Introduction to R - Basic Statistics with R

Natàlia Vilor-Tejedor
natalia.vilor@isglobal.org

Barcelona Institute for Global Health (ISGlobal)
http://www.isglobal.org

Introduction to R
2nd Edition
May 31th 2017

ISGlobal Barcelona Institute for Global Health
Generalitat de Catalunya
Ajuntament de Barcelona
Parc de Salut Barcelona MAR
upf. Universitat Pompeu Fabra Barcelona
Unió Europea Fons social europeu L'FSE inverteix en el teu futur

Outline

Required Packages

These are the required packages for this session:

```
> library(MASS)
```

These are the required data for this session:

```
> data(birthwt)
```

Inspecting data

```
> names(birthwt)
 [1] "low"   "age"   "lwt"   "race"  "smoke" "ptl"   "ht"
 [8] "ui"    "ftv"   "bwt"
> ?birthwt
```

Manipulate data

Subsetting data:

```
> b.s <- birthwt[,c("bwt", "smoke", "race", "age", "low")]
```

Inspecting data:

```
> head(b.s)
```

```
    bwt smoke race age low
85 2523     0    2  19   0
86 2551     0    3  33   0
87 2557     1    1  20   0
88 2594     1    1  21   0
89 2600     1    1  18   0
91 2622     0    3  21   0
```

Outline

Descriptive Statistics

Descriptive Statistics with R

R provides a wide range of functions for obtaining summary statistics.

Type of variables

Generally, statistical variables can be:

- Continuous variables that are numeric. They represent a measurable quantity. (i.e. *Age, BMI, ...*)
- Categorical variables that take on values that are names or labels (i.e. *Sex, Colors, ...*)

Type of variables

Generally, statistical variables can be:

- Continuous variables that are numeric. They represent a measurable quantity. (i.e. *Age, BMI, ...*)
- Categorical variables that take on values that are names or labels (i.e. *Sex, Colors, ...*)

Descriptive analysis for Continuous variables

Example: Simulating 300 observations under a normal distribution.

```
> x<-rnorm(300,sd=10,mean=100)
> head(x)

[1] 122.89340  92.25365  93.84143  98.25029  97.45983
[6] 104.29076
```

Calculating some statistics of interest:

```
> mean(x); median(x)

[1] 99.46846

[1] 99.11464

> quantile(x)

      0%       25%       50%       75%      100%
 69.62631  92.26574  99.11464 106.55742 123.97740
```

Best option to obtain descriptive information:

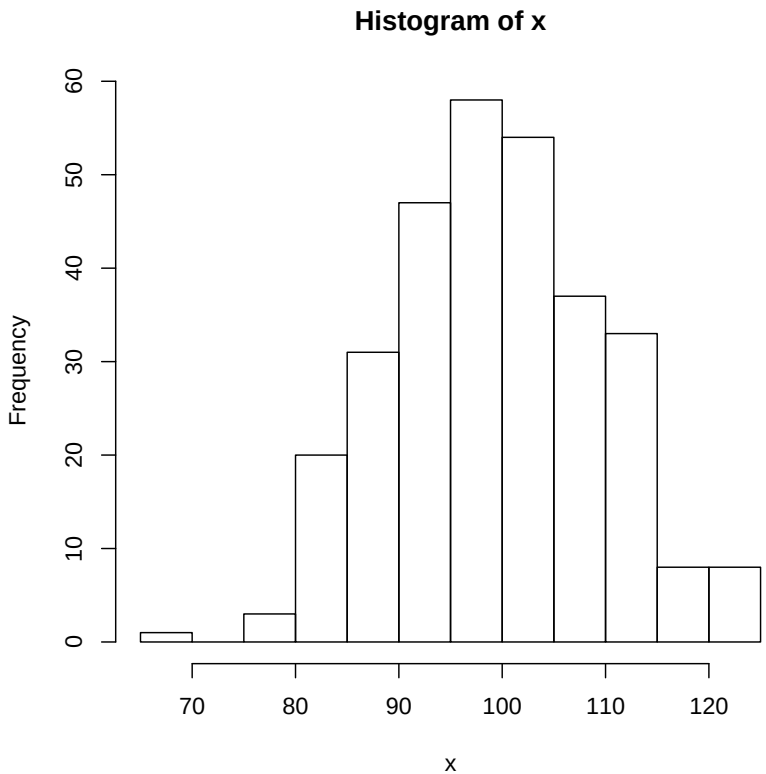
```
> summary(x)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  69.63   92.27   99.11   99.47  106.60  124.00
```

Descriptive analysis for Continuous variables

Histogram:

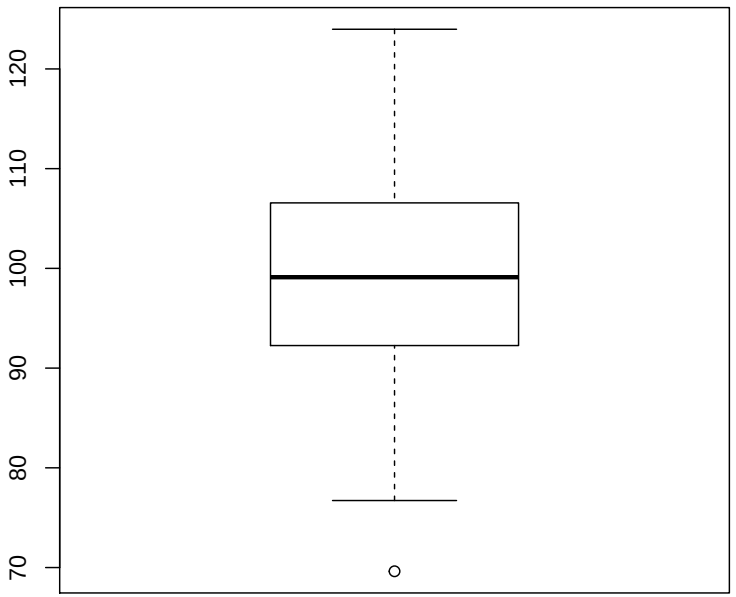
```
> hist(x)
```



Descriptive analysis for Continuous variables

Boxplot:

```
> boxplot(x)
```



Type of variables

Generally, statistical variables can be:

- Continuous variables are numeric. They represent a measurable quantity. (i.e. *Age, BMI, ...*)
- Categorical variables take on values that are names or labels (i.e. *Sex, Colors, ...*)

Descriptive analysis for Categorical variables

Example: Defining a dichotomous variable.

```
> var<-c(0,1,0,0,0,1,1,0,0,1,1,1,1,1);var
 [1] 0 1 0 0 0 1 1 0 0 1 1 1 1 1
```

Recategorizing variable's values:

```
> var<-factor(var,labels=c("control","case"));var
 [1] control case    control control control case    case
 [8] control control case    case    case    case    case
Levels: control case
```

Computing basic recount:

```
> recvar<-table(var); recvar
var
control    case
      6       8
```

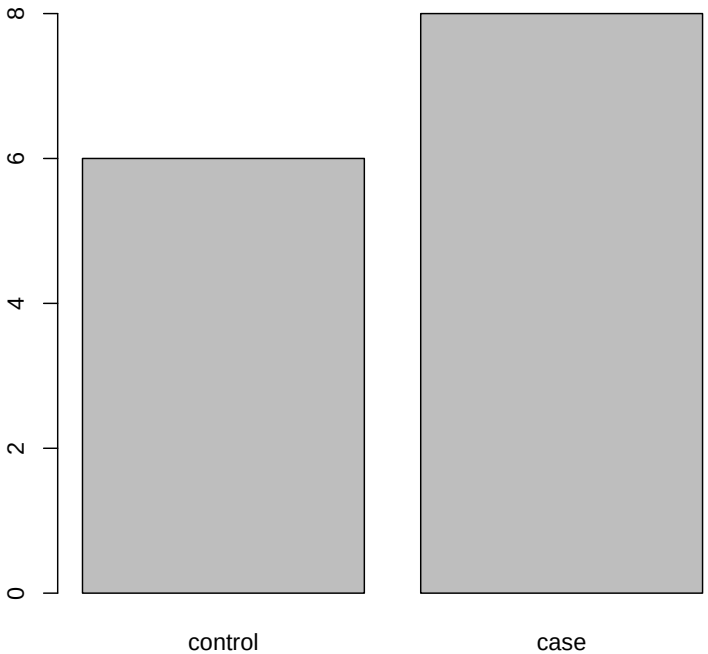
Computing relative frequencies:

```
> frelvar<-prop.table(table(var)); frelvar
var
  control      case
0.4285714 0.5714286
```

Descriptive analysis for Categorical variables

Barplot:

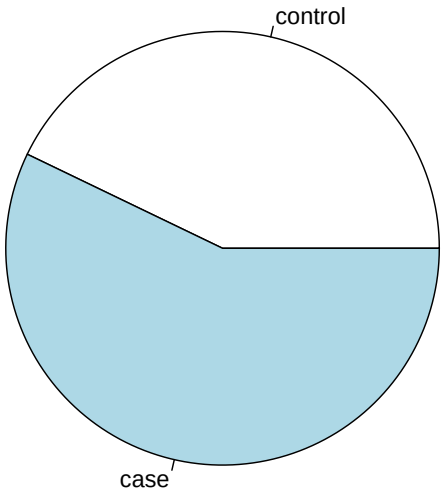
```
> barplot(recvar)
```



Descriptive analysis for Categorical variables

Pie chart:

```
> pie(recvar)
```



Descriptive analysis for `birthwt` data

Remember:

```
> head(b.s)
```

	bwt	smoke	race	age	low
85	2523	0	2	19	0
86	2551	0	3	33	0
87	2557	1	1	20	0
88	2594	1	1	21	0
89	2600	1	1	18	0
91	2622	0	3	21	0

	bwt	smoke	race	age	low
Type of Variable:	Continuous	Categorical	Categorical	Continuous	Categorical

Descriptive analysis for `birthwt` data

Remember:

```
> head(b.s)
```

```
    bwt smoke race age low
85 2523     0    2  19   0
86 2551     0    3  33   0
87 2557     1    1  20   0
88 2594     1    1  21   0
89 2600     1    1  18   0
91 2622     0    3  21   0
```

	bwt	smoke	race	age	low
Type of Variable:	Continuous	Categorical	Categorical	Continuous	Categorical

Descriptive analysis for `birthwt` data

```
> summary(b.s)
```

```
      bwt             smoke             race      
 Min.   : 709   Min.   :0.0000   Min.   :1.000  
 1st Qu.:2414   1st Qu.:0.0000   1st Qu.:1.000  
 Median :2977   Median :0.0000   Median :1.000  
 Mean   :2945   Mean   :0.3915   Mean   :1.847  
 3rd Qu.:3487   3rd Qu.:1.0000   3rd Qu.:3.000  
 Max.   :4990   Max.   :1.0000   Max.   :3.000  
      age              low        
 Min.   :14.00   Min.   :0.0000  
 1st Qu.:19.00   1st Qu.:0.0000  
 Median :23.00   Median :0.0000  
 Mean   :23.24   Mean   :0.3122  
 3rd Qu.:26.00   3rd Qu.:1.0000  
 Max.   :45.00   Max.   :1.0000  
```

Exercise 1:

Convert `race`, `low` and `smoke` in factor variables.

Descriptive analysis for `birthwt` data

Solution 1:

```
> b.s$race <- factor(b.s$race,
+                    labels = c("white", "black", "other"))
> b.s$low <- factor(b.s$low, labels = c("normal", "< 2.5kg"))
> b.s$smoke <- factor(b.s$smoke,
+                     labels = c("non smoker", "smoker"))
```

Exercise 2:

Now, inspect the `summary()` function.

Descriptive analysis for `birthwt` data

Solution 2:

```
> summary(b.s)
```

```
      bwt                 smoke        race          age       
 Min.   : 709   non smoker:115   white:96   Min.   :14.00  
 1st Qu.:2414   smoker    : 74   black:26   1st Qu.:19.00  
 Median :2977                    other:67   Median :23.00  
 Mean   :2945                               Mean   :23.24  
 3rd Qu.:3487                               3rd Qu.:26.00  
 Max.   :4990                               Max.   :45.00  
      low        
 normal :130  
 < 2.5kg: 59  
```

Descriptive analysis for `birthwt` data

Exercise 3:

Compute relative frequencies and percentages for categorical variables.

Descriptive analysis for `birthwt` data

Solution 3: Frequencies

```
> frsmoke<-prop.table(table(b.s$smoke));frsmoke

non smoker     smoker
 0.6084656  0.3915344

> frrace <- prop.table(table(b.s$race)); frrace

    white     black     other
0.5079365 0.1375661 0.3544974

> frlow <- prop.table(table(b.s$low));frlow

   normal    < 2.5kg
0.6878307 0.3121693
```

Descriptive analysis for `birthwt` data

Solution 3: Percentages

```
> frsmoke*100

non smoker     smoker
  60.84656   39.15344

> frrace*100

   white    black    other
50.79365 13.75661 35.44974

> frlow*100

  normal   < 2.5kg
68.78307 31.21693
```

Descriptive analysis for `birthwt` data

Exercise 4:

Make a plot for each variable.

Descriptive analysis for `birthwt` data

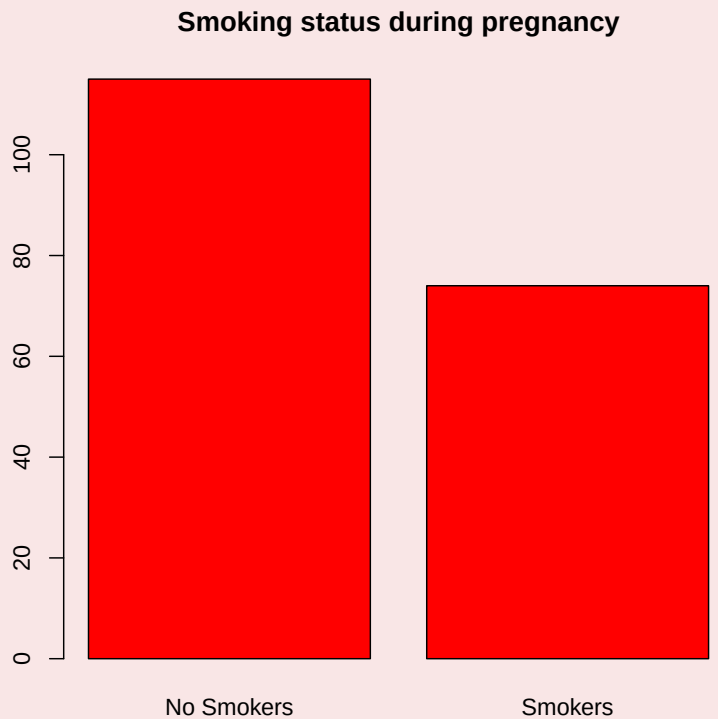
Solution 4:

```
> b.c <- b.s$bwt
> hist(b.c, main="Histogram of birth weight",
+      xlab=c("birth weight in grams"), col="red")
```

Descriptive analysis for `birthwt` data

Solution 4:

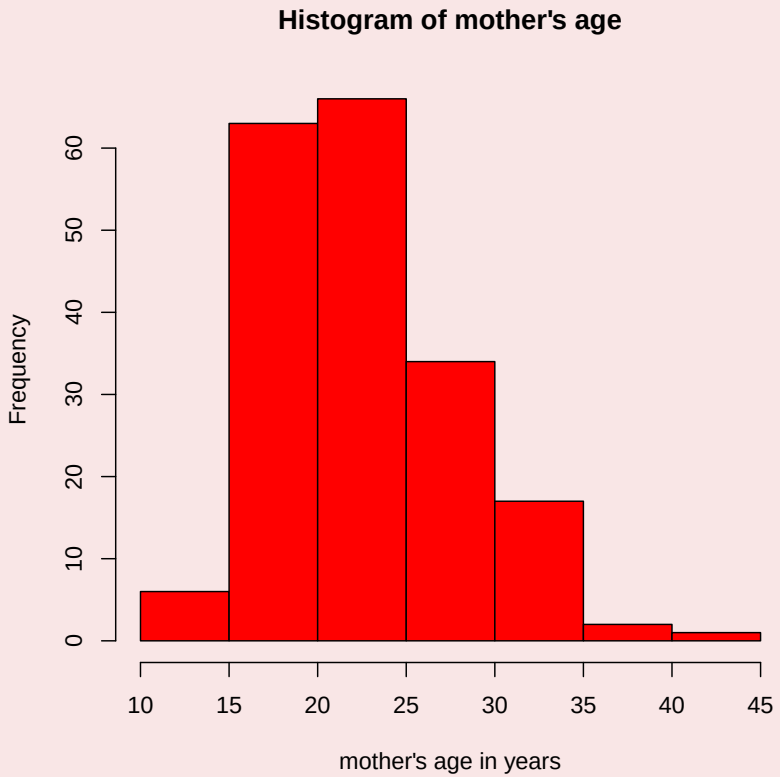
```
> s.c <- table(b.s$smoke)
> barplot(s.c, names.arg=c("No Smokers", "Smokers"),
+         main="Smoking status during pregnancy", col="red")
```



Descriptive analysis for `birthwt` data

Solution 4:

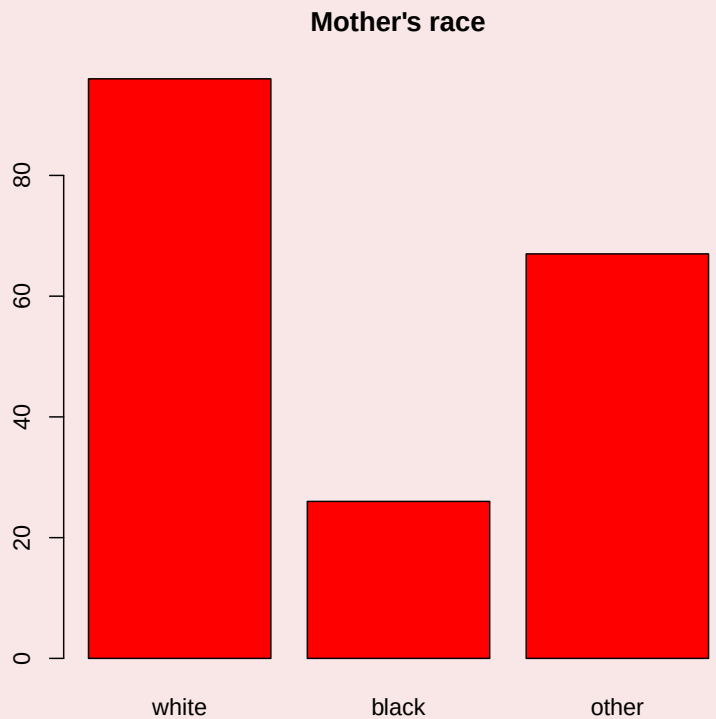
```
> a.c <- b.s$age
> hist(a.c, main="Histogram of mother's age",
+      xlab=c("mother's age in years"), col="red")
```



Descriptive analysis for `birthwt` data

Solution 4:

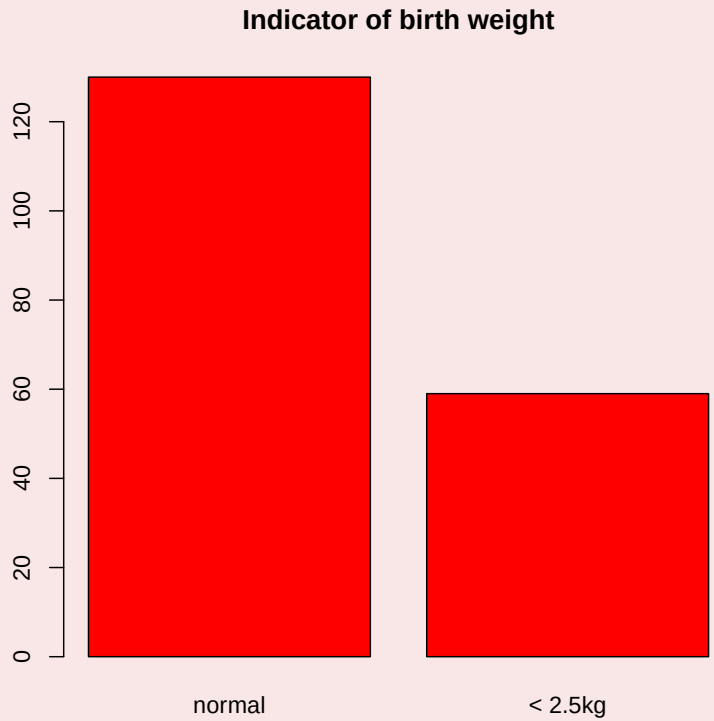
```
> s.r <- table(b.s$race)
> barplot(s.r, names.arg=c("white", "black", "other"),
+         main="Mother's race", col="red")
```



Descriptive analysis for `birthwt` data

Solution 4:

```
> s.l <- table(b.s$low)
> barplot(s.l, names.arg=c("normal", "< 2.5kg"),
+         main="Indicator of birth weight", col="red")
```



Outline

Bivariate analysis

- Categorical variable - Categorical Variable (i.e. *Sex, Race*)
- Continuous variable - Continuous variable (i.e. *Age, BMI*)

Bivariate analysis for categorical variables

- Categorical variable - Categorical Variable (i.e. *Sex, Race*)
- Continuous variable - Continuous variable (i.e. *Age, BMI*)

Bivariate analysis for categorical variables

Contingency table:

```
> smoke.race<-table(b.s$smoke, b.s$race); smoke.race
```

```
            white black other
  non smoker    44    16    55
  smoker        52    10    12
```

Cell proportions:

```
> prop.table(smoke.race)
```

```
                 white      black      other
  non smoker 0.23280423 0.08465608 0.29100529
  smoker     0.27513228 0.05291005 0.06349206
```

Row proportions:

```
> smoke.race.r<-prop.table(smoke.race,1); smoke.race.r
```

```
                 white     black     other
  non smoker 0.3826087 0.1391304 0.4782609
  smoker     0.7027027 0.1351351 0.1621622
```

Bivariate analysis for categorical variables

Column proportions:

```
> smoke.race.c<-prop.table(smoke.race,2); smoke.race.c
```

```
                white     black     other
  non smoker 0.4583333 0.6153846 0.8208955
  smoker     0.5416667 0.3846154 0.1791045
```

Frequencies:

```
> margin.table(smoke.race, 1) # smoke frequencies (summed over
```

```
non smoker      smoker
       115          74
```

```
> margin.table(smoke.race, 2) # race frequencies (summed over
```

```
white black other
   96    26    67
```

Bivariate analysis for categorical variables

Chi-squared Test of Independence:

```
> smoke.race

             white black other
  non smoker    44    16    55
  smoker        52    10    12

> chisq.test(smoke.race)

        Pearson's Chi-squared test

data:  smoke.race
X-squared = 21.779, df = 2, p-value = 1.865e-05
```

Interpretation:

Since the P-value (1.865e-05) is less than the significance level (0.05), we cannot accept the null hypothesis. Thus, we conclude that there is a relationship between race and smoking.

Bivariate analysis for categorical variables

Relative Risk: Probability of having the disease for people who were exposed to the treatment or environmental factor, divided by the probability of having the disease for people who were not exposed to that treatment or environmental factor.

```
> smoke.low <-table(b.s$smoke, b.s$low); smoke.low

            normal < 2.5kg
  non smoker     86      29
  smoker         44      30

> smoke.low <- smoke.low[,c(2,1)]; smoke.low

            < 2.5kg normal
  non smoker      29     86
  smoker          30     44

> smoke.low <- smoke.low[c(2,1),]; smoke.low

            < 2.5kg normal
  smoker          30     44
  non smoker      29     86
```

Bivariate analysis for categorical variables

```
> source("Riskfunctions.R")
> calcRelativeRisk(smoke.low)

[1] "category = smoker , relative risk =  1.607642124883 5"
[1] "category = smoker ,  95 % confidence interval = [ 1.05781
```

Interpretation

Smokers have 60% the chance to have a child with low birth weight than non smokers.

Bivariate analysis for categorical variables

With a retrospective case-control data, direct calculations of the relative risk should not be performed, as the results are not meaningful. In these cases we use the **Odds ratio** measure.

```
> calcOddsRatio(smoke.low)
```

```
[1] "category = smoker , odds ratio =  2.021943573667 71"
[1] "category = smoker ,  95 % confidence interval = [ 1.08065
```

You can interpret this odds ratio as a relative risk.

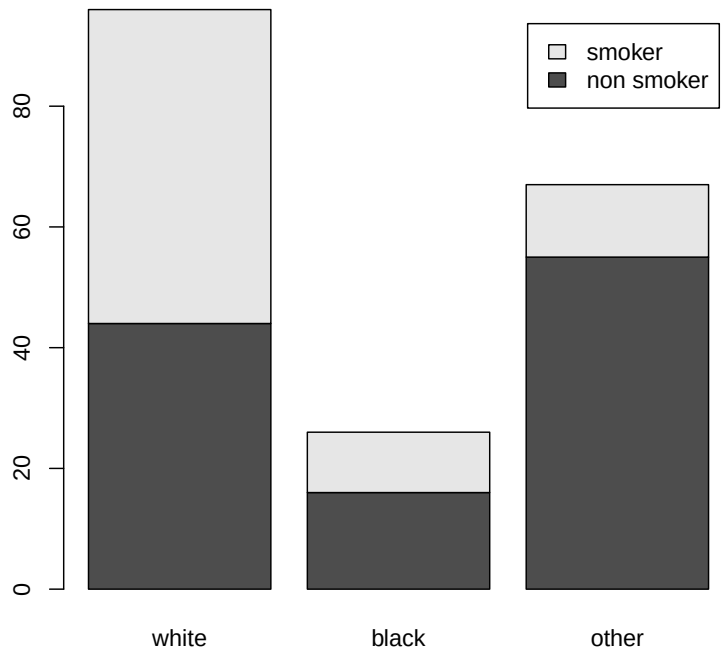
Interpretation

The risk of a smoker to have a child with low birth weight is about two times the risk of a non-smoker.

Bivariate analysis for categorical variables

Barplot:

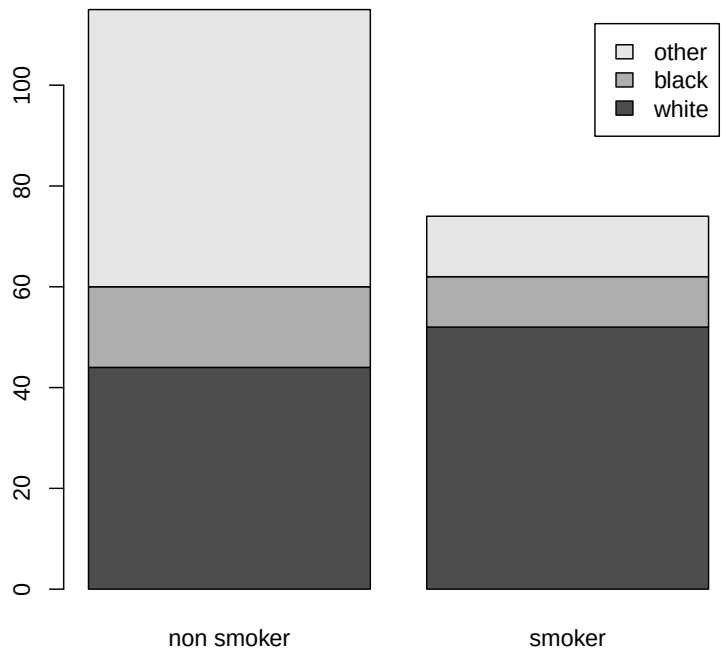
```
> barplot(smoke.race, legend=rownames(smoke.race))
```



Bivariate analysis for categorical variables

Barplot:

```
> barplot(t(smoke.race), legend=colnames(smoke.race))
```



Bivariate analysis for continuous variables

- Categorical variable - Categorical Variable (i.e. *Sex, Race*)
- Continuous variable - Continuous variable (i.e. *Age, BMI*)

Bivariate analysis for continuous variables

Correlation

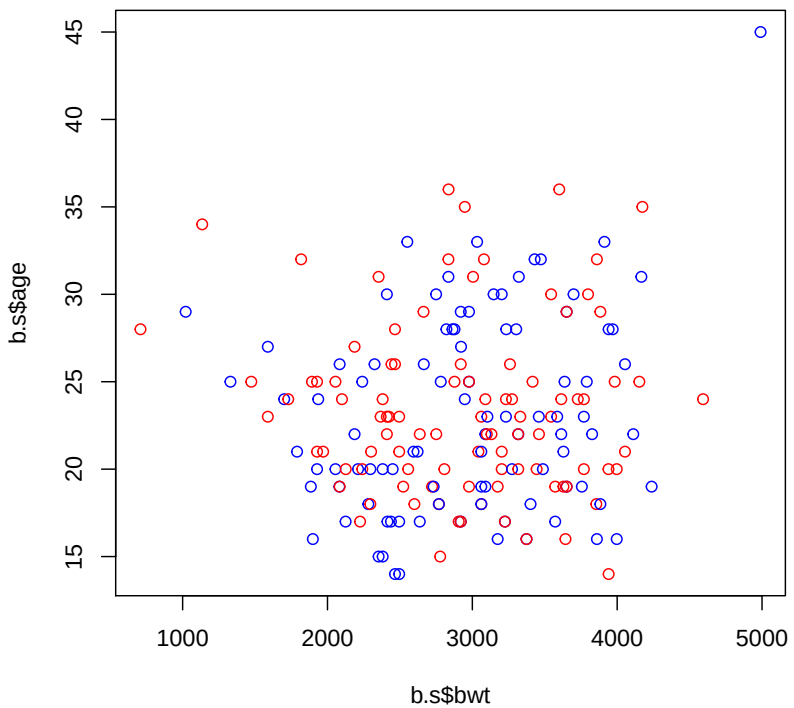
Correlation coefficients measure the strength of association between two variables.The sign and the absolute value of a correlation coefficient describe the direction and the magnitude of the relationship between two variables.

- The value of a correlation coefficient, ρ, ranges between -1 and 1.
- A positive correlation means that if one variable gets bigger, the other variable tends to get bigger ($\rho \sim 1$).
- A negative correlation means that if one variable gets bigger, the other variable tends to get smaller ($\rho \sim -1$).
- The weakest linear relationship is indicated by a correlation coefficient equal to 0.

Bivariate analysis for continuous variables

How to calculate a correlation coefficient with R:

```
> cor(b.s$bwt, b.s$age)
[1] 0.09031781
> plot(b.s$bwt, b.s$age, col=c("red", "blue"))
```



Bivariate analysis for continuous variables

Test for correlation:

```
> cor(b.s$bwt, b.s$age)
[1] 0.09031781
> cor.test(b.s$bwt, b.s$age)

        Pearson's product-moment correlation

data:  b.s$bwt and b.s$age
t = 1.2401, df = 187, p-value = 0.2165
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.05309694  0.23008208
sample estimates:
       cor
0.09031781
```

Interpretation:

Since the P-value (0.2165) is greater than the significance level (0.05), we cannot reject the null hypothesis. Hence, there are not statistically correlation between birth weight and mother's age.

Bivariate analysis for continuous variables

Linear regression

Represents a cause and effect relationship where the independent variable is the cause, and the dependent variable is the effect.

Least squares regression line

Linear regression finds the straight line, called the least squares regression line that best represents observations in a bivariate data set (minimizes the sum of squared differences between observed values and predicted values):

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Bivariate analysis for continuous variables

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- The regression constant, β_0, is the intercept of the regression line.
- The regression coefficient, β_1, is the average change in the dependent variable, Y, for a 1-unit change in the independent variable, X. It is the slope of the regression line.

Bivariate analysis for continuous variables

```
> m1 <- lm(bwt ~ age, data=b.s)
> summary(m1)

Call:
lm(formula = bwt ~ age, data = b.s)

Residuals:
     Min       1Q   Median       3Q      Max
-2294.78  -517.63    10.51   530.80  1774.92

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2655.74     238.86   11.12   <2e-16 ***
age            12.43      10.02    1.24    0.216
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 728.2 on 187 degrees of freedom
Multiple R-squared:  0.008157,        Adjusted R-squared:  0.0
F-statistic: 1.538 on 1 and 187 DF,  p-value: 0.2165
```

Bivariate analysis for continuous variables

Coefficient of determination R^2

It is interpreted as the proportion of the variance in the dependent variable that is predictable from the independent variable.
(i.e. an R^2 of 0.10 means that 10 percent of the variance in Y is predictable from X)

```
> ans <- summary(m1)
> names(ans)
```

```
 [1] "call"          "terms"         "residuals"
 [4] "coefficients"  "aliased"       "sigma"
 [7] "df"            "r.squared"     "adj.r.squared"
[10] "fstatistic"    "cov.unscaled"
```

```
> ans$adj.r.squared
```

```
[1] 0.002853336
```

Outline

Inference analysis

Test to Compare Two Variances:

```
> b.s.w <- subset(b.s, race=="white")
> b.s.b <- subset(b.s, race=="black")
> var.test(b.s.w$bwt, b.s.b$bwt)

        F test to compare two variances

data:  b.s.w$bwt and b.s.b$bwt
F = 1.2988, num df = 95, denom df = 25, p-value =
0.4621
alternative hypothesis: true ratio of variances is not equal t
95 percent confidence interval:
 0.6493981 2.3095510
sample estimates:
ratio of variances
          1.298838
```

Interpretation:

Since the P-value (0.4621) is greater than the significance level (0.05), we cannot reject the null hypothesis. Hence, variances are statistically equal.

Inference analysis

Two sample t-test:

```
> b.s.w <- subset(b.s, race=="white")
> b.s.b <- subset(b.s, race=="black")
> t.test(b.s.w$bwt, b.s.b$bwt,
+        var.equal=ifelse(var.test(b.s.w$bwt, b.s.b$bwt)$p.val
+                         > 0.05,TRUE,FALSE))

        Two Sample t-test

data:  b.s.w$bwt and b.s.b$bwt
t = 2.4393, df = 120, p-value = 0.01618
alternative hypothesis: true difference in means is not equal
95 percent confidence interval:
  72.13796 693.91493
sample estimates:
mean of x mean of y
 3102.719  2719.692
```

Interpretation:

Since the P-value (0.0161) is less than the significance level (0.05), we can reject the null hypothesis. Hence, means are statistically different.

Inference analysis

Normality Test:

```
> shapiro.test(b.s.w$bwt)

        Shapiro-Wilk normality test

data:  b.s.w$bwt
W = 0.98727, p-value = 0.4861

> shapiro.test(b.s.b$bwt)

        Shapiro-Wilk normality test

data:  b.s.b$bwt
W = 0.97696, p-value = 0.8038
```

Interpretation:

Since the P-value (0.4861) is greater than the significance level (0.05), we cannot reject the null hypothesis. Hence, birth weight is normally distributed.

Inference analysis

Example:

```
> aa <- runif(100,0,1); bb <- runif(100,1,5); cc <- runif(100,
> shapiro.test(aa)

        Shapiro-Wilk normality test

data:  aa
W = 0.94608, p-value = 0.0004619

> shapiro.test(bb)

        Shapiro-Wilk normality test

data:  bb
W = 0.94989, p-value = 0.0008157

> shapiro.test(cc)

        Shapiro-Wilk normality test

data:  cc
W = 0.94613, p-value = 0.0004654
```

Outline

Simple linear regression

```
> m1 <- glm(bwt ~ age, data=b.s)
> summary(m1)
Call:
glm(formula = bwt ~ age, data = b.s)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2294.78   -517.63     10.51    530.80   1774.92

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2655.74     238.86   11.12   <2e-16 ***
age            12.43      10.02    1.24    0.216
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 530236.2

    Null deviance: 99969656  on 188  degrees of freedom
Residual deviance: 99154173  on 187  degrees of freedom
```

Simple logistic regression

```
> m2 <- glm(low ~ age, data=b.s, family=binomial)
> ans2 <- summary(m2)
> ans2$coefficients
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.38458192	0.73212479	0.5252956	0.5993777
age	-0.05115294	0.03151376	-1.6231937	0.1045480

Multiple linear regression

```
> mli <- glm(bwt ~ -1+age+smoke+race, data=b.s)
> summary(mli)

Call:
glm(formula = bwt ~ -1 + age + smoke + race, data = b.s)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2322.6   -447.3     28.4    502.2   1612.3

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
age                2.134      9.771   0.218 0.827326
smokenon smoker 3281.673    260.664  12.590  < 2e-16 ***
smokesmoker     2855.579    247.404  11.542  < 2e-16 ***
raceblack       -444.069    156.194  -2.843 0.004973 **
raceother       -447.858    119.017  -3.763 0.000226 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 476133.9
```

Multiple logistic regression

```
> mlo <- glm(smoke ~ age+race, data=b.s, family=binomial)
> summary(mlo)
Call:
glm(formula = smoke ~ age + race, family = binomial, data = b.

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4558  -1.0468  -0.6041   1.0641   2.0303

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.26703    0.77840   1.628   0.1036
age         -0.04521    0.03082  -1.467   0.1424
raceblack   -0.76998    0.46499  -1.656   0.0977 .
raceother   -1.79049    0.38842  -4.610 4.03e-06 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

Outline

Final Exercise

In this last exercise we will use a simulated data set which contains:

- `sex`: Sex status variable
- `bmi`: Body Mass Index variable
- `age`: Individual's Age variable
- `exposure`: Continuous exposure variable

Final Exercise - Preliminary

- Read and Inspect data (`myData.txt`). Save the data in a R object named `myData`.
- Transform to factor the variable `sex` (`male` as reference level).
- Add to `myData` a dichotomization of `exposure`, named `Ebin`, as a factor with levels `low`(reference level) and `high` with the threshold in the median.

Final Exercise - Univariate descriptive

- For quantitative variables compute:
  `min`, `pct 2.5%`, `median`, `mean`, `pct 97.5%`, `max`, `sd`, `N` and histograms.
- For qualitative variables compute:
  relative frequencies, percentages and Barplots or Pie charts.

Final Exercise - Bivariate descriptive

Perform:

- A contingence table for `Ebin` and `sex`.
- A scatter plot between `exposure` and `bmi`.

Final Exercise - Inference

Reproduce and Interpret:

- A correlation test between `exposure` and `bmi`.
- A normality test for `exposure`.

Final Exercise - GLM

- Fit a GLM with `bmi` as the response and `exposure` as the explanatory variable, adjusting for `sex` and `age` as possible confounders. Save this model in a R object named `mod1`.
- Explore objects in `mod1` and in `summary(mod1)`. Interpret the results.

Final Exercise - GLM

- Fit a GLM with dichotomic `Ebin` as the response (logistic regression) and `bmi` as the explanatory variable, adjusting for `sex` and `age` as possible confounders. Save this model in a R object named `mod2`.
- Explore objects in `mod2` and in `summary(mod2)`. Interpret the results.