Essential Statistics

John Hamre III

August 10, 2017

- Here I present essential statistics that are generated by using R code
- Many of these are critical for planning a clinical trial
- Further, many can be used in analysis of clinical trial data
- Slides are presented mainly as a title of the statistical test, code and results

Student's t-Test

```
problem1 <- read.csv("data/prob1data.csv")
t.test(problem1[,"X2011"],problem1[,"X2012"], paired=TRUE)
```

```
##
##  Paired t-test
##
## data:  problem1[, "X2011"] and problem1[, "X2012"]
## t = 2.1119, df = 24, p-value = 0.04529
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##    642.1576 55808.6424
## sample estimates:
## mean of the differences
##                 28225.4
```

```
#Reject the null hypothesis
#The number of crabs has gone down
```

Student's t-Test continued

```
problem2 <- read.csv("data/prob2data.csv")
prob2_matrix <- as.matrix(problem2[1:dim(problem2)[1], 2:dim(problem2)[2]])
t.test(prob2_matrix[c(1,5,14,17,19), 1:4], prob2_matrix[c(1,5,14,17,19),5:8], paired=FALSE, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  prob2_matrix[c(1, 5, 14, 17, 19), 1:4] and prob2_matrix[c(1, 5, 14, 17, 19), 5:8]
## t = -2.5091, df = 38, p-value = 0.01649
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -71.820544  -7.679456
## sample estimates:
## mean of x mean of y
##     46.75     86.50
```

```
# Reject the null hypothesis
```

F test to compare two variances, no significance

```
##
##  F test to compare two variances
##
## data:  prob3_matrix[c(1), 1:4] and prob3_matrix[c(1), 5:8]
## F = 0.19184, num df = 3, denom df = 3, p-value = 0.2083
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.01242543 2.96182924
## sample estimates:
## ratio of variances
##          0.1918385
```

Wilcox test

```
wilcox.test(prob3_matrix[c(1), 1:4] , prob3_matrix[c(1),5:8])
```

```
##
##  Wilcoxon rank sum test
##
## data:  prob3_matrix[c(1), 1:4] and prob3_matrix[c(1), 5:8]
## W = 8, p-value = 1
## alternative hypothesis: true location shift is not equal to 0
```

```
# This gene is not significanly different than control using wilcox test
```

Paired Wilcox test

```
problem5 <- read.csv("data/prob5data.csv")
wilcox.test(problem5[,"August"],problem5[,"November"], paired=TRUE)
```

```
##
##  Wilcoxon signed rank test
##
## data:  problem5[, "August"] and problem5[, "November"]
## V = 16, p-value = 0.03979
## alternative hypothesis: true location shift is not equal to 0
```

```
# According to a paired Wilcox, there is 95% confidence that we reject Null
```

Chi-squared distribution

```
O = c(57,330,2132,4584,4604,2119,659,251)
E = c(77.9,547.1,2126.7,4283.3,4478.5,2431.1,684.1,107.2)
x2 = sum((O-E)^2/E)
1-pchisq(x2,5)
```

```
## [1] 0
```

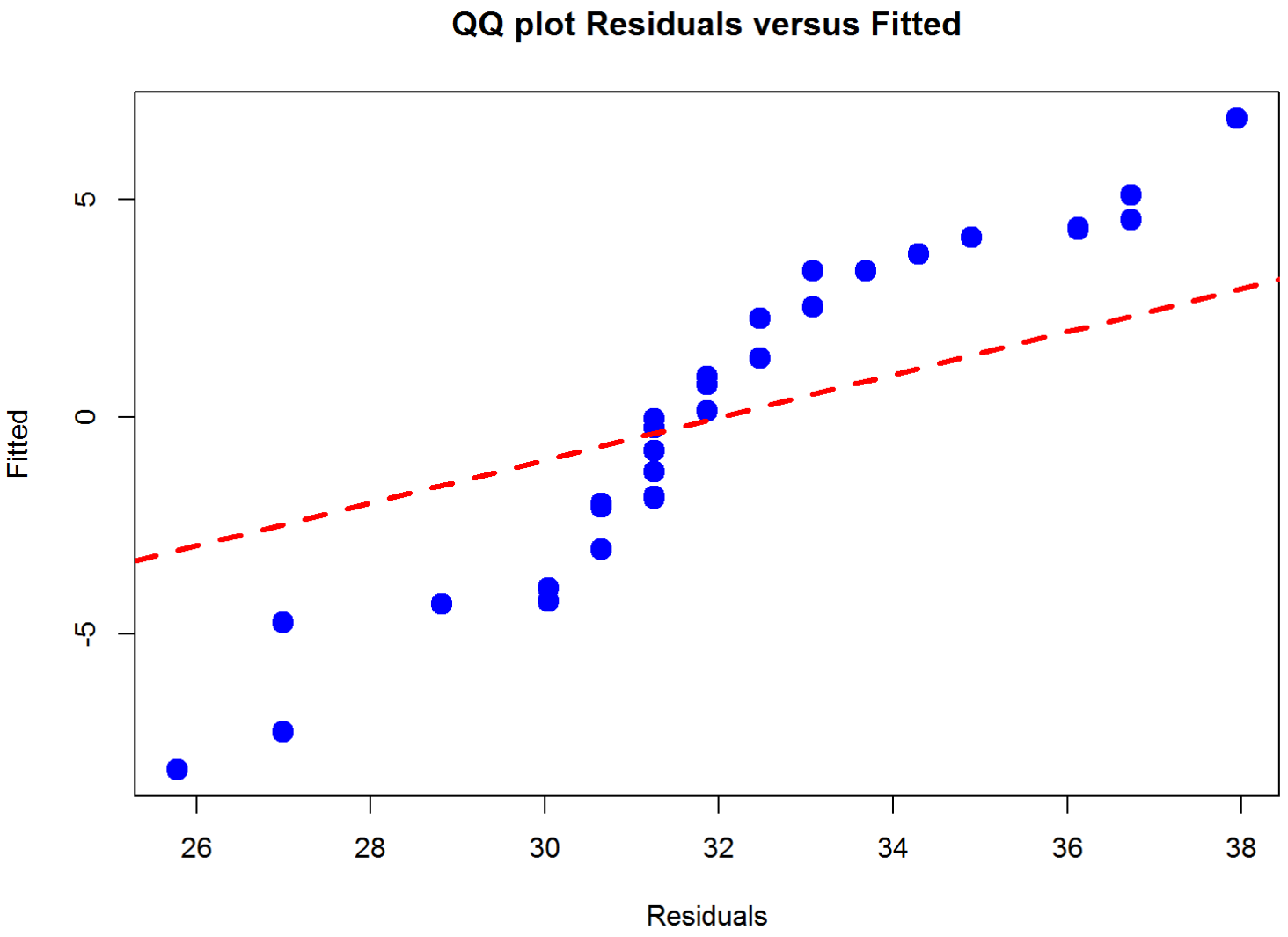
```
#Extreme statistical significance
#A small P value, as we have here, is evidence that the data are not sampled from the distribution you expected.
```

Least squares

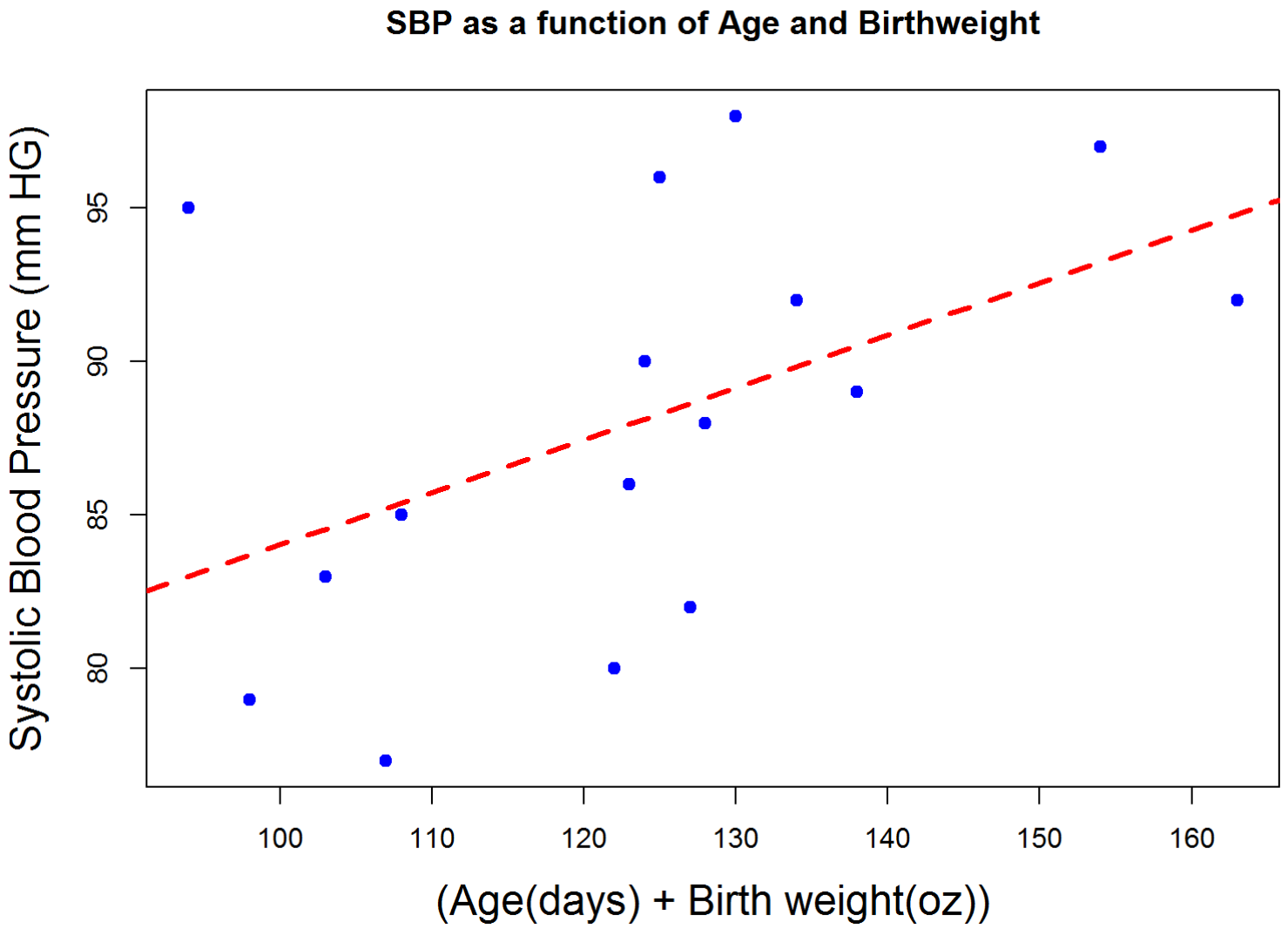
```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.1200 -2.0381 -0.0381  3.3537  6.8800
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  21.5234     2.6204   8.214 4.68e-09 ***
## x             0.6082     0.1468   4.143 0.000271 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.821 on 29 degrees of freedom
## Multiple R-squared:  0.3718, Adjusted R-squared:  0.3501
## F-statistic: 17.16 on 1 and 29 DF,  p-value: 0.0002712
```

QQ plot

```
## Warning in if (datax) {: the condition has length > 1 and only the first
## element will be used
```


QQ plot Residuals versus Fitted
Fitted
5
0
-5
26
28
30
32
34
36
38
Residuals

Least squares plot



knn prediction

```
library(class)
library(ISLR)
```

```
## Warning: package 'ISLR' was built under R version 3.2.4
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.2.5
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 3.2.5
```

```
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 3.2.5
```

```
knn.pred <- knn(Khan$xtrain,Khan$xtest,Khan$ytrain,k=1)
knn.pred
```

```
##  [1] 3 4 4 2 1 3 4 4 4 1 4 4 1 2 2 2 4 4 4 4
## Levels: 1 2 3 4
```

```
table(knn.pred, Khan$ytest)
```

```
##
## knn.pred 1 2 3 4
##        1 3 0 0 0
##        2 0 4 0 0
##        3 0 0 2 0
##        4 0 2 4 5
```

Density, distribution function

```
sigma = 40
n = 200
mu0 = 190
z=sqrt(n)*(181.52-mu0)/sigma
z
```

```
## [1] -2.998133
```

```
2 * (1-pnorm(abs(z)))
```

```
## [1] 0.002716393
```

```
# the Null is rejected
```

Probability of success in a Bernoulli experiment

```
binom.test(400, 10000, p = 0.02, alternative = c("greater"),  conf.level = 0.95)
```

```
##
##  Exact binomial test
##
## data:  400 and 10000
## number of successes = 400, number of trials = 10000, p-value <
## 2.2e-16
## alternative hypothesis: true probability of success is greater than 0.02
## 95 percent confidence interval:
##  0.03682635 1.00000000
## sample estimates:
## probability of success
##                   0.04
```

Fisher's exact test

```
data <- matrix(c(2,5,23,30),2,byrow=TRUE)
tab <- t(matrix(data, nrow=2,ncol=2))
fisher.test(tab)
```

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tab
## p-value = 0.6882
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.04625243 3.58478157
## sample estimates:
## odds ratio
##   0.527113
```