Statistical Applications using

Bharat Bhushan Verma

https://www.linkedin.com/in/dearbharat/

Agenda

Statistical Applications using R

- Basic Statistics
- Correlation
- Linear Regression
- Multiple Regression
- 2 Sample T-Test
- 1 Sample T-Test
- ANOVA
- Clustering

Mean Syntax

mean(x, trim = 0, na.rm = FALSE, ...)

trim is used to drop some observations from both end of the sorted vector

na.rm is used to remove the missing values from the input vector.

Eg:

mean(x)

mean(x,trim=0.3)

Median Syntax

median(x, na.rm = FALSE)

x is the input vector.
na.rm is used to remove the missing values from the input vector.

Summary

summary(x)

Min, 1st Q, Median, Mean, 3rd Q, Max

Correlation

It provides a measure of strength and direction of linear relationship between two variables

Correlation Example - MTCars

```
corMat <- cor(mtcars[,c(1,3,6)])
round(corMat,2)
```

	mpg	disp	wt
mpg	1.00	-0.85	-0.87
disp	-0.85	1.00	0.89
wt	-0.87	0.89	1.00

Linear Regression

Linear Regression

- Finding a straight line that best describes the data
- The best fit line is then used for making prediction

Linear Regression

```
# Fitting linear model
m<-lm(y ~ x,data)

# Intercept and Slope
coef(m)
summary(m)

# Prediction
p <- predict(m, data.frame(x = ..))
```

LR Example - MTCars

```
m <- lm(mpg ~ wt, data = mtcars)

coef(m)
(Intercept)          wt
 37.285126          -5.344472
summary(m)

p <- predict(m, data.frame(wt = 3))
```

Linear Regression - MTCars

```
plot(mgp~wt,data=mtcars)
abline(m)
```

Challenge: Linear Regression

1. Do a linear regression for the quake dataset - mag vs stations
2. Predict the quake mag when there are 100 stations receiving the quake signal.

Time : 5 mins

Multiple Regression

```
m<-lm(y ~ x1+x2+x1*x2...,data)

coef(m)
summary(m)

p<- predict(m, data.frame(x1=..,x2=..,...))
```

Multiple Regression - MTCars

```
m <- lm(mpg ~ wt+hp, data = mtcars)

coef(m)
(Intercept)          wt        hp
37.22                -3.87     -0.03
summary(m)

p<- predict(m, data.frame(wt = 3, hp = 200))
```

MR with Interaction - MTCars

```
m <- lm(mpg ~ wt*hp, data = mtcars)


coef(m)
(Intercept)                    hp          wt       hp*wt
49.80                          -0.12       -8.21      0.02
summary(m)

p<- predict(m, data.frame(wt = 3, hp = 200))
```

Logistics Regression

Logistics regression is a nonlinear regression that apply to response (y) that is binary

```
m<-glm(y ~ x1+x2+x3...,data)
```

```
coef(m)
summary(m)
```

```
p<- predict(m, data.frame(x1=..,x2=..,...))
```

Logistics Regression - MTCars

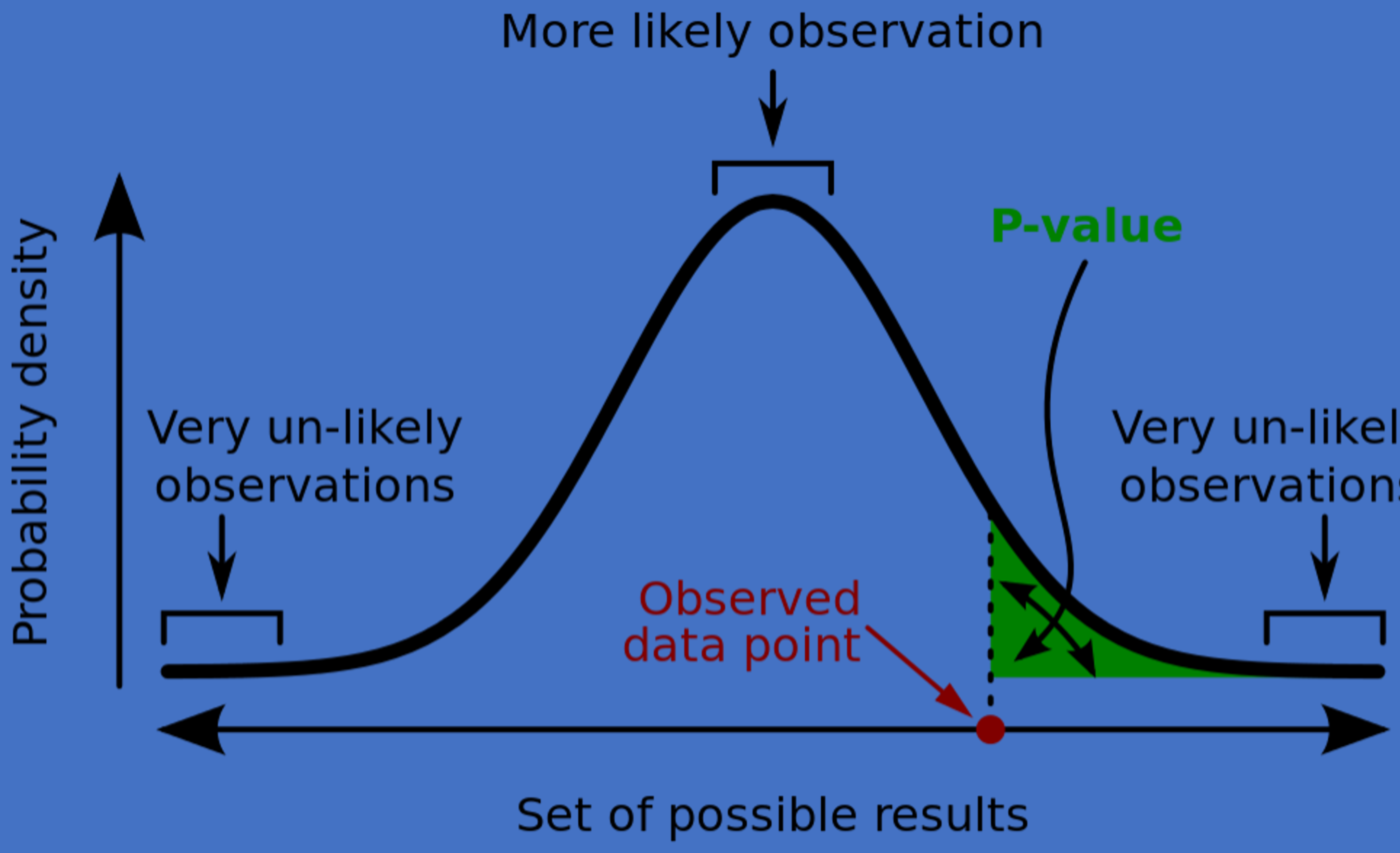
```
m <- glm(am ~ mpg+wt+hp, data = mtcars)


coef(m)
(Intercept)         mpg          wt          hp
 0.195756590  0.036308683 -0.338756898
  0.003891913
summary(m)
```

Hypothesis Testing

Hypothesis Testing

- Assume Null Hypothesis is true - No difference or no effect
- Compute the p-value
- If the p-value is small, then reject Null Hypothesis, else do not reject Null Hypothesis

P-Value



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

Guideline for p-value

If p value > .10 → “not significant”

If p value ≤ .10 → “marginally significant”

If p value ≤ .05 → “significant”

If p value ≤ .01 → “highly significant.”

2 Sample t-Test (2 Sided)

```
data(sleep)
extra<-sleep$extra
group<-sleep$group

t.test(extra~group,data=sleep)
```

2 Sample t-Test (One Sided)

```
data(sleep)
extra<-sleep$extra
group<-sleep$group

t.test(extra~group,data=sleep,,alternative="less")
```

1 Sample t-Test

```
t.test(extra ~ group, sleep, paired=TRUE)
```

Challenge: t-Test

Do a 2 sample t-test to compare the performance of chickwts feed - casein vs horsebean

Time : 5 mins

ANOVA

- t-test can do test 2 groups.
- When there are more than 2 groups then use ANOVA to test if there are statistically significant difference between the means.
- ANOVA does this by analysing the variance in the data set.
- ANOVA is similar to multiple regression

ANOVA

```
m<-aov(y~x1+x2+x1*x2....,data)
summary(m)
```

ANOVA Example - Chickwts

```
m<-aov(weight~feed,data=chickwts)
summary(m)
```

Alternatively can use Linear Regression Method

```
m <- lm(weight~feed,data=chickwts)
summary(m)
```

Challenge: ANOVA

Perform an ANOVA to determine any difference between the test scores of 3 teaching methods

Method A	Method B	Method C
79	71	82
86	77	68
94	81	70
89	83	76