Design of Experiments: Lab 1

Exercise 1. [Adapted from Lloyd (1999)]. A surgery for a condition is performed in two hospitals: hospital A and hospital B. We are interested in comparing the survival rates of the hospitals. The code below will read in the data we will use for this exercise.

```
cond = matrix(c("A", "critical", "survived",
                "A", "critical", "died",
                "A", "noncritical", "survived",
                "A", "noncritical", "died",
                "B", "critical", "survived",
                "B", "critical", "died",
                "B", "noncritical", "survived",
                "B", "noncritical", "died"), ncol = 3, byrow = T)
colnames(cond) = c("Hospital", "Status", "Outcome")
df = as.data.frame(cond)
df$Total = c(17, 101, 100, 3, 2, 36, 175, 8)
```

a) Compare the (marginal) survival rates of hospitals A and B, ignoring the status of the patients. Comment on what you see.

Solution: The survival rate of hospital A is about 53%, whereas the survival rate of hospital B is about 80%. At first glance, it would seem like hospital B is "better" than hospital A.

```
library(tidyverse)
hosp = df %>% uncount(Total)
100*prop.table(table(hosp$Hospital, hosp$Outcome), 1)
```

```
##
##         died survived
##   A 47.05882 52.94118
##   B 19.90950 80.09050
```

b) Now, find the mortality rates by hospital for critical and noncritical patients, separately. Compare the results to what you found in part 1. Why is this happening?

Solution: For patients in critical condition, the survival rate is 14.4% in hospital A and 5% in hospital B. For patients in noncritical conditions, the survival rate is 97% in hospital A and 95% in hospital B. This "contradicts" what we saw in part a). However, in part a) we missed a critical part: 53% of the patients in hospital A are in critical condition, whereas only 17% of the patients in hospital B are in critical condition. The `status` of the patients is a confounding variable.

```
100*prop.table(table(hosp$Hospital, hosp$Outcome, hosp$Status), c(1, 3))
```

```
## , ,  = critical
##
##
##          died  survived
##   A 85.593220 14.406780
##   B 94.736842  5.263158
##
## , ,  = noncritical
##
##
##          died  survived
##   A  2.912621 97.087379
##   B  4.371585 95.628415
```

```
100*prop.table(table(hosp$Hospital, hosp$Status), 1)
```

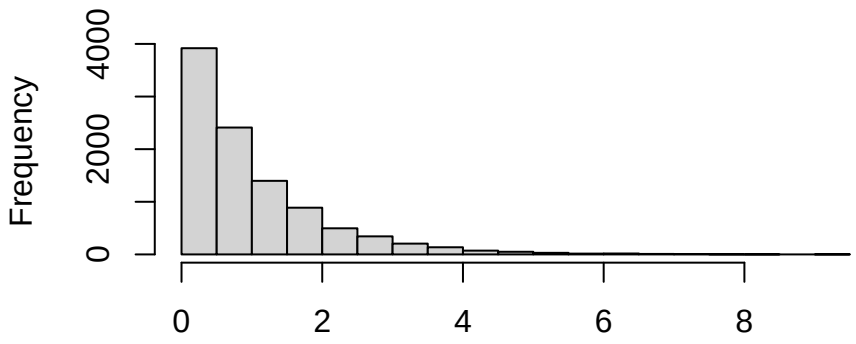
```
##
##     critical noncritical
##   A 53.39367    46.60633
##   B 17.19457    82.80543
```

Exercise 2. The goal of this exercise is reviewing the central limit theorem, which states that, under mild conditions, sample averages are approximately normal.

a) Draw a sample of size 10000 from an Exponential(1) distribution and create a histogram. Does it look approximately normal?

Solution: The Exponential(1) distribution is positive and skewed. It doesn't look like the normal distribution.

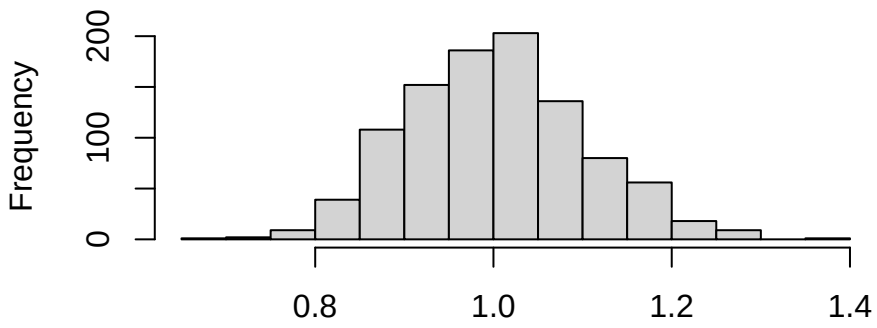
```
nsim = 1e4
samp = rexp(nsim)
hist(samp, main = "", xlab = "")
```



b) Draw 1000 samples of size 100 from Exponential(1). For each of the 1000 samples, find the sample mean. Create a histogram of the sample means. Do they look normal?

Solution: Yes, the sample means look approximately normal: this is a consequence of the central limit theorem.

```
n = 100
nsim = 1000
samps = matrix(rexp(n*nsim, 1), nrow = nsim, ncol = n)
samp_means = rowMeans(samps)
hist(samp_means, main = "", xlab = "")
```



c) [Optional] Draw 1000 samples of size 100 from Exponential(1). For each of the 1000 samples, find a confidence interval for the population mean μ assuming normality. Check whether each interval contains the expected value of the Exponential(1). Comment on your results.

Solution: The expected value of an Exponential(1) random variable is 1. Approximately 95% of the intervals contain the expected value, even if the distribution of the data itself is far from being normal.

```
check_CI = function(x, mu = 1) {
  interval = t.test(x)$conf.int
  (interval[1] < mu) & (interval[2] > mu)
}

n = 100
nsim = 1000
samps = matrix(rexp(n*nsim, 1), nrow = nsim, ncol = n)
100*mean(apply(samps, 1, check_CI))
```

```
## [1] 93.8
```

Exercise 3. In previous courses, you saw the two-sample t-test for comparing the population means of two groups. In class, we have seen a test (based on the F distribution) that lets us compare the populations means of k groups, where $k \geq 2$. The goal of this exercise is checking that when $k = 2$, the tests are equivalent. We will use `data(hsb2)` from `library(openintro)`. If you don't have `library(openintro)` installed, you can install it with the command `install.packages("openintro")`.

Social scientists are interested in checking whether average `math` scores depend on whether go to public or private schools.

a) Assume that the variances of the scores in public and private schools are the same. Find a 90% confidence interval for the difference in `math` averages between public and private schools.

Solution: The interval is [-5.4119377, 0.4000329].

```
library(openintro)
```

```
## Loading required package: airports
```

```
## Loading required package: cherryblossom
```

```
## Loading required package: usdata
```

```
data(hsb2)
t.test(math ~ schtyp, data = hsb2, var.equal = TRUE, conf.level = 0.90)
```

```
##
##  Two Sample t-test
##
## data:  math by schtyp
## t = -1.3901, df = 198, p-value = 0.1661
## alternative hypothesis: true difference in means between group public and group private is not equal
## 90 percent confidence interval:
##  -5.4851691  0.4732644
## sample estimates:
##  mean in group public mean in group private
##              52.24405              54.75000
```

b) Use the `t.test` function to test at the significance level $\alpha = 0.05$ whether average `math` scores are different in public and private schools.

Solution: The p-value is 0.1661 (see part a), so we don't reject the null hypothesis that the population means are equal.

c) Do the same with `aov`. Compare the p-values. Are they the same? Take the square of the observed T statistic you found in part a) and compare it to the F statistic you find with `aov`.

Solution: The p-value in the ANOVA table only has 3 decimal places, but it's the same. The square of the T statistic is the F statistic.

```
mod = aov(math ~ schtyp, data = hsb2)
summary(mod)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## schtyp        1    169  168.80   1.932  0.166
## Residuals   198  17297   87.36
```

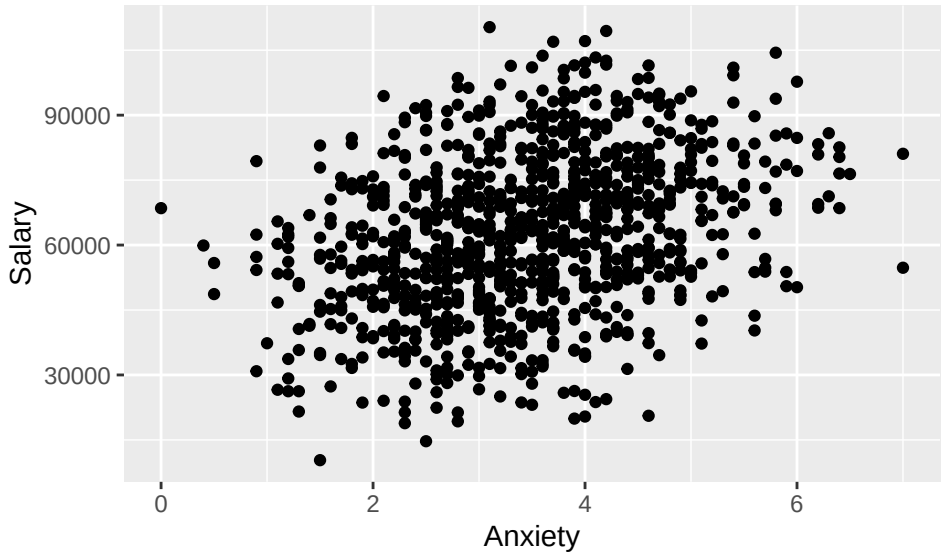
Exercise 4. The dataset `salary.csv` contains salaries (in USD), anxiety levels on a scale that goes from 0 (no anxiety) to 7 (very anxious), and education level. You can read in the dataset with the command:

```
salary = read.csv("http://vicpena.github.io/sta9750/salary.csv")
```

a) Create a plot that only shows the relationship between anxiety and salary. Comment on what you see.

Solution: It would seem that there is a positive association between anxiety and salary.

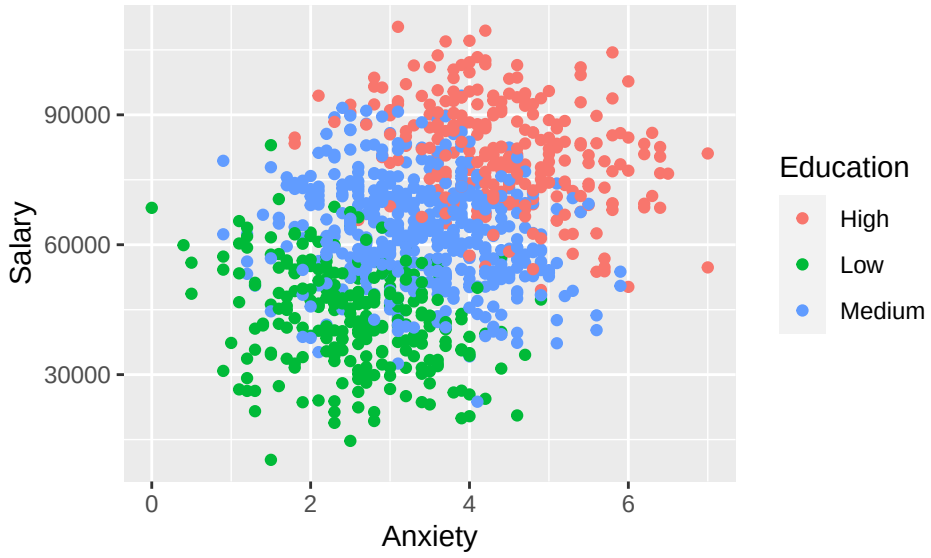
```
qplot(x = Anxiety , y = Salary,  data = salary)
```



b) Now create a plot that displays salary, anxiety, and education. Comment on what you see now.

Solution: After controlling by `Education` we see that the relationship is reversed: for any given level of `Education`, the relationship between `Anxiety` and `Salary` is negative.

```
qplot(x = Anxiety , y = Salary, color = Education, data = salary)
```



c) An article claims that higher salaries come at the cost of higher anxiety levels. Do your findings agree with this claim? Explain why or why not keeping the technical considerations to the minimum.

Solution: This is another instance of the phenomenon we observed in Exercise 1. In this case, the confounder is `Education`: higher `Education` levels are associated with higher `Anxiety`.

Exercise 5. In this exercise, we will use `data(hsb2)` from `library(openintro)`.

a) Create a variable called `average` that has the final average scores the students got combining their results in `read`, `write`, `math`, `science`, and `socst`.

Solution: Here's one way to create the variable:

```
hsb2$average = (hsb2$read+hsb2$write+hsb2$math+hsb2$science+hsb2$socst)/5
```

b) Find a 95% confidence interval for the population mean of `average`.

Solution: Here's the interval

```
t.test(hsb2$average)$conf.int
```

```
## [1] 51.24162 53.52038
## attr(,"conf.level")
## [1] 0.95
```

c) Is there evidence at the $\alpha = 0.05$ significance level that the `average` scores are greater than 50 points at the population level?

Solution: Be careful: this is a one-sided test. The p-value is significant.

```
t.test(hsb2$average, mu = 50, alternative = "greater")
```

```
##
##  One Sample t-test
##
## data:  hsb2$average
## t = 4.1209, df = 199, p-value = 2.768e-05
## alternative hypothesis: true mean is greater than 50
## 95 percent confidence interval:
##  51.42617      Inf
## sample estimates:
## mean of x
##    52.381
```

Exercise 6. Let's keep on using `data(hsb2)`.

a) Find a 99% confidence interval for the difference in averages in `math` scores between `male` and `female` students.

Solution: Here's the interval

```
t.test(math ~ gender, data = hsb2, conf.level = 0.99)$conf.int
```

```
## [1] -4.036787  2.935668
## attr(,"conf.level")
## [1] 0.99
```

b) Is there evidence at the $\alpha = 0.01$ significance level that `math` scores differ between `male` and `female` students?

Solution: No, the p-value of the test isn't significant:

```
t.test(math ~ gender, data = hsb2)
```

```
##
##  Welch Two Sample t-test
##
## data:  math by gender
## t = -0.41097, df = 187.58, p-value = 0.6816
## alternative hypothesis: true difference in means between group female and group male is not equal to
## 95 percent confidence interval:
##  -3.193325  2.092206
## sample estimates:
## mean in group female   mean in group male
##             52.39450             52.94505
```

Exercise 7. [From Montgomery (1986).] During cooking, doughnuts absorb fat in various amounts. A scientist wished to learn if the amount absorbed depends on the type of fat used. For each of four fats, six batches of doughnuts were prepared. The data in the table below are the grams of fat absorbed per batch, coded by deducting 100g to give simpler figures.

T1	T2	T3	T4
64	78	75	55
72	91	93	66
68	97	78	49
77	82	71	64
56	85	63	70
95	77	76	68

a) Read the data into R and plot it. By looking at your plot, do you think that there will be significant differences between the types of fat?

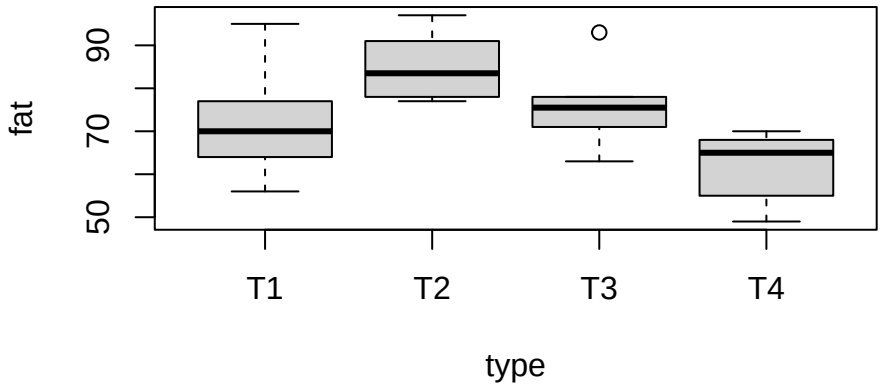
Solution:

```
mat = matrix(c(64, 78, 75, 55,
               72, 91, 93, 66,
               68, 97, 78, 49,
               77, 82, 71, 64,
               56, 85, 63, 70,
               95, 77, 76, 68), byrow = T, ncol = 4)
colnames(mat) = c("T1", "T2", "T3", "T4")
mat = as.data.frame(mat)
df = mat %>% pivot_longer(cols = c(T1, T2, T3, T4),
                          names_to = "type",
                          values_to = "fat")
df$type = factor(df$type)
```

b) Check that the assumptions of the one-way ANOVA model are satisfied.

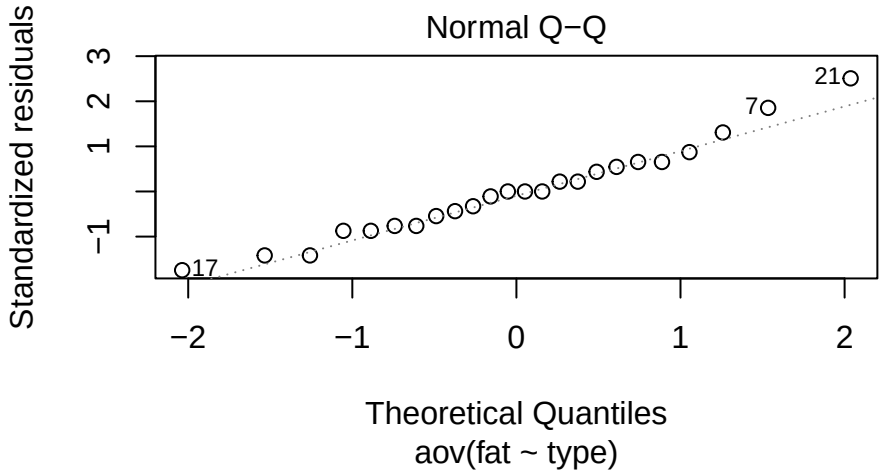
Solution: The variance of the groups doesn't look too different:

```
boxplot(fat ~ type, data = df)
```



Most points in the *qq*-plot are near the dashed line, so it seems that the residuals are approximately normal:

```
mod = aov(fat ~ type, data = df)
plot(mod, which = 2)
```



We don't have an easy way to check the assumption of independence. We can proceed with the analysis.

c) Fit a one-way ANOVA model with the sum-to-zero constraint. Find point estimates and confidence intervals for the grand mean and the treatment effects. Comment on the results.

Solution: It seems like there will be differences between treatments. Type 2 has a big positive effect, whereas Type 4 has a strong negative effect.

```
options(contrasts = c("contr.sum", "contr.poly"))
mod = aov(fat ~ type, data = df)
dummy.coef(mod)
```

```
## Full coefficients are
##
## (Intercept):     73.75
## type:               T1     T2     T3     T4
##                  -1.75  11.25   2.25 -11.75
```

```
confint(mod)
```

```
##                 2.5 %    97.5 %
## (Intercept) 69.472927 78.027073
## type1       -9.158108  5.658108
## type2        3.841892 18.658108
## type3       -5.158108  9.658108
```

d) Is there evidence to claim that there are differences between types of fat at the $\alpha = 0.01$ significance level?

Solution: Yes, there is; the p-value of the global test is $< \alpha$.

```
summary(mod)
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## type         3   1636   545.5   5.406 0.00688 **
## Residuals   20   2018   100.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

e) Run `TukeyHSD` to perform pairwise comparisons and comment on the results.

Solution: The only p-value that is small is that comparing type 2 to type 4.

```
TukeyHSD(mod)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = fat ~ type, data = df)
##
## $type
##       diff        lwr       upr     p adj
## T2-T1   13  -3.232221 29.232221 0.1461929
## T3-T1    4 -12.232221 20.232221 0.8998057
## T4-T1  -10 -26.232221  6.232221 0.3378150
## T3-T2   -9 -25.232221  7.232221 0.4270717
## T4-T2  -23 -39.232221 -6.767779 0.0039064
## T4-T3  -14 -30.232221  2.232221 0.1065573
```