Jakub Štenc

Using Statistics in Ecology 2019, homework part two

„Generate a data set" means that you should also describe briefly a biologically realistic setting which may have yielded such data, i.e. the variables must have names and the results should be meaningfully interpreted. Everything which is in the figures must be readable. Everyone will do his/her work independently, you are allowed to ask for advice but direct similarities will be punished. Excessive numerical accuracy is moderately punishable. Please copy the text of each task before you answers. Display your "a" value.

A ← 8 ,
my code is available on github: (do not hesitate to ask me how)

1. Generate biologically relevant data in which chi-squared test would reveal a statistically significant association in a 2x2 (two times two) frequency table. In one of the cells there should be value 6a. Present **the table and the result b**oth as copy-paste from R and as such a **sentence** which you would be OK for a research paper. Present a **table** with identical marginal distributions in which there is no significant association between the variables (this can you do by hand at home).

Data description: I would like to compare frequency of flower phenotypes, which have 2 colour forms (red vs. blue) and forms with different number of anthers (five vs. three) in population.

```
(petal_colour.anther_number    <-    matrix(c(a*6,    10,18,40),2,2,dimnames    =    list(c("five",
"three"),c("red", "blue"))) )
       red blue
five   48   18
three  10   40

chisq.test(petal_colour.anther_number)
          Pearson's Chi-squared test with Yates' continuity correction

data:  petal_colour.anther_number
X-squared = 29.562, df = 1, p-value = 5.414e-08
```

	Red	Blue	**Row marginals**
Five	48	18	**66**
Three	10	40	**50**
Collum marginals	**58**	**58**	**N =116**

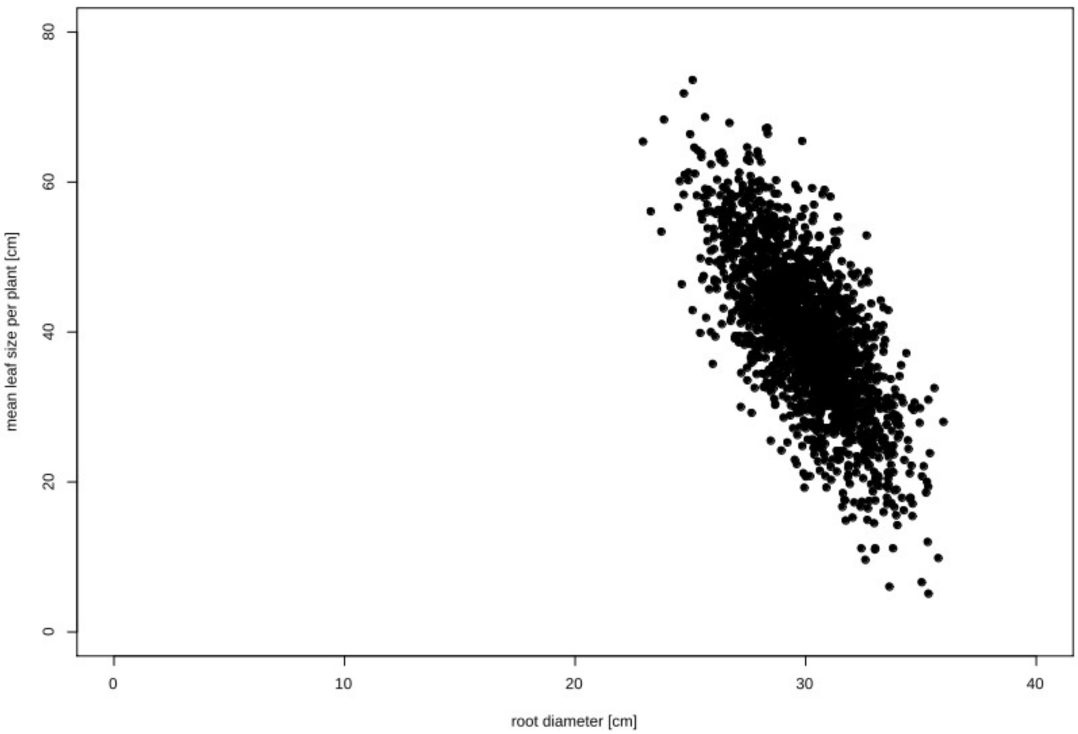
A Chi- square test of independence was calculated comparing the frequency of flower phenotypes in population. A significant interaction was found (χ2 (1) = 29.6, p < .001). Red flowers were more likely to have five anthers (41%) than blue flowers (16%) and blue flowers were more likely to have three anthers (34%) than red flowers (9%).

```
(petal_colour.anther_number.nonsignificat    <-    matrix(c(a*6,    34,42,40),2,2,dimnames    =
list(c("five", "three"),c("red", "blue"))))
       red blue
five   48   42
three  34   40

chisq.test(petal_colour.anther_number.nonsignificat)

	Pearson's Chi-squared test with Yates' continuity correction

data:  petal_colour.anther_number.nonsignificat
X-squared = 0.61562, df = 1, p-value = 0.4327
```

	Red	Blue	**Row marginals**
Five	48	42	**90**
Three	34	40	**74**
Collum marginals	**82**	**82**	**N =164**

A Chi- square test of independence was calculated comparing the frequency of flower phenotypes in population. A interaction was not significant (χ2 (1) = .62, p = .4).

2. Generate a data set in which two continuous independent variables are negatively correlated (-0.65>r>-0.85, **present a graph, and a result of correlation analysis)** and the dependent variable (sample mean between 5a and 6a) depends on both of these variables, when examined one at a time (present graphs). Perform type I analyses with both orders of the variables (do not include the interaction), present results as raw tables (copy-paste from R). Explain the difference in the interpretation of the results. Present the results also for type III analysis but this time as an edited ANOVA table (as you would present it in a research paper). Please do not draw more lines than it is customary for tables in research papers.



Graph of relationship between two independent variables mean leaf size per plant and root diameter. There is a significant strong negative correlation ($r_{(1998)}$= -.7, p = < .001, Pearson).

```
Pearson's product-moment correlation

data:  df$x and df$y
t = -43.814, df = 1998, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.7216883 -0.6769387
sample estimates:
 cor
-0.7
```

There is a significant strong negative correlation between two independent variables mean leaf size per plant and root diameter ($r_{(1998)}$= -.7, p = < .001, Pearson).

```
anova(lm(herbivory~root_dia+m_leaf_size, data=df))
Analysis of Variance Table

Response: herbivory
            Df Sum Sq Mean Sq   F value Pr(>F)    
root_dia     1 233739  233739 7718.1305 <2e-16 ***
m_leaf_size  1     35      35    1.1704 0.2794    
Residuals 1997  60478      30                     
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> # Type I linear model - changed order
> anova(lm(herbivory~m_leaf_size+root_dia, data=df))
Analysis of Variance Table

Response: herbivory
            Df Sum Sq Mean Sq F value    Pr(>F)    
m_leaf_size 1 111673  111673  3687.5 < 2.2e-16 ***
root_dia    1 122102  122102  4031.8 < 2.2e-16 ***
Residuals 1997  60478      30                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

There is difference in ANOVA Type I with different order. It is because predictors (independent variables) are correlated. It means, that we can not answer the question which variable has a direct effect and which has an indirect (mediated) effect. In this example we cannot claim a direct (not mediated by root diameter) effect of mean leaf size on herbivory of plant.

ANOVA type III table for herbivory rate

Source	*df*	*Type III SS*	*F*	*p*
Mean leaf size	1	35	1.2	non-significant
Root diameter	1	122102	4031.9	<.001

And for variables in opposite order:

ANOVA type III table for herbivory rate

Source	*df*	*Type III SS*	*F*	*p*
Root diameter	1	122102	4031.9	<.001
Mean leaf size	1	35	1.2	non-significant

3. Generate data in which the results of a type I ANOVA with two discrete independent variables do not depend at all on the order of the variables in the model (recall **which assumption should be met** to achieve this). Do not include an interaction in the model. Please present the results with the two different orders as raw tables. Calculate the residuals and present graphically frequency distributions of the **residuals for both levels** of one of the independent variables. Are the assumptions of ANOVA met? (Tell why. No worries if not).

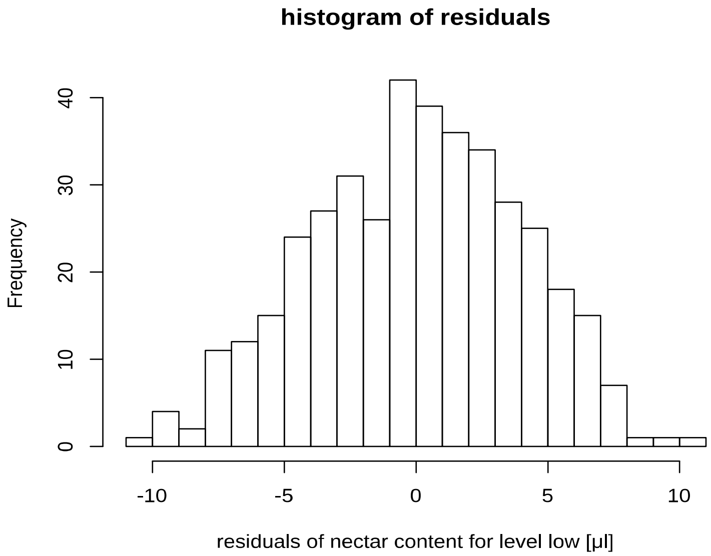
1. Which assumption should be met – there should not be association between independent variables – in case of two discrete variables, they should have equal frequency in (see table).

```
        high low
  blue  200 200
  red   200 200
```

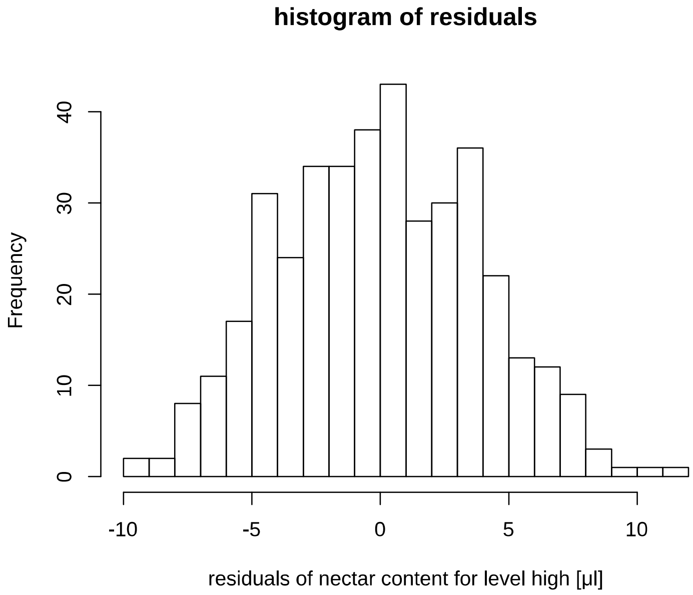
Generate data:
I generate data with scenarios when I want to compare rate of flower visitation (measured as number of flower visitors per hour) on flowers which differs in a) petal colour (red, blue) and nectar content (low, high).

Output of I ANOVA:

```
> mod1 <- (lm(f_visi~nectar_content+petal_colour, data = fl_visitors))
> anova(mod1)
Analysis of Variance Table
Response: f_visi
                Df  Sum Sq Mean Sq F value    Pr(>F)    
nectar_content   1  5269.7  5269.7  341.70 < 2.2e-16 ***
petal_colour     1  5630.2  5630.2  365.07 < 2.2e-16 ***
Residuals      797 12291.4    15.4                      

> mod2 <- (lm(f_visi~petal_colour+nectar_content, data = fl_visitors))
> anova(mod2)
Analysis of Variance Table
Response: f_visi
                Df  Sum Sq Mean Sq F value    Pr(>F)    
petal_colour     1  5630.2  5630.2  365.07 < 2.2e-16 ***
nectar_content   1  5269.7  5269.7  341.70 < 2.2e-16 ***
Residuals      797 12291.4    15.4                      
```



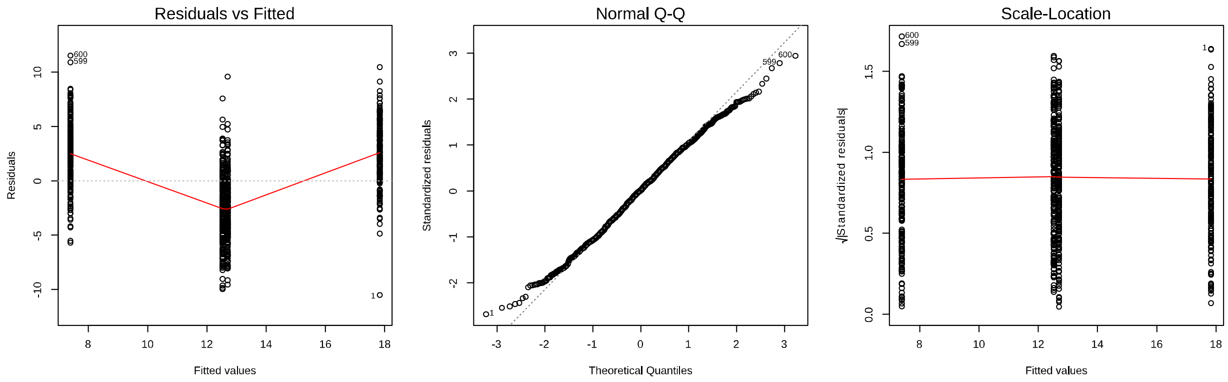
Histogram of residuals of nectar content for level low [µl] from the second ANOVA model.



Histogram of residuals of nectar content for level high [µl] from the second ANOVA model.

Assumptions of ANOVA are met:

1. *Independence of observations - I measured randomly selected flowers across the galaxy (trust me).*
2. *Normality in distribution of residuals - residuals are in this case normally distributed - as you can see on the second diagnostic plot or on histogram of residuals.*
3. *Homoscedasticity - variance in levels of predictors should be more or less equal.*



Diagnostic plots of the second ANOVA model. It is clear that assumptions of ANOVA are met.

4. Examine section „Analysis of covariance – one important application“ (topic 7) and present an example from the ecology of fish (what was studied and why?) in which the described logic applies. Discuss the setting and explain, how would you interpret the results of the analyses in which the covariate 1) is not, 2) is included in the model and in the last case 1) will or 2) will not attain significance. No need to perform the analysis but present a hypothetical table of results.

Our study investigate influence of nitrogen concentration in water on reproduction of fish. Nitrogen concentration in water was for long time in suspicion for negative impact on fish populations. In this study was selected one species and number of eggs laid per female was measured by using a diving cylinder in artificial ponds, which do not differ in anything else than in nitrogen content. Length of females was used as covariate (all females in analysis was measured underwater in order to avoid any harm). Fortunately is length of females normally distributed in all ponds.

Results of ANOVA with covariate is not included:

Table
ANOVA Summary Table for Variable 1

Source	*df*	MS	*F*	*p*
Nitrogen conc.	1	220.22	2.56	.06
Total	99			

R^2 = .18

There is not a significant relationship between fish reproduction (number of eggs per female) and nitrogen concentration (F(1,99) = 2.56, p = .06).

#Results of ANOVA with included covariate:
Table
ANOVA Summary Table for Variable 1

Source	*df*	MS	*F*	*p*
Female length	1	420.02	7.03	.01
Nitrogen conc.	1	30.22	3.56	.02
Total	98			

R^2 = .29

There is a significant effect of nitrogen concentration in water on fish reproduction after controlling for the effect of fish length (F (1,98) = 3.56, p = 0.2)

#Results of ANOVA with included covariate:

Table

ANOVA Summary Table for Variable 1

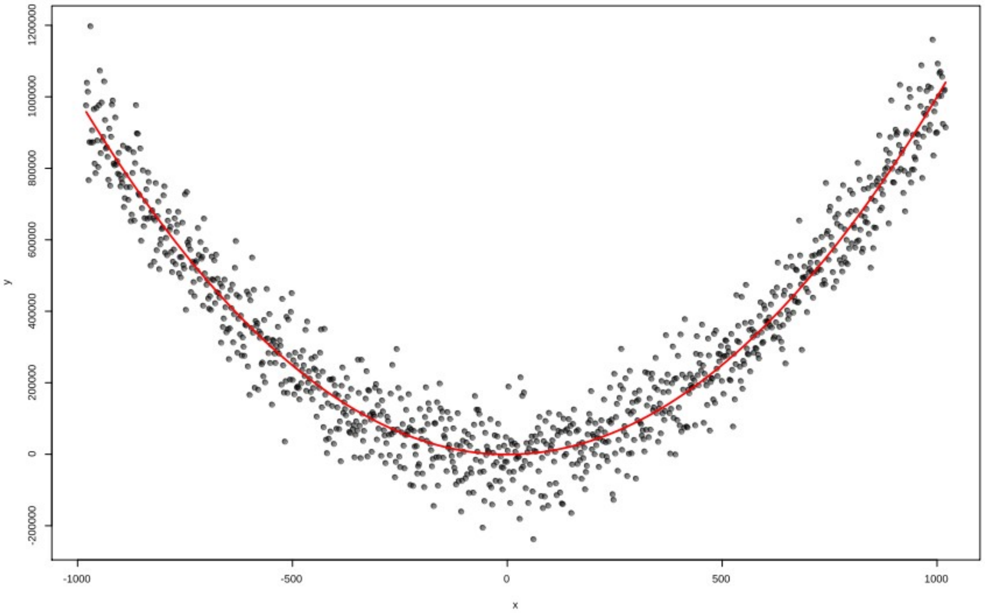
Source	*df*	MS	*F*	*p*
Female length	1	120.02	1.03	.1
Nitrogen conc.	1	230.22	1.56	.2
Total	98			

R^2 = .29

There is not a significant effect of nitrogen concentration in water on fish reproduction after controlling for the effect of fish length (F (1,98) = 1.56, p = 0.2)

5. Generate data in which the relationship between two variables deviates significantly from linearity and the relationship is convex. The mean of the sample of the independent variable should be between 2a and 3a. Present the test of non-linearity (raw table), the graph and the equation of the relationship (edited).

```
Analysis of Variance Table

Response: y
           Df     Sum Sq    Mean Sq   F value    Pr(>F)    
x           1 5.7149e+11 5.7149e+11    90.981 < 2.2e-16 ***
I(x^2)      1 8.9097e+13 8.9097e+13 14184.367 < 2.2e-16 ***
Residuals 997 6.2625e+12 6.2813e+09                        
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



Coefficients:

```
 (Intercept)            x       I(x^2) 
-528.2689326    1.3977949    0.9991705 
```

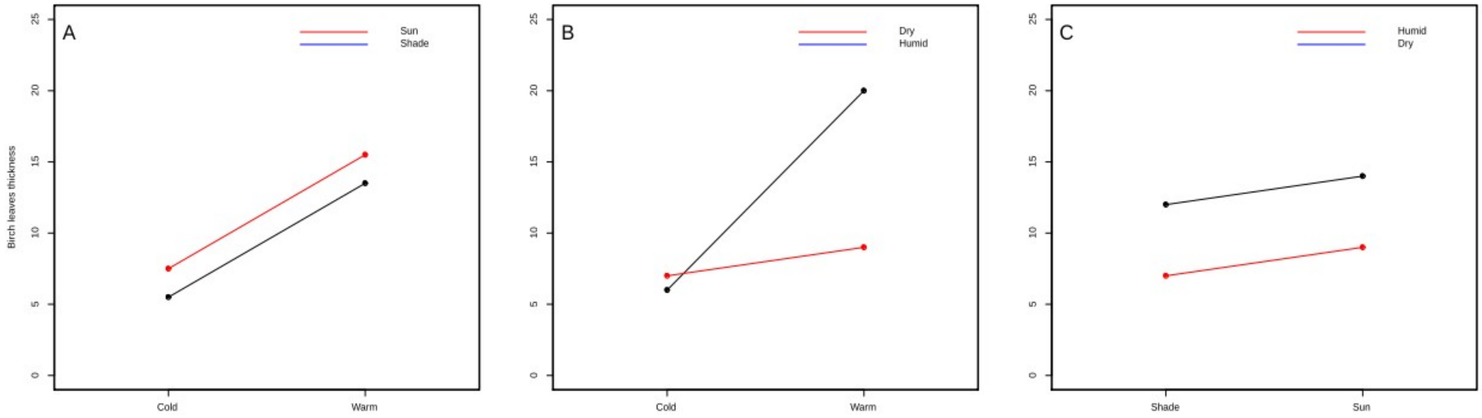
Equation:

$y = -528.2689 + 1.3977949 x – 0.9991705x^2$

6. The thickness of birch leaves in shade, humid and warm conditions is a. Please propose the values of leaf thickness for all combinations of illumination (shade/sun), humidity (humid/dry) and temperature (warm/cold) so that there were one and only one two-factor interaction in the data, a three-way interaction should not be there either. Present the situation graphically (you may draw by hand). How would you verbally formulate a three-way interaction if it still was there?

Mean values of birch leaves thickness in different conditions

	Shade		Sun	
	humid	dry	humid	dry
Warm	8	19	10	21
Cold	6	5	8	7



Graphs describing relations between variables. Only graph B shows two-way interaction between humidity and temperature.

How to formulate three-way interaction:
Thickness is dependent on levels of all three predictors. Ergo: thickness in different shade condition dependents on levels of humidity and temperature, in different humidity levels dependents on levels of temperature and shade and in different temperature levels dependents on levels of humidity and shade.

7. Describe a fictional situation from bird ecology in which you would use a mixed ANOVA with one random (not "brood", invent something different) and one fixed factor. Explain, why are the factors treated as fixed/random. Generate the data and present the result as a raw table. Would it be reasonable to alternatively treat one of the random factors as fixed. If not, then why? If yes, how would it change the interpretation of the results?

Birds of two closely related sympatric species were observed in Spain and amount of mites per bird was counted. Aim of the study was investigate if there is difference in bird species infestation. For the analyses was used species as fixed effect and study site identity was used as random effect.

Species identity was treated as fixed factor, because aim of the analyses was investigated difference between species.
Study site identity was treated as random effect, because birds from different sites differs in the number of mites (probably because some sites were more suitable for mites of species A and some for mites of species B).

Analysis with fixed and random factors

```
> anova(  lme(mites~spec , random = ~1|site, data=bi))
            numDF denDF  F-value p-value
(Intercept)     1   395 453.9841  <.0001
spec            1   395 625.0603  <.0001
> anova(  lm(mites~spec +site, data=bi))
Analysis of Variance Table
```

Analysis with both factors as fixed

```
Response: mites
           Df Sum Sq Mean Sq F value    Pr(>F)    
spec        1  47415   47415 625.060 < 2.2e-16 ***
site        3   9439    3146  41.477 < 2.2e-16 ***
Residuals 395  29963      76                      
---
```

It could be reasonable to treat one of the random factor as fixed – in scenario where we want to test difference between sites. If there is difference between sites

8. There were 6 turtles, 3 were fed with snails, 3 were not (two treatments). Shell thickness on the turtles was measured in one, two, three and four years starting from the beginning of the treatment. The researchers are interested whether there is some overall effect of the treatment on shell thickness, and whether the dynamics of shell thickness in time differs between the treatments. Please find the data as an excel sheet, first please analyse the data with an ordinary ANOVA (treatment and time) not accounting for the fact that the same individuals were repeatedly measured, and then as a correct repeated measurements ANOVA. Present the results of **both analyses as raw tables,** find the differences (also in degrees of freedom) and explain what do they come from. Please note that in repeated measurements ANOVA, the number of error degrees of freedom may differ for different effects. Formulate the result of the repeated measurements ANOVA as a biologically meaningful sentence.

Analysis without repeated measuremens

```
 anova(turmod)
Analysis of Variance Table

Response: shell
               Df Sum Sq Mean Sq F value Pr(>F)
treatment       1  2.042  2.0417  1.0515 0.3174
time            1  1.008  1.0083  0.5193 0.4795
treatment:time  1  0.075  0.0750  0.0386 0.8462
Residuals      20 38.833  1.9417
```

Analysis with repeated measuremens

```
> anova(turmod2)
               numDF denDF   F-value p-value
(Intercept)        1    16 142.79091  <.0001
treatment          1     4   2.96350  0.1603
time               1    16   0.54053  0.4729
treatment:time     1    16   0.01608  0.9007
```

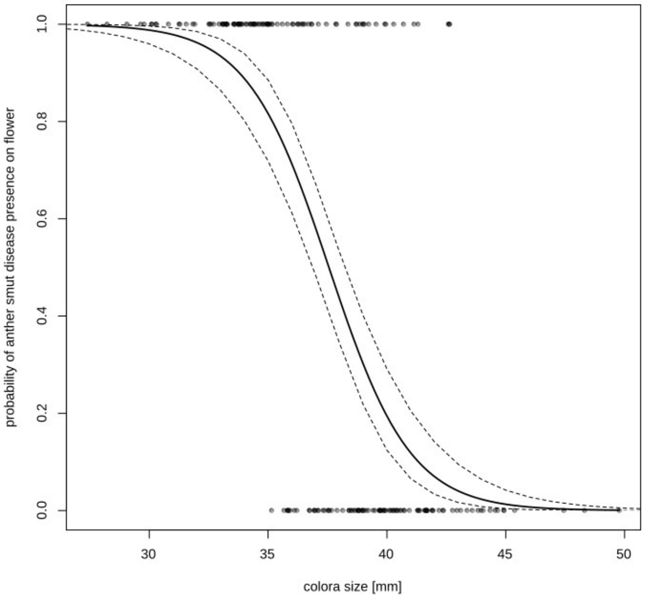
Effect of turtle diet does not change in time (p=.9). Furthermore, there is no difference between turtle diet (p=.16) and shell thickness does not change in time (p=.47).

The differences comes from the fact, that measurements of individuals are not treated as independent observations, because they would be affected from previous state.

9. Generate data which you would analyse using the method of logistic regression and where the parameter b is in the range -0.7<b<-0.5. Perform the analysis, present the figure, significance test in the raw form and the equation of the logistic function (edited). Present the result as a biologically informative sentence as you would do it in a research paper.

How is presence of anther smut disease in flowers of Dianthus carthusianorum affected by corolla size?

```
glm(formula = y ~ x1, family = binomial)
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.80250  -0.67773   0.01658   0.55096   2.45444
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 21.90020    2.97940   7.351 1.97e-13 ***
x1          -0.58309    0.07909  -7.373 1.67e-13 ***
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 277.26  on 199  degrees of freedom
Residual deviance: 166.18  on 198  degrees of freedom
AIC: 170.18
Number of Fisher Scoring iterations: 5

Analysis of Deviance TableModel: binomial, link: logit
Response: y
Terms added sequentially (first to last)

     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                   19     27.726
x1    1   11.862        18     15.864 0.000573 ***
 m5
Call:  glm(formula = y ~ x1, family = binomial)
Coefficients:
(Intercept)           x1
    -12.039        1.082
Degrees of Freedom: 19 Total (i.e. Null);  18 Residual
Null Deviance:      27.73
Residual Deviance: 15.86        AIC: 19.86
```

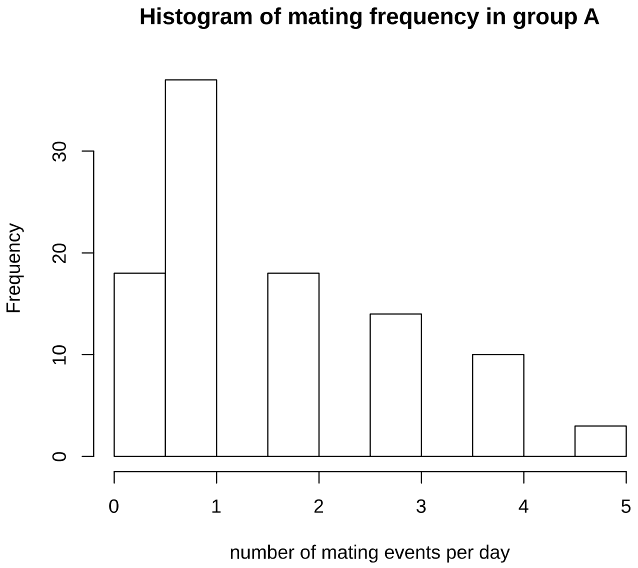


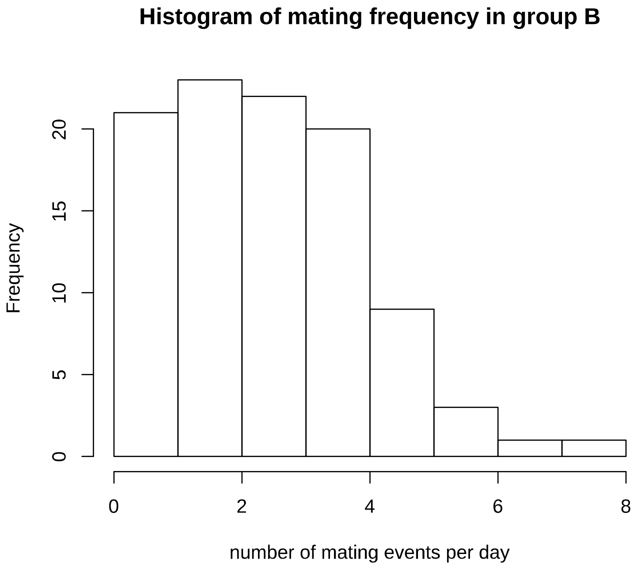
Flower corolla size had significant effect on presence of anther smut disease in flowers of Dianthus carthusianorum (p < .001)

10. Present a fictional case in which you would compare a Poisson distributed variable between two populations of monkeys (in one of the groups, let it be μ = 1 + a/10 (± 10%)) Generate respective data and present these distributions graphically. Present tests to show that the generated distributions do not deviate significantly from Poisson distribution. Present a test for the among-population differences in the values of these variables. Present the results of the test both as raw output and as a meaningful sentence.

```
  gf = goodfit(ska[ska$group == "B",]$mat_freq,type= "poisson",method= "ML")
> gf.summary = capture.output(summary(gf))[[5]]
> pvalue = unlist(strsplit(gf.summary, split = " "))
> pvalue = as.numeric(pvalue[length(pvalue)]); pvalue
[1] 0.09868836 – I do not rejected H0

 gf = goodfit(ska[ska$group == "A",]$mat_freq,type= "poisson",method= "ML")
> gf.summary = capture.output(summary(gf))[[5]]
> pvalue = unlist(strsplit(gf.summary, split = " "))
> pvalue = as.numeric(pvalue[length(pvalue)]); pvalue
[1] 0.9112508 – I do not rejected H0


Model:
mat_freq ~ group
       Df Deviance    AIC    LRT  Pr(>Chi)
<none>      224.45 705.78
group   1   253.84 733.16 29.384 5.938e-08 ***
```

There is a significant difference between apes groups in mating frequency per day (p < .001).


Histogram of mating frequency in group A


Histogram of mating frequency in group B

11. Describe a situation when, studying domestic cats, you should „merge“ three variables into one principal component (please avoid overlap with the example presented in the lecture, invent something by yourself). Present the interpretation of the new variable. Why cannot it be measured directly? Generate the data and find out whether the new variable correlates with some fourth one (present raw results of the correlation analysis). Present a table showing the values of the original variables and the calculated PCA component score.

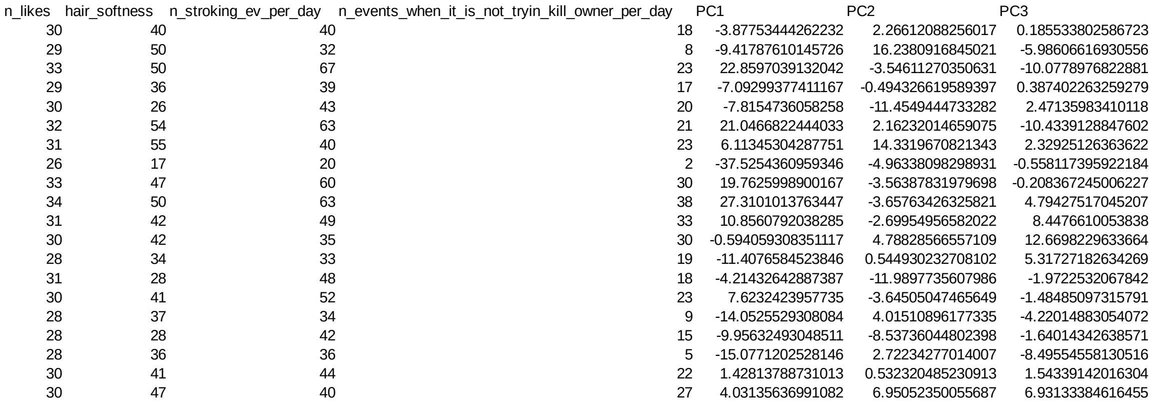
I studied relation between number of positive values obtained by cat’s pictures on internet and their cuteness which was described by using principal component from their hair softness score, number of stroking events per day and number of events when they are not trying kill their owner.

```
shapiro.test(cuteness$x[,1])

        Shapiro-Wilk normality test

data:  cuteness$x[, 1]
W = 0.95939, p-value = 0.5317

shapiro.test(df$n_likes)

        Shapiro-Wilk normality test

data:  df$n_likes
W = 0.96017, p-value = 0.5473
```

- it is possible to use Pearson correlation test, because both variables are not significantly different from normal distribution.

```
> cor.test(cuteness$x[,1], df$n_likes)

        Pearson's product-moment correlation

data:  cuteness$x[, 1] and df$n_likes
t = 12.905, df = 18, p-value = 1.552e-10
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8755055 0.9803668
sample estimates:
      cor
0.9499772
```

Highly significant and strong correlation between numbers of positive marks on internet and PC1 which describes cuteness of cats (number of likes) (r(df = 18) = .9, p < .001)

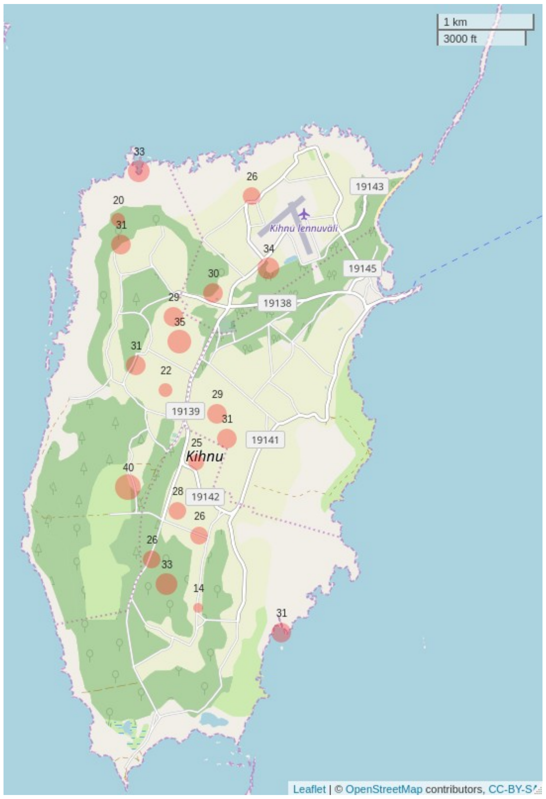


n_likes	hair_softness	n_stroking_ev_per_day	n_events_when_it_is_not_tryin_kill_owner_per_day	PC1	PC2	PC3
30	40	40	18	-3.87753444262232	2.26612088256017	0.185533802586723
29	50	32	8	-9.41787610145726	16.2380916845021	-5.98606616930556
33	50	67	23	22.8597039132042	-3.54611270350631	-10.0778976822881
29	36	39	17	-7.09299377411167	-0.494326619589397	0.387402263259279
30	26	43	20	-7.8154736058258	-11.4549444733282	2.47135983410118
32	54	63	21	21.0466822444033	2.16232014659075	-10.4339128847602
31	55	40	23	6.11345304287751	14.3319670821343	2.32925126363622
26	17	20	2	-37.5254360959346	-4.96338098298931	-0.558117395922184
33	47	60	30	19.7625998900167	-3.56387831979698	-0.208367245006227
34	50	63	38	27.3101013763447	-3.65763426325821	4.79427517045207
31	42	49	33	10.8560792038285	-2.69954956582022	8.4476610053838
30	42	35	30	-0.594059308351117	4.78828566557109	12.6698229633664
28	34	33	19	-11.4076584523846	0.544930232708102	5.31727182634269
31	28	48	18	-4.21432642887387	-11.9897735607986	-1.9722532067842
30	41	52	23	7.6232423957735	-3.64505047465649	-1.48485097315791
28	37	34	9	-14.0525529308084	4.01510896177335	-4.22014883054072
28	28	42	15	-9.95632493048511	-8.53736044802398	-1.64014342638571
28	36	36	5	-15.0771202528146	2.72234277014007	-8.49554558130516
30	41	44	22	1.42813788731013	0.532320485230913	1.54339142016304
30	47	40	27	4.03135636991082	6.95052350055687	6.93133384616455

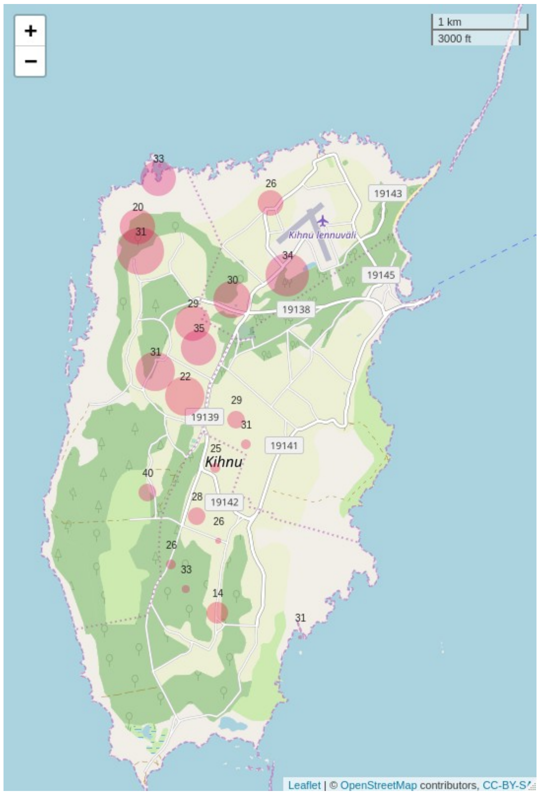
Raw table with original data and PCA output.

12. Describe a fictional ecological study in which you have studied the dependence of one continuous variable on another, so that the values of both variables have recorded from 20 sites on Kihnu island. There should be spatial autocorrelation (SAC) in the independent variable but not in the dependent variable. Present a map displaying the values of both variables for all study sites. Present a verbal interpretation of the SAC. What may have caused the SAC?

We studied relation between plant seed production and nutrient content on 20 sites on Kihnu island. The nutrient content is strongly spatially correlated, possibly because there is some historical influence in management of sites and the northern part of the Kihnu island is more nutrient rich. Seed production is independent on nutrient content, because it is strongly affected by other factors (number of mates, herbivory rate etc.).



Map of the Kihnu island. Points represent sites, numbers represent mean number of seeds and size of points represent number of seeds on the site.



Map of the Kihnu island. Points represent sites, numbers represent mean number of seeds and size of points represent number nutritient content on sites.

** end of tasks ******************************