Preliminaries

The theory related to this assignment is in chapter 5, sections 1-4 of the statistics textbook.

**Histograms and empirical densities.** Let $x$ be a vector containing observations of a random variable $X$. The histogram of $x$ can be regarded as a rough, or coarse grained representation of the probability density function of $X$. A smoother estimate of the density function of $X$ based on the data set $x$ is provided by R through the function `density()`. For example, let $x$ consist of 1000 sample values from a Gamma random variable $X \sim \text{Gamma}(r, \lambda)$. Say, for concreteness, that $r = 3$ and $\lambda = 1$. We can simulate $x$ in R thus: $x \leftarrow \text{rgamma}(1000, 3, 1)$.

Rather than give a histogram, we may prefer to plot the empirical density function of the data using `density()`. We do this in the below script and compare the result with the theoretical density function in dashed line, producing the graph shown above.

```r
# I'll first plot the theoretical density, # in dashed lines, for reference: a = seq(from=0, to=10, length.out=1000) plot(a, dgamma(a, 3, 1), type='l', lty='dashed', xlim=range(c(0, 10)), xlab='x', ylab='density')
```

```r
# Now I generate 1000 values of a Gamma random variable # with parameters r=3 and lambda=1:
```
\begin{verbatim}
x=rgamma(1000,3,1) #and add the empirical density plot on top of the previous graph: lines(density(x),type='l') abline(h=0) #This adds a horizontal line at y=0 grid()

The empirical quantile function. Just as we can produce in R an approximate density function for the given data, we can also obtain the quantiles using the R-function quantile(). Here is an example, where we obtain the first quartile (or 25th percentile) of data vector x:

> x=rgamma(1000,3,1)
> quantile(x,0.25)

   25%
1.636593

The main sampling distributions. For a given sequence of independent and identically distributed (i.i.d.) random variables $X_1, X_2, \ldots, X_n$, representing the values of independent observations of some quantity of interest, and having mean $\mu$ and variance $\sigma^2$, one is often interested in the distribution of various statistics (that is, random variables) associated to the $X_i$ and the probability distributions of those statistics. First, we have the sample mean

$$\bar{X} = \frac{X_1 + \cdots + X_n}{n}$$

which is used to estimate $\mu$, and the $Z$-score

$$Z = \frac{\bar{X} - \mu}{\sigma \sqrt{n}}.$$

The central limit theorem implies that $Z$ is approximately (or exactly, if the $X_i$ are normal) normally distributed for large enough $n$. We are also interested in the sample variance

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2,$$

used to estimate $\sigma^2$, which follows the so-called $\chi^2$ (or Chi-square) distribution. There are other related statistics such as

$$T = \frac{\bar{X} - \mu}{S / \sqrt{n}},$$

which follows the Student's $t$-distribution (named after Sealy Gosset, who published under the pseudonym “Student”), and the so-called Snedecor-Fisher's distribution, or $F$-distribution, for the random variable

$$W = \frac{U / \nu_1}{V / \nu_2},$$

where $U$ is $\chi^2$-distributed with $\nu_1$ degrees of freedom and $V$ is $\chi^2$-distributed with $\nu_2$ degrees of freedom. These random variables and distributions are explained in chapter 5, sections 1-4 of the statistics textbook.

We are already familiar with normal distributions and their associated R functions. In this assignment we explore the Chi-square, Student’s $t$, and Snedecor-Fisher’s $F$ distributions using R. The R-functions of main interest here are listed in the next table.
\end{verbatim}
In the above, \( x \) is a real number (or a vector of real numbers) in the range of values of the respective random variables; namely, the full real line for the normal or the Student’s t distributions, the positive half-line for Chi-squared or the Fisher’s F distributions. The variable \( p \) is a real number (or a vector of real numbers) in the interval \([0, 1]\) and the variable \( n \) is a positive integer number (or a vector of positive integers). The parameters \( \text{df} \), \( \text{df1} \), and \( \text{df2} \) are positive integers, called the number of degrees of freedom of the family of distributions.

Here are some examples of how these functions are used:

1. **Find the value of the Chi-squared density with 8 degrees of freedom at \( x = 7.34 \).**
   
   **Solution:**
   
   \[
   > \text{dchisq}(7.34, 8) \\
   \text{[1]} \quad 0.1049437
   \]

2. **Find the upper \( \alpha \)-critical point \( \chi^2_{\text{8}, \alpha} \) of the \( \chi^2 \) distribution for \( \alpha = 0.10 \).** (See page 177 of the statistics textbook for the definition and Table A.5 for the tabulated values of \( \chi^2_{\nu, \alpha} \).)
   
   **Solution:** According to table A.5, this number is \(13.362\). To obtain the same number in R observe that, by the definition of the symbol \( \chi^2_{\text{8}, \alpha} \)
   
   \[
   \alpha = P\left(\chi^2 > \chi^2_{\text{8}, \alpha}\right) = 1 - P\left(\chi^2 \leq \chi^2_{\text{8}, \alpha}\right) = 1 - F\left(\chi^2_{\text{8}, \alpha}\right)
   \]
   
   where \( F(x) \) is the c.d.f. of \( \chi^2_{\text{8}} \). Therefore,
   
   \[
   \chi^2_{\text{8}, \alpha} = F^{-1}(1 - \alpha).
   \]
   
   In other words, the \( \alpha \)-critical point is the \( 1 - \alpha \) quantile because the inverse of the c.d.f. \( F(x) \) is the quantile function. With this in mind, we may obtain \( \chi^2_{\text{8}, 0.10} \) in R thus:
   
   \[
   > \text{qchisq}(0.9, 8) \\
   \text{[1]} \quad 13.36157
   \]
3. Simulate 10000 values of $Z_1^2 + Z_2^2 + Z_3^2$, where the $Z_i$ are i.i.d. standard normal random variables, and draw a histogram. Then compare it (by superimposing the graphs) with the p.d.f. of the Chi-squared distribution with 3 degrees of freedom.

Solution. Here is the graph:

It was obtained through the following script:

```r
n=10000
x=matrix(0,1,n)
for (i in 1:n){
  x[i]=sum(rnorm(3,0,1)^2)
}
hist(x,30,prob=TRUE,ylim=range(c(0,0.25)),main='Histogram of Chi-squared and density plot')
a=seq(from=0,to=20,length.out=100)
lines(a,dchisq(a,3),type='l')
```

4. Draw the graphs of the p.d.f. of the following distributions:

(a) The standard normal p.d.f.
(b) The Chi-squared p.d.f. with 5 degrees of freedom.
(c) The F distribution p.d.f. with degrees of freedom 8 and 15.
(d) The Student's t p.d.f. with 5 degrees of freedom.

Solution: For this purpose we use the following script:

```r
par(mfrow=(c(2,2)))
#Standard normal density
x=seq(from=-4,to=4,length.out=1000)
plot(x,dnorm(x),type='l',main='Standard normal p.d.f')
```
The resulting graph is the following.
Problems

1. **Illustrating the central limit theorem.** Let $X$ be a random variable having the uniform distribution over the interval $[1,2]$. Denote by $X_1, X_2, X_3, \ldots$ a sequence of independent random variables with the same distribution as $X$. Define the sample mean $\overline{X}$ by
   \[ \overline{X} = \frac{X_1 + \cdots + X_n}{n}. \]
   The central limit theorem applied to this particular case implies that the probability distribution of
   \[ \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \]
   converges to the standard normal distribution for certain values of $\mu$ and $\sigma$.
   
   (a) For what values of $\mu$ and $\sigma$ does the convergence hold? (This is to be done by hand.)
   
   (b) For each of the four values $n = 1, 2, 3$ and 10, do the following: Obtain 10000 independent values of the sample mean random variable $\overline{X}$, where the $X_i$ are generated from the uniform distribution over the interval $[1,2]$; then plot the empirical density of $\frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$ with the graph of the standard normal density (in dashed line) superimposed.

2. **Testing the normal approximation with Q-Q plots.** We have seen in HW 4 how normal plots (which are a special case of Q-Q plots) provide another way to decide whether given data are roughly normally distributed. For the same $\frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$ as in the above first problem, and for each $n = 1, 2, 3$ and 10, obtain the normal plots (with the reference straight line as shown in that homework set). Does it look like the sample values are becoming more normally distributed as $n$ increases? Suggestion: when plotting the `qqnorm` graph for a large data set $x$, indicate graph points with small dot characters. For example, once $x$ is generated, use:

   ```r
   qqnorm(x, cex=0.1, main='normal q-q plot, n=1')
   qqline(x)
   ```

   Note: Although using 10000 points as in the first problem produces nice and convincing Q-Q plots, printing the graphs may take a while. Because of this, you may prefer to reduce the number of points to 1000 or fewer.

3. **Exercise 5.22, page 192** (slightly modified.) This exercise uses simulation to illustrate definition (5.3) that the sum of
   \[ X = Z_1^2 + \cdots + Z_\nu^2 \]
   is distributed as $\chi_\nu^2$ where $Z_1, \ldots, Z_\nu$ are i.i.d. $N(0,1)$ r.v.’s.
   
   (a) Generate 1000 random samples of size four, $Z_1, \ldots, Z_4$, from an $N(0,1)$ distribution and calculate
   \[ X = Z_1^2 + Z_2^2 + Z_3^2 + Z_4^2 \sim \chi_4^2. \]
   Display the result by plotting the empirical density and the $\chi_4^2$ density, as in the plot shown on page 1 of this assignment.
   
   (b) Find the 25th, 50th, and 90th percentiles of the simulated sample. How do these percentiles compare with the corresponding percentiles of the $\chi_4^2$ distribution?
4. **Exercise 5.28, page 193** (Slightly modified.) In this exercise we simulate the $t$-distribution with 4 d.f. using the general definition (5.13)

\[
T = \frac{Z}{\sqrt{U/\nu}}
\]

where $Z \sim N(0, 1)$ and $U \sim \chi^2_\nu$.

(a) Generate 1000 i.i.d. standard normal r.v.’s (Z’s) and 1000 i.i.d. $\chi^2_4$ (U’s) and compute

\[
T = 2Z / \sqrt{U}.
\]

Display the result by plotting the empirical density and the $t_4$ density, as in the plot shown on page 1 of this assignment.

(b) Find the 25th, 50th, and 90th percentiles of the simulated sample of the $T$ values. How do these percentiles compare with the corresponding percentiles of the $t_4$-distribution?