TMA4267 - Linear Statistical Models
Exercise 2 - V2014

Problem 1: The chi-square, t and F-distribution

a) Let $U \sim N(0,1)$. Find the pdf of U^2. Also find the moment generating function of U^2.

b) Let $V \sim \chi^2_p$. Show that the pdf of V equals

$$f_V(v) = \frac{1}{\Gamma(p/2)2^{p/2}} v^{(p/2)-1} e^{-v/2}$$

Hint: first use the result from a) to find the MGF for the χ^2_p. Then find the MGF of V from $f_V(v)$. Show that the two coincide.

c) Use the multivariate transformation formula (see Bingham&Fry (2010), section 2.2) to find the pdf of the F-distribution.
Hint: Let $V \sim \chi^2_p$ and $W \sim \chi^2_q$, where V and W are independent. Let then $F = \frac{V/p}{W/q}$, and $G = W$, and use the multivariate transformation formula to find the joint pdf of F and G. Find the marginal distribution of F from this joint distibution. For the last part it will help you to recognice the integral of a χ^2 pdf.

d) Let $U \sim N(0,1)$ and $V \sim \chi^2_p$, and U and V are independent. Find the pdf of the random variable $T = \frac{U}{\sqrt{V/p}}$.
Hint: First find the joint pdf of U and V, then use the multivariate transformation formula for T and $W = V$ to find the joint pdf of T and W, and then find the marginal pdf of T. For the last part it will help you to recognice the integral of gamma-pdf with parameters $\alpha = (p+1)/2$ and $\beta = 2/(1+t^2/p)$. That is if X is gamma(α,β), then

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}$$

e) Let $T \sim t_q$ (t-distribution with q degrees of freedom). Then show that $T^2 \sim F_{1,q}$ (Fisher-distribution with 1 and q degrees of freedom).

f) Make a map visualizing these results.

Problem 2: N, chi-square, t and F by simulation - in R

Let `dist` denote a given distribution, e.g. `norm`, `chisq`, `t`, or `f`. In R we have functions to sample from a distibution (prefix `r`), calculate the pdf (prefix `d`), calculate the cdf (prefix `p`), and calculate a critical value (prefix `q`).

Let $B = 10000$ and $n = 10$.

a) Find out more about the combinations of prefix and distribution names by typing `?rnorm`, and some of the other combinations.

b) Start with the normal distribution. Make a plot of the standard normal pdf. Then add vertical lines at the critical values for the 0.05 and 0.95 quantiles. Color the tails, e.g. by using the `polygon` function.

c) Move to the χ^2. Simulate B data from the standard normal distribution and square the data. Plot a histogram of the data. Add the pdf of the χ_1 to the histogram. Then add vertical lines for the the critical values for the 0.1 and 0.9 quantiles.

d) Now to the t. First simulate B data from the standard normal distribution, and then simulate B data independently from the chi-square distribution with $n-1$ degrees of freedom. Make the t-ratio (see 1d) and plot a histogram of the data. Add the pdf of the t_{n-1} pdf to the histogram. Then add vertical lines for the the critical values for the 0.15 and 0.85 quantiles. Repeat this for other values of n.

e) Then, the F-distribution. Take the B t-ratios in d) and square them. Plot a histogram of the squared t-ratios. Add the pdf of the $F_{1,n-1}$ to the histogram. Then add vertical lines for the the critical values for the 0.05 and 0.95 quantiles.

f) The F-distribution can also be constructed as a ratio of chi-square variables. Let $n_1 = 5$ nad $n_2 = 40$. Simulate B data independently from the chi-square distributions with n_1 and n_2 degrees of freedom. Make the F-ratio in 1c. Plot a histogram of the data. Add the pdf of the F_{n_1,n_2} to the histogram. Then add vertical lines for the the critical values for the 0.05 and 0.95 quantiles.

Problem 3: Teorem 2.4 of Bingham & Fry

Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$. Further, let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$.
Then the following holds:

- $\bar{X}$ and S^2 are independent.
- $\bar{X} \sim N(\mu, \sigma^2/n)$.
- $nS^2/\sigma^2 \sim \chi^2_{n-1}$.

We have shown this in theory in great detail in the lectures. Now you may study this in practice. Do so by generating data in R and show empirically that the three properties hold.

Problem 4: ANOVA

To compare three different types of investment funds (called A, B and C), 2000 NOK was invested in 18 funds (6 funds of types A, B and C) each year in a period of five years. For each investment the pay-off (in NOK) was registered and is given in the table below.

A	B	C
13288	15738	14790
12782	14249	13827
12812	12369	13860
11713	12822	13150
11201	12117	12669
12233	12605	14267

First read the data into R. You may either read them in by hand, or directly from the course www-page.

By hand (not with the ... - that must be substituted by numbers),

```
x=c(13288,...,14267)
funds=c("A",...,"C")
ds=data.frame(x,funds)
```

or directly

```
ds=read.table("http://www.math.ntnu.no/~mettela/TMA4267/Data/funddat.txt",header=T)
```

Assume the following model for the data

$$X_{ij} = \mu_i + \varepsilon_{ij}$$

where $i = 1, 2, 3$ refers to the three fund types, and $j = 1, \ldots, 6$ are the observations for each fund type. Further, ε_{ij} are assumed to be i.i.d. $N(0, \sigma^2)$.

a) Explain what this model means.

Explore the data by making boxplots, and calculating the group means and group sample variances (we may used the unbiased version). Comment on what you observe.

R-commands:

```
?boxplot, ?ave
```

b) As an alternative model to the one previously assumed, let σ_1^2 denote the variance for the observations of fund type A, σ_2^2 the variance for the observations of fund type B and finally σ_3^2 the variance for the observations of fund type C.

We would like to test if the variance for fund type A and fund type B are the same.

$$H_0 : \frac{\sigma_1^2}{\sigma_2^2} = 1 \text{ vs. } H_1 : \frac{\sigma_1^2}{\sigma_2^2} \neq 1$$

Let $S_i^2 = \frac{1}{n_i - 1} \sum_{j=1}^{n_1} (X_{ij} - \bar{X}_i)^2$ be the (unbiased) sample variance for fund type i. Explain why the test statistic $\frac{S_1^2}{S_2^2}$ might be a sensible test statistics for testing the given hypothesis. What is the distribution of this test statistics under the null hypothesis? Carry out the hypothesis test. What is the p-value of the test? What is your conclusion on a 5% level of significance?

c) We now assume that $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma^2$.

Perform a one-way analysis of variance. What is the null hypothesis tested in a one-way ANOVA?

What is the unbiased estimator for σ^2? Call this S^2. What is the relationship between S_1^2, S_2^2, S_3^2 and S^2?

Perform the analysis in R by directly calculating SS, SSA and SSE. Write down the ANOVA table (SSs, MSs, F and p-value).

d) Now use the R-function `aov` to perform the one-way ANOVA. Compare with your results in c).

e) The assumption of normality of the error terms ε can be assessed by studying the residuals (estimates of the errors), $e_{ij} = y_{ij} - \hat{\mu}_i$, where $\hat{\mu}_i$ is the group i sample mean. Construct residuals and assess normality of errors by producing normal QQ-plots of the residuals. Normal probability plots and quantile plots are explained in Chapter 8 of WMMY (TMA4245 textbook) and in Example 5.9 in Bingham & Fry (TMA4267 textbook).

R-commands: `qqnorm, qqline`.

f) The assumption of equal variances has partly been studied in b), but there also exist other methods to compare many variances. One such method is the Bartlett test. We have previously defined S_i^2 and S^2, and let $i = 1, ..., r$, and $n = \sum_{i=1}^{r} n_i$. The Bartlett test statistics is defined as (WMMY)

$$B = \frac{[(S_1^2)^{n_1-1}(S_2^2)^{n_2-1} \cdots (S_r^2)^{n_r-1}]^{1/(n-r)}}{S^2}$$

where B is related to a so-called Bartlett distribution under H_0. The Bartlett test is found to be sensitive to departures from normality - if your groups come from non-normal distributions, then Bartlett's test may simply be testing for non-normality. Bartletts test is available as `bartlett.test` in R.

A competing method is the Levene test. Look at the absolute value of the residuals from the one-way ANOVA, $\text{abs}(e_{ij})$, or the absolute value of a robust version of the residuals, $\text{abs}(X_{ij} - \text{median}_{j=1}^{n_i}(X_{ij}))$. Is there a difference in means for these absolute residuals wrt the factor under study? Find out by performing a one-way ANOVA with the residuals as input. If the ANOVA is significant (that is, the means of the absolute residuals differ), at level $\alpha = 0.01$, reject the null hypothesis of homogeneity of variances. Levens test is robust to departures from normality. Levene's test is availble in the library `car` in the function `leveneTest`.

Try out the two tests in R on the funds data. What is your conclusion?

g) Finally, we want to test if the mean for fund type A can be seen as the average of the means for fund types B and C. That is, we want to test the null hypothesis:

$$H_0 : \mu_1 = (\mu_2 + \mu_3)/2 \text{ vs. } H_1 : \mu_1 \neq (\mu_2 + \mu_3)/2$$

Define the test statistic $U = \bar{X}_1 - \frac{\bar{X}_2 + \bar{X}_3}{2}$, where $\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} X_{ij}$, n_i is the number of observations for group i.

Show that under the null hypothesis $U \sim N(0, \sigma^2/4)$. Explain why $T = \frac{2U}{S}$ is a natural test statistic for this hypothesis situation. What is the distribution of T assuming the null hypotesis is true? What is the conclusion of this test given the data?