

Study 2: data analysis

Example analysis using R

Steps for data analysis

- Install software on your computer or locate computer with software
 - (e.g., R, systat, SPSS)
- Prepare data for analysis
 - Subjects (rows) x variables (columns)
 - first row contains labels
- Read data
 - either copy and read.clipboard, or read a file

Using R

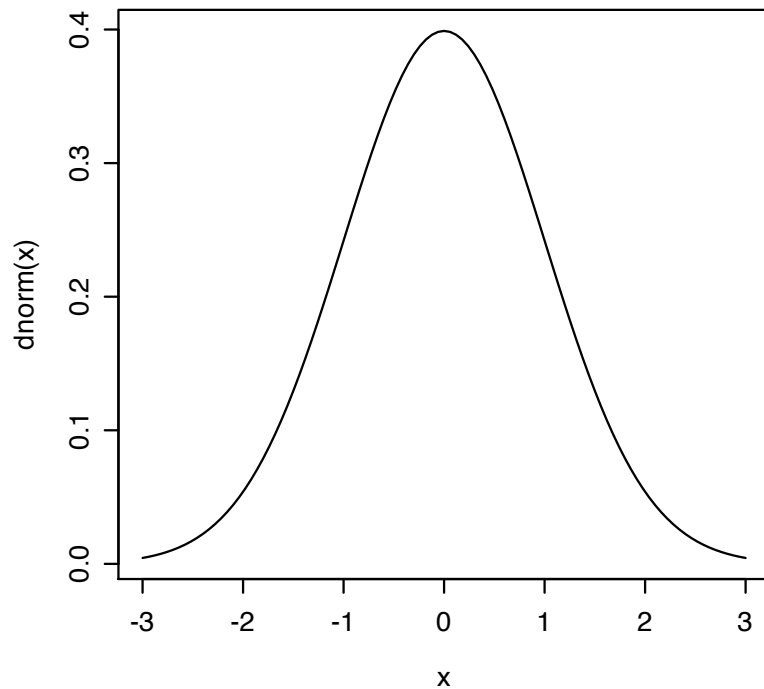
- Install R (download from <http://cran.r-project.org>)
- add psych package using the package installer
 - on a PC, click on package options and install packages
 - on a Mac, go to package installer and get list
- run R

R: a fancy calculator

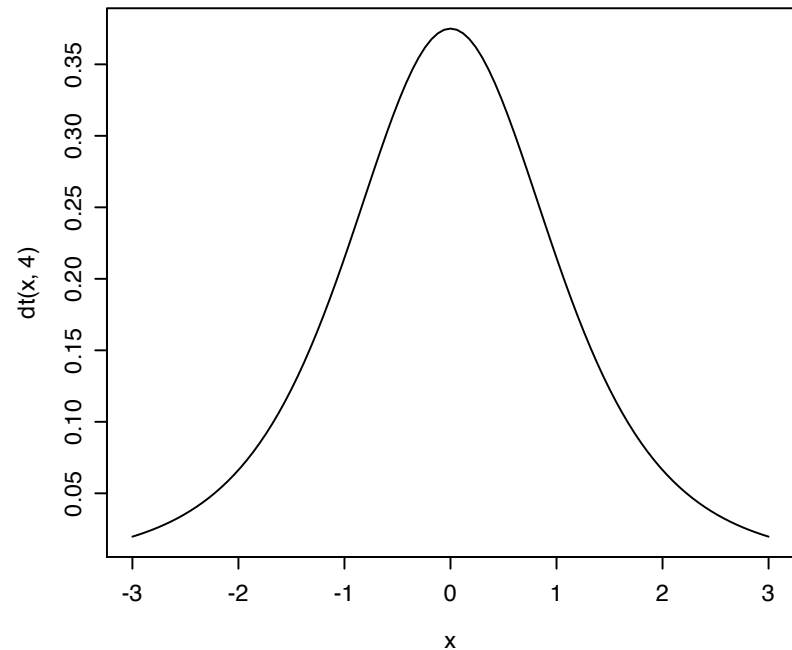
```
Addition      > 2+3
                [1] 5
Multiplication  > 3 * 4
                [1] 12
Exponentiation > 2^5
                [1] 32
Vectors        > exp(1:4)
                [1] 2.718282 7.389056 20.085537 54.598150
output         > round(exp(1:4),2)
control        [1] 2.72 7.39 20.09 54.60
vectors        > x <- 2:6
                > y <- 3:7
                > x
                [1] 2 3 4 5 6
                > y
                [1] 3 4 5 6 7
                > x *y
                [1] 6 12 20 30 42
```

R: A graphical device

`curve(dnorm(x),-3,3)`



`curve(dt(x,4),-3,3)`



R: A statistical look up table

```
> pnorm(1.96) # Probability of a Normal < 1.96
[1] 0.9750021
> pt(1.96,22) # Probability of a t df=22 < 1.96
[1] 0.9686083
> pt(2,40) # Probability of t (40) < 2
[1] 0.9738388
> pf(4,1,40) # Probability of F(1,40) < 4
[1] 0.9476777
dbinom(0:10,10,.5) * 1024 #binomial distribution
[1] 1 10 45 120 210 252 210 120 45 10 1
```

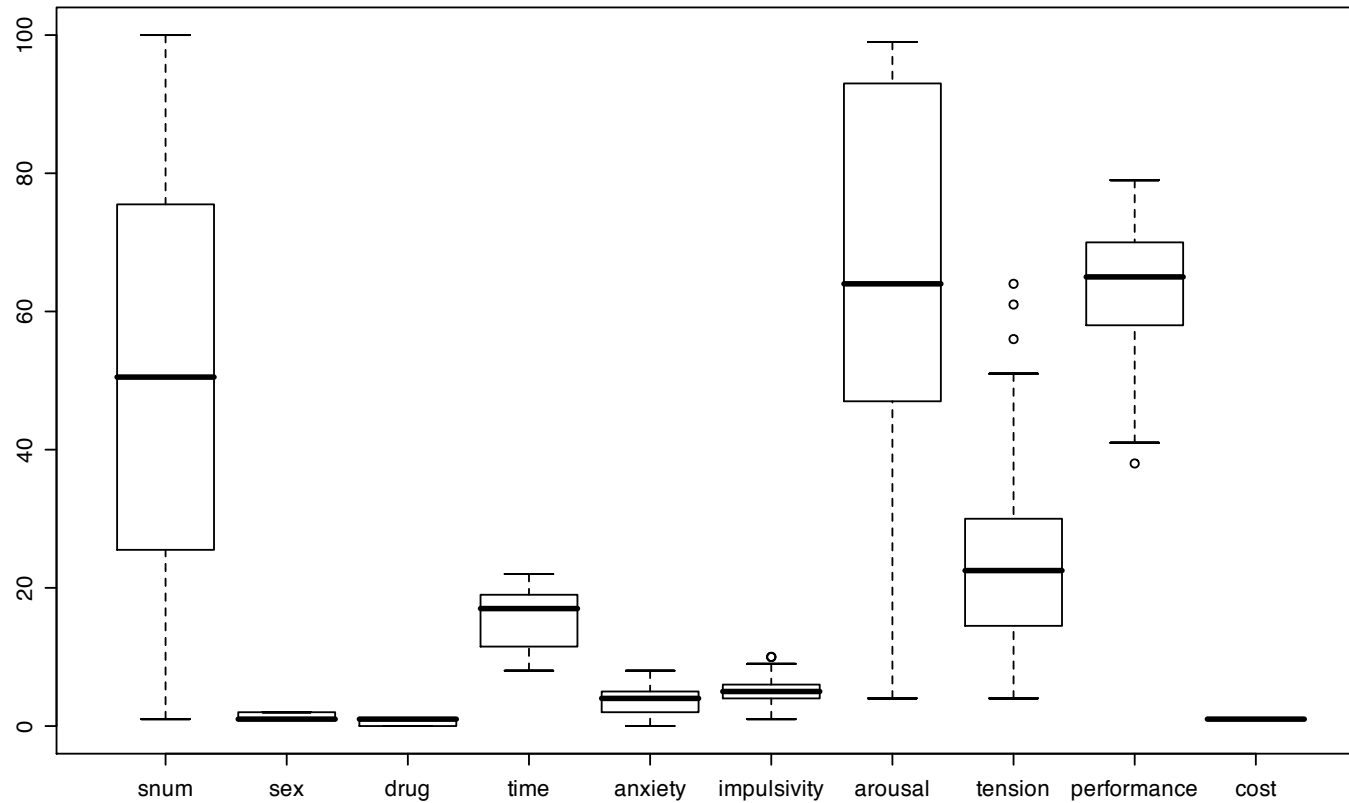
Data analysis with R

```
> library(psych)           #Make this package active
> sim.data <- read.clipboard() #read the data from clipboard
> describe(sim.data)      #basic descriptive statistics
```

	var	n	mean	sd	median	mad	min	max	range	skew	kurtosis	se
snum	1	100	50.50	29.01	50.5	37.06	1	100	99	0.00	-1.24	2.90
sex	2	100	1.48	0.50	1.0	0.00	1	2	1	0.08	-2.01	0.05
drug	3	100	0.55	0.50	1.0	0.00	0	1	1	-0.20	-1.98	0.05
time	4	100	15.57	4.32	17.0	4.45	8	22	14	-0.28	-1.29	0.43
anxiety	5	100	3.90	1.91	4.0	1.48	0	8	8	0.03	-0.73	0.19
impulsivity	6	100	5.19	2.14	5.0	1.48	1	10	9	-0.06	-0.31	0.21
arousal	7	100	63.53	27.88	64.0	35.58	4	99	95	-0.40	-0.85	2.79
tension	8	100	23.63	12.46	22.5	11.12	4	64	60	0.87	0.79	1.25
performance	9	100	63.55	9.05	65.0	8.90	38	79	41	-0.47	-0.43	0.91
cost	10	100	1.00	0.00	1.0	0.00	1	1	0	NaN	NaN	0.00

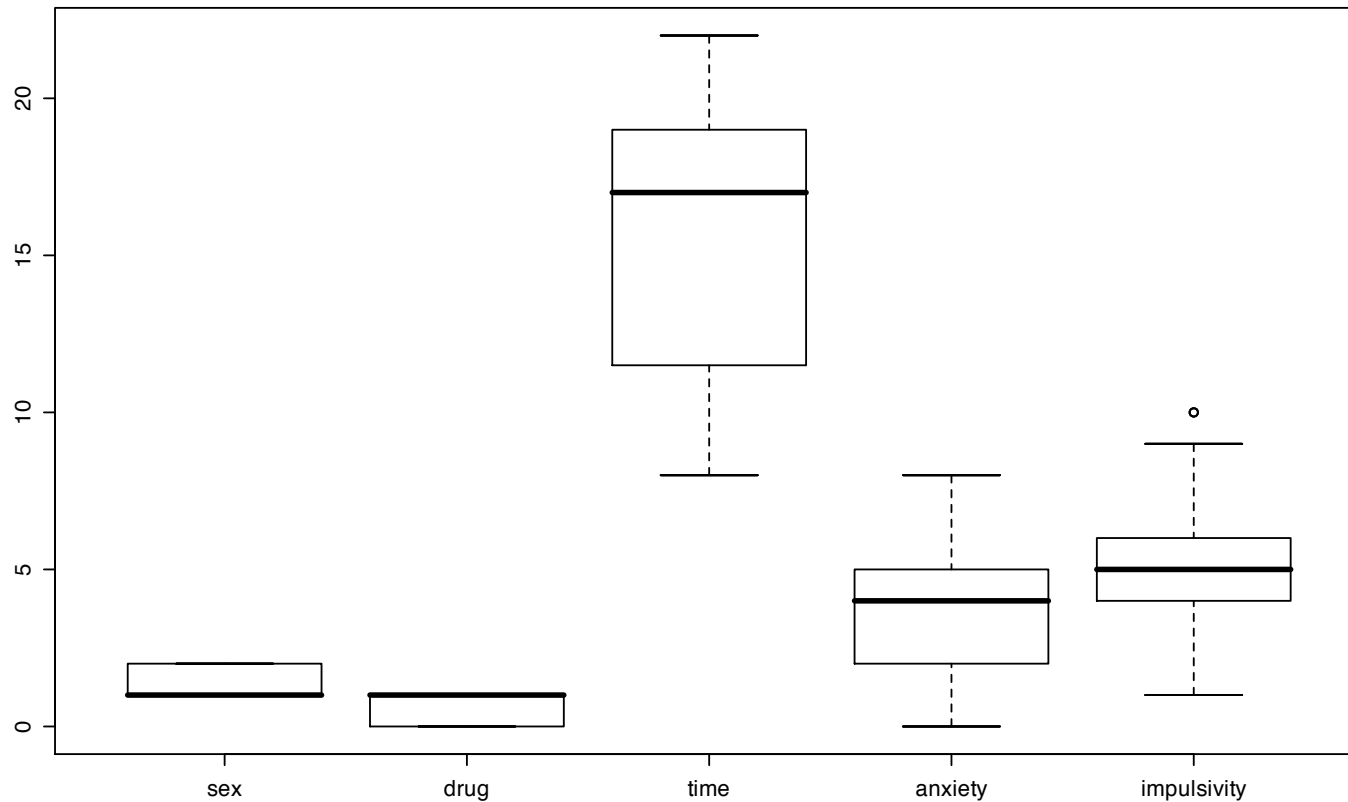
Simple descriptives

`boxplot(sim.data)`



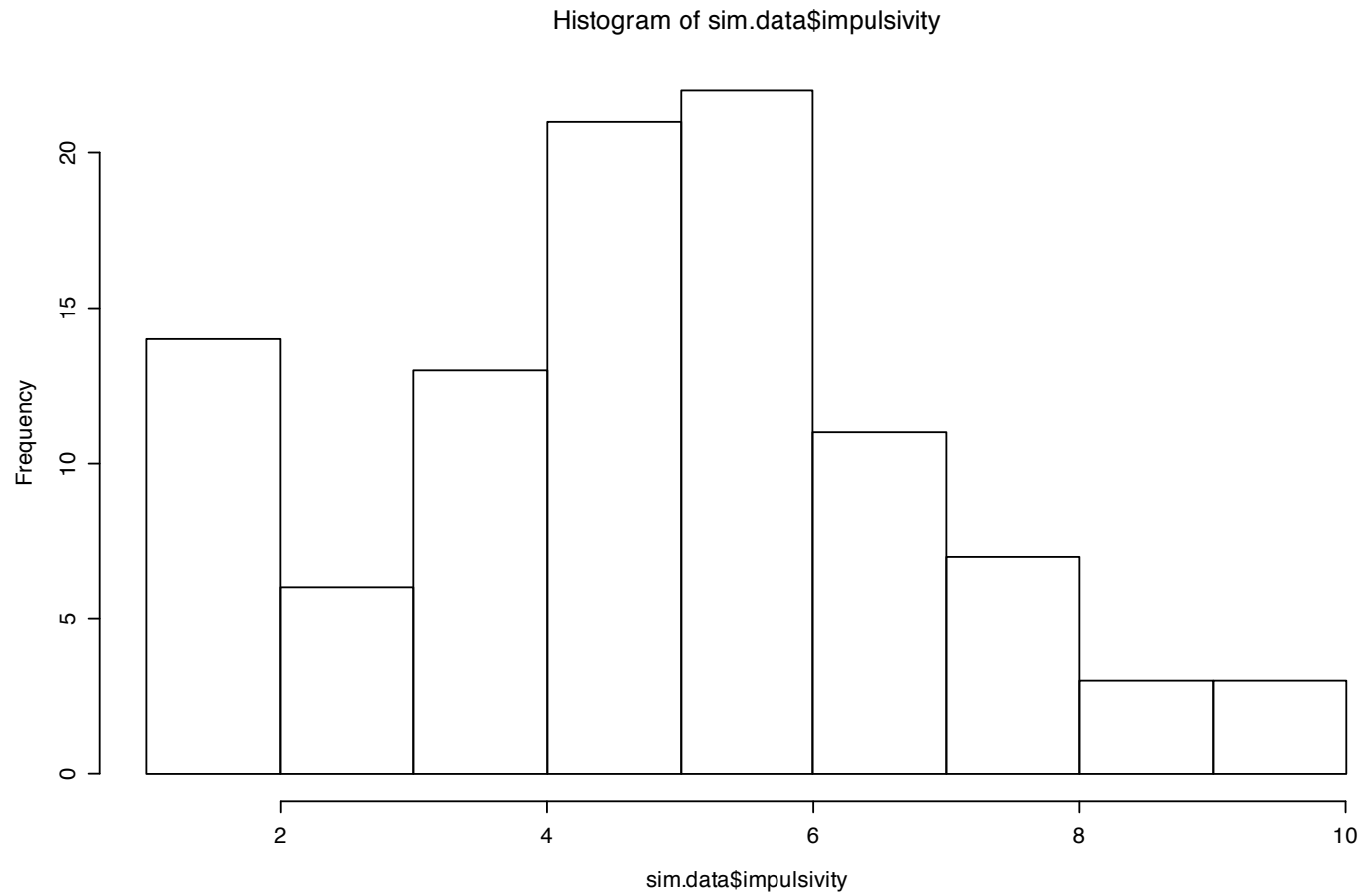
Even more descriptives

```
boxplot(sim.data[2:6])
```



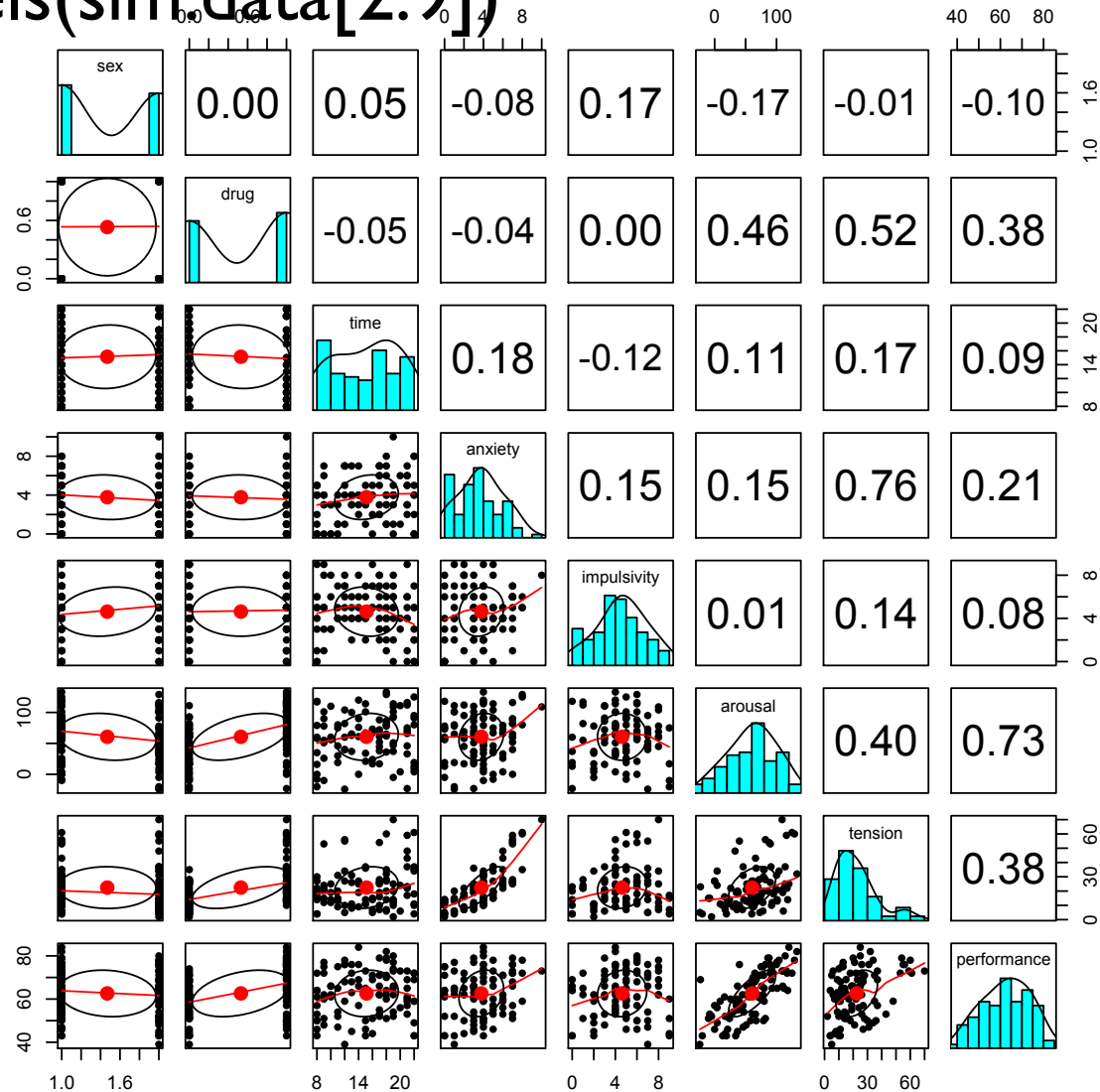
Histograms

```
hist(sim.data$impulsivity)
```



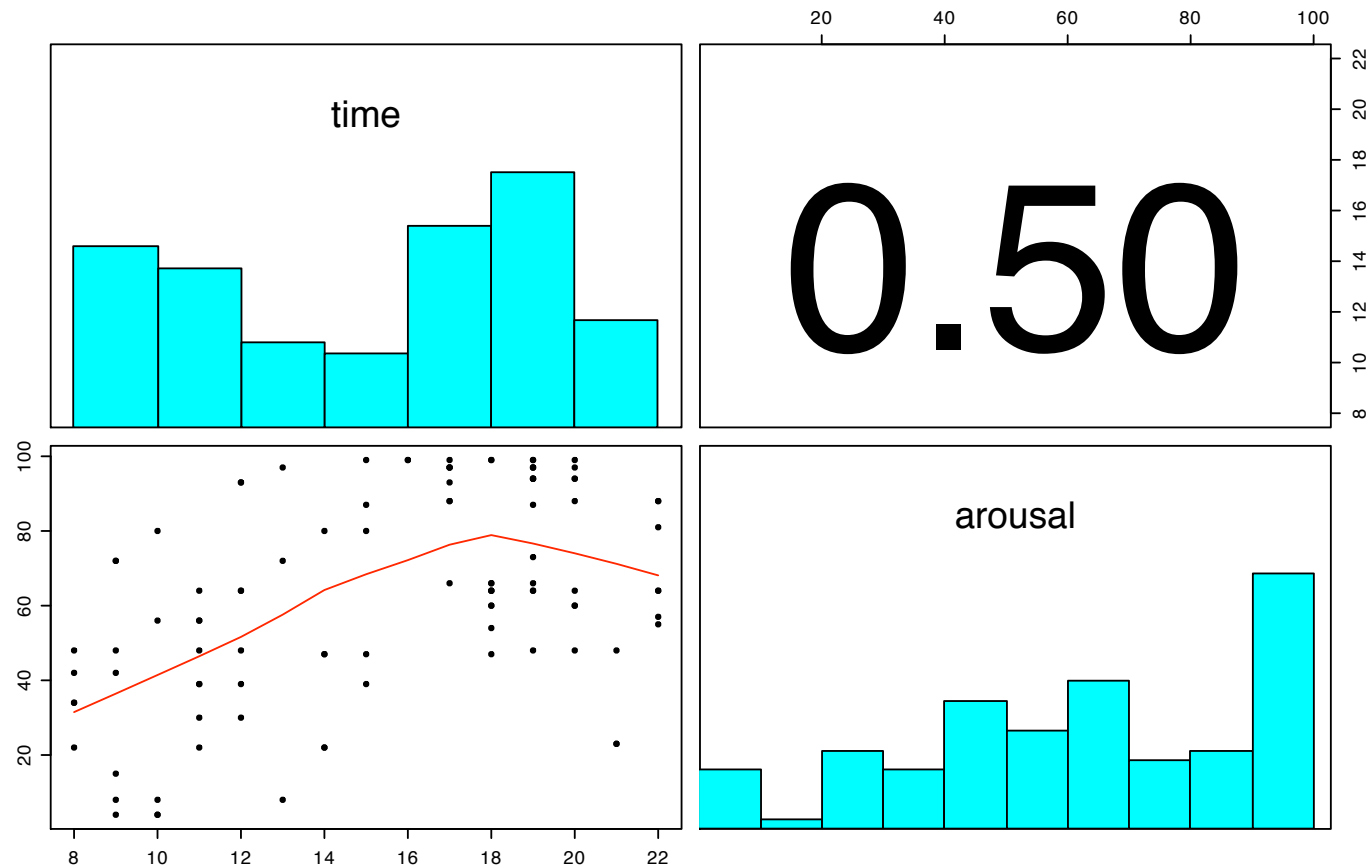
Graphical summary

`pairs.panels(sim_data[2:9])`

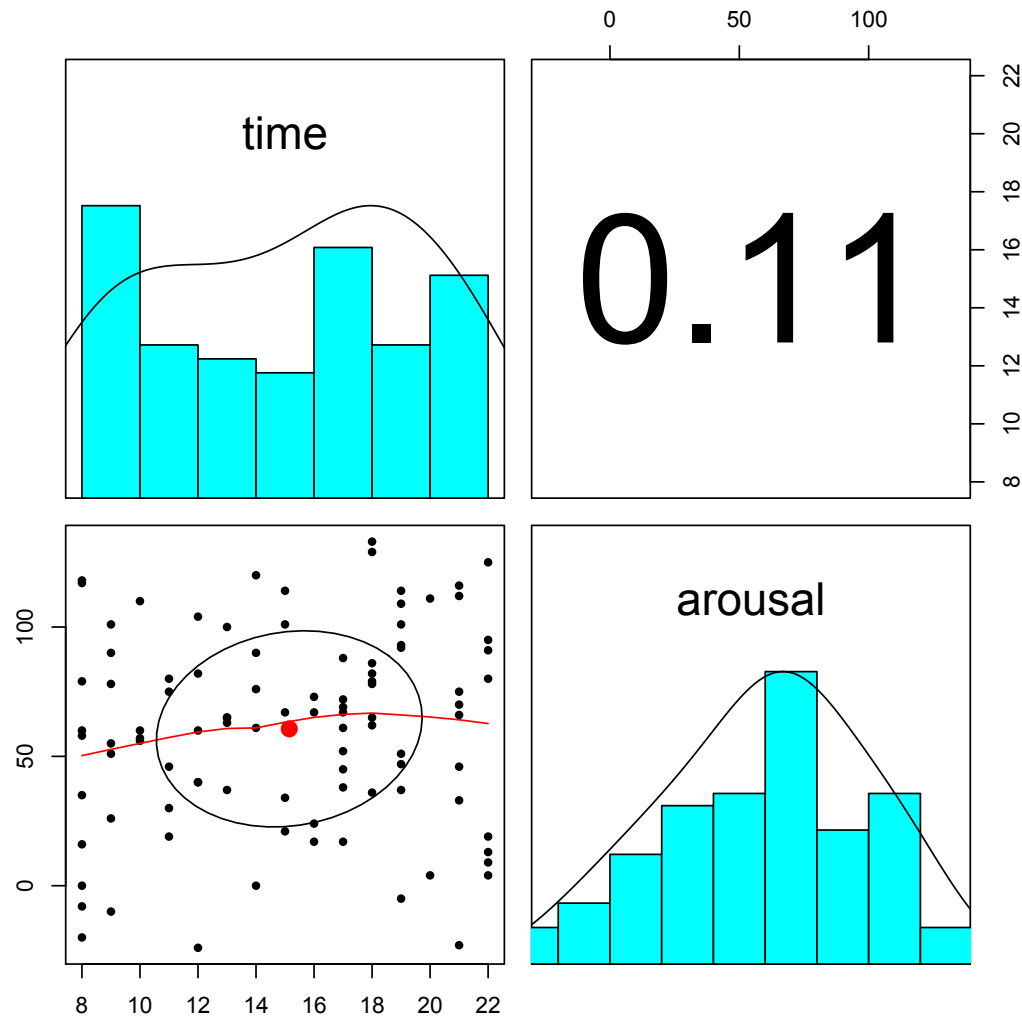


Detailed plots

```
pairs.panels(sim.data[c(4,7)])
```



But, in a replication



Centering the data for regression analysis

```
> cen.data <- data.frame(scale(sim.data,scale=FALSE)) #use the scale function  
> describe(cen.data) #note how the means are now 0
```

	var	n	mean	sd	median	mad	min	max	range	skew	kurtosis	se
snum	1	100	0	29.01	0.00	37.06	-49.50	49.50	99	0.00	-1.24	2.90
sex	2	100	0	0.50	-0.48	0.00	-0.48	0.52	1	0.08	-2.01	0.05
drug	3	100	0	0.50	0.45	0.00	-0.55	0.45	1	-0.20	-1.98	0.05
time	4	100	0	4.32	1.43	4.45	-7.57	6.43	14	-0.28	-1.29	0.43
anxiety	5	100	0	1.91	0.10	1.48	-3.90	4.10	8	0.03	-0.73	0.19
impulsivity	6	100	0	2.14	-0.19	1.48	-4.19	4.81	9	-0.06	-0.31	0.21
arousal	7	100	0	27.88	0.47	35.58	-59.53	35.47	95	-0.40	-0.85	2.79
tension	8	100	0	12.46	-1.13	11.12	-19.63	40.37	60	0.87	0.79	1.25
performance	9	100	0	9.05	1.45	8.90	-25.55	15.45	41	-0.47	-0.43	0.91
cost	10	100	0	0.00	0.00	0.00	0.00	0.00	0	NaN	NaN	0.00

Linear modeling

```
> model1 <- lm(arousal ~ drug * time,data=cen.data)
> summary(model1)
```

Call:

```
lm(formula = arousal ~ drug * time, data = cen.data)
```

Residuals:

Min	1Q	Median	3Q	Max
-45.388	-9.813	1.909	10.036	34.465

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-0.01501	1.44334	-0.010	0.992
drug	38.87331	2.90131	13.399	<2e-16 ***
time	3.41692	0.33596	10.171	<2e-16 ***
drug:time	-0.20428	0.67673	-0.302	0.763

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 14.42 on 96 degrees of freedom

Multiple R-Squared: 0.7405, Adjusted R-squared: 0.7323

F-statistic: 91.29 on 3 and 96 DF, p-value: < 2.2e-16

Data transforms

- Sometimes we want to take continuous variables and refer to them by groups (e.g., high and low impulsivity).
- In order to break continuous variables into discrete categories, we use the cut function. e.g.
 - `imp2 <- cut(impulsivity,c(-1,5,10)) # 2 levels`
 - `tod3 <- cut(time,c(7,12,18,24)) #3 levels`

Range of cut()

- cut takes values greater than lower and up to and including upper value:
- > imp2 <- cut(impulsivity,c(-1,5,12))
- > table(imp2)
- imp2
- (-1,5] (5,12]
- 54 46

But better not to make it discrete

- Using continuous variables as predictors is more powerful than using dichotomous variables.
- This is done using the linear modeling function (lm)

The linear model

- ANOVA is a special case of the linear model with categorical values of an IV
- Continuous values of IV may be considered in regression models.
- Note that if we want to interpret interactions we need to standardize (or at least zero center) the IVs

Analysis of Variance

- $Y = bX + c$ is the linear model
- ANOVA is a special case of the linear model where the predictor variables are categorical “factors” rather than continuous variables.

Make new variables

```
> imp2 <- as.factor(cut(impulsivity,c(-2,5,12)))  
> tod3 <- as.factor(cut(time,c(7,12,18,24)))  
> caff <- as.factor(drug)  
> my.data <- data.frame(sim.data,imp2,tod3,caff)
```

ANOVA in R

```
> model <- aov(arousal ~ caff * tod3, data=my.data)
```

```
> summary(model)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
caff	1	35462	35462	162.0883	< 2.2e-16	***
tod3	2	20659	10329	47.2134	6.387e-15	***
caff:tod3	2	276	138	0.6312	0.5342	
Residuals	94	20566	219			

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

But what does it look like?

```
> print(model.tables(model, "means", digits=2))
```

```
Tables of means
```

```
Grand mean
```

```
63.53
```

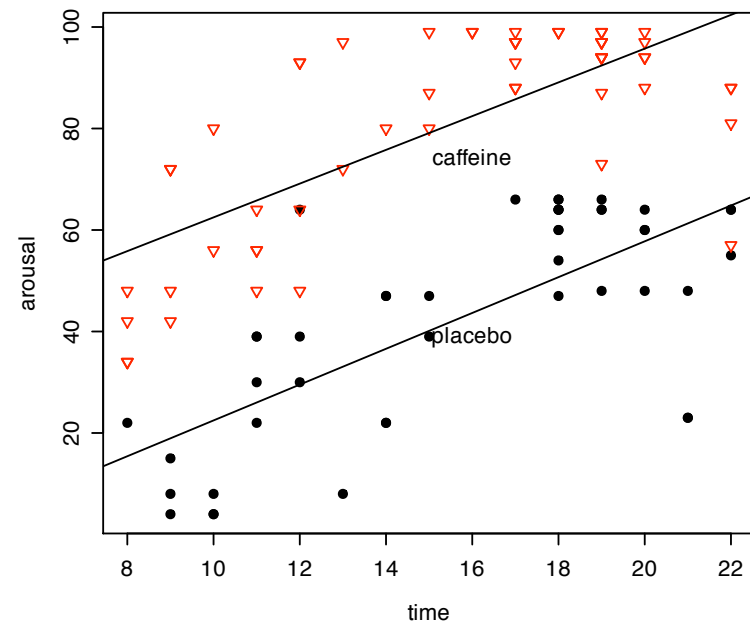
```
caff    0      1
      42.71 80.56
rep 45.00 55.00
tod3
      (7,12] (12,18] (18,24]
      42.59  72.86   73.9
rep 32.00  34.00   34.0
caff:tod3
      tod3
caff  (7,12] (12,18] (18,24]
  0   23.43  49.59  53.64
  rep 14.00  17.00  14.00
  1   58.33  92.35  90.55
  rep 18.00  17.00  20.00
```

Graphics: an example

```
symb=c(19,25,3,23)  =#choose some nice plotting symbols  
colors=c("black","red","green","blue") #choose some nice colors  
attach(my.data)     #make this the active data set
```

```
plot(time,arousal,pch = symb[caff],col=colors[caff])  
by(my.data,caff,function(x) abline(lm(arousal~time, data= x)))  
text(16,40,"placebo")  
text(16,75,"caffeine")
```

Don't use colors for
publications, just slides



Data Analysis

- Exploratory analysis describes the data