Exercise 5. Parameter estimation

1. Generate 100 samples of size 100 from a Poisson distribution with mean 4. Store the 100 sample means in a separate column (Stat > Basic Statistics > Store Descriptive Statistics; Manip > Stack > Stack Rows). What distribution does the Central Limit Theorem say will approximate the distribution of these means? Assess this claim with a suitable plot.

2. Suppose that the daily mean temperature at the Plato Cave weather station is precisely a sequence of independent Normal random variables with expectation 10°C and standard deviation 5°C. Generate a sequence of 100 daily mean temperatures. Now pretend that you do not know the true mean of the Normal distribution, and calculate a point estimate for it. By hand, also compute 90%, 95% and 99% confidence intervals. Do any of your intervals contain the true mean? What is the correct interpretation of the intervals?

   Point estimate =
   90% confidence interval =
   95% confidence interval =
   99% confidence interval =

3. Simulate 99 more samples of 100 temperatures and store the 100 sample means in a separate column. Calculate the standard deviation of your sample means and compare it to the standard error that you would expect theoretically. Compute a 90% confidence interval from each sample, storing the lower and upper limits in separate columns. What proportion of the intervals contain the true value, 10°C? What proportion would you expect?

   Sample standard deviation =
   Theoretical standard error =
   Proportion of intervals =
   Expected proportion =

4. Store the 100 sample medians in another column. Compare the sampling distributions of the sample means and medians. What do you conclude about properties of the two estimators? Which estimator do you prefer?

5. Carefully work through the demonstrations 'Sampling Distribution Simulation' and 'Confidence Intervals' in the 'Rice Virtual Lab in Statistics'.

Exercise 6. Hypothesis testing

1. The file `cloud.txt` contains data collected from a U.S. experiment in the early 1970s that dropped silver nitrate crystals from aircraft to 'seed' clouds and make rain. The data are the rainfall amounts (in acre-feet!) from a sample of 26 unseeded clouds and a sample of 26 seeded clouds. Explore the distributions of the two samples, noting any similarities and differences.

2. The experimenters want to know if seeding a cloud affects the amount of rainfall. One way to do this is to assess whether or not the two samples come from populations with different means. Write down the null and alternative hypotheses. Which statistical test can be used to test these hypotheses? Select an appropriate level of significance for the test.

3. What assumptions does your chosen test make about the data? Are these assumptions reasonable for the rainfall data? If not, transform the data (by taking square roots for example) so that the test will be appropriate.

4. Compute the test statistic for your test by hand. Write down the distribution of your test statistic and sketch its probability density. Use your statistical tables to determine critical regions for your test and add them to your sketch. Use the tables to obtain an approximate p-value for your test statistic. Now perform the test using MINITAB (Stat > Basic Statistics), note the p-value and compare it with your chosen level of significance. Write down clearly what conclusion you draw from the test and what this tells the experimenters about cloud seeding.

5. Does the result change if you perform the test without transforming the data?

6. Shortly after the introduction of the euro coins in 2002, BBC on-line news published this article:

   'Meanwhile, two Polish statisticians have discovered something about euro coins that should gladden the hearts of confidence tricksters. The coin apparently favours heads. When Tomas Gliszczynski and Waclaw Zawadowski of the Podlaska Academy spun one Belgian euro coin 250 times, it came up with King Albert's head 140 times. "The euro is struck asymmetrically," Mr Gliszczynski told Germany's Die Welt newspaper. He said he hoped to experiment with German euro coins at a maths conference next month. "I know the phenomenon from other coins like the two zloty piece, which we have thrown more than 10,000 times," he said.'

   Write a short note describing a statistical test for the hypothesis that the Belgian euro is biased given these results, and comment on Mr Gliszczynski's claim.

Exercise 7. Linear regression

1. Read the data in the file `xy.txt` into MINITAB. Compute the sample correlation between x_1 and y_1 (Stats > Basic Statistics > Correlation) then repeat for the other three pairs. Record the values in the table below. Also write down what these values tell you about the association between the four pairs of x and y variables. Plot y_1 against x_1 (Graph > Plot) then repeat for the other three pairs. Do these plots change your ideas about the associations? What implications does this have for the interpretation of correlations?

Pair	1	2	3	4
Correlation				

2. Write down a mathematical representation for the simple linear regression of each y-variable on the corresponding x-variable. Make sure that you know which is the response variable and which is the explanatory variable. Using a calculator instead of MINITAB, estimate the slope and intercept parameters in each of the four cases using the formulae in the lecture notes.

Dataset	$\hat{\beta}_0$	$\hat{\beta}_1$	R^2	p
1				
2				
3				
4				

3. Now perform the linear regressions with MINITAB (Stat > Regression > Regression) and check that the parameter estimates agree with your calculations. What are the values of the coefficient of determination? What does this tell you about the linear models? What is the p-value for testing whether or not the slope is zero? What do you conclude about the explanatory power of the x variable in each case?

4. Compute the residuals from the fitted models and assess the model fits by making diagnostic plots. Are any of the model assumptions inappropriate? Do these plots change your conclusion about the explanatory power of the x variable in each case?

Exercise 8. Multiple regression

In this exercise we look at the Central England Temperature (CET) dataset in the file `atldjf.txt`. The column 'cet' contains the mean winter central England temperature (°C) from 1866 to 1997. Columns C2–C5 contain mean sea-level pressure (SLP) measurements (hPa) at four locations. As you work through this exercise you should compile a short report of your analysis, including the regression equation for each model that you fit.

1. Plot CET against year and plot a histogram of the temperatures. What features do you notice in the series? What does the distribution look like?

2. One way to assess evidence for a time trend in CET is to fit a linear regression of temperature on year. Fit this model using MINITAB and interpret the results. What do you conclude about the changes in CET through time? Examine any diagnostic plots that you consider appropriate and comment on their consequences for your model.

3. Now regress CET on just the Iceland SLP and assess the model fit. What does the estimate of the slope parameter say about the relationship between CET and Iceland SLP? Can you explain this result scientifically?

4. The plot of residuals against observation order indicates a slight, increasing time trend. Fit the multiple regression of CET on Iceland SLP and year. Examine the significance of the two explanatory variables and assess the model fit. What conclusions do you draw now about any time trend in CET?

5. If you have time, experiment by including and excluding different SLP series from the regression and compare the fitted models. Also try adding a polynomial time trend by storing squared year in a new column.

Exercise 9. Time series analysis

1. The Darwin SLP data discussed in the lectures is in `darwin.txt`. Plot the data against time. What can you tell about any cycles or long-term time trends?

2. Apply moving average filters of different lengths to obtain a clearer view of the long-term trends (Stat > Time Series > Moving Average).

3. Compute the autocorrelation function up to lag 36 months (Stat > Time Series > Autocorrelation). The cyclical behaviour indicates the seasonality of the data. If X_t is the SLP for month t, plot X_t against X_{t-12} to visualise this dependence.

4. Apply the backward difference filter with lag 12 to remove the seasonality, then plot the differenced series to obtain a clearer view of the inter-annual variations (Stat > Time Series > Differences).

5. (The final two questions are optional.) In addition, apply the backward difference filter with lag 1 to remove the month-to-month dependence and plot the resulting series, which we shall call Y_t. This looks like white noise. If this were so, what would you expect the autocorrelation function to look like? Compute the autocorrelation function up to lag 36 and interpret the result.

6. The autocorrelation function for Y_t suggests that a particular seasonal ARIMA model, with moving average components at lags 1 and 12, will be a good description of the Darwin SLP series. Fit this model (Stat > Time Series > ARIMA) to the Darwin SLP series by selecting a seasonal model with period 12, one nonseasonal and one seasonal difference, one nonseasonal and one seasonal moving average term, and no constant term. Plot the autocorrelation function and histogram of the residuals. Do the residuals look like white noise?