-E,E)
CI_M
```

```
## [1] 1.713718 1.742487
```

Hypothesis Testing

- Statistical hypotheses
  - Two-tailed hypotheses: $H_0 : \mu_m = 1.728$, $H_1 : \mu_m \neq 1.728$
  - Lower-tailed hypotheses: $H_0 : \mu_m \geq 1.728$, $H_1 : \mu_m < 1.728$
- Hypothesis testing
  - Obtains information in a random sample from the population
  - If the information is inconsistent with the null hypothesis H_0, we reject the null hypothesis. Otherwise, we fail to reject H_0.

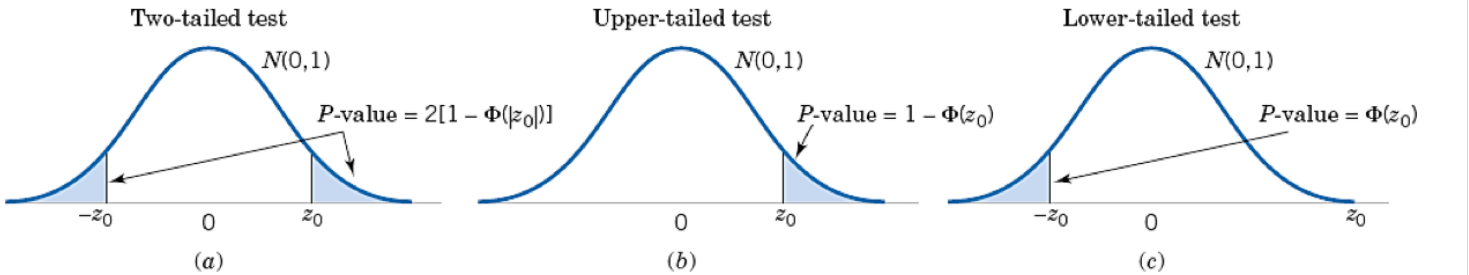


Figure 9-7 The P-value for a z-test. (a) The two-sided alternative $H_1 : \mu \neq \mu_0$. (b) The one-sided alternative $H_1 : \mu > \mu_0$. (c) The one-sided alternative $H_1 : \mu < \mu_0$.

* Source: Applied Statistics and Probability for Engineers, Montgomery and Runger

Two Tailed Z-Test

- Two-tailed test with known variance: $H_0 : \mu_m = 1.728$, $H_1 : \mu_m \neq 1.728$
- Test statistics $z_0 = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma}$
  - Fail to reject H_0 if $-z_{\alpha/2} \leq z_0 \leq z_{\alpha/2}$

Two Tailed Z-Test in R

- Test the hypothesis with $\alpha = 0.05$

```
z = sqrt(n)*(Mbar-1.728)/sigma
z
```

```
## [1] 0.2406529
```

```
z.half.alpha = qnorm(1-0.05/2)
# check if z is in the interval
c(-z.half.alpha, z.half.alpha)
```

```
## [1] -1.959964  1.959964
```

Lower Tailed Z-Test

- Two-tailed test with known variance: $H_0 : \mu_m \geq 1.728$, $H_1 : \mu_m < 1.728$
- Test statistics $z_0 = \frac{\sqrt{n}(\bar{X} - \mu_0)}{\sigma}$
  - Fail to reject H_0 if $z_{-\alpha} \leq z_0$

Lower Tailed Z-Test in R

- Test the hypothesis with $\alpha = 0.05$

```
z = sqrt(n)*(Mbar-1.728)/sigma
z
```

```
## [1] 0.2406529
```

```
z.alpha = -qnorm(1-0.05)
z.alpha
```

```
## [1] -1.644854
```

Two Tailed T-Test

- Two-tailed test with unknown variance:
$$H_0 : \mu_m = 1.728$$
$$H_1 : \mu_m \neq 1.728$$
- Test statistics $t_0 = \frac{\sqrt{n}(\bar{X} - \mu_0)}{S}$
  - Fail to reject H_0 if $-t_{\alpha/2} \leq t_0 \leq t_{\alpha/2}$

Two Tailed T-Test in R

```
t = sqrt(n)*(Mbar-1.728)/sd(m)
t
```

```
## [1] 0.01425566
```

```
t.half.alpha = qt(1-0.05/2, df=n-1)
c(-t.half.alpha, t.half.alpha)
```

```
## [1] -1.994437  1.994437
```

One-Sided T-test in R

```
t.test(m,mu=1.728,alternative="less")
```

```
##
##  One Sample t-test
##
## data:  m
## t = 0.014256, df = 70, p-value = 0.5057
## alternative hypothesis: true mean is less than 1.728
## 95 percent confidence interval:
##      -Inf 1.740125
## sample estimates:
## mean of x
##  1.728103
```

Multivariate Hypothesis Testing

- Two-tailed hypotheses $H_0 : \mu_m = \mu_f$, $H_1 : \mu_m \neq \mu_f$
- Paired vs. Unpaired (Independent)
- Equal variance assumtion vs. Welch's t-test

Paired T-test in R

```
t.test(m,f,paired=T)
```

```
##
##  Paired t-test
##
## data:  m and f
## t = 51.601, df = 70, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  0.1165859 0.1259605
## sample estimates:
## mean of the differences
##               0.1212732
```

Independent T-test in R

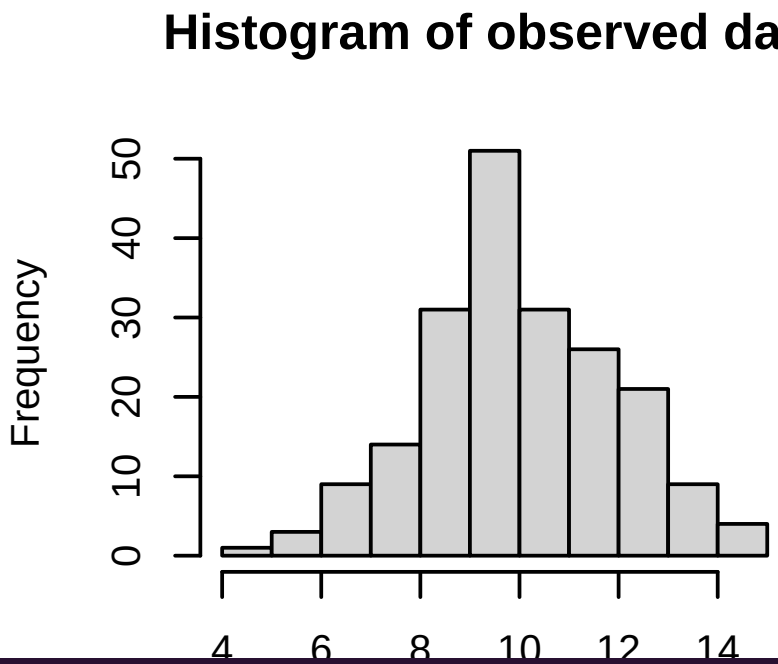
```
t.test(m,f,var.equal=T)
```

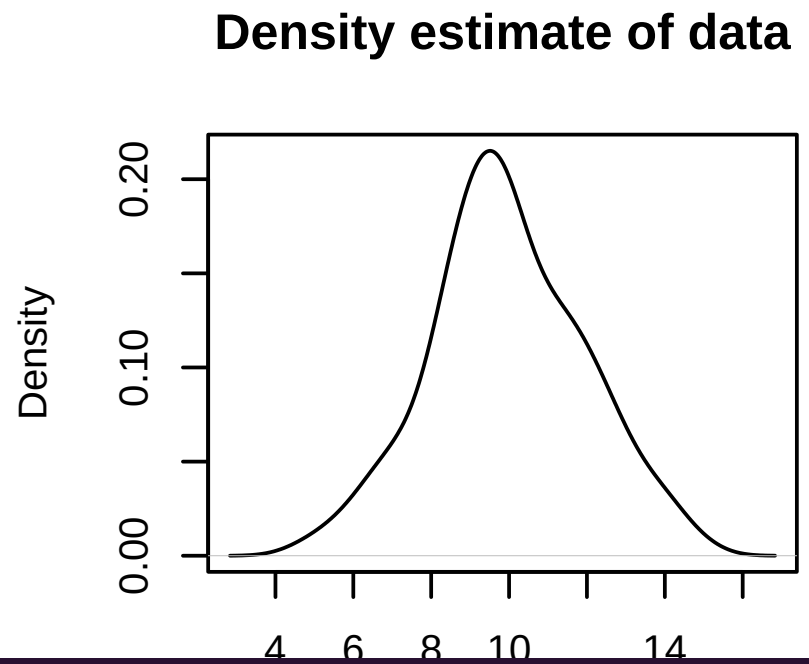
```
##
##  Two Sample t-test
##
## data:  m and f
## t = 12.331, df = 140, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not
## 95 percent confidence interval:
##  0.1018290 0.1407175
## sample estimates:
## mean of x mean of y
##  1.728103  1.606830
```

Distribution Fitting in R

- Histogram and Density estimation

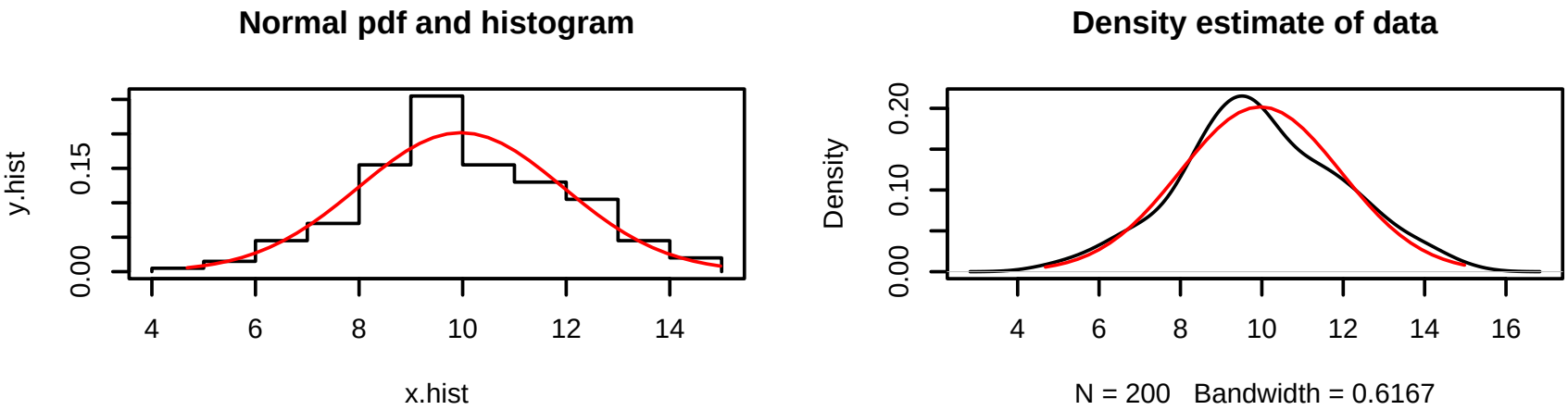
```
# Generate 200 random variates from N(10,2^2)
x.norm = rnorm(n=200,m=10,sd=2)
par(mfrow=c(1,2), cex=0.7)
hist(x.norm,breaks=10,main="Histogram of observed data")
plot(density(x.norm),main="Density estimate of data")
```





Distribution Fitting in R

```
par(mfrow=c(1,2), cex=0.5)
h = hist(x.norm, breaks=10, plot=F)
x.hist = c(min(h$breaks),h$breaks)
y.hist = c(0,h$density,0)
x.fit = seq(min(x.norm),max(x.norm),length=40)
y.fit = dnorm(x.fit, mean=mean(x.norm), sd=sd(x.norm))
plot(x.hist, y.hist, type="s", ylim=c(0,max(y.hist,y.fit)),
     main="Normal pdf and histogram")
lines(x.fit, y.fit, col="red")
plot(density(x.norm), main="Density estimate of data")
lines(x.fit,y.fit, col="red")
```



Distribution Fitting using FitDistrPlus

```
library("fitdistrplus")
```

```
f_weibull = fitdist(x.norm, "weibull")
summary(f_weibull)
```

```
## Fitting of the distribution ' weibull ' by maximum likelihood
## Parameters :
##         estimate Std. Error
## shape  5.536753  0.2957003
## scale 10.771345  0.1454363
## Loglikelihood:  -422.5133   AIC:  849.0266   BIC:  855.6233
## Correlation matrix:
##           shape     scale
## shape 1.0000000 0.3242938
## scale 0.3242938 1.0000000
```

Distribution Fitting using FitDistrPlus

```
f_normal = fitdist(x.norm, "norm")
summary(f_normal)
```

```
## Fitting of the distribution ' norm ' by maximum likelihood
## Parameters :
##      estimate Std. Error
## mean 9.964265 0.13945035
## sd   1.972126 0.09860617
## Loglikelihood:  -419.6101   AIC:  843.2202   BIC:  849.8168
## Correlation matrix:
##              mean           sd
## mean 1.000000e+00 5.862261e-10
## sd   5.862261e-10 1.000000e+00
```

Q-Q Plot

- Q-Q Plot
  - A graphical method for comparing two probability distributions by plotting their quantiles against each other. If the two sets come from a population with the same distribution, the points should fall approximately along a 45° reference line.

Q-Q Plot in R

```
z.norm = (x.norm-mean(x.norm))/sd(x.norm)
qqnorm(x.norm)
#Use qqplot(data1, data2) to compare distributions of 2 datasets
abline(mean(x.norm),b=sd(x.norm),col="red") #slope & intercept
```