UKSta18_Liang_Ex09

August 22, 2018

1 Student's t-test

On Moodle you find a copy of Student's original paper in which he introduced the t-statistic and the t-test. On page 20 he starts giving illustrations of his method. Have a look at his *Illustration I*, in particular the table (it obviously contains a typo in the 5th row). * a: What is apparently his null hypothesis? Can you reproduce the odds values that he gives? For that, implement the one-sample t-test yourself, and perform a one-sided test.

His H_0 is that the treatment(1) has positive effect on increasing sleeping hours.

- b: In which cases can you reject the null hypothesis with a 1% significance level? What are the critical values of the t-statistics? What are the corresponding values in terms of x?
- c: Can you reproduce your result with R's *t.test()* function? Hint: when considering the differences, try the "paired t-test".

```
In [ ]: # a) t-test
        slp1 <- c(0.7, -1.6, -0.2, -1.2, -0.1, 3.4, 3.7, 0.8, 0.0, 2.0)
        slp2 <- c(1.9, 0.8, 1.1, 0.1, -0.1, 4.4, 5.5, 1.6, 4.6, 3.4)
```

2 Correlation test: Hubble expansion

Hubble's 1929 discovery of the proportionality between the distance and the recession velocity of galaxies marked the beginning of the era of observational cosmology.

- a: Generate a perfectly correlated dataset x_i, y_i (e.g. using *seq()*), and a totally uncorrelated dataset (e.g., using *rnorm()*). Compute the Spearman rank correlation coefficient r_S for both datasets.
- b: Load the Hubble dataset (file *Hubble.dat*) into R and compute r_S.
- c: For all three datasets, carry out a test of the null hypthesis that the x and y values are not correlated (use *cor.test(x,y,method="spearman")*). In the case of the Hubble dataset you get a Warning message - why?
- d: Carry out a test of the null hypothesis that the distances of the 24 galaxies in the Hubble dataset are not correlated with the first 24 Fibonacci numbers. What do you find, and where does this result come from? **Hint**: To calculate the Fibonacci sequence, e.g. start with a vector c(0,1) and use a *for*-loop to repeatedly *append(...)* the sum of the last two entries. Alternatively, instead of the Fibonacci sequence you could also use the prime numbers from the first problem sheet.

```
In [1]: # a) c)
        xi <- seq(0, 100, by=0.1)
        yi <- seq(-50, 50, by=0.1)
        cor.test(xi, yi, alternative="greater", method="spearman")
        cat("r_S = ",1-6*sum((rank(xi)-rank(yi))^2)/(length(xi)^3-length(xi)), "\n")
        xir <- rnorm(100)
        yir <- rnorm(100)
        cor.test(xir, yir, alternative="greater", method="spearman")
        cat("r_S = ", 1-6*sum((rank(xir)-rank(yir))^2)/(length(xir)^3-length(xir)), "\n")
        # b) c)
        hub <- read.table("Hubble.dat",header=TRUE)
        dist <- hub$dist
        vel <- hub$vel
        cor.test(dist, vel, alternative="greater", method="spearman")
        cat("Hubble dataset: r_S = ", 1-6*sum((rank(dist)-rank(vel))^2)/(length(dist)^3-length
        cat("We get the warning message because there are identical numbers in the value of di
        so that the ranks are not unique anymore \n")
        # d)
        fibb <- c(0,1)
        for(j in 3:length(dist)){
            fibb <- c(fibb, fibb[j-2]+fibb[j-1])
        }
        cor.test(dist, fibb, alternative="greater", method="spearman")
        cat("The test result shows us that the distances of the 24 galaxies are highly correla
        \n")
```

```
Spearman's rank correlation rho

data:  xi and yi
S = 0, p-value < 2.2e-16
alternative hypothesis: true rho is greater than 0
sample estimates:
rho
  1
```

```
r_S =  1
```

```
Spearman's rank correlation rho

data:  xir and yir
S = 165250, p-value = 0.4669
alternative hypothesis: true rho is greater than 0
sample estimates:
        rho
0.008412841
```

```
r_S =  0.008412841
```

```
Warning message in cor.test.default(dist, vel, alternative = "greater", method = "spearman"):
"Cannot compute exact p-value with ties"
```

```
Spearman's rank correlation rho

data:  dist and vel
S = 295.45, p-value = 1.477e-08
alternative hypothesis: true rho is greater than 0
sample estimates:
      rho
0.8715442
```

```
Hubble dataset: r_S =  0.8717391
We get the warning message because there are identical numbers in the value of distance, which
so that the ranks are not unique anymore
```

```
Warning message in cor.test.default(dist, fibb, alternative = "greater", method = "spearman"):
"Cannot compute exact p-value with ties"
```

```
Spearman's rank correlation rho

data:  dist and fibb
S = 18.527, p-value < 2.2e-16
alternative hypothesis: true rho is greater than 0
sample estimates:
      rho
0.9919449
```

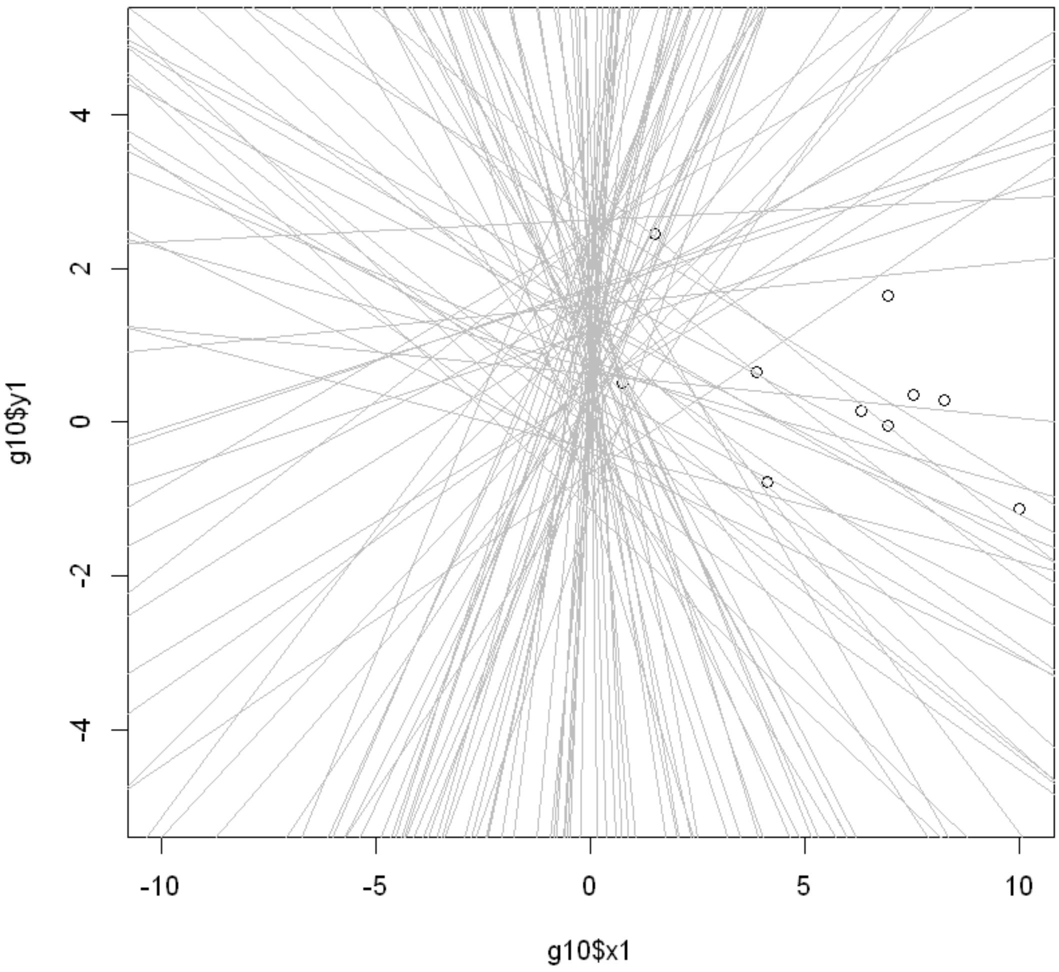
```
The test result shows us that the distances of the 24 galaxies are highly correlated to 24 Fib
```

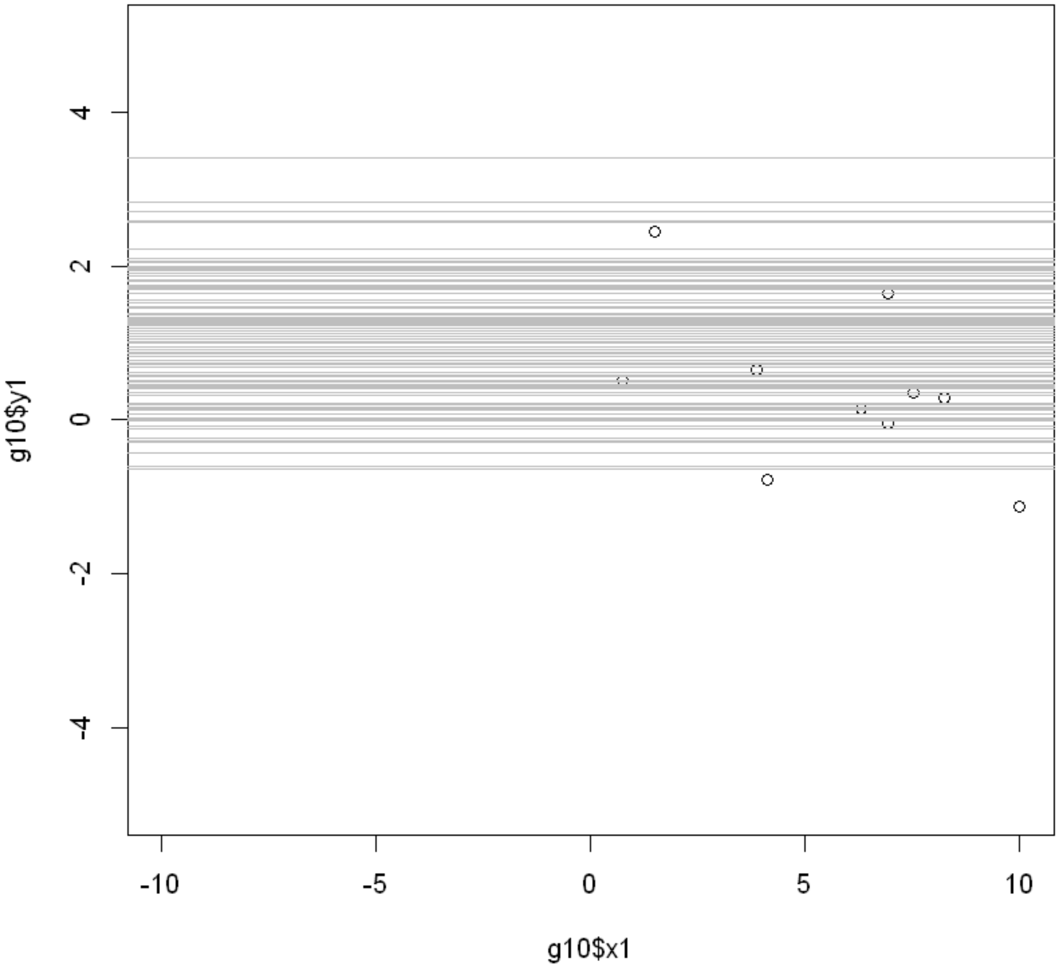
3 Bayesian model comparison: Is there evidence for a non-zero gradient?

This exercise is a copy of the problem discussed in section 11.3 in the book of Coryn Bailer-Jones (see book chapter on Moodle). Read the section, and re-do the model comparison described there! Different from the problem in the section, the files_ gradient10.dat_ and *gradient50.dat* contain new (x,y) data sets for the (in total five each) lines to be analyzed. The coefficients of the straight lines have changed a little bit, and are kept secret here. However, the priors that Coryn species are still applicable, except for $a_0$where $P(a_0)N(1,1)$ should be used. You are expected to provide your own R-implementation of the problem, but – of course – you can study the code given in the book for inspiration. The result should be similar to what is given in the book. Can you confirm this?

```
In [2]: g10 <- read.table("gradient10.dat", header=TRUE)
        g50 <- read.table("gradient50.dat", header=TRUE)
        # define log likelihood
        # M_2: y = b_0 + x tan(alpha) + epsilon, epsilon ~ U(0, sigma)
        # P(b_0) = N(1,1), P(alpha) = U(0, 2pi), P(log(epsilon)) = U(log(0.5),log 2)
        # M_1: y = b_0 + epsilon    (alpha = 0)

        ### Define likelihood

        # Return log10(likelihood), a scalar.
        # theta is the vector of model parameters, here c(b_0, alpha, log10(ysig)).
        # data is the two-column matrix [x,y].
        # dnorm(..., log=TRUE) returns log base e, so multiply by 1/ln(10) = 0.434 to get log
        log.like <- function(theta, data) {
            # convert alpha to b_1 and log10(ysig) to ysig
            theta[2] <- tan(theta[2])
            theta[3] <- 10^theta[3]
            # likelihood
            modPred <- drop( theta[1:2] %*% t(cbind(1,data$x)) )
            # Dimensions in mixed vector/matrix products: [Ndat] = [P] %*% [P x Ndat]
            logLike <- (1/log(10))*sum( dnorm(modPred - data$y, mean=0, sd=theta[3], log=TRUE)
            return(logLike) }

        Nsamp <- 1e5
        priorSamp <- cbind(rnorm(n=Nsamp, mean=1, sd=1),
                           runif(n=Nsamp, min=-pi/2, max=pi/2),
                           runif(n=Nsamp, min=log10(0.5), max=log10(2)))
        sel <- sample.int(n=Nsamp, size=100)
        plot(g10$x1, g10$y1, xlim=c(-10,10), ylim=c(-5,5))
        for(j in sel) {
            abline(a=priorSamp[j,1], b=tan(priorSamp[j,2]), col="grey") }
        priorSamp2 <- cbind(rnorm(n=Nsamp, mean=1, sd=1),
                           0,
                           runif(n=Nsamp, min=log10(0.5), max=log10(2)))
        sel <- sample.int(n=Nsamp, size=100)
        plot(g10$x1, g10$y1, xlim=c(-10,10), ylim=c(-5,5))
```

```
for(j in sel) {
    abline(a=priorSamp2[j,1], b=tan(priorSamp2[j,2]), col="grey") }
```





4 Laterality among physicists

Combining a poll from 2015, 2016, 2017, and 2018 on the laterality of the participants of UKSta we obtained the following contingency table:

	left-handed	right-handed
left-footed	2	13
right-footed	6	38

The sample is not big but what can one say about the connection between handedness and footedness based on the given data?

- a: Perform Fisher's exact test. Is there a connection between handedness and footedness?

  The test result is to be seen as below. As we can see, the odds ratio is less than 1, so we reject the null hypothesis of independence. Thus there exists correlation between handedness and footedness.

- b: Can one say something about preferences if one restricts oneself to the subgroups of left-handed and right-handed persons?

  In both subgroups one tends to be right-footed.

```
In [3]: # a)
        fisher.test(rbind(c(2,13),c(6,38)), alternative="less")
```

```
Fisher's Exact Test for Count Data

data:  rbind(c(2, 13), c(6, 38))
p-value = 0.6734
alternative hypothesis: true odds ratio is less than 1
95 percent confidence interval:
 0.000000 5.041291
sample estimates:
odds ratio
 0.9747855
```

5 Central limit theorem and Kolmogorov–Smirnov

During the lectures you checked the convergence towards a normal distribution according the central limit theorem by a χ^2-test. Do this now with the Kolmogorov–Smirnov test (making use of *ks.test()* available in R!

```
In [4]: X <- replicate(10000, rnorm(25,mean=60,sd=40))
        X_bar <- colMeans(X)
        ks.test(X_bar[sample(1:length(X_bar),1)], "pnorm")
```

```
One-sample Kolmogorov-Smirnov test

data:  X_bar[sample(1:length(X_bar), 1)]
D = 1, p-value < 2.2e-16
alternative hypothesis: two-sided
```