Introduction in Statistics
- using R -

Anna Vinkhuyzen

March 2016

Outline Lecture

- Descriptive statistics
  - 'First stage' statistics, summarizing the sample data
  - Descriptive statistics include proportions, means, variances, covariances, correlations
  - Used to characterize a population
- Inferential statistics
  - Statistics to generalize the information obtained form the measured sample to the population
- Analysis methods
  - Numerous methods to analyse the data
  - Choice of method depends on question asked as well as on the data available
- Analyses tools
  - R

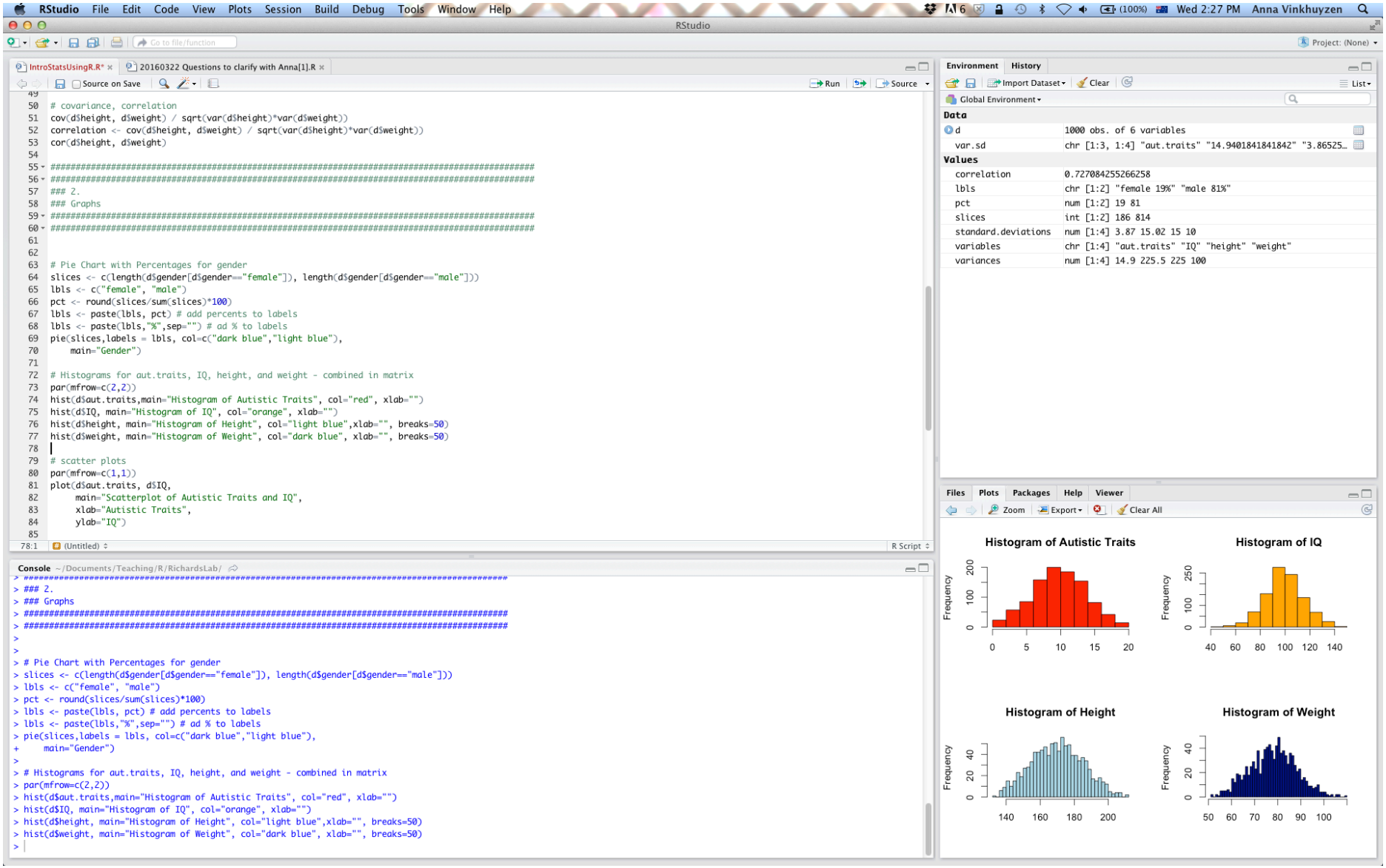
Outline Lecture

- Descriptive statistics
  - 'First stage' statistics, summarizing the sample data
  - Descriptive statistics include proportions, means, variances, covariances, correlations
  - Used to characterize a population
- Inferential statistics
  - Statistics to generalize the information obtained form the measured sample to the population
- Analysis methods
  - Numerous methods to analyse the data
  - Choice of method depends on question asked as well as on the data available
- **Analyses tools**
  - **R**

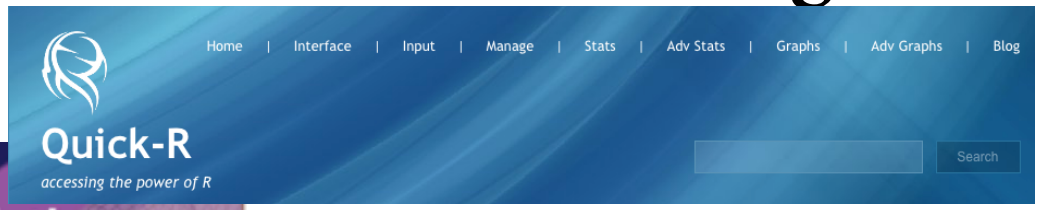
What is R?

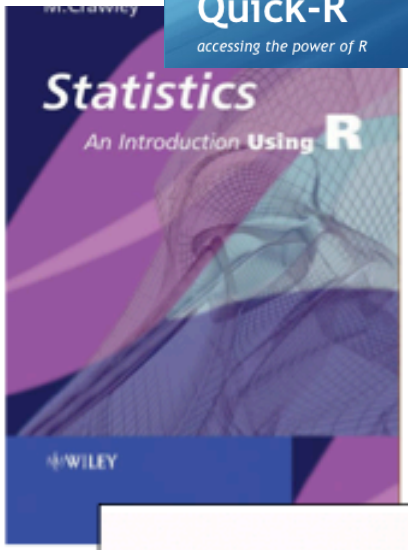
- Statistical programming environment with a command interface
- SAS, SPSS, STATA – box standard software with pre-defined algorithms
- Typical R session
  - R console
  - R script
  - R graphic window

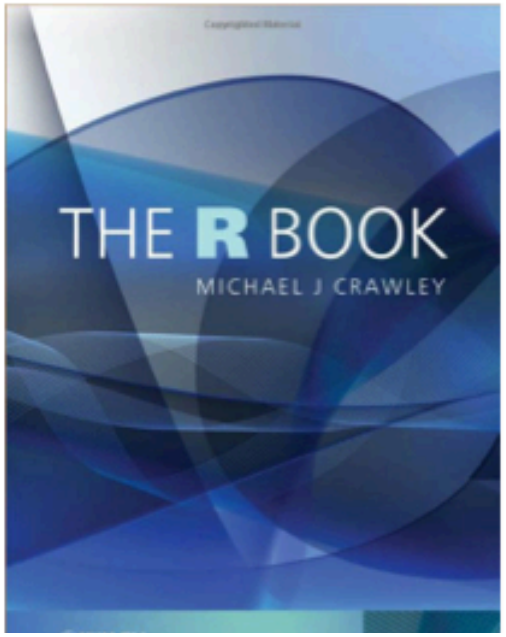
R Environment

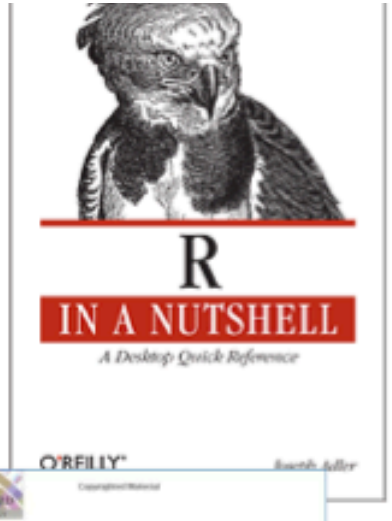


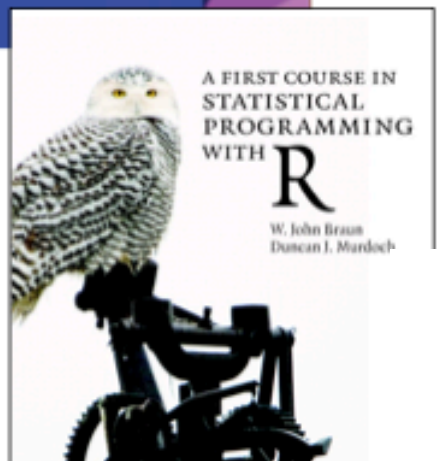
R Reading

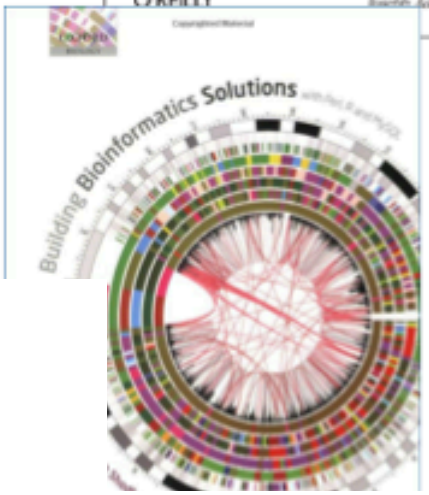

Home | Interface | Input | Manage | Stats | Adv Stats | Graphs | Adv Graphs | Blog
Quick-R
accessing the power of R
Search


Statistics
An Introduction Using R
WILEY


THE R BOOK
MICHAEL J CRAWLEY


R
IN A NUTSHELL
A Desktop Quick Reference
O'REILLY


A FIRST COURSE IN
STATISTICAL
PROGRAMMING
WITH R
W. John Braun
Duncan J. Murdoch


Building Bioinformatics Solutions


Google
Australia

Managing your R session?

```
> getwd()                    # Where does R save/retrieve files?
> dir()                      # What files are there?
> setwd()                    # Where would you like to save/retrieve files?
> ls()                       # What is in the current environment?
> a <- 1:10                  # Make an object
> installed.packages()       # R lists all installed packages
> help.start                 # Main help vignette
> ?t.test                    # Query a function
> install.packages("psych")  # Install package
> library(psych)             # Make package available
> help(package='psych')#Query whole package
> rm()                       # Remove unwanted objects
```

Reading In Data

- # helpful packages
  - "foreign"
  - "RODBC"

```
> data <- read.table()    # general import
> data <- read.delim()    # tab delimited file
> data <- read.csv()      # comma separated values
> data <- read.spss()     # SPSS
> data <- read.dta()      # STATA
```

- # Other useful functions

```
> dim()          # returns the number of rows and columns
> head()         # quick look at first 6 rows of the data
> tail()         # quick look at last 6 rows of the data
```

Managing data

- R has data structures: vectors, matrices, arrays, data frames
- Structures can be operated on through functions that perform statistical analyses and create graphs
- > head(d)

```
   ID aut.traits  IQ height weight gender
1 P.1         14 129 148.96  64.03   male
2 P.2         11 109 150.61  69.29   male
3 P.3         10 135 172.98  75.91   male
4 P.4         10  88 183.02  83.11 female
5 P.5          8 104 174.68  80.05   male
6 P.6         12  79 164.64  65.25   male
```

- > class(d)

```
[1] "data.frame"
```

Make new variable

```
> head(d)
```

	ID	aut.traits	IQ	height	weight	gender
1	P.1	14	129	148.96	64.03	male
2	P.2	11	109	150.61	69.29	male
3	P.3	10	135	172.98	75.91	male
4	P.4	10	88	183.02	83.11	female
5	P.5	8	104	174.68	80.05	male
6	P.6	12	79	164.64	65.25	male

- Make new variable

```
> d$bmi <- NA   # new variable has all missings
> d$bmi <- d$weight/(d$height/100)^2
> head(d)
```

	ID	aut.traits	IQ	height	weight	gender	asd	bmi
1	P.1	14	129	148.96	64.03	male	0	28.85653
2	P.2	11	109	150.61	69.29	male	0	30.54660
3	P.3	10	135	172.98	75.91	male	0	25.36923
4	P.4	10	88	183.02	83.11	female	0	24.81168
5	P.5	8	104	174.68	80.05	male	0	26.23463
6	P.6	12	79	164.64	65.25	male	0	24.07187

Subsetting

```
> head(d)
```

	ID	aut.traits	IQ	height	weight	gender
1	P.1	14	129	148.96	64.03	male
2	P.2	11	109	150.61	69.29	male
3	P.3	10	135	172.98	75.91	male
4	P.4	10	88	183.02	83.11	female
5	P.5	8	104	174.68	80.05	male
6	P.6	12	79	164.64	65.25	male

- Subset data.frame (columns)

```
> vars <- c("ID", "IQ", "gender")
> d.2 <- d[vars]
```

	ID	IQ	gender
1	P.1	129	male
2	P.2	109	male
3	P.3	135	male
4	P.4	88	female
5	P.5	104	male
6	P.6	79	male

- Subset data.frame (rows)

```
> d.3 <- d[which(d$gender=="female"),]
```

	ID	aut.traits	IQ	height	weight	gender
4	P.4	10	88	183.02	83.11	female
9	P.9	11	120	162.49	65.27	female
10	P.10	12	100	187.20	91.92	female
17	P.17	9	85	149.39	56.07	female
37	P.37	8	86	172.91	74.83	female
39	P.39	7	112	164.89	73.56	female

Graphics

?plot() # low level plotting function

Usage

plot(x, y, ...)

Arguments

x the coordinates of points in the plot. Alternatively, a single plotting structure, function or *any* ***R*** *object with a plot method* can be provided.

y the y coordinates of points in the plot, *optional* if x is an appropriate structure.

... Arguments to be passed to methods, such as graphical parameters (see par). Many methods will accept the following arguments:

Graphics

```
# Selection of useful parameters
```

`main` an overall title for the plot

`xlab` a title for the x axis

`ylab` a title for the y axis

`las` numeric in {0,1,2,3} parallel, horizontal, perpendicular, vertical

`cex` A numerical value giving the amount by which plotting text and symbols should be magnified relative to the default

`col` colour of the points that are being plotted

`lty` The line type. Line types can either be specified as an integer (0=blank, 1=solid (default), 2=dashed, 3=dotted, 4=dotdash, 5=longdash, 6=twodash) or as one of the character strings "blank", "solid", "dashed", "dotted", "dotdash", "longdash", or "twodash", where "blank" uses ‘invisible lines’ (i.e., does not draw them).

`lwd` The line width, a positive number, defaulting to 1. The interpretation is device-specific, and some devices do not implement line widths less than one. (See the help on the device for details of the interpretation.)

Outline Lecture

- **Descriptive statistics**
  - **'First stage' statistics, summarizing the sample data**
  - **Descriptive statistics include proportions, means, variances, covariances, correlations**
  - **Used to characterize a population**
- Inferential statistics
  - Statistics to generalize the information obtained form the measured sample to the population
- Analysis methods
  - Numerous methods to analyse the data
  - Choice of method depends on question asked as well as on the data available
- Analyses tools
  - R

Descriptive statistics

- Used to check whether the sample and variable characteristics are representative of the population
- Provide information about the distribution of the data
- Check whether statistical assumptions that are necessary to test hypotheses are met
- Explore data
  - differences between subgroups
  - outliers
  - trends
  - correlations

Scale of measurement

- Nominal
  - Gender (male, female)
  - Religion (Christianity, Islam, Baha'i, Hinduism)
  - Political affiliation (liberal, labor, republican)
- Ordinal
  - Educational level (primary, secondary, university)
  - Rank orders (1^{st}, 2^{nd}, 3^{rd})
  - Political orientation (left, centre, right)
- Interval
  - Temperature on Celsius scale (-3°C, 10°C, 25°C)
  - IQ scores (70, 80, 110, 145)
- Ratio
  - Ruler (inches or centimeters)
  - Years (of work experience)
  - Income (in $$ per year)

Descriptive statistics

- **Proportion/probability** $\hat{p} = \frac{X}{n}$
  > var.x/n
- **Mean** $\bar{x} = \sum \frac{x_i}{n}$
  > mean(var.x)
- **Variance** $var = \sigma^2 = \sum \frac{(x_i - \bar{x})^2}{n-1}$
  > var(var.x)
- **Standard deviation** $sd = \sigma = \sqrt{\sum \frac{(x_i - \bar{x})^2}{n-1}}$
  > sd(var.x)
- **Covariance** $cov_{x,y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{N-1}$
  > cov(var.x,var.y)
- **Correlation** $cor_{x,y} = \frac{cov_{x,y}}{\sigma_x \sigma_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(N-1)\sigma_x \sigma_y}$
  > cor(var.x,var.y)

Outline Lecture

- Descriptive statistics
  - 'First stage' statistics, summarizing the sample data
  - Descriptive statistics include proportions, means, variances, covariances, correlations
  - Used to characterize a population
- **Inferential statistics**
  - **Statistics to generalize the information obtained form the measured sample to the population**
- Analysis methods
  - Numerous methods to analyse the data
  - Choice of method depends on question asked as well as on the data available
- Analyses tools
  - R

Hypothesis testing

- Dealing with sample, not population data
- Aim to generalize the information obtained from the sample to the population
- H_0 = null hypothesis, H_1 = alternative hypothesis
- Statistical tests used to keep or reject the H_0
- Dealing with sample, so possibility of errors must be considered
- α is the maximum allowable probability of incorrectly rejecting the H_0 (false positive result)

Hypothesis testing

		Conclusion drawn	
		Accept H_0	Reject H_0
True state	H_0 True	CORRECT	Type I Error α
	H_0 False	Type II Error β	CORRECT $(1-\beta)$

Probability distributions

- Describe how the values of a random variable are distributed
  - Binomial distribution: E.g., collection of all possible outcomes of a sequence of coin tossing
  - Normal distribution: E.g., the means of sufficiently large samples of a population
- Characteristics of these theoretical distributions are well understood
- Can be used to make statistical inferences on the entire data population

Binomial distribution

- Discrete probability distribution that describes the outcome of *n* independent trials in an experiment.
- Each trial has 2 outcomes (e.g., heads or tails)

$$f(x) = \binom{n}{x} p^x (1-p)^{(n-x)}, where\ x = 0,1,2,\dots,n$$

- Example: exam with 12 multiple choice questions, each question 5 possible answers, only 1 is correct.
  - What is the probability of having exact four correct answers by chance?

  > dbinom(4,size=12,prob=0.2) A: 0.1329

  - What is the probability of having four or less correct answers by chance?

  > pbinom(4,size=12,prob=0.2) A. 0.93

Poisson distribution

- The probability distribution of independent event occurrences in an interval.

$$f(x) = \frac{\lambda^x e^{-x}}{x!}, where\ x = 0,1,2,3, \ldots.$$

- Example: There are 12 cars crossing a bridge per minute on average.
  - What is the probability of having 17 or more cars crossing the bridge in a particular minute?

```
> ppois(16, lambda=12)  # lower tail (default in R), A=0.90
> ppois(16,lambda=12, lower=FALSE), A=0.10
```

Normal distribution

- The probability density function of the normal distribution is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

- If a random variable follows a normal distribution: $X \sim N(\mu, \sigma^2)$
- Example: Test scores of an exam fit a normal distribution. Mean test score is 72 and standard deviation is 15.2.
  - What is the percentage of students scoring 84 or more?

> pnorm(84, mean=72, sd=15.2, lower.tail=FALSE). A=0.21

Sampling distributions

- Theoretical distribution of the expected values of a test statistic
- What we would observe if we repeatedly collected random samples form a population and computed the value of the statistic for each sample
- E.g., normal distribution, chi-squared distribution, binomial distribution

Sampling Distribution - step-by-step

- Consider normal distribution: $N(\mu, \sigma^2)$
- Repeatedly take samples and calculate $\bar{x}$ for each sample
- Calculated means follow a normal distribution
- Standard deviation of the sampling distribution of the mean is called the *standard error*
- Standard error of the mean: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Sampling Distribution - example

- Assume two populations with test scores
- $\mu_1 = 32$ and $\mu_2 = 22$
- $\sigma^2_1 = 60$ and $\sigma^2_2 = 70$
- $n_1 = 10$ and $n_2 = 14$
- **What is the probability that the mean of the sample of population 1 exceeds the mean of the sample of population 2 by 5 *or more*?**
- ***Sampling distribution of the difference between two means***

Sampling Distribution - example

- Compute means of two samples (M_1 and M_2)
- Compute difference between means (M_1 - M_2)
- Mean of sampling distribution of difference between means is:

$$\mu_{M_1-M_2} = \mu_1 - \mu_2$$

- Compute variance of the sampling distribution

$$\sigma^2{}_{M_1-M_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

  - Remember: variance of the sampling distribution of the mean is $\sigma^2{}_M = \frac{\sigma^2}{N}$

- Thus, standard error of the difference between means is

$$\sigma_{M_1-M_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Sampling Distribution - example

- $\mu_1 = 32$ and $\mu_2 = 22$
- $\sigma^2_1 = 60$ and $\sigma^2_2 = 70$
- $n_1 = 10$ and $n_2 = 14$

$$\mu_{M_1 - M_2} = \mu_1 - \mu_2 = 32 - 22 = 10$$

$$\sigma_{M_1 - M_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{60}{10} + \frac{70}{14}} = 3.317$$

- Thus, the sampling distribution is normally distributed with a mean of 10 and a standard deviation of 3.317

Sampling Distribution - example

- **What is the probability that the mean of the sample of population 1 exceeds the mean of the sample of population 2 by *5 or more*?**
- Using Z-scores and the Z-table, we can look up the probability:

$$Z = \frac{x-(\mu_{M_1-M_2})}{\sigma_{M_1-M_2}} = \frac{5-10}{3.317} = -1.507$$

- Shaded area is the probability that the mean of the sample of population 1 will exceed the mean of sample of population 2 *by 5 or more*
- Probability is 0.934

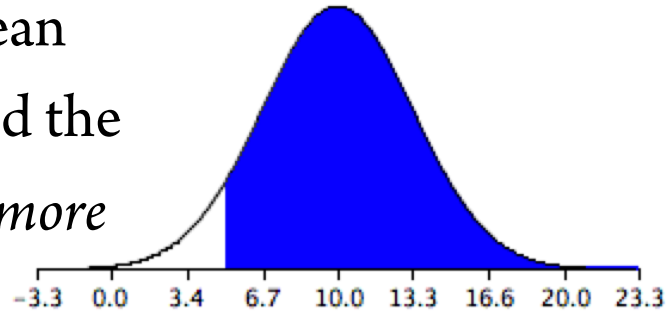


Figure 1. The sampling distribution of the difference between means.

> pnorm(5, mean=10, sd=3.317, lower.tail=FALSE). A=0.934

>1-pnorm(-1.507)

Outline Lecture

- Descriptive statistics
  - 'First stage' statistics, summarizing the sample data
  - Descriptive statistics include proportions, means, variances, covariances, correlations
  - Used to characterize a population
- Inferential statistics
  - Statistics to generalize the information obtained form the measured sample to the population
- **Analysis methods**
  - **Numerous methods to analyse the data**
  - **Choice of method depends on question asked as well as on the data available**
- Analyses tools
  - R

Difference in means

- t-test
  - $H_0 : \mu_1 = \mu_2$
  - $H_1 : \mu_1 \neq \mu_2$ or $H_1 : \mu_1 > \mu_2$

$$t = \frac{\bar{y}_1 - \bar{y}_2}{s\sqrt{1/n_1 + 1/n_2}}$$

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

```
> t.test (var.x, var.y)
```

Difference in means

```
> head(d)
   ID aut.traits  IQ height weight gender
1 P.1         14 129 148.96  64.03   male
2 P.2         11 109 150.61  69.29   male
3 P.3         10 135 172.98  75.91   male
4 P.4         10  88 183.02  83.11 female
5 P.5          8 104 174.68  80.05   male
6 P.6         12  79 164.64  65.25   male
> mean(d$aut.traits[which(d$gender=="male")],)
[1] 10.5688
> mean(d$aut.traits[which(d$gender=="female")],)
[1] 8.94086
> hist(d$aut.traits[which(d$gender=="male")],breaks=50)
> hist(d$aut.traits[which(d$gender=="female")],breaks=50)
> sd(d$aut.traits[which(d$gender=="male")],)
[1] 3.814204
> sd(d$aut.traits[which(d$gender=="female")],)
[1] 3.817704
```

Difference in means

```
> t.test (d$aut.traits[which(d$gender=="female")],d$aut.traits[which(d$gender=="male")])

        Welch Two Sample t-test

data:  d$aut.traits[which(d$gender == "female")] and d$aut.traits[which(d$gender == "male")]
t = -5.2478, df = 275.75, p-value = 3.075e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.238622 -1.017250
sample estimates:
mean of x mean of y
  8.94086  10.56880

> t.test (d$aut.traits[which(d$gender=="female")], d$aut.traits[which(d$gender=="male")], alternative="less")

        Welch Two Sample t-test

data:  d$aut.traits[which(d$gender == "female")] and d$aut.traits[which(d$gender == "male")]
t = -5.2478, df = 275.75, p-value = 1.537e-07
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -1.115961
sample estimates:
mean of x mean of y
  8.94086  10.56880
```

Simple regression

- Linear model in which we predict outcome variable y from explanatory variable x

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

- Using least squares estimation:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \text{ and } \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

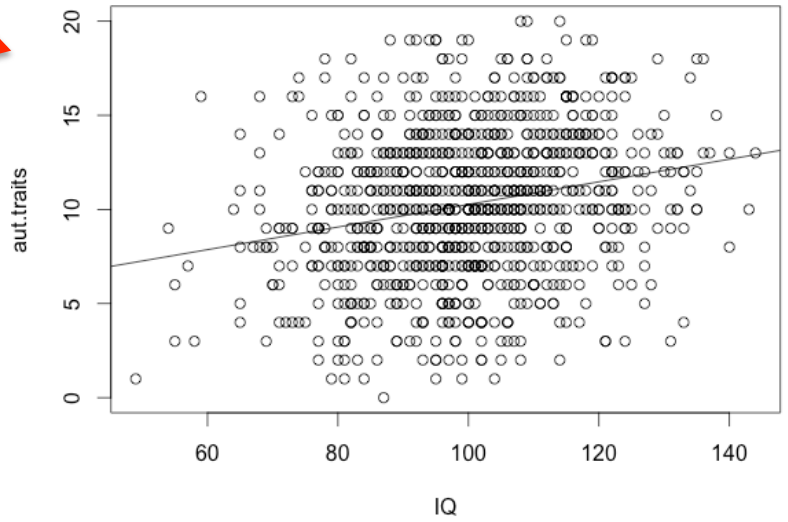
```
> head(d)
   ID aut.traits  IQ height weight gender
1 P.1         14 129 148.96  64.03   male
2 P.2         11 109 150.61  69.29   male
3 P.3         10 135 172.98  75.91   male
4 P.4         10  88 183.02  83.11 female
5 P.5          8 104 174.68  80.05   male
6 P.6         12  79 164.64  65.25   male
> m.1 <- lm(aut.traits~IQ, data=d) ## test whether IQ can predict aut.traits
> plot(aut.traits~IQ, data=d)
> abline(m.1) ## regression slope
> summary(m.1)

Call:
lm(formula = aut.traits ~ IQ, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5070 -2.4912  0.0124  2.6131  9.4540

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.260242   0.801140   5.318 1.30e-07 ***
IQ          0.060065   0.007924   7.580 7.85e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.76 on 998 degrees of freedom
Multiple R-squared:  0.05444,   Adjusted R-squared:  0.0535
F-statistic: 57.46 on 1 and 998 DF,  p-value: 7.852e-14
```



Multiple regression

- Linear model in which we predict outcome variable y from q explanatory variables $x_1 \ldots x_q$

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_q x_{iq} + \varepsilon_i$$

To test H_0: $\beta_1 = \cdots = \beta_q = 0$, we use the Mean Square Ratio:

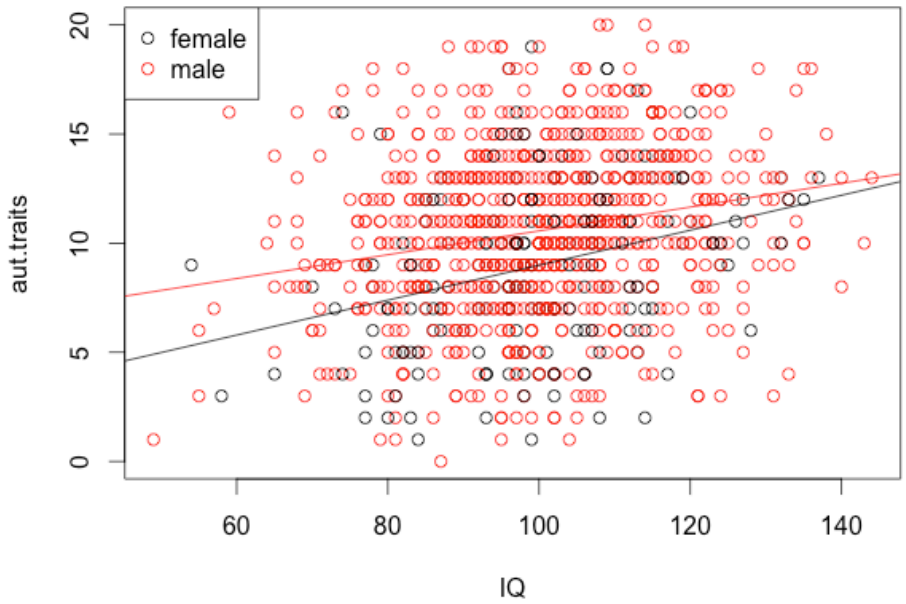
$$F = \frac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 / q}{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 / n - q - 1}$$

- Individual regression coefficients can be *assessed* using the t-statistics (est/std.error)

Exploring the data

```
> head(d)
   ID aut.traits  IQ height weight gender
1 P.1         14 129 148.96  64.03   male
2 P.2         11 109 150.61  69.29   male
3 P.3         10 135 172.98  75.91   male
4 P.4         10  88 183.02  83.11 female
5 P.5          8 104 174.68  80.05   male
6 P.6         12  79 164.64  65.25   male
```

```
> library(car)
> plot(aut.traits~ IQ, pch=1, col=gender, data=d)
> abline(lm(aut.traits~IQ, data=d,
+           subset=gender=="female"), col="black")
> abline(lm(aut.traits~IQ, data=d,
+           subset=gender=="male"),col="red")
> legend("topleft",legend=c("female","male"),
+        pch=1:1,col=c("black","red"))
```



Fitting the model

```
> head(d)
   ID aut.traits  IQ height weight gender
1 P.1         14 129 148.96  64.03   male
2 P.2         11 109 150.61  69.29   male
3 P.3         10 135 172.98  75.91   male
4 P.4         10  88 183.02  83.11 female
5 P.5          8 104 174.68  80.05   male
6 P.6         12  79 164.64  65.25   male
> m.1 <- lm(aut.traits~IQ+gender+IQ:gender, data=d) ## test whether IQ can predict aut.traits
> summary(m.1) ## F-statistic can be calculated from the ANOVA table,

Call:
lm(formula = aut.traits ~ IQ + gender + IQ:gender, data = d)

Residuals:
    Min      1Q  Median      3Q     Max
-9.8525 -2.4253  0.0112  2.5115 10.0863

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.98492    1.83455   0.537   0.5915
IQ             0.08009    0.01826   4.385 1.28e-05 ***
gendermale     4.12312    2.03306   2.028   0.0428 *
IQ:gendermale -0.02556    0.02021  -1.264   0.2063
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.71 on 996 degrees of freedom
Multiple R-squared:  0.08126,   Adjusted R-squared:  0.07849
F-statistic: 29.36 on 3 and 996 DF,  p-value: < 2.2e-16
```

```
> anova(m.1) ## here, the ((812.59 +378.17+ 22.01)/13.77 )/3
Analysis of Variance Table

Response: aut.traits
           Df  Sum Sq Mean Sq F value    Pr(>F)
IQ          1   812.6  812.59  59.022 3.721e-14 ***
gender      1   378.2  378.17  27.468 1.949e-07 ***
IQ:gender   1    22.0   22.01   1.599    0.2063
Residuals 996 13712.5   13.77
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Logistic regression

- Generalized Linear Model (GLM) in which we predict outcome variable y (binary) from q explanatory variables $x_1 \ldots .. x_q$
- Using *logit* link function:

$$logit(p) = \log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q$$

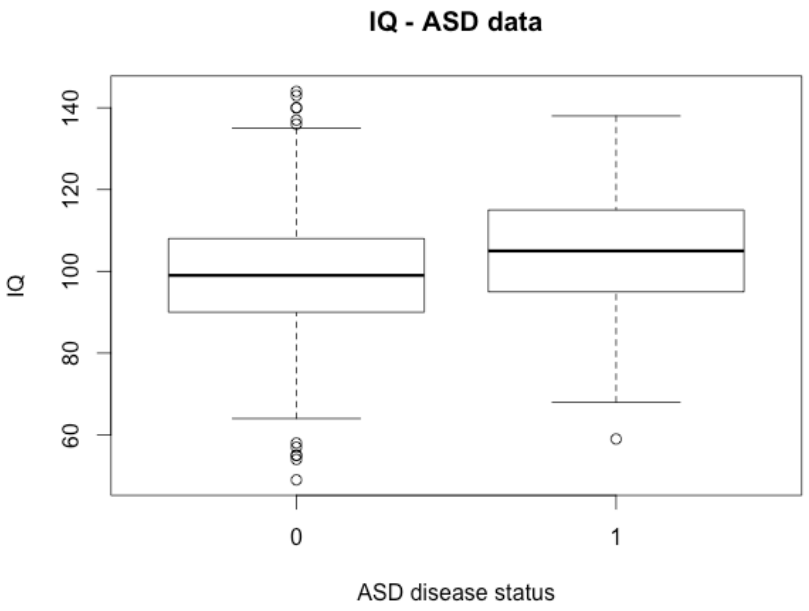
- The logit of a probability is the log of the odds of the response taking the value one

$$p(x_1, x_2, \ldots, x_q) = \frac{\exp\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q\right)}{1 + \exp\left(\beta_0 + \beta_1 x_1 + \cdots + \beta_q x_q\right)}$$

Creating disease variable

```
> #### general linear model (logistic regression)
> head(d)
   ID aut.traits  IQ height weight gender
1 P.1         14 129 148.96  64.03   male
2 P.2         11 109 150.61  69.29   male
3 P.3         10 135 172.98  75.91   male
4 P.4         10  88 183.02  83.11 female
5 P.5          8 104 174.68  80.05   male
6 P.6         12  79 164.64  65.25   male
> d$asd <- NA ## make new variable with autistic traits recoded as disease
> summary(d$aut.traits)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    8.00   10.00   10.27   13.00   20.00
> d$asd [which(d$aut.traits>14)] <- 1
> d$asd [which(d$aut.traits<=14)] <- 0
> d$asd <- as.factor(d$asd)
> table(d$asd)

  0   1
862 138
> boxplot(IQ~asd,data=d, main="IQ - ASD data",
+         xlab="ASD disease status", ylab="IQ")
```



```
> m.f <- glm(asd~ IQ+gender, data=d,
+ family=binomial())
> summary(m.f) ## increase of 1 IQ unit increases the log-odds of asd by an estimated 0.02

Call:
glm(formula = asd ~ IQ + gender, family = binomial(), data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8381  -0.5873  -0.5203  -0.4051   2.4908

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -4.544280   0.701248  -6.480 9.16e-11 ***
IQ           0.020113   0.006242   3.222  0.00127 **
gendermale   0.782504   0.295573   2.647  0.00811 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 802.63  on 999  degrees of freedom
Residual deviance: 783.51  on 997  degrees of freedom
AIC: 789.51

Number of Fisher Scoring iterations: 5
```

Increase of 1 IQ point increases the log-odds of ASD by an estimated 0.02

easier interpretation: convert to odds-ratio

```
> exp(coef(m.f)["IQ"])
      IQ
1.020316
> exp(confint(m.f,parm="IQ"))
Waiting for profiling to be done...
   2.5 %   97.5 %
1.007983 1.032977
```

Comparing full and reduced model using ANOVA

```
> m.f <- glm(asd~ IQ+gender, data=d,
+ family=binomial())
```

```
> ## fitting a reduced model to test for significance of IQ
> m.r <- glm(asd~ gender, data=d,
+ family=binomial())
> anova(m.r, m.f, test="Chisq")
Analysis of Deviance Table

Model 1: asd ~ gender
Model 2: asd ~ IQ + gender
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       998     794.07
2       997     783.51  1   10.565 0.001152 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```