introduction to stats in R and Rbrul

PRIFYSGOL
BANGOR
UNIVERSITY

10:00 – 11:10		Basic descriptive and inferential statistics
	short break	
11:20 – 12:30		Intro to R: graphics and model-building
	lunch break	
13:30 – 14:40		Rbrul: a front end for regression analysis
	short break	
14:50 – 15:00		Mixed-effects models: why and how?

Daniel Ezra Johnson, Lancaster University

danielezrajohnson@gmail.com

www.danielezrajohnson.com/bangor_workshop.pdf (or .pptx)

what are statistics?

- turn a large amount of observations (data) into a smaller amount of numbers
- use complex data to answer simpler questions
- descriptive statistics
  - wh-questions
  - answers in numbers
- inferential statistics
  - yes-no questions
  - sample -> population

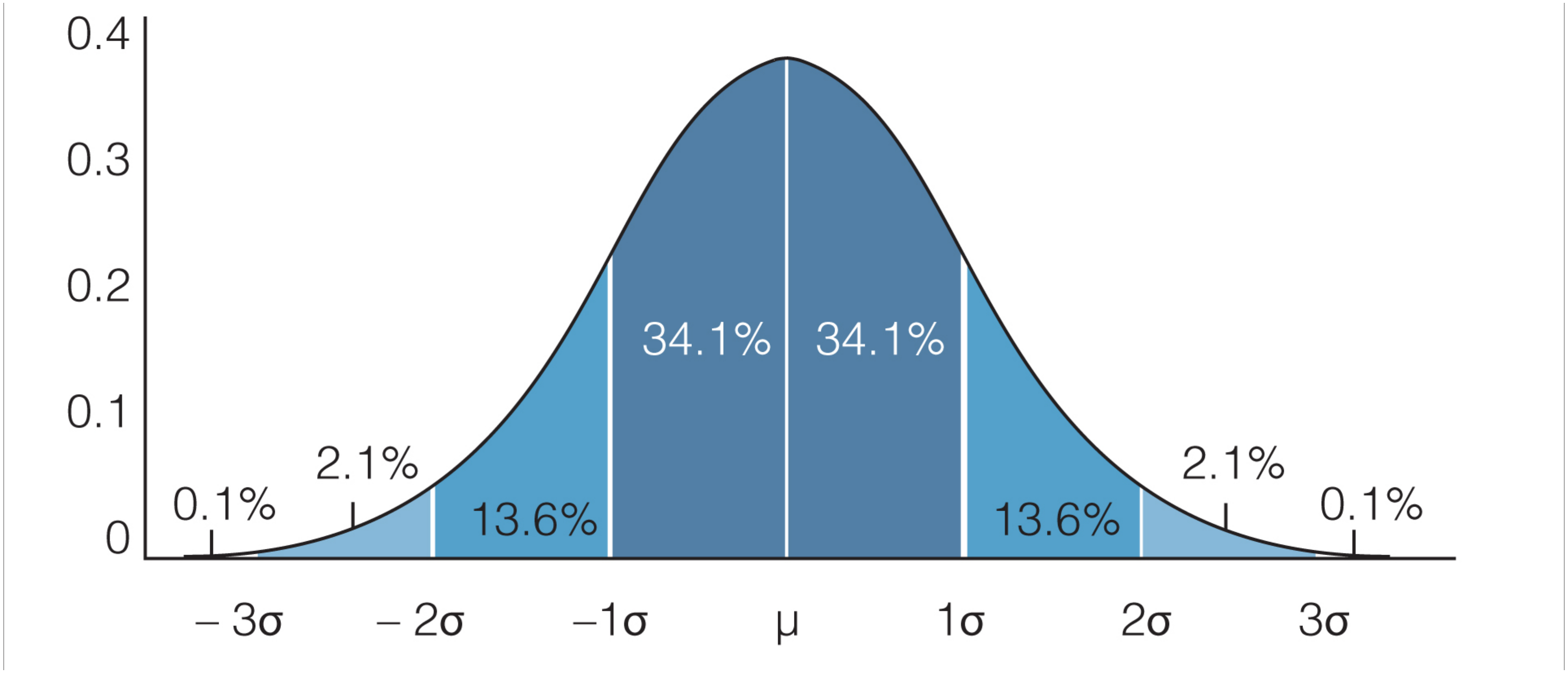
```
   what is the pattern of
  bilingual clauses among
    the speakers in the
       Siarad corpus?

     do women use more
     bilingual clauses
         than men?
```

descriptive statistics – one variable

- data types
  - nominal, ordinal
  - interval, ratio
  - categorical, continuous
- distributions
  - normal, skewed
- central tendency
  - mean, median, mode
- dispersion
  - standard deviation

```
nominal: unordered
         categories

ordinal: ordered categories

interval/ratio: numeric

normal: bell curve
skewed: one tail is longer

mean: sum / # of items
median: middle item
mode: most frequent item

standard deviation:
(approximately) average
distance to mean
```

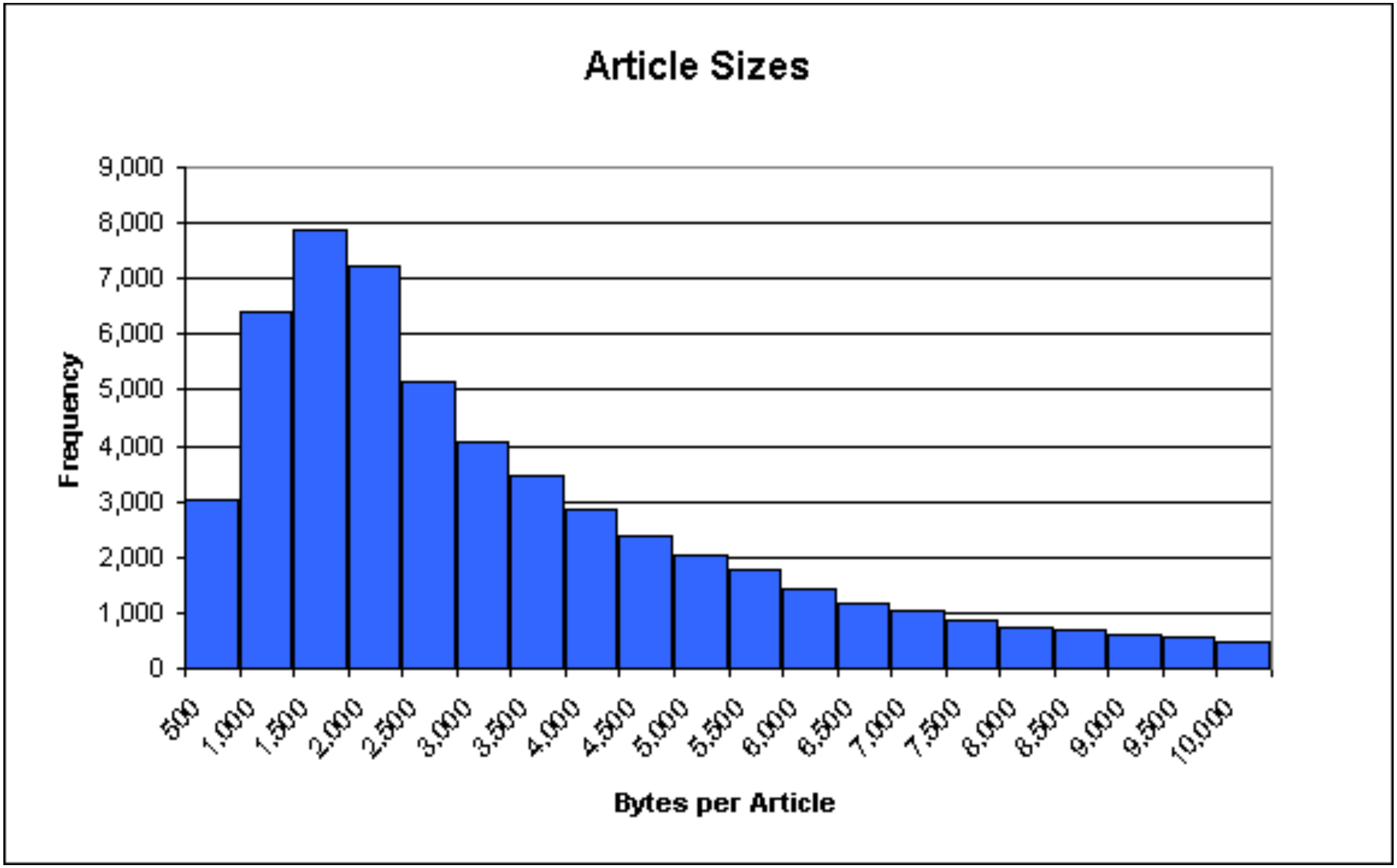
normal distribution: “bell curve”



-3 s.d. -2 s.d. -1 s.d. mean +1 s.d. +2 s.d. +3 s.d.

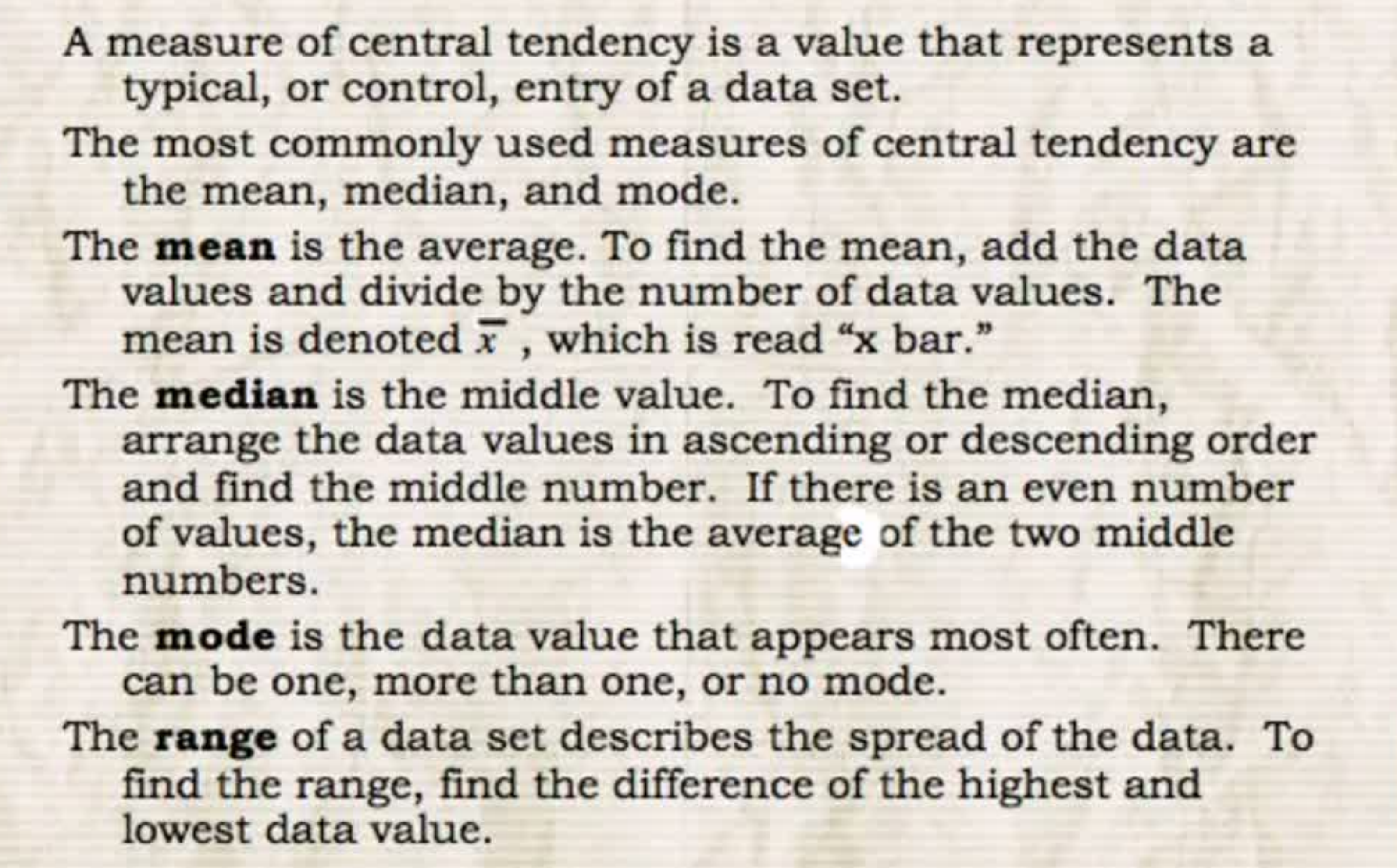
|--------- 95% ---------|

one-variable statistics: histogram



central tendency

mean, median, mode; range



A measure of central tendency is a value that represents a typical, or control, entry of a data set.

The most commonly used measures of central tendency are the mean, median, and mode.

The **mean** is the average. To find the mean, add the data values and divide by the number of data values. The mean is denoted $\overline{x}$, which is read "x bar."

The **median** is the middle value. To find the median, arrange the data values in ascending or descending order and find the middle number. If there is an even number of values, the median is the average of the two middle numbers.

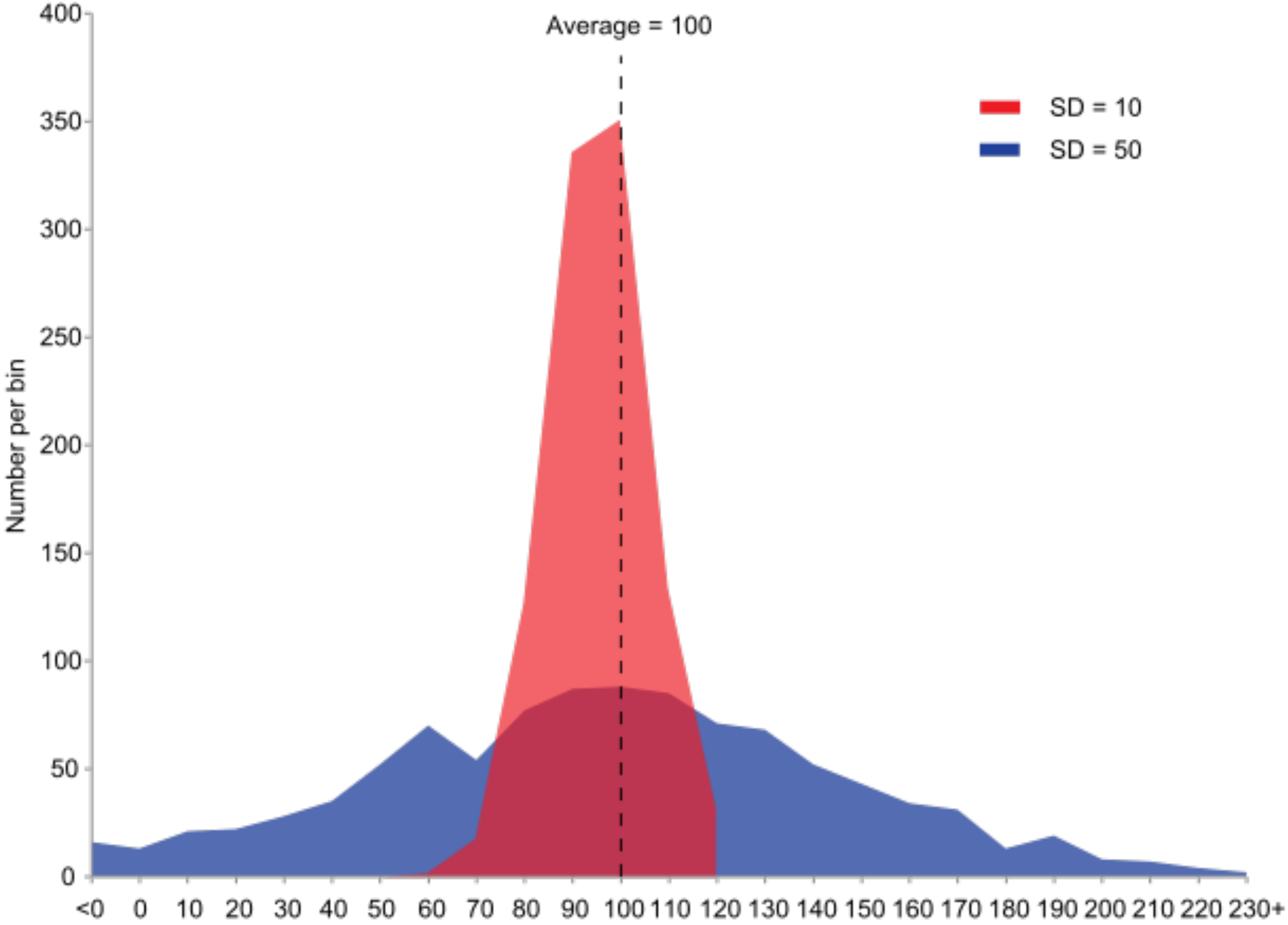
The **mode** is the data value that appears most often. There can be one, more than one, or no mode.

The **range** of a data set describes the spread of the data. To find the range, find the difference of the highest and lowest data value.

DATA:
10
8
4
3
3
3
2
2
1
sum=36
N = 9

mean =   median =   mode =   range =

dispersion: standard deviation



descriptive statistics – 2+ variables

- association
- correlation
- regression
  - linear regression:
    $y = a + b*x\ldots$
  - logistic regression:
    $\text{log-odds}(p) = a + b*x\ldots$
    $\ln(p/(1-p)) = a + b*x\ldots$
  - multiple regression
  - multivariate regression

association: lack of
independence between
variables (one helps
predict the other)

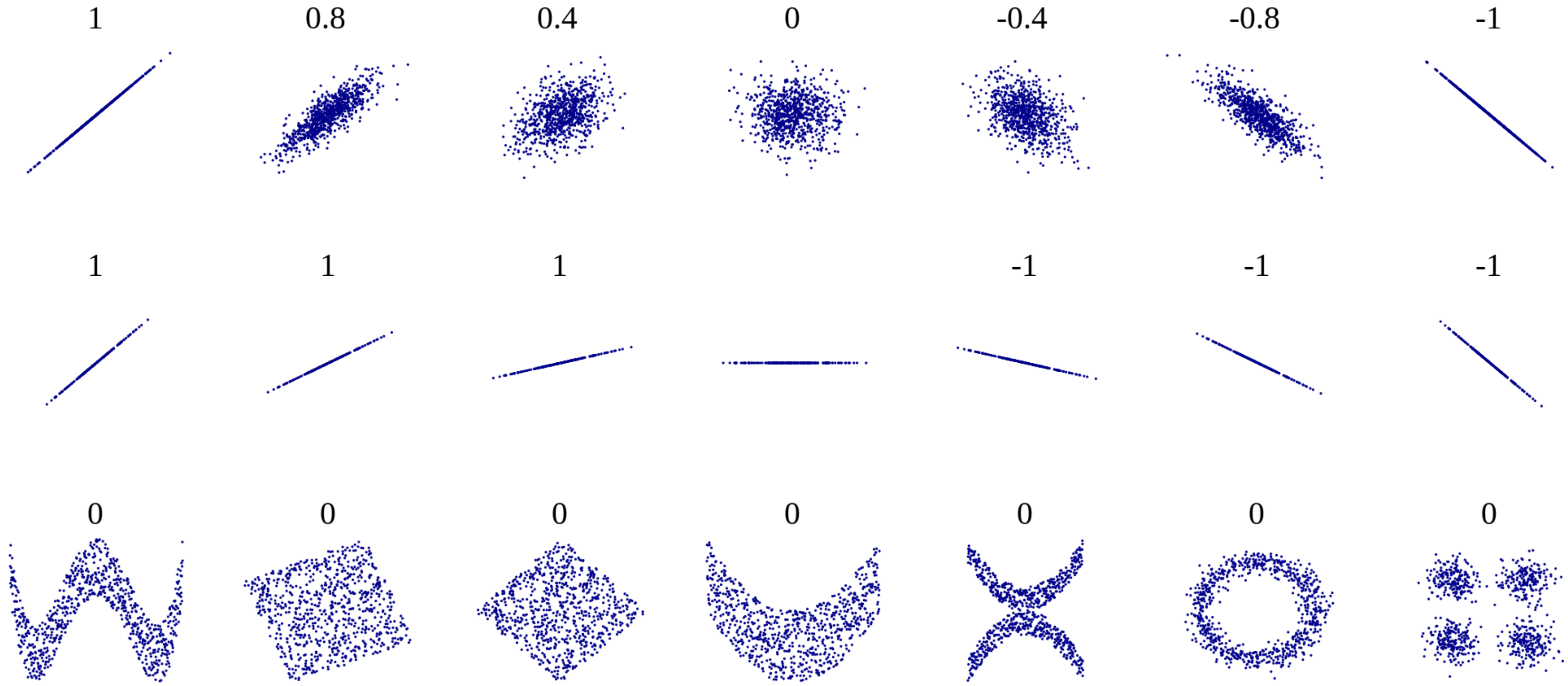
correlation: from -1 to 1
how "tight" association is
not how "strong" effect is

linear: predicts a value
logistic: a probability

multiple: > 1 indep. var.

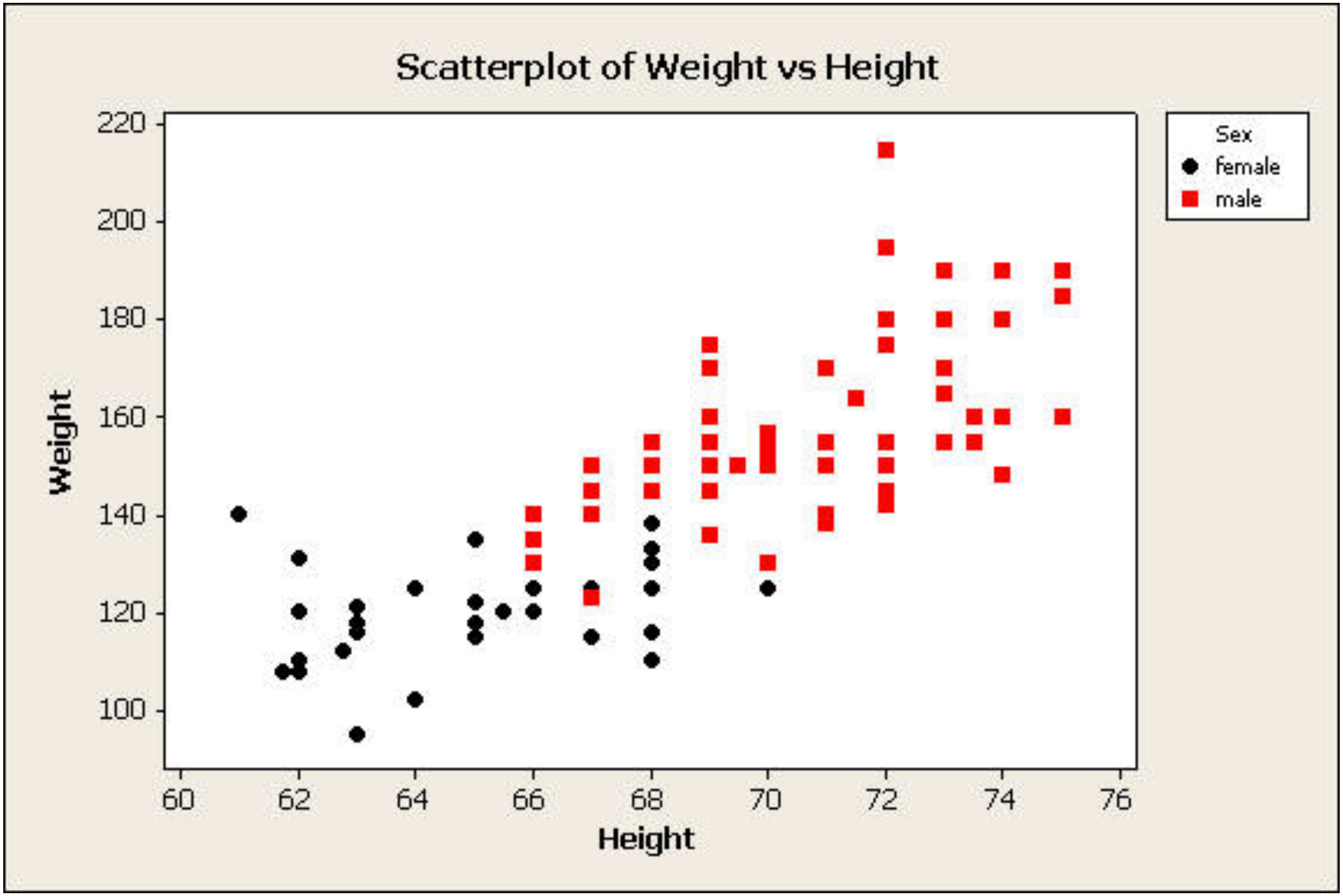
multivariate: > 1 dep. var.

two variables: correlation



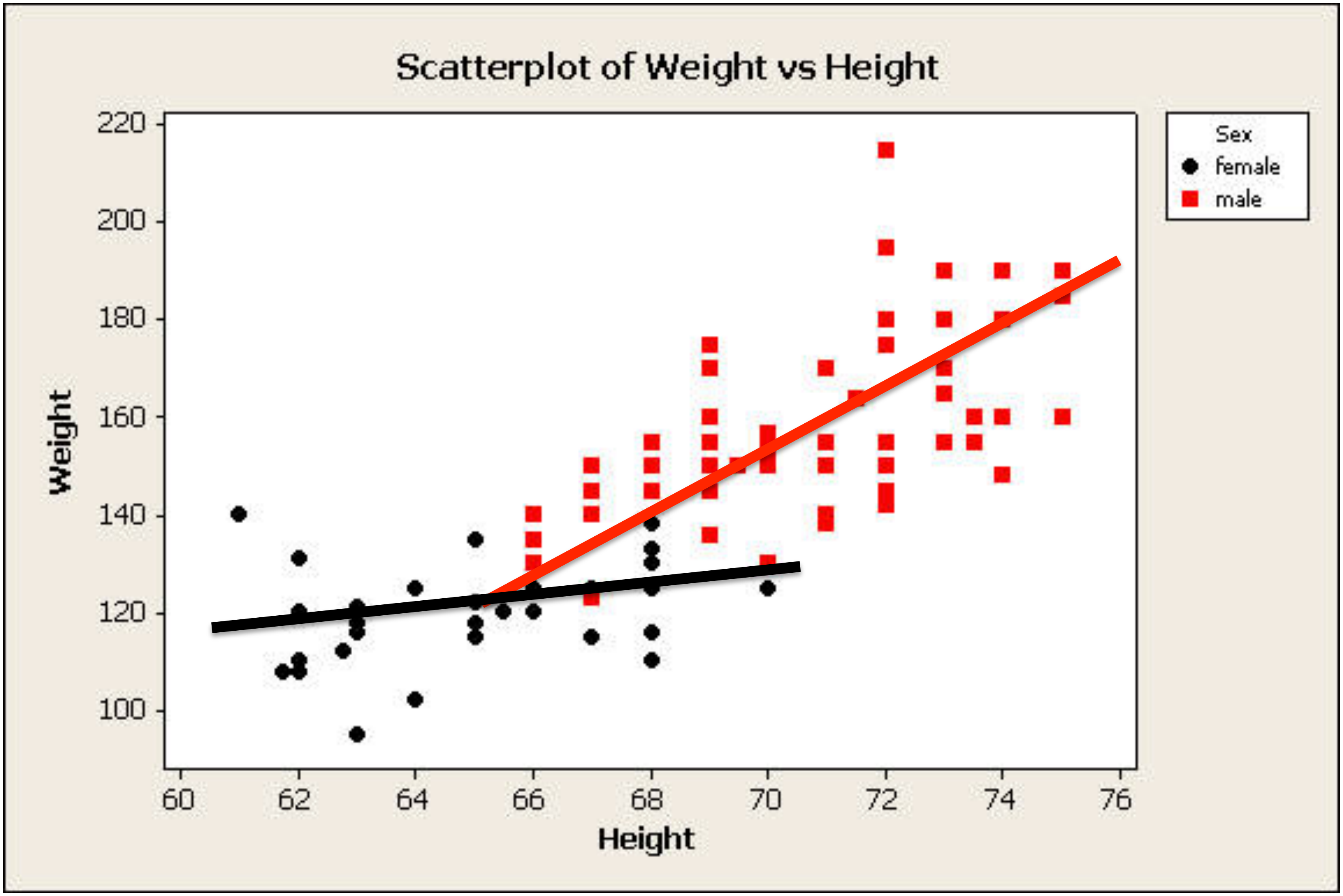
two (or more) variables*: scatterplot

*dependent and independent

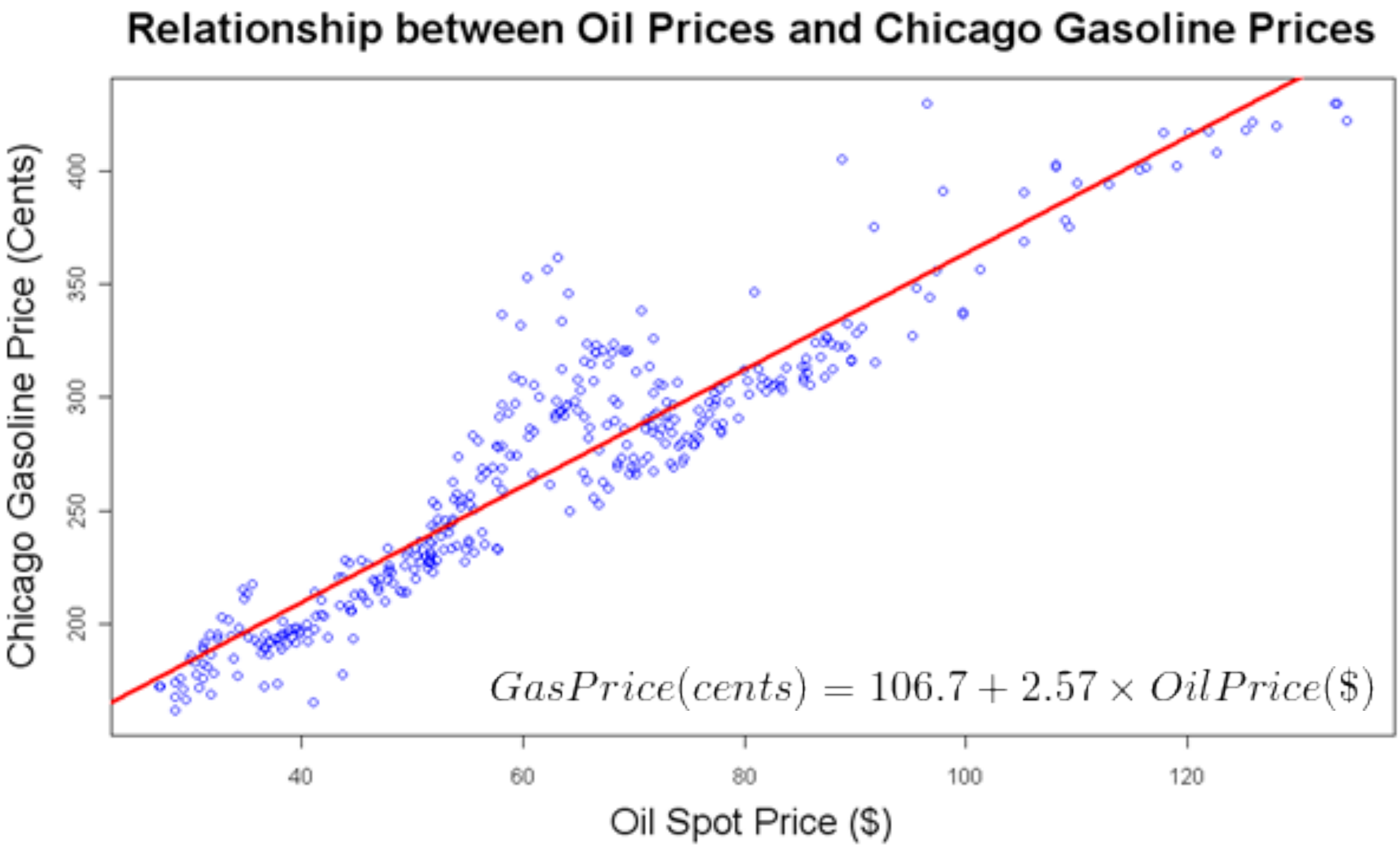


two (or more) variables*: scatterplot

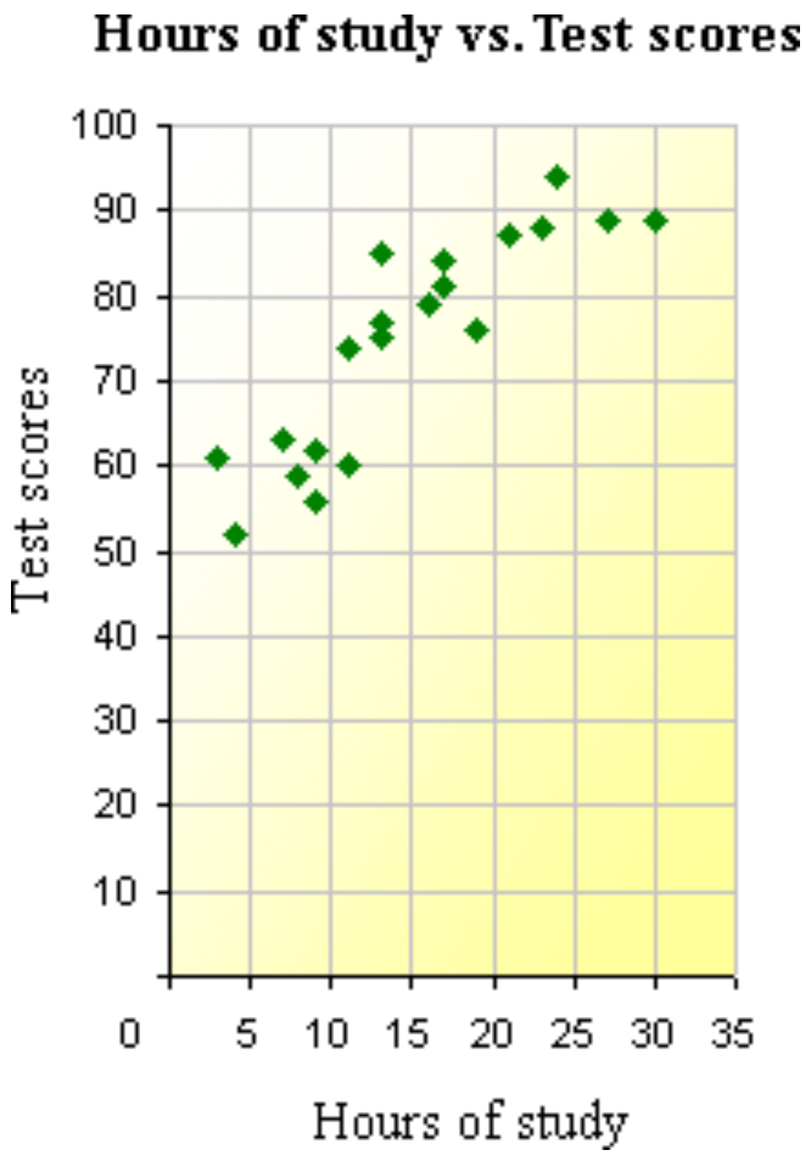
*dependent and independent



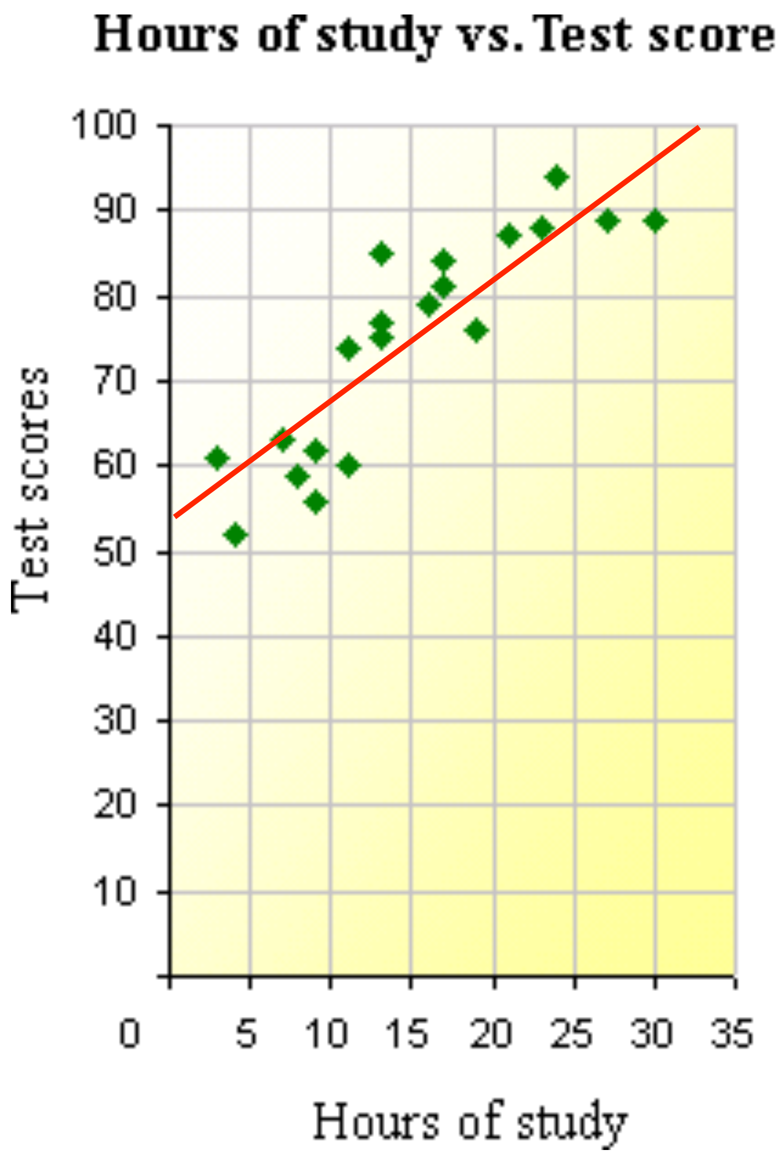
(linear) regression



does a continuous IV affect the DV?
how strong is the association between IV and DV?



does a continuous IV affect the DV?
how strong is the association between IV and DV?



linear regression

Test score = 55 + 1.2 * Hours of Study

$R^2 = 0.81$

inferential statistics

- yes-no questions
- drawing conclusions
- making predictions
- frequentist vs. Bayesian
- (hypothesis) <u>tests</u>
- null hypothesis
- p-values
- alpha
- significance

descriptive stats describe a sample (or population)

inferential stats use sample to make inferences about pop.

frequentist hypothesis testing is not very fashionable

null hypothesis is often:
no difference between groups
no effect of x on y (in pop.)

p-value is chance of observing sample effect if no pop. diff.

p < .05 is totally arbitrary but very well-entrenched significance threshold

when do we need a significance test?

when do we need a significance test?

when do we need a significance test?

difference in means / std. dev. of population = t

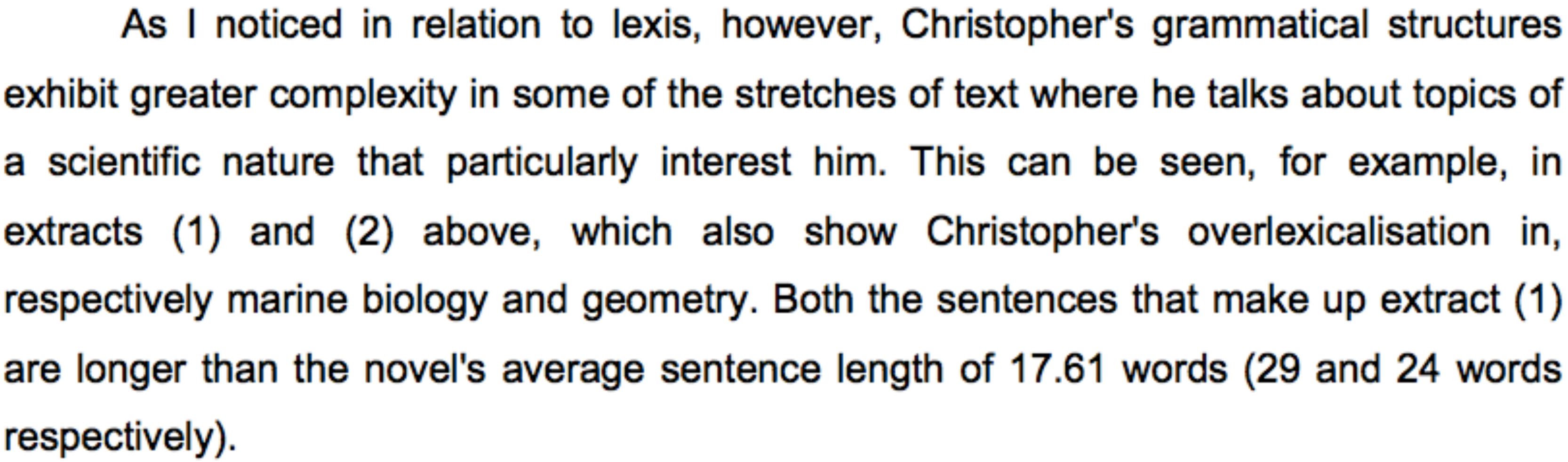
t-test: large t = significant difference

when do we need a significance test?

difference in means / std. dev. of pop. = t

t-test: small t = non-significant difference

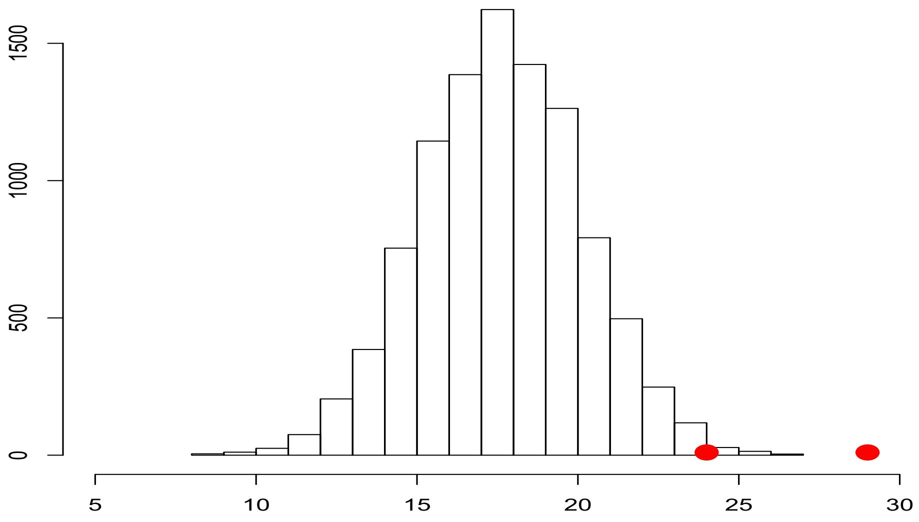
when is a difference significant?



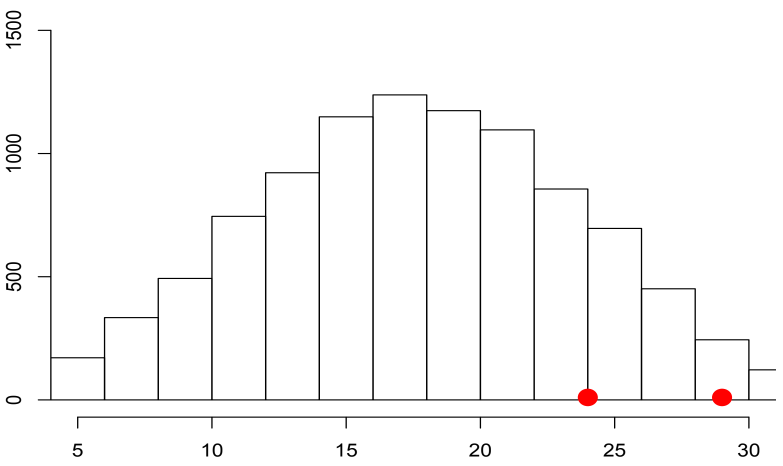
As I noticed in relation to lexis, however, Christopher's grammatical structures exhibit greater complexity in some of the stretches of text where he talks about topics of a scientific nature that particularly interest him. This can be seen, for example, in extracts (1) and (2) above, which also show Christopher's overlexicalisation in, respectively marine biology and geometry. Both the sentences that make up extract (1) are longer than the novel's average sentence length of 17.61 words (29 and 24 words respectively).

- the novel's mean (average) sentence length is 17.61 words
- two sentences about marine biology are 24 and 29 words
- is this a significant difference?
- we don't know! we don't know the standard deviation…

when is a difference significant?



histogram of 10,000 sentences: mean 17.61, std. dev. 2.5



histogram of 10,000 sentences: mean 17.61, std. dev. 6.5

- the novel's mean (average) sentence length is 17.61 words
- two sentences about marine biology are 24 and 29 words
- is this a significant difference? perform test, is p < .05?
- on the left: standard deviation 2.5 p < .001 (significant)
- on the right: standard deviation 6.5 p = .06 (n.s.)

recommended reading

- my chapter on descriptive statistics:
  - http://www.danielezrajohnson.com/johnson_descriptive_stats.pdf

- a good chapter about regression basics:
  - http://people.stern.nyu.edu/wgreene/Statistics/MultipleRegressionBasicsCollection.pdf

- websites for statistical computation
  - http://www.vassarstats.net (and others)

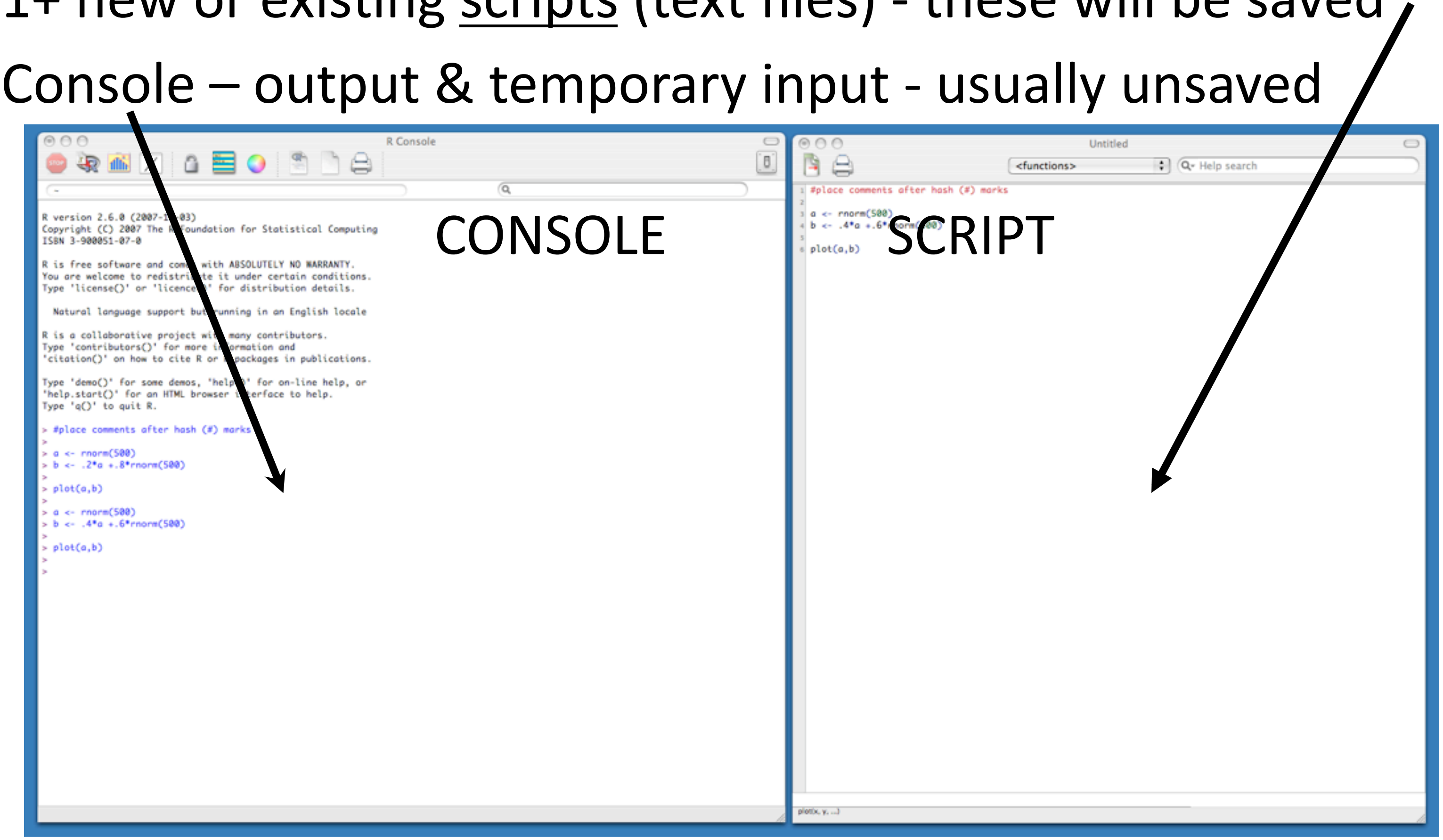
what is R?

- a free programming language for statistics
  - open-source
  - user-contributed packages
- basic operation
  - console window
    - (input and) output
  - script window
    - save commands
    - copy into console

```
people tend to love or hate

can be very fiddly to use

start R now

you will see the console

> is the prompt where you
can enter all commands

a much better way is to
work in a script window

store sequences of commands
there and run as needed
```

typical R session

- Start up R via the GUI or favorite text editor
- Two windows:
  - 1+ new or existing <u>scripts</u> (text files) - these will be saved
  - Console – output & temporary input - usually unsaved



CONSOLE

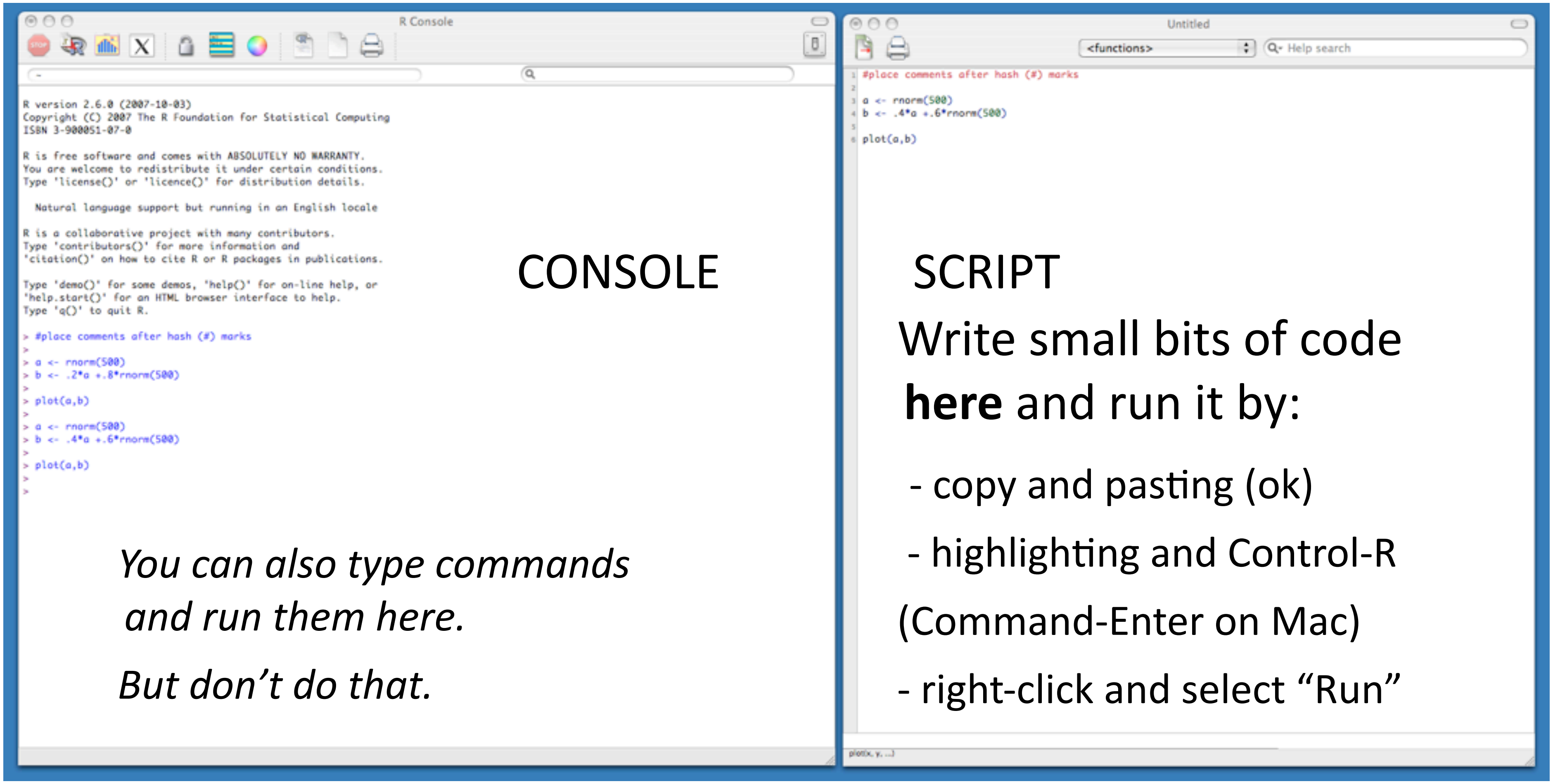
```
R version 2.6.0 (2007-10-03)
Copyright (C) 2007 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> #place comments after hash (#) marks
>
> a <- rnorm(500)
> b <- .2*a +.8*rnorm(500)
>
> plot(a,b)
>
> a <- rnorm(500)
> b <- .4*a +.6*rnorm(500)
>
> plot(a,b)
>
>
```

SCRIPT

```
#place comments after hash (#) marks

a <- rnorm(500)
b <- .4*a +.6*rnorm(500)

plot(a,b)
```

typical R session

- R sessions are *interactive*



```
R version 2.6.0 (2007-10-03)
Copyright (C) 2007 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> #place comments after hash (#) marks
>
> a <- rnorm(500)
> b <- .2*a +.8*rnorm(500)
>
> plot(a,b)
>
> a <- rnorm(500)
> b <- .4*a +.6*rnorm(500)
>
> plot(a,b)
>
>
```

CONSOLE

You can also type commands and run them here.

But don't do that.

```
#place comments after hash (#) marks

a <- rnorm(500)
b <- .4*a +.6*rnorm(500)

plot(a,b)
```

SCRIPT

Write small bits of code **here** and run it by:

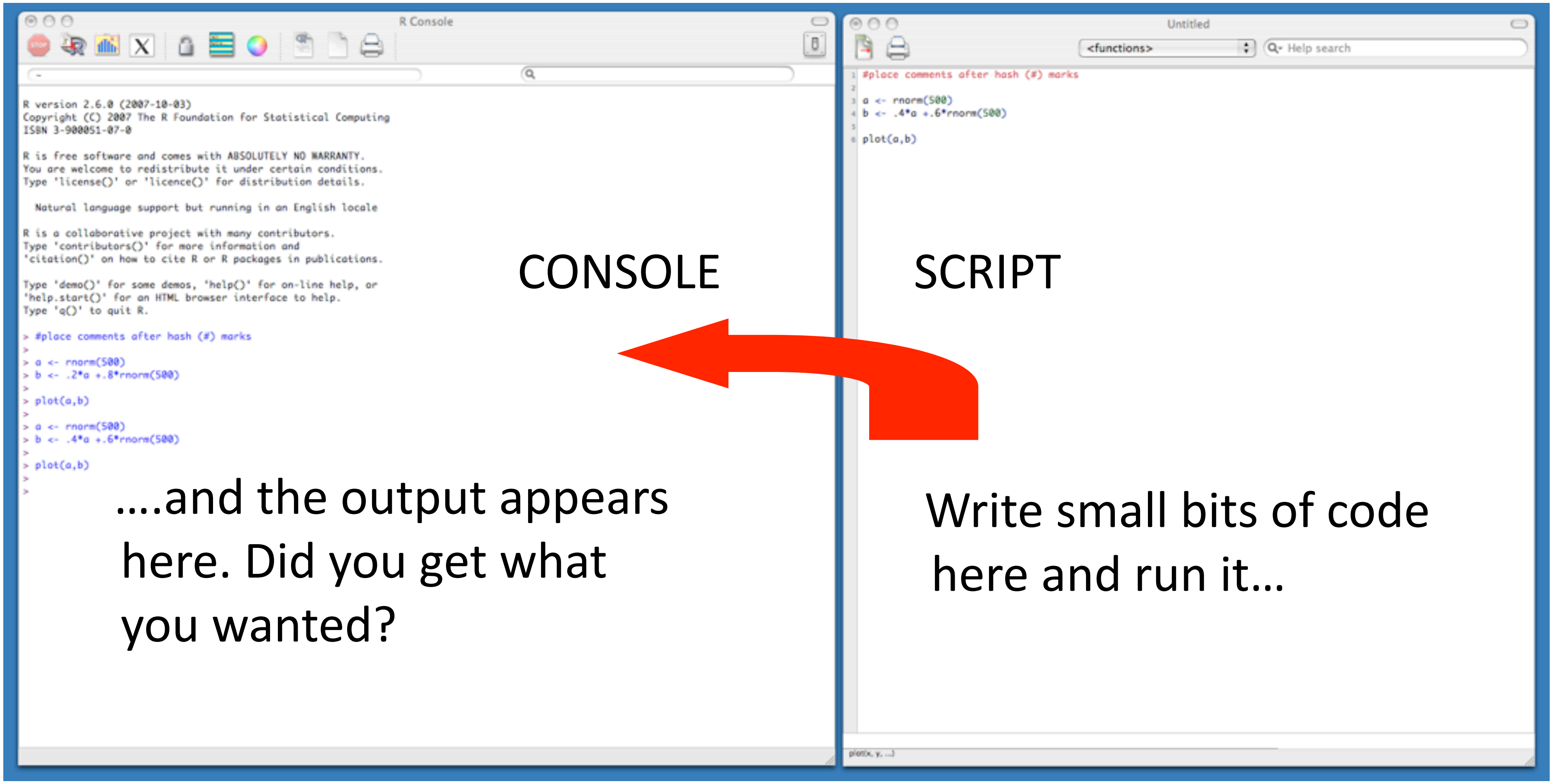
- copy and pasting (ok)

- highlighting and Control-R (Command-Enter on Mac)

- right-click and select "Run"

typical R session

- R sessions are *interactive*



```
R version 2.6.0 (2007-10-03)
Copyright (C) 2007 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> #place comments after hash (#) marks
>
> a <- rnorm(500)
> b <- .2*a +.8*rnorm(500)
>
> plot(a,b)
>
> a <- rnorm(500)
> b <- .4*a +.6*rnorm(500)
>
> plot(a,b)
>
>
```

```
#place comments after hash (#) marks

a <- rnorm(500)
b <- .4*a +.6*rnorm(500)

plot(a,b)
```

CONSOLE

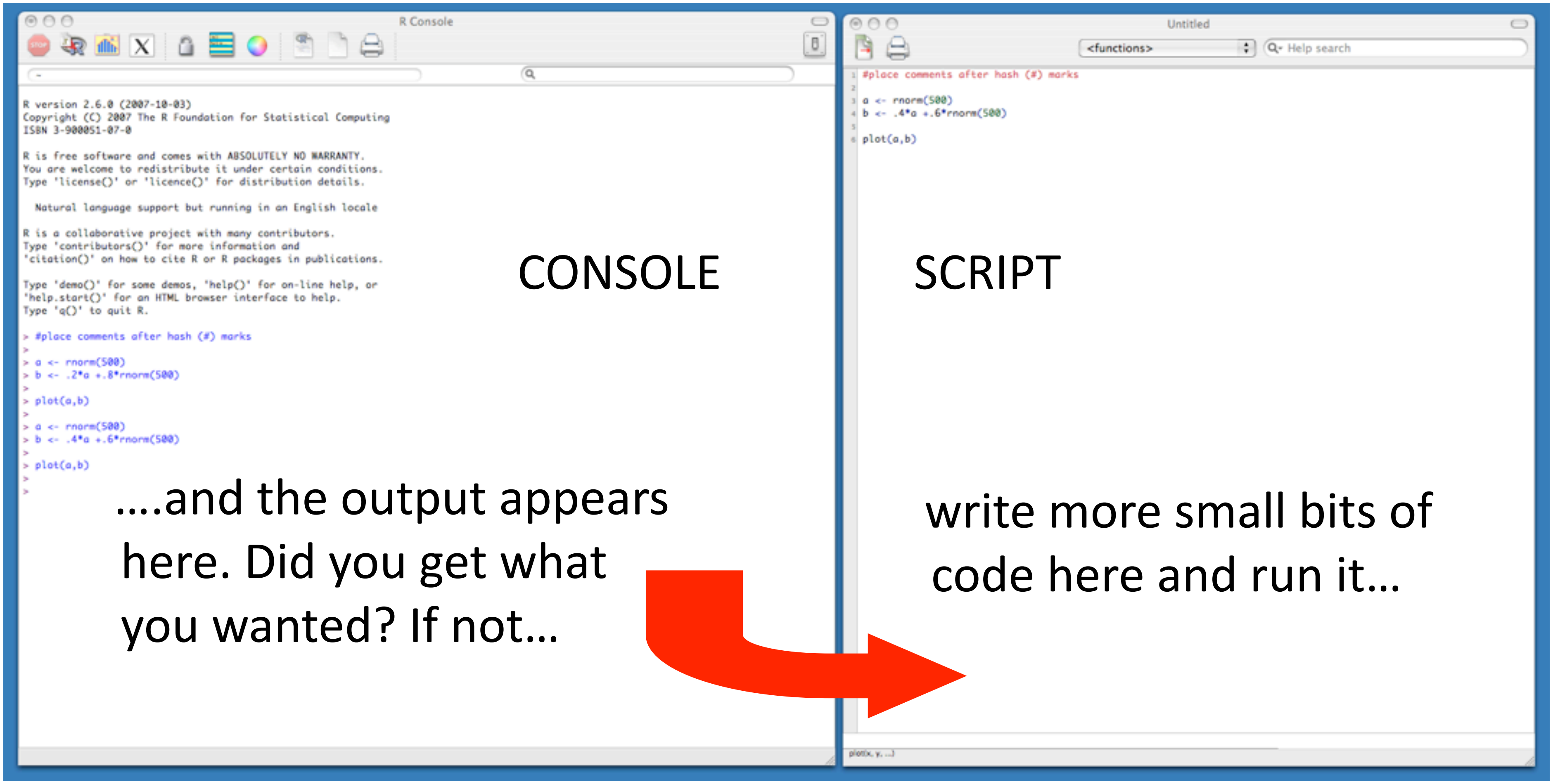
SCRIPT

Write small bits of code here and run it…

….and the output appears here. Did you get what you wanted?

typical R session

- R sessions are *interactive*



```
R version 2.6.0 (2007-10-03)
Copyright (C) 2007 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> #place comments after hash (#) marks
>
> a <- rnorm(500)
> b <- .2*a +.8*rnorm(500)
>
> plot(a,b)
>
> a <- rnorm(500)
> b <- .4*a +.6*rnorm(500)
>
> plot(a,b)
>
>
```

```
#place comments after hash (#) marks

a <- rnorm(500)
b <- .4*a +.6*rnorm(500)

plot(a,b)
```

CONSOLE

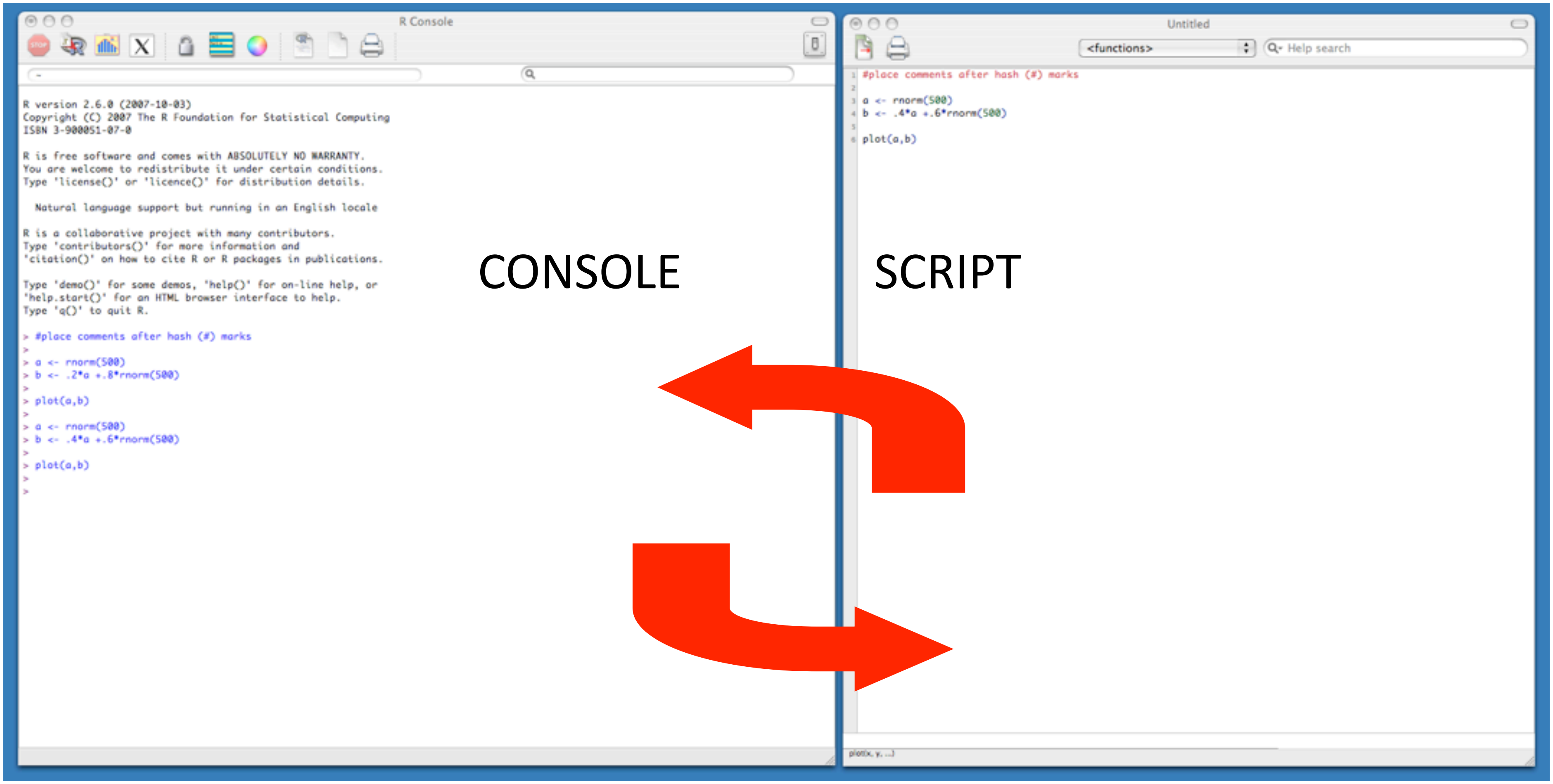
SCRIPT

write more small bits of code here and run it…

….and the output appears here. Did you get what you wanted? If not…

typical R session

- R sessions are *interactive*



CONSOLE

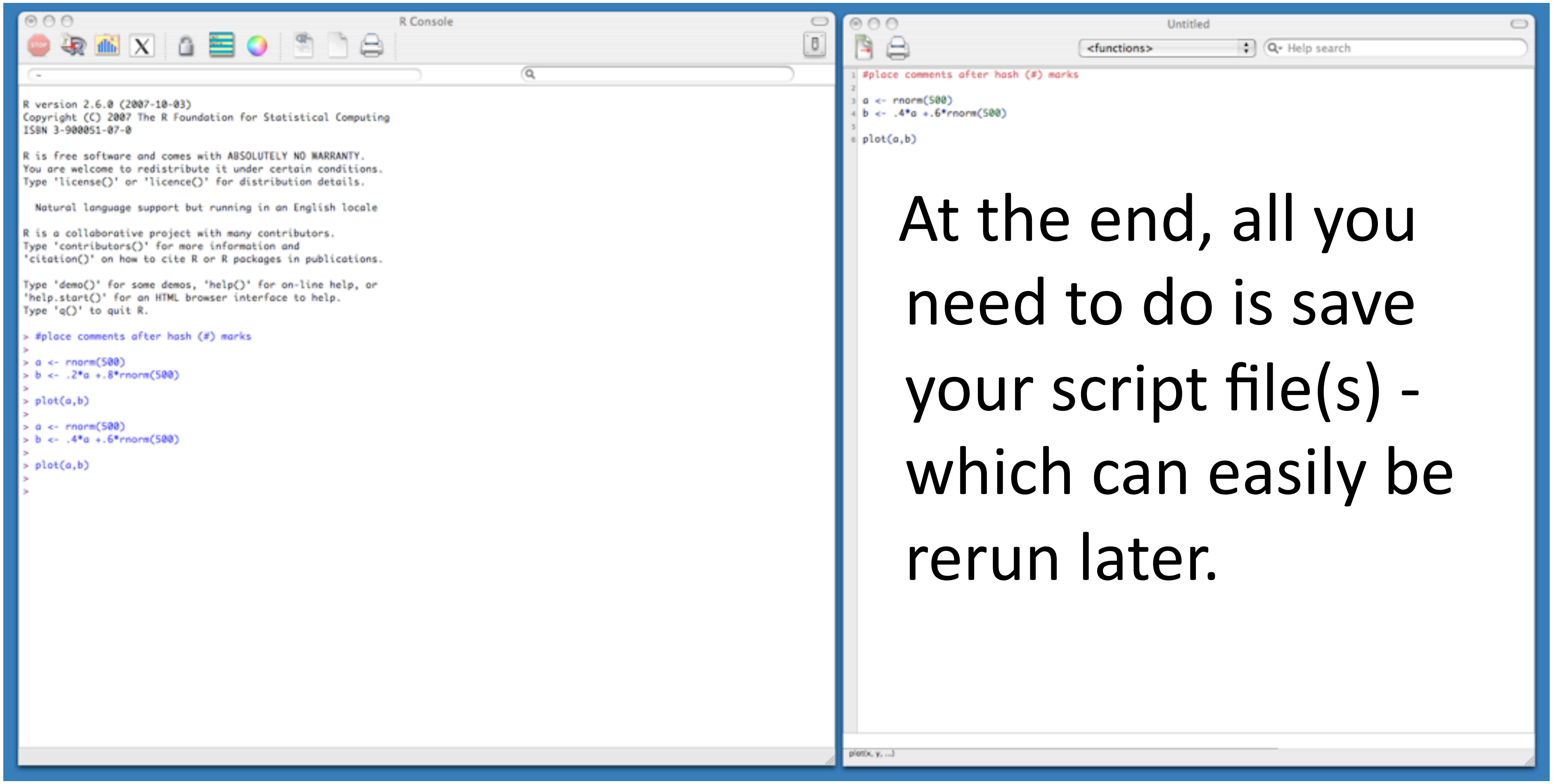
```
R version 2.6.0 (2007-10-03)
Copyright (C) 2007 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> #place comments after hash (#) marks
>
> a <- rnorm(500)
> b <- .2*a +.8*rnorm(500)
>
> plot(a,b)
>
> a <- rnorm(500)
> b <- .4*a +.6*rnorm(500)
>
> plot(a,b)
>
>
```

SCRIPT

```
#place comments after hash (#) marks

a <- rnorm(500)
b <- .4*a +.6*rnorm(500)

plot(a,b)
```

typical R session

- R sessions are *interactive*



```
R version 2.6.0 (2007-10-03)
Copyright (C) 2007 The R Foundation for Statistical Computing
ISBN 3-900051-07-0

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> #place comments after hash (#) marks
>
> a <- rnorm(500)
> b <- .2*a +.8*rnorm(500)
>
> plot(a,b)
>
> a <- rnorm(500)
> b <- .4*a +.6*rnorm(500)
>
> plot(a,b)
>
>
```

```
#place comments after hash (#) marks

a <- rnorm(500)
b <- .4*a +.6*rnorm(500)

plot(a,b)
```

At the end, all you need to do is save your script file(s) - which can easily be rerun later.

how do I format my data for R?

- start in Excel
- use a header row
- don’t leave any gaps or partial rows
- save as .csv text file (comma-separated values)
- open in R with:

```
> dat <- read.csv("path/file")
> dat <- read.csv("url")
```

- data frame (dat$x)

```
> siarad <-
read.csv("http://
www.danielezrajohnson.com/
siarad.csv")

some ways to overview data:
> head(siarad)
> str(siarad)
> names(siarad)

looking at a row or column:
> siarad[1, ]
> siarad[, "Age"]
> siarad$Age

[1] 58 16 53 73 52 65 71 25
42 32 36 . . .
```

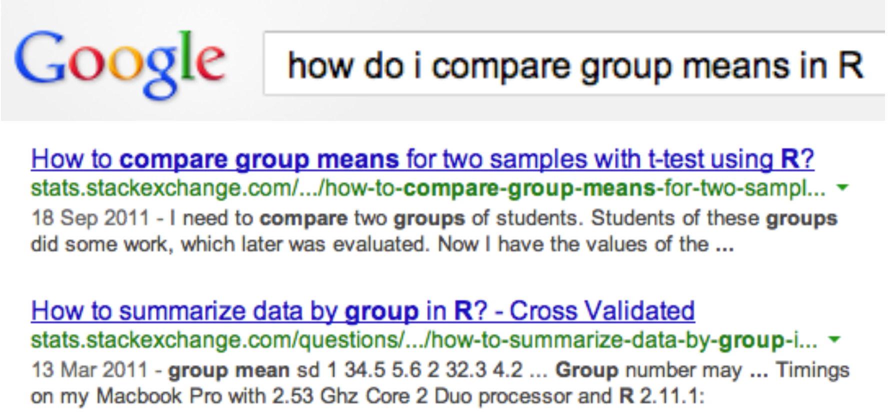
how do I do ________ in R?

- use books or tutorial websites
- adapt existing code
- just ask Google!
  - someone has asked a similar question
- to install a package:

```
> install.packages("package")
> library(package)
```

- for R documentation:

```
> ?function
> ??keyword
```



how do i compare group means in R

How to **compare group means** for two samples with t-test using **R**?
stats.stackexchange.com/.../how-to-**compare**-**group**-**means**-for-two-sampl...
18 Sep 2011 - I need to **compare** two **groups** of students. Students of these **groups** did some work, which later was evaluated. Now I have the values of the ...

How to summarize data by **group** in **R**? - Cross Validated
stats.stackexchange.com/questions/.../how-to-summarize-data-by-**group**-i...
13 Mar 2011 - **group mean** sd 1 34.5 5.6 2 32.3 4.2 ... **Group** number may ... Timings on my Macbook Pro with 2.53 Ghz Core 2 Duo processor and **R** 2.11.1:

```
answer 2: descriptive
> tapply(siarad$Age,
   siarad$Sex, mean)
          F        M
   40.14103 43.38571

answer 1: inferential test
> t.test(Age ~ Sex, siarad)

  p-value = 0.335
  mean in group F mean in group M
         40.14103        43.38571
```

some R functions/operators

abline
abs
anova
as.character
as.factor
as.numeric
c
cat
cbind
class
coef
cor
data.frame
else
exp
head
if
ifelse
fisher.test

fixef
for
function
getwd
glm
glmer
grep
head
image
install.packages
is.na
ks.test
length
library
log
logLik
max
mean
median

min
mosaicplot
names
paste
pchisq
pf
plogis
plot
print
qlogis
ranef
range
rbind
read.csv
rep
repeat
rnorm
round
runif

sample
seq
setwd
set.seed
shapiro.test
signif
sqrt
str
summary
table
tail
t.test
vector
which
wilcox.test
write.csv
xtabs
xyplot
lm

() [] {} + - * / ^ ! & | %in% %% : = <- == # ? ??

more at http://statmaster.sdu.dk/bent/courses/ST501-2011/Rcard.pdf

basic descriptive statistics in R

- central tendency
  - mean()
  - median(), mode()
- dispersion
  - sd(), range()
- other
  - summary(), xtabs()
  - max(), min(), c()
- correlation
  - cor()

```
> mean(siarad$PerWelsh)
[1] 87.28262
> median(siarad$PerWelsh)
[1] 90
> range(siarad$PerWelsh)
[1] 13.24503 99.67532
> sd(siarad$PerWelsh)
[1] 11.165
> summary(dat2$PerWelsh)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  13.25   82.93   90.00   87.28   95.11   99.68

> cor(siarad$PerWelsh,
   siarad$PerEng)
[1] -0.7727451

> xtabs(~Balance + Sex,
   siarad)
```

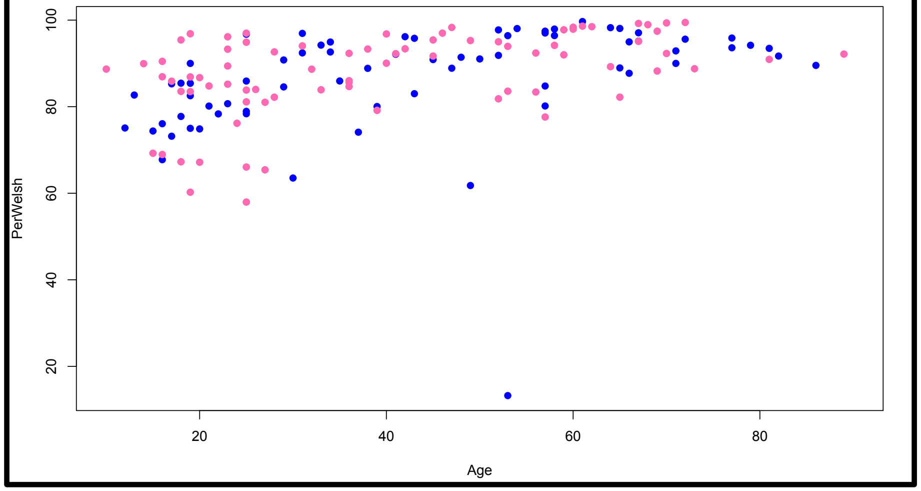
basic graphics in R (should do first)

- formulas in R
  - y ~ x
  - y ~ x1 + x2…
- plot()

```
> plot(y ~ x)
> plot(dat$y ~ dat$x)
```

- points(), lines(), abline()
- plots are customizable
- other graphics types
  - xyplot(), ggplot()
- boxplot(), hist(), etc.

```
> s <- siarad
> plot(PerWelsh ~ Age, s)

> plot(PerWelsh ~ Age,
  subset(s, Sex == "M"),
  col = "blue")
> points(PerWelsh ~ Age,
  subset(s, Sex == "F"),
  col = "hotpink")
```



basic regression in R

- linear regression: lm()
  - continuous dependent variable
- logistic regression: glm()
  - binary dependent variable
  - d.v. of 3+ categories: difficult!
- model-building

```
> m1 <- lm(y ~ x1, dat)
> m2 <- lm(y ~ x1 + x2, dat)
```

- hypothesis testing
  - model fit vs. model complexity

```
> anova(m1, m2, test="Chisq")
```

- illustration in R
  - http://www.danielezrajohnson.com/bangor_regression.R

```
   do men use more all-Welsh
      clauses than women?

> m0 <- lm(PerWelsh ~ 1, s)
> m1 <- lm(PerWelsh ~ Sex, s)
> anova(m0, m1)
                 p = .49

  do older speakers use more
      all-Welsh clauses?

> m0 <- lm(PerWelsh ~ 1, s)
> m1 <- lm(PerWelsh ~ Age, s)
> anova(m0, m1)
                 p = .0000003
```

effect size vs. significance

- in regression, the size or importance of an effect can mean two different things
- effect size
  - regression coefficient
  - slope / size of difference
- significance
  - expressed as p-value
  - could this be chance?
- related but distinct

given the same sample size
a larger effect size
is more significant

but with a small sample
large effects may not be
"significant"

and with a large sample
very small effects may be
"significant"

statistically significant
doesn't mean
practically significant

multiple regression: a “real” example

- dependent variable:
  - – % of all-Welsh clauses
- associated with:
  - – age (10 to 89)
  - – relative ability (W, =, E)
- questions:
  - – is each association significant on its own?
  - – is each one significant <u>on top of the other</u>?

```
> m.0 <- lm(PerWelsh ~ 1, s)
> m.a <- lm(PerWelsh ~ Age, s)
> anova(m.0, m.a)
               p = .0000003

> tapply(s$PerWelsh, s$Balance, mean)
   English    Equal    Welsh
  81.74150 87.39258 89.99533

> m.b <- lm(PerWelsh ~ Balance, s)
> anova(m.0, m.b)
               p = .033

> m.ab <- lm(PerWelsh ~ Age +
   Balance, s)
> anova(m.a, m.ab)
               p = .019
> anova(m.b, m.ab)
               p = .0000002
```

recommended reading

- an entertaining and thorough printed textbook:
  - http://www.amazon.co.uk/Discovering-Statistics-Using-Andy-Field/dp/1446200469/

- a free textbook on probability and statistics:
  - http://cran.r-project.org/web/packages/IPSUR/vignettes/IPSUR.pdf

- a great resource for all kinds of information about R:
  - http://statistics.ats.ucla.edu/stat/r/

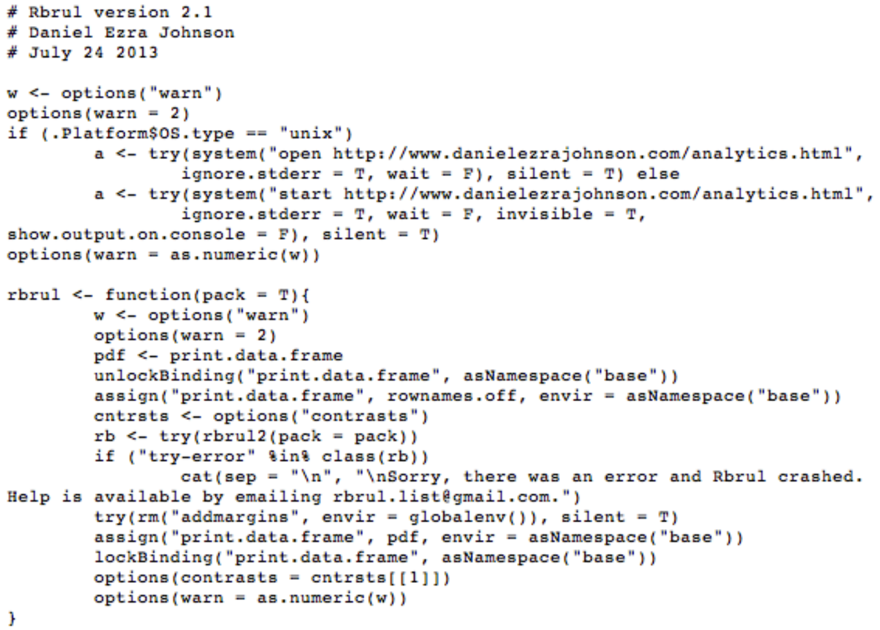
- a series of video tutorials from Google:
  - http://www.youtube.com/playlist?list=PLOU2XLYxmsIK9qQfztXeybpHvru-TrqAP

what is Rbrul?

- today: most statistical analyses can be done with: SPSS, SAS, R, etc.
- 1970’s: VARBRUL developed for sociolinguists
  - now called GoldVarb
  - limited in several ways
- Rbrul is an R program
  - text file, paste or source()
  - familiar to GoldVarb users
  - more flexible regressions

"Rbrul offers a compromise of the old and new that I believe will be widely used in the near future."

"I've been finding it so much easier than trying to do the same in R."

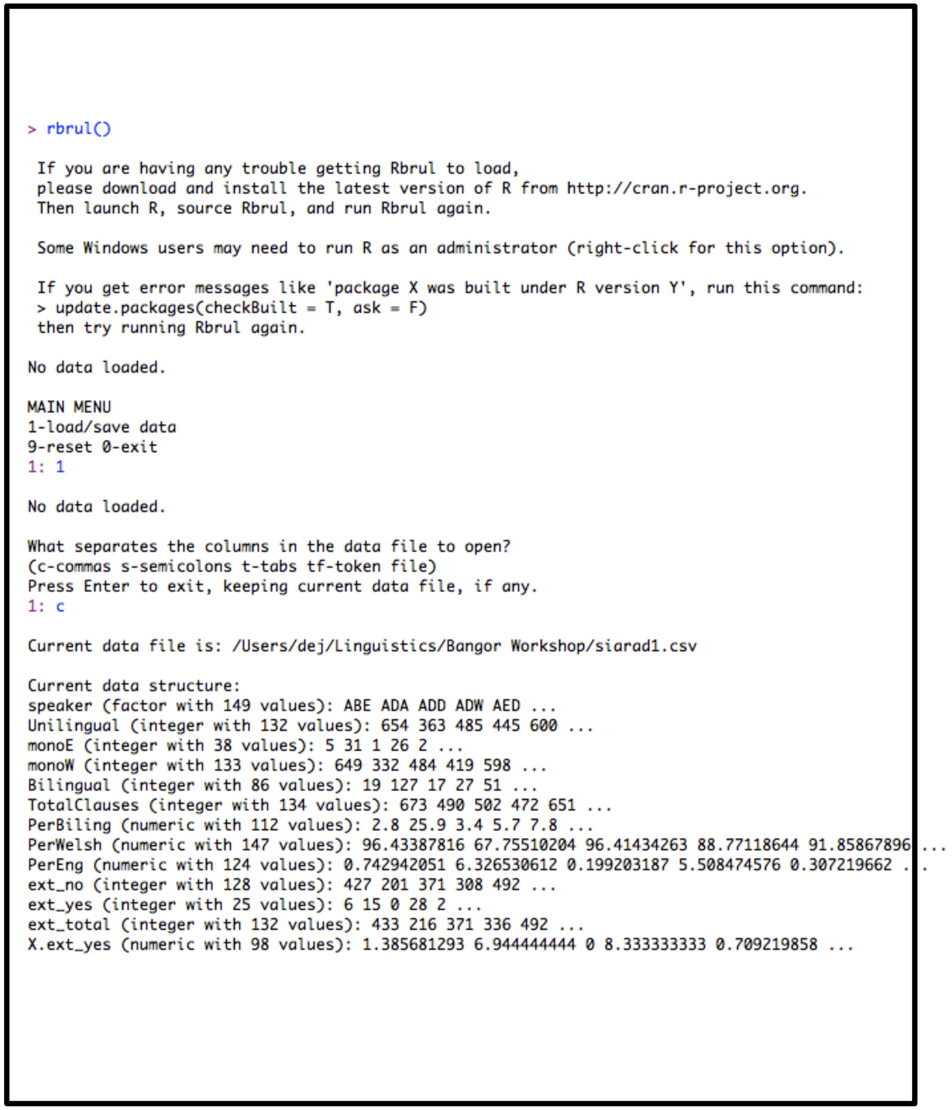


```
# Rbrul version 2.1
# Daniel Ezra Johnson
# July 24 2013

w <- options("warn")
options(warn = 2)
if (.Platform$OS.type == "unix")
        a <- try(system("open http://www.danielezrajohnson.com/analytics.html",
                ignore.stderr = T, wait = F), silent = T) else
        a <- try(system("start http://www.danielezrajohnson.com/analytics.html",
                ignore.stderr = T, wait = F, invisible = T,
show.output.on.console = F), silent = T)
options(warn = as.numeric(w))

rbrul <- function(pack = T){
        w <- options("warn")
        options(warn = 2)
        pdf <- print.data.frame
        unlockBinding("print.data.frame", asNamespace("base"))
        assign("print.data.frame", rownames.off, envir = asNamespace("base"))
        cntrsts <- options("contrasts")
        rb <- try(rbrul2(pack = pack))
        if ("try-error" %in% class(rb))
                cat(sep = "\n", "\nSorry, there was an error and Rbrul crashed.
Help is available by emailing rbrul.list@gmail.com.")
        try(rm("addmargins", envir = globalenv()), silent = T)
        assign("print.data.frame", pdf, envir = asNamespace("base"))
        lockBinding("print.data.frame", asNamespace("base"))
        options(contrasts = cntrsts[[1]])
        options(warn = as.numeric(w))
}
```

http://www.danielezrajohnson.com/johnson_compass_final.pdf

how does Rbrul work?

- not a command-line interface like R
- not a graphical interface
- text-based interface
  - questions, menu options
  - can’t go backwards
- stepwise regression
  - step-up, step-down
  - not recommended!
- “one-level”
  - tests variables like drop1()



```
> rbrul()

 If you are having any trouble getting Rbrul to load,
 please download and install the latest version of R from http://cran.r-project.org.
 Then launch R, source Rbrul, and run Rbrul again.

 Some Windows users may need to run R as an administrator (right-click for this option).

 If you get error messages like 'package X was built under R version Y', run this command:
 > update.packages(checkBuilt = T, ask = F)
 then try running Rbrul again.

No data loaded.

MAIN MENU
1-load/save data
9-reset 0-exit
1: 1

No data loaded.

What separates the columns in the data file to open?
(c-commas s-semicolons t-tabs tf-token file)
Press Enter to exit, keeping current data file, if any.
1: c

Current data file is: /Users/dej/Linguistics/Bangor Workshop/siarad1.csv

Current data structure:
speaker (factor with 149 values): ABE ADA ADD ADW AED ...
Unilingual (integer with 132 values): 654 363 485 445 600 ...
monoE (integer with 38 values): 5 31 1 26 2 ...
monoW (integer with 133 values): 649 332 484 419 598 ...
Bilingual (integer with 86 values): 19 127 17 27 51 ...
TotalClauses (integer with 134 values): 673 490 502 472 651 ...
PerBiling (numeric with 112 values): 2.8 25.9 3.4 5.7 7.8 ...
PerWelsh (numeric with 147 values): 96.43387816 67.75510204 96.41434263 88.77118644 91.85867896 ...
PerEng (numeric with 124 values): 0.742942051 6.326530612 0.199203187 5.508474576 0.307219662 ...
ext_no (integer with 128 values): 427 201 371 308 492 ...
ext_yes (integer with 25 values): 6 15 0 28 2 ...
ext_total (integer with 132 values): 433 216 371 336 492 ...
X.ext_yes (numeric with 98 values): 1.385681293 6.944444444 0 8.333333333 0.709219858 ...
```

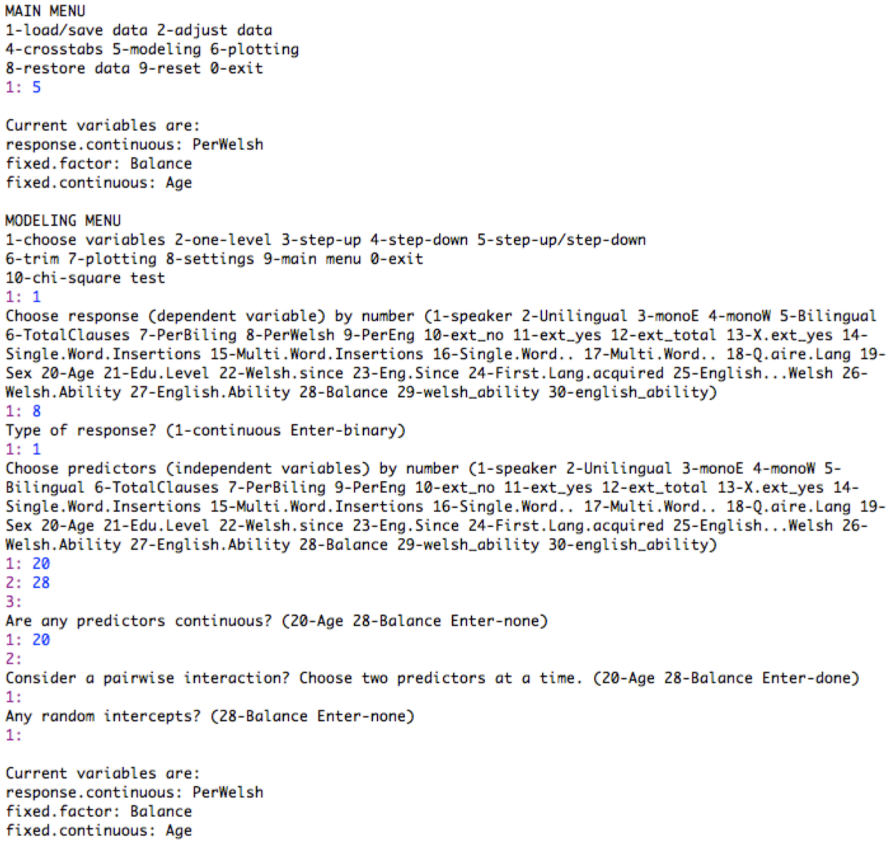
http://www.danielezrajohnson.com/stepwise.pdf

choosing variables in Rbrul

- response
  - dependent variable
  - continuous or binary
- predictors
  - independent variables
  - any continuous?
  - any interactions?
  - *random effects?*
  - random effects not needed if 1 obs./spkr.

```
R: PerWelsh ~ Age + Balance
```

```
Rbrul:
```



```
MAIN MENU
1-load/save data 2-adjust data
4-crosstabs 5-modeling 6-plotting
8-restore data 9-reset 0-exit
1: 5

Current variables are:
response.continuous: PerWelsh
fixed.factor: Balance
fixed.continuous: Age

MODELING MENU
1-choose variables 2-one-level 3-step-up 4-step-down 5-step-up/step-down
6-trim 7-plotting 8-settings 9-main menu 0-exit
10-chi-square test
1: 1
Choose response (dependent variable) by number (1-speaker 2-Unilingual 3-monoE 4-monoW 5-Bilingual
6-TotalClauses 7-PerBiling 8-PerWelsh 9-PerEng 10-ext_no 11-ext_yes 12-ext_total 13-X.ext_yes 14-
Single.Word.Insertions 15-Multi.Word.Insertions 16-Single.Word.. 17-Multi.Word.. 18-Q.aire.Lang 19-
Sex 20-Age 21-Edu.Level 22-Welsh.since 23-Eng.Since 24-First.Lang.acquired 25-English...Welsh 26-
Welsh.Ability 27-English.Ability 28-Balance 29-welsh_ability 30-english_ability)
1: 8
Type of response? (1-continuous Enter-binary)
1: 1
Choose predictors (independent variables) by number (1-speaker 2-Unilingual 3-monoE 4-monoW 5-
Bilingual 6-TotalClauses 7-PerBiling 9-PerEng 10-ext_no 11-ext_yes 12-ext_total 13-X.ext_yes 14-
Single.Word.Insertions 15-Multi.Word.Insertions 16-Single.Word.. 17-Multi.Word.. 18-Q.aire.Lang 19-
Sex 20-Age 21-Edu.Level 22-Welsh.since 23-Eng.Since 24-First.Lang.acquired 25-English...Welsh 26-
Welsh.Ability 27-English.Ability 28-Balance 29-welsh_ability 30-english_ability)
1: 20
2: 28
3:
Are any predictors continuous? (20-Age 28-Balance Enter-none)
1: 20
2:
Consider a pairwise interaction? Choose two predictors at a time. (20-Age 28-Balance Enter-done)
1:
Any random intercepts? (28-Balance Enter-none)
1:

Current variables are:
response.continuous: PerWelsh
fixed.factor: Balance
fixed.continuous: Age
```

Rbrul output compared to R

- Rbrul's output is more user-friendly than R's
- for categorical predictors (factors), Rbrul includes redundant information, e.g.:
  - men: +15
  - women: -15
- R might just say:
  - Sex1: +15

in R:

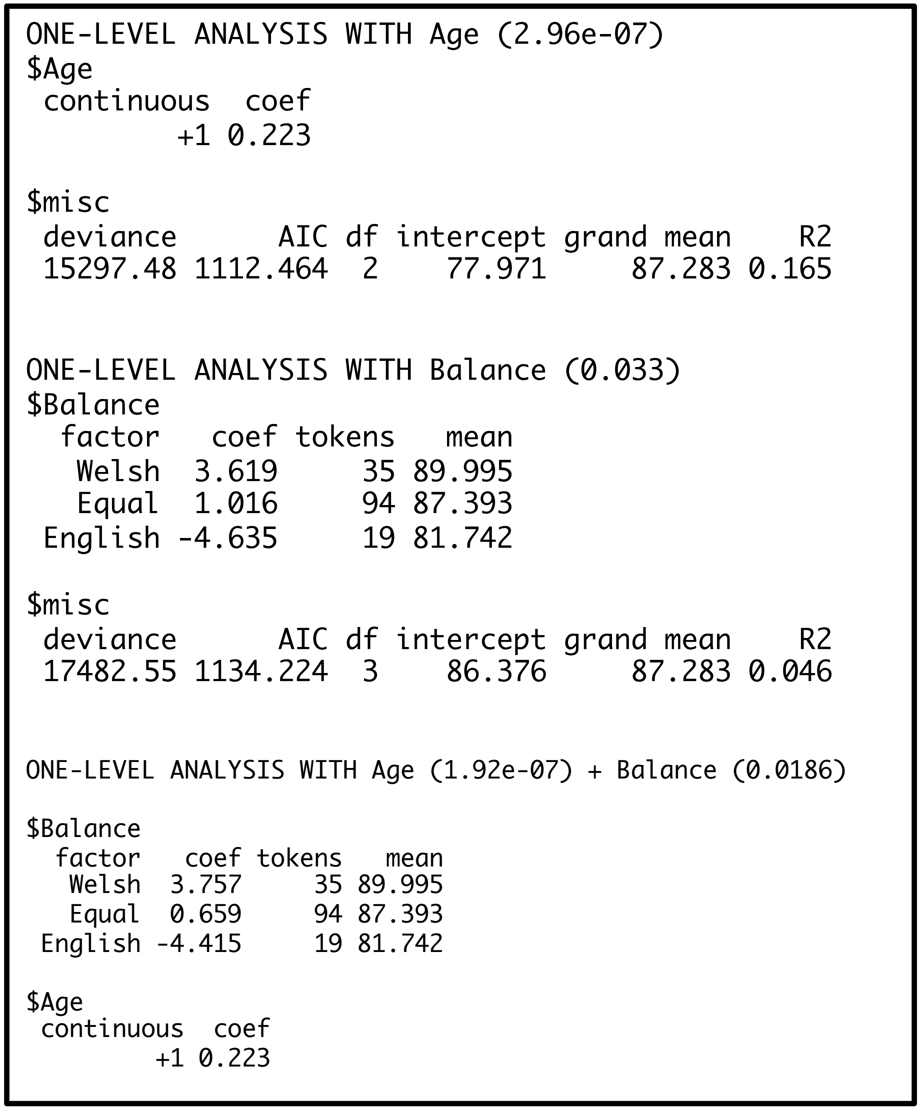
```
> m.ab

Call:
lm(PerWelsh ~ Age + Balance, data = s)

Coefficients:
Intercept    Age   Balance1    Balance2
    77.24   0.22      -4.42        0.66
```

in Rbrul:

```
ONE-LEVEL ANALYSIS WITH Age (1.92e-07) + Balance (0.0186)

$Balance
  factor   coef tokens   mean
   Welsh  3.757     35 89.995
   Equal  0.659     94 87.393
 English -4.415     19 81.742

$Age
 continuous  coef
         +1 0.223

$misc
 deviance      AIC df intercept grand mean   R2
  14473.4 1108.268  4    77.244     87.283 0.21
```

multiple regression: a “real” example

- dependent variable:
  - – % of all-Welsh clauses
- associated with:
  - – relative ability (W, =, E)
  - – age (10 to 89)
- questions:
  - – is each association significant on its own?
  - – is each one significant on top of the other?



```
ONE-LEVEL ANALYSIS WITH Age (2.96e-07)
$Age
 continuous  coef
         +1 0.223

$misc
 deviance      AIC df intercept grand mean    R2
 15297.48 1112.464  2    77.971     87.283 0.165


ONE-LEVEL ANALYSIS WITH Balance (0.033)
$Balance
  factor   coef tokens   mean
   Welsh  3.619     35 89.995
   Equal  1.016     94 87.393
 English -4.635     19 81.742

$misc
 deviance      AIC df intercept grand mean    R2
 17482.55 1134.224  3    86.376     87.283 0.046


ONE-LEVEL ANALYSIS WITH Age (1.92e-07) + Balance (0.0186)

$Balance
  factor   coef tokens   mean
   Welsh  3.757     35 89.995
   Equal  0.659     94 87.393
 English -4.415     19 81.742

$Age
 continuous  coef
         +1 0.223
```

recommended reading

- a “bible” for regression analysis:
  - http://www.amazon.co.uk/Regression-Modeling-Strategies-Applications-Statistics/dp/0387952322

- for help with Rbrul and/or to report errors:
  - please email me!
  - it usually helps to send your data file as well
  - danielezrajohnson@gmail.com

what are mixed-effects models?

- to be able to work with mixed models was the main reason for creating Rbrul
- because of a common structure of natural speech data sets, ordinary fixed-effects regression models are prone to error

Three Types of Error

Errors about significance:

Type I error: you reject the null hypothesis when you shouldn't (false +).

Type II error: you accept the null hypothesis when you shouldn't (false -).

Errors about effect size:

You misestimate the effect.

what is this special structure?

- grouping (nesting)
- imagine 10,000 tokens (obs.) of a variable
- 100 individual spkrs.
- 100 tokens from each
- for certain purposes, you have a sample of 10,000…
- but often, only 100

if you only care about between-speaker (external) effects, you might average over speakers, which solves this problem!

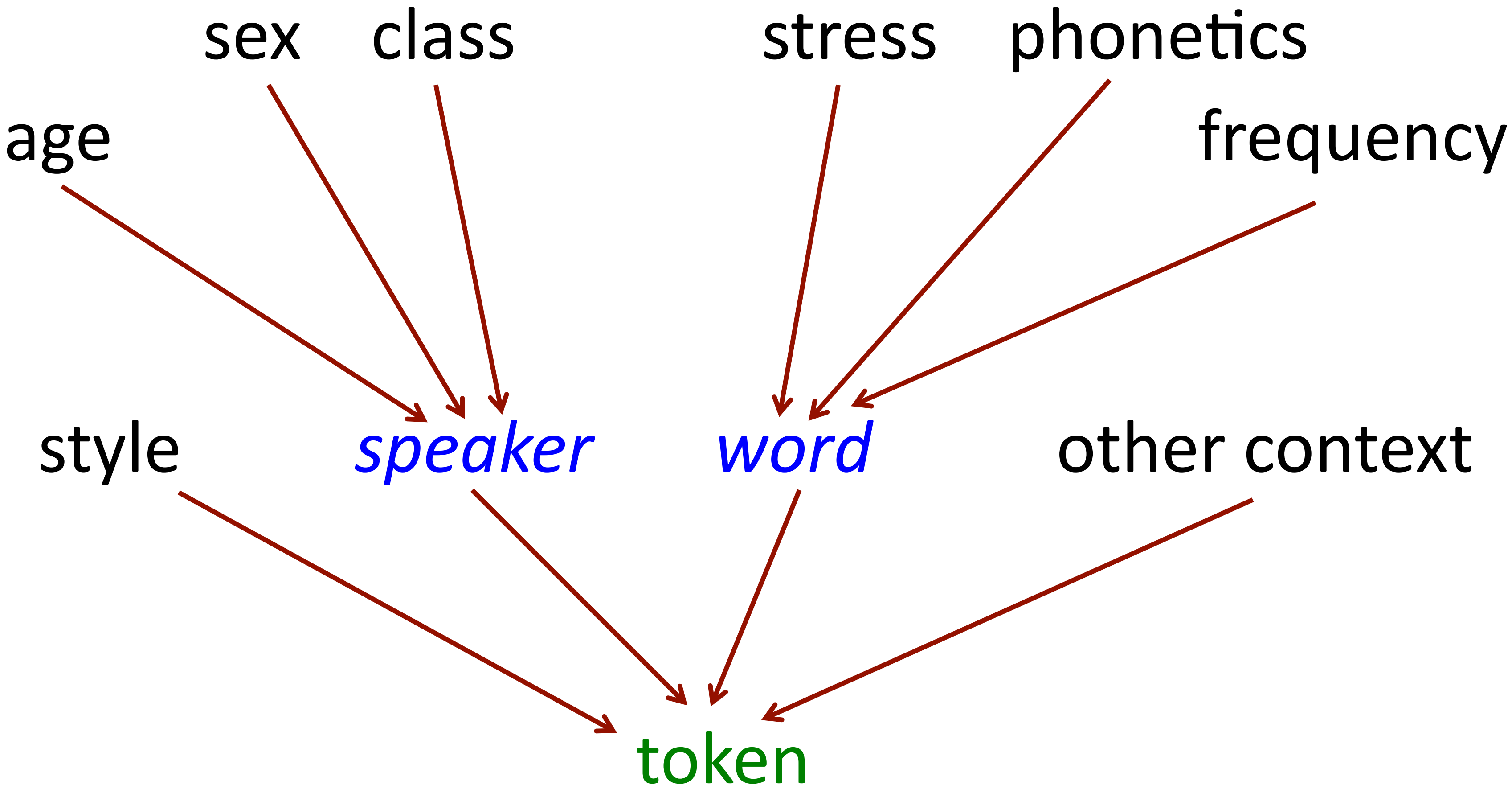
if you also care about within-speaker (internal) effects, you must analyze individual tokens: you have this problem!

if individual speakers vary, must account for speaker

if individual words vary, must account for word in model

VARBRUL method got this wrong

architecture of variables



sex class stress phonetics

age frequency

style *speaker* *word* other context

token

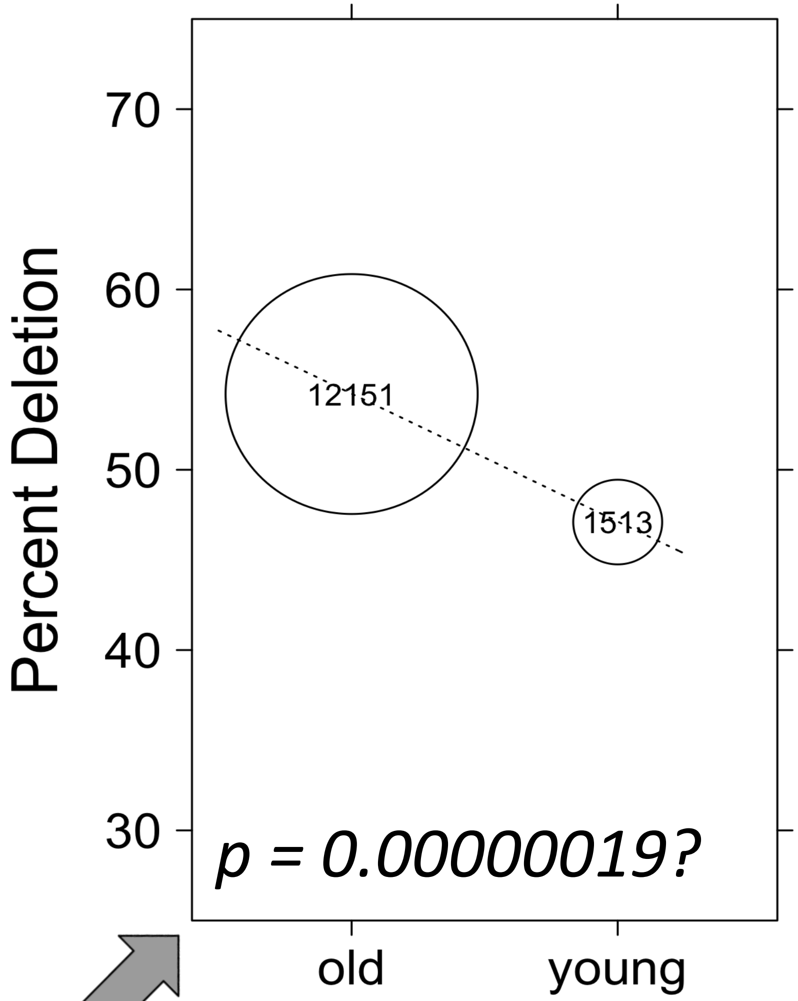
fixed effect *random effect*

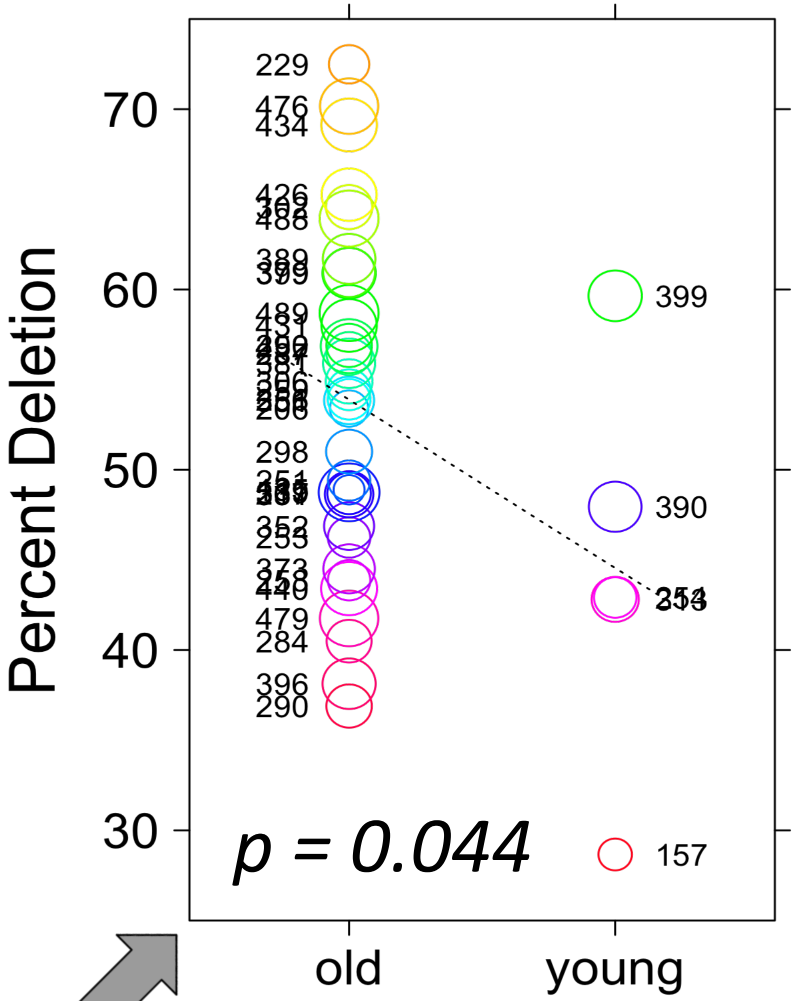
test script 2: http://www.danielezrajohnson.com/york_four.R

why do mixed models work better?

- capture variation among grouping units (e.g. speaker, word)
- handle unbalanced data better
- often conservative
  - less Type I error
  - (chance effects)
- http://dejonedge.blogspot.co.uk/2013/07/random-slopes-now-rbrul-has-them-you.html

```
in R, use lmer() function

add random intercepts like
   y ~ x1 + (1 | speaker)

add random slopes* like
   y ~ x2 + (x2 | speaker)

in Rbrul, straightforward

*important, but slow, may
not work at all!
```





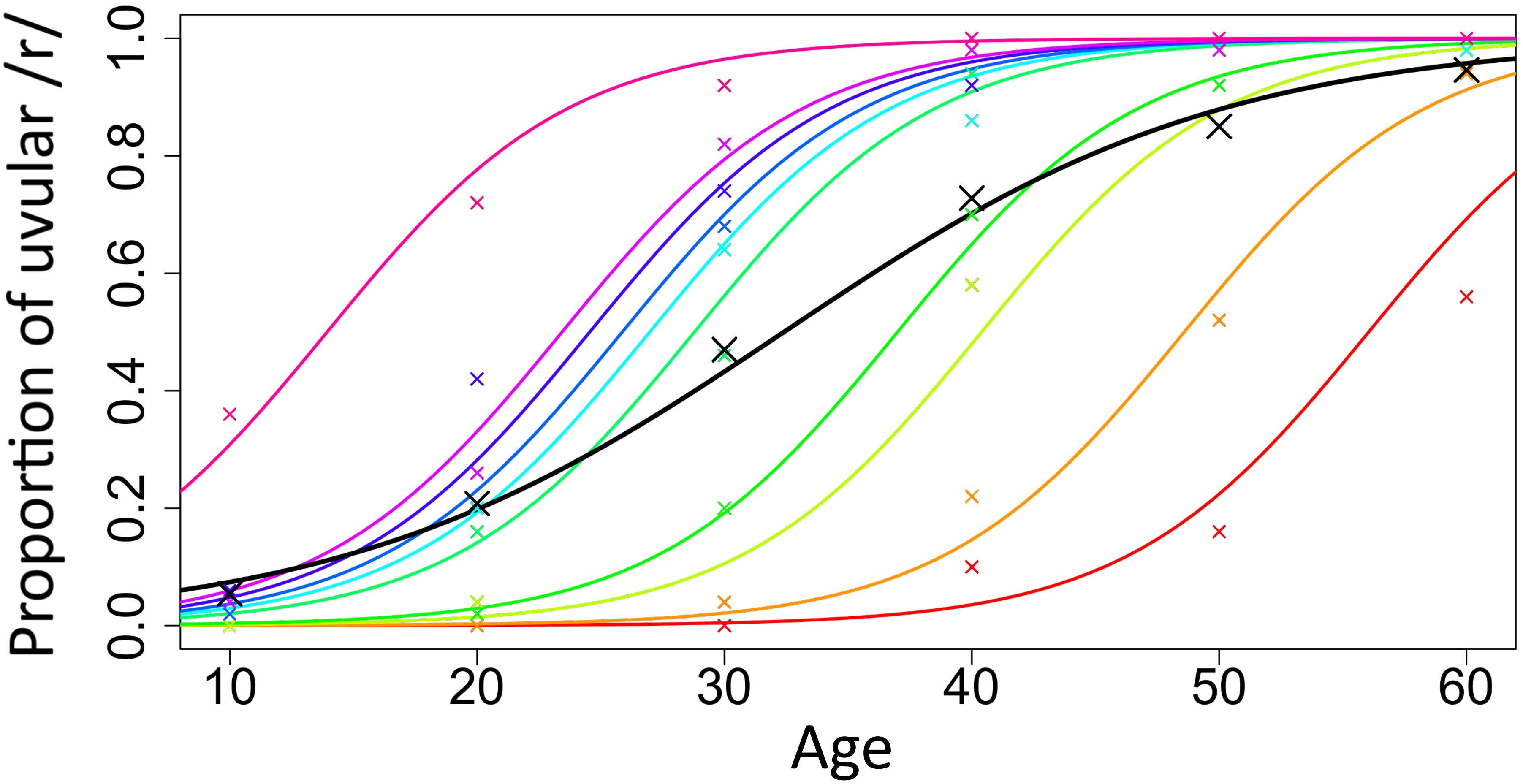
0.535 0.465

0.547 0.453

age w/ no random effect

age + random intercept: speaker

Significance of between-speaker predictor



age coefficient w/ no random effect: 0.113 log-odds/year
age coeff. w/ speaker random effect: 0.205 log-odds/year

Effect size of within-speaker predictor (logistic regression only)

recommended reading

- a “bible” for mixed-effects modeling:
  - http://www.amazon.co.uk/Mixed-Effects-Models-S-PLUS-Statistics-Computing/dp/1441903178
- unfinished book by same author (Doug Bates):
  - http://lme4.R-forge.R-project.org/book/
- R-sig-ME and R-Lang listservs
  - https://stat.ethz.ch/mailman/listinfo/r-sig-mixed-models
  - https://mailman.ucsd.edu/mailman/listinfo/ling-r-lang-l

any questions?

- some audiences comfortable with regression
- particularly interested in mixed models
- many sociolinguists deal mostly with binary data

- what is your data like?
- what are your concerns?

- thank you for coming, I hope this was useful
- email me any time with more questions