Worksheet 2 - Basic statistics

Basic statistics references

- Quinn & Keough (2002) - Chpt 1, 2, 3 & 4

Question 1 - Population parameters

The little spotted kiwi (*Apteryx owenii*) is a very rare flightless bird that is extinct on mainland New Zealand and survives as 1000 individuals on Kapiti Island. In order to monitor the population, researches in the recovery team systematically captured all of the individuals in the population over a two week period. Each individual was weighed, banded, assessed and released. The file *.csv lists the weights of each individual male little spotted kiwi in the population.

Format of kiwi.csv data files

Band	Weight
64955	1.749
65318	2.551
64612	1.768
64393	2.327
64092	2.127
...	...

Band Unique bird identification band number

Weight Weight (grams) of the individual male birds

Open the kiwi data file. HINT.

Generate a **frequency histogram** of male kiwi weights. HINT. This distribution represents the population (all possible observations) of male kiwi weights. Note that this is the statistical population and not a biological population - obviously a biological population entirely lacking in females would not last long!

Q1-1. Describe the shape of the distribution?

Since we have the weights of all male kiwi in the population, is possible to calculate population parameters (such as population **mean and standard deviation**) directly!

Q1-2. What is the mean (a location measure) and standard deviation (a measure of spread) of the population?

a. Mean HINT

b. SD HINT

Assuming, the population is normally distributed, it is possible to calculate the probability that a randomly recaptured male kiwi will weigh greater than a particular value, less than a particular value, or weigh between a range of weights. This probability is just the area under a particular region of a normal distribution and can be calculated using the **normal probabilities**.

Q1-3. Assuming that the population is normally distributed, what is the probability of recapturing a male little spotted kiwi that weighs greater than 2.9 kg? HINT

For data sets with large numbers of observations, the distribution of observations can be examined via a histogram - as demonstrated above. However, histograms are only meaningful for summarizing large data sets. For smaller data sets other exploratory tools (such as **boxplots**) are necessary. To appreciate the relationship between boxplots and the underlying distribution of data, **construct a boxplot** of male kiwi weights. HINT

Question 2 - Samples as estimates of populations

Here is a modified example from Quinn and Keough (2002). Lovett et al. (2000) studied the chemistry of forested watersheds in the Catskill Mountains in New York State. They had 38 sites and recorded the concentrations of ten chemical variables, averaged over three years. We will look at two of these variables, dissolved organic carbon (DOC) and hydrogen ions (H).

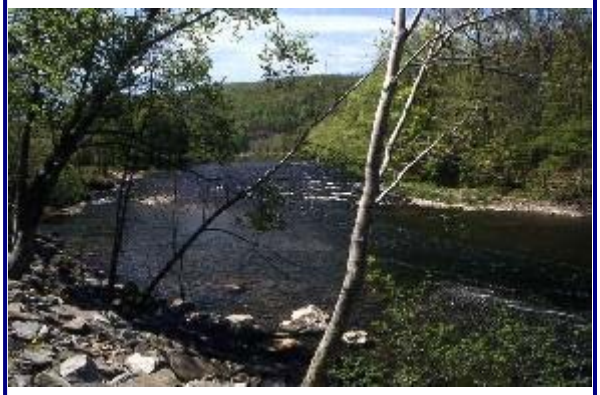
Format of lovett.csv data files

STREAM	DOC	H
Hunter	180.4	0.48
West Kill	108.8	0.24
Mill	104.7	0.47
Kelly Hollow	84.5	0.23
Pigeon	82.4	0.37

STREAM Name of the site (stream) from which observations were collected

DOC Dissolved oxygen concentration (mmol.L^{-1})

H Hydrogen concentration (mmol.L^{-1})



Open the lovett data file.

Q2-1. What is the purpose of sampling?

Before continuing, make sure you are clear on what the **observations, variables and populations** are.

Construct a **boxplot** of dissolved organic carbon (DOC) from the sample observations. HINT

Q2-2. How would you describe the **boxplot?**

Q2-3. Are there any outliers? (Y or N)

Provided the data were collected without bias (ideally random) and with adequate replication, the sample should reflect the entire population. Therefore **sample statistics** should be good estimates of the population parameters.

Q2-4. Calculate the sample mean HINT

The mean of a sample is considered to be a location characteristic of the sample. Along with the mean, it is often desirable to characterize the spread of data in a sample - that is to determine how variable the sample is.

Q2-5. Calculate the sample standard deviation HINT

For most purposes, the sample itself is of little interest - it is purely used to estimate the population. Therefore it is necessary to be able to estimate how well the sample mean estimates the true population mean. The Standard error (SE) of the mean is a measure of the **precision** of the mean.

Q2-6. Calculate the standard error of the mean HINT

Following on from the idea of precision of the mean, is the concept of **confidence intervals**, by which an interval is calculated that we are 95% confident will contain the true population mean.

Q2-7. Calculate the 95% confidence interval of the mean HINT +/-

Construct a **boxplot** of hydrogen concentration (H) from the sample observations HINT

Q2-8. How would you describe the boxplot?

Many statistical analyses assume that the population from which the sample was collected is normally distributed. However, biological data is not always normally distributed. To normalize the data, try **transforming** to **logs.** HINT

Q2-9. Does the transformation **successfully normalize** these data? (Y or N)

Earlier we identified the presence of an outlier, to investigate the impact of this outlier on a range of summary statistics, calculate the following measures of **location** (mean and median) and **spread** (standard deviation and interquartile range) for DOC, **with and without the outlying observation** and complete the table below.

Summary Statistic	DOC	Modified DOC
Mean	HINT	HINT
Median	HINT	HINT

Variance	HINT ☐	HINT ☐
Standard deviation	HINT ☐	HINT ☐
Inter-quartile range	HINT ☐	HINT ☐

Q2-10. Which measures of location and spread are most robust to inclusion and exclusion of a single unusual observation?

Question 3 - Exploratory data analysis

Sánchez-Piñero & Polis (2000) studied the effects of seabirds on tenebrionid beetles on islands in the Gulf of California. These beetles are the dominant consumers on these islands and it was envisaged that seabirds leaving guano and carrion would increase beetle productivity. They had a sample of 25 islands and recorded the beetle density, the type of bird colony (roosting, breeding, no birds), % cover of guano and % plant cover of annuals and perennials.

Format of sanchez.csv data files

COLTYPE	BEETLE96	GUANO	PLANT96
..	..	..	..
..	..	..	..
..	..	..	..

COLTYPE	Type of bird colony (N = no birds, R = roosting, B = breeding
BEETLE96	Abundance of beetles (number per carrion trap) in 1996
GUANO	% cover of guano on island in 1995 and 1996
PLANT96	% cover of total plants (annual and perennial) on island in 1996

Open the sanchez data file.

Q3-1. For percentage plant cover, **Calculate the following summary statistics separately for each colony type** and complete the table below.

Summary Statistic	No colonies	Roosting colony	Breeding colony
Mean HINT	☐	☐	☐
Variance HINT	☐	☐	☐
Standard deviation HINT	☐	☐	☐

a. Which colony type has the greatest variance? (N, R or B)

Normality

Before proceeding, make sure you are familiar with the significance of normally distributed sample data and thus why it is necessary to examine the **distribution of sample data** as part of routine exploratory data analysis (EDA) prior to any formal data analysis.

Q3-2. Construct a **boxplot** for total 1996 beetle abundance for each colony type separately. HINT

a. Are there any outliers identified? (Y or N)

b. Describe the shape of each distribution.

c. Now **transform** the response variable to logs and redraw the boxplots HINT. Does this change (improve?) the shape of the distributions? (Y or N)

Linearity

Often it is necessary to examine the **nature of the relationship or association between variables** as part of routine exploratory data analysis (EDA) prior to any formal data analysis. The nature of relationships/associations between continuous data is explored using **scatterplots**.

Q3-3. Construct a **scatterplot** for beetle abundance **against** total 1996 plant cover (HINT).

a. Is there any evidence of non-linearity? (Y or N)

b. Note, that the boxplots also enable us to explore the normality of both variables (populations). Is there any evidence of non-normality? (Y or N)

Sánchez-Piñero & Polis (2000) measured a number of continuous variables (% cover of guano, % cover or plants and abundance of beetles. Therefore, they might be interested in exploring the relationships between each of these variables. That is, the relationship between guano and plants, guano and beetles, and beetles and plants. While it is possible to create separate scatterplots for each pair (in this case three separate scatterplots), a scatterplot matrix is usually more informative and efficient.

Q3-4. Construct a **scatterplot matrix or SPLOM** for % of guano, % of plant cover and beetle abundance HINT. Are there any obvious relationships?

Homogeneity of variance

Many statistical hypothesis tests assume that populations are equally varied. For hypothesis tests that compare populations (such as t-tests - see Question 4), it is important that one of the populations is not substantially more or less variable than the other population(s). Thus, such tests assume **homogeneity of variance**.

Q3-5. Construct an examine **boxplots** of beetle abundance for each of the three colony types. HINT

a. Firstly, is there any evidence of non-normality? (Y or N) _____ ☐

b. Try square-root transforming (preferred over log transformation when applying to count data, since log(0) is not legal) the beetle variable (function is sqrt) and using this transformed variable to reconstruct the boxplots. Note that it may be necessary to perform a forth-root transformation (which performing the square-root transformation twice) in order to normalize this highly skewed data. This can be done using the expression to compute as sqrt(sqrt(BEETLE96)) HINT or HINT. If this successfully normalizes the data, focus on whether there is any evidence that the populations are equally varied. Was a forth-root transformation successfull? (Y or N) _____ ☐

c. Try calculating the **variance or standard deviation** of beetle abundance for each colony type separately (remember to use the transformed data, as the raw data was obviously non-normal and non-normality often results in unequal variances). Do these values provide any evidence for unequally varied populations? (Y or N) _____ ☐

Question 4 - Hypothesis testing

Furness & Bryant (1996) studied the energy budgets of breeding northern fulmars (*Fulmarus glacialis*) in Shetland. As part of their study, they recorded the body mass and metabolic rate of eight male and six female fulmars.

Format of furness.csv data files

SEX	METRATE	BODYMASS
MALE	2950	875
FEMALE	1956	765
MALE	2308	780
MALE	2135	790
MALE	1945	788

SEX	Sex of breeding northern fulmars (*Fulmarus glacialis*)
METRATE	Metabolic rate (hJ/day)
BODYMASS	Body mass (g)

Open the furness data file.

Q4-1. The researchers were interested in testing whether there is a difference in the metabolic rate of male and female breeding northern fulmars. In light of this, list the following:

a. The biological hypotheses of interest

b. The biological null hypotheses

c. The statistical null hypotheses (H_0)

The appropriate statistical test for testing the null hypothesis that the means of two independent populations are equal is a **t-test**

Before proceeding, make sure you understand what is meant by **normality** and **equal variance** as well as the **principles of hypothesis testing using a t-test**.

Q4-2. For the null hypothesis test of interest (that the mean population metabolic rate of males and females were the same), calculate the **Degrees of freedom**

Q4-3. Calculate the critical t-values for the following null hypotheses (&alpha = 0.05)

a. The metabolic rate of males is higher than that females **(one-tailed test)** HINT

b. The metabolic rate of males is the same as that of females **(two-tailed test)** HINT

Since most hypothesis tests follow the same basic procedure, confirm that you understand the **basic steps of hypothesis tests**.

Q4-4. In the table below, list the assumptions of a t-test along with how violations of each assumption are diagnosed and/or the risks of violations are minimized.

Assumption	Diagnostic/Risk Minimization
I.	
II.	
III.	

So, we wish to investigate whether or not male and female fulmars have the same metabolic rates, and that we intend to use a t-test to test the null hypothesis that the population mean metabolic rate of males is equal to the population mean metabolic rate of females. Having identified the important assumptions of a t-test, use the samples to evaluate whether the assumptions are likely to be violated and thus whether a t-test is likely to be reliability.

Q4-5 Is there any evidence that; HINT

a. The assumption of normality has been violated?

b. The assumption of homogeneity of variance has been violated?

Q4-6. Perform a t-test to examine the effect of sex on the mass of fulmars using either (which ever is most appropriate) a **pooled variance t-test** (for when population variances are very similar HINT) or **separate variance t-test** (for when the variance of one population is likely to be up to 2.5 times greater or less than the other population HINT). Ensure that you are familiar with the **output of a t-test**.

a. What is the t-value? (Excluding the sign. The sign will depend on whether you compared males to females or females to males, and thus only indicates which group had the higher mean).

b. What is the df (degrees of freedom).

c. What is the p value.

Q4-7. Write the results out as though you were writing a research paper/thesis. For example (select the phrase that applies and fill in gaps with your results):
The mean metabolic rate of male fulmars was (choose correct option)
(choose correct option) (t = , df = , P =)
the mean metabolic rate of female fulmars.

Q4-8. Construct a **bar graph** showing the mean metabolic rate of male and female fulmars and an indication of the precision of the means with error bars.HINT

Question 5 - Paired data

Here is a modified example from Quinn and Keough (2002). Elgar et al. (1996) studied the effect of lighting on the web structure or an orb-spinning spider. They set up wooden frames with two different light regimes (controlled by black or white mosquito netting), light and dim. A total of 17 orb spiders were allowed to spin their webs in both a light frame and a dim frame, with six days `rest' between trials for each spider, and the vertical and horizontal diameter of each web was measured. Whether each spider was allocated to a light or dim frame first was randomized. The H_0's were that each of the two variables (vertical diameter and horizontal diameter of the orb web) were the same in dim and light conditions. Elgar et al. (1996) correctly treated these as paired comparisons because the same spider spun her web in a light frame and a dark frame.

Format of elgar.csv data files

PAIR	VERTDIM	HORIZDIM	VERTLIGH	HORIZLIGH
..	..	..	..	..
..	..	..	..	..
..	..	..	..	..

PAIR	Name given to each pair of webs spun by a particular spider
VERTDIM	The vertical dimension or height (mm) of webs spun in dim conditions
HORIZDIM	The horizontal dimension or width (mm) of webs spun in dim conditions
VERTLIGH	The vertical dimension or height (mm) of webs spun in light conditions
HORIZLIGH	The horizontal dimension or width (mm) of webs spun in light conditions

Note:for paired t-tests, it is traditional for categories to be column labels rather than entries in a categorical variable. Compare the structure of the elgar data (paired t-test) set with that of the furness (standard t-test) data set.

Open the elgar data file.

Q5-1. What is an appropriate statistical test for testing an hypothesis about the difference in dimensions of webs spun in light versus dark conditions? Explain why?

Q5-2. The actual H_0 is that the mean of the differences between the pairs (light and dim for each spider) equals zero. Use a **paired t-test** to test the H_0 that the mean of the differences in vertical diameter (HINT) and separately, in horizontal diameter (HINT) of the web between the pairs (light and dim for each spider) equal zero.

Q5-3. Write the results out as though you were writing a research paper/thesis. For example (select the phrase that applies and fill in gaps with your results):
The mean vertical diameter of spider webs in dim conditions was (choose correct option)
(choose correct option) ☐ (t = ☐, df = ☐, P = ☐)
the vertical dimensions in light conditions.
The mean horizontal diameter of spider webs in dim conditions was (choose correct option)
(choose correct option) ☐ (t = ☐, df = ☐, P = ☐)
the horizontal dimensions in light conditions.

Question 6 - Non-parametric tests

We will now revisit the data set of Furness & Bryant (1996) that was used in Question 4 to investigate the effects of gender on the metabolic rates of breeding northern fulmars (*Fulmarus glacialis*). Furness & Bryant (1996) also recorded the body mass of the eight male and six female fulmars they captured.

Since the males and female fulmars were all independent of one another, a t-test would be appropriate to test the null hypothesis of no difference in mean body weight of male and female fulmars.

Q6-1. Are the assumptions underlying this test met? (Y or N) Hint: check the relative sizes of the two sample variances and the distribution of body weight for each sex. ☐

When the distributional assumptions are violated, parametric tests are unreliable. Under these circumstances, **non-parametric tests** can be very useful.

Q6-2. The Wilcoxon-Mann-Whitney test is described as a non-parametric test for comparing two groups.

a. What null hypothesis does this test actually evaluate?

b. What are the underlying assumptions of a Wilcoxon-Mann-Whitney test?

Q6-3. If the assumptions are met, test the null hypothesis of no difference in body weight between male and female fulmars using a **Wilcoxon test** HINT. Based on this outcome, what are your conclusions?

a. Statistical:

b. Biological (include trend):

Q6-4. Construct a **bar graph** showing the mean mass of male and female fulmars and an indication of the precision of the means with error bars. HINT

Question 7 - Randomization (permutation) test

A wildlife ecologist responsible for the management of a significant population of southern brown bandicoot, *Isoodon obesulus*, was interested in determining the impacts that picnickers were having on the health of bandicoots in the park. In particular, he was interested in determining whether bandicoots that occupied areas frequented by picnickers were heavier (and thus fatter) than bandicoots that occurred in other woodland areas. Fifty adult male bandicoots were captured from picnic and woodland areas and the weights of all individuals were measured.

Format of bandicoots.csv data files

AREA	WEIGHT
PICNIC	875
PICNIC	765
...	780
WOODLAND	790
WOODLAND	788

AREA Area from which bandicoots were captured (Picnic or Woodland)

WEIGHT Capture weight (g) of bandicoots

Open the bandicoots data file.

Q7-1. Are the assumptions underlying this test met? (Y or N) Hint: check the relative sizes of the two sample variances and the distribution of body weight for each sex.

Q7-2. There is a clear problem with non-normality and this problem cannot be fixed by a transformation. Why?

When the assumptions of the t-test have been violated, the distribution of all possible t-values cannot reliably be assumed to follow a mathematical t distribution. So, when the null hypothesis is true, and there is no effect of AREA on the WEIGHT of bandicoots, what t-values would we expect.

Q7-3. In this case, it is likely that observations (individual bandicoots) were collected via random sampling. It would be logistically impossible to do so. Therefore, since the variances are not wildly unequal, a **randomization (or permutation) test** is appropriate. Such a test, repeatedly shuffles the sample data, each time calculating a specific statistic. So while the assumptions of the t-test have been violated, and therefore the distribution of all possible t-values cannot reliably be assumed to follow a mathematical t distribution, we can generate a distribution of possible t-statistics from the randomized t-statistics.

Q7-4. A randomization test involves the following sequence of tasks:

a. Define a new function that accepts a data set and returns an appropriate statistic. In this case, since we are comparing two populations a t-statistic is appropriate. **Define an appropriate function** .

b. Next we need to define a function that alters the structure of the data. In this case, we need to define a function that randomly shuffles the categorical variable (group labels) around. **Define an appropriate function** .

c. Then we use the 'boot()' function to repeatedly calculate the statistic, each time on the randomly altered data. In this case, we want to repeatidly (100 times) calculate the t-statistic from the data set in which the group labels have been randomly shuffled. **Perform the bootstrapping.**

Q7-5. What is the actual sample t-value in this case? HINT

Q7-6. Having performed the randomization proceedure to calculate a set of t-values that we might expect to obtain when the null hypothesis is true, we can now explore the distribution of these t-values. This distribution is in effect a t-distribution. However, rather than being a general, theoretical t-distribution that can be applied to all populations (provided assumptions are met), this distribution is specific to the populations that we are interested in.

a. Examine the distribution of these t-values (the t-distribution). HINT

b. **Determine what proportion of resampled t-values are as great or greater than our actual sample t-value**. This then represents the probability of obtaining our sample t-value when the null hypothesis is true, and is thereby interpreted as any other p-value. What is the p-value?.

Q7-7. What conclusions would you draw from the analysis?

Q7-8. Given that generated t-distribution is specific to the population(s) that we are interested in, and is more robust than parametric statistical analyses, we might wonder why parametric analyses are prefered over randomization procedures. Why are parametric analyses prefered?

Question 8 - Power analysis

An ornithologist studying various populations of the Australian magpie, *Gymnorhina tibicen*, was primarily interested in whether the growth of urban magpies might be stunted as a result of the increased consumption of processed foods. To investigate this hypothesis she intended to measured the total body lengths in centimeters of a number of birds from both urban and rural locations. The null hypothesis of interest is that the population mean length of urban magpies is equal to that of rural magpies and thus a t-test is an appropriate test. Previous research had indicated that the mean body length of rural magpies was 36.87cm with a standard deviation of 2.

Q8-1. If the ornithologist considered a 10% decrease in mean body length to be of biological significance

a. What **effect size** is she interested in detecting? HINT
b. In order to have an 80% chance of detecting such an effect (if one really exists), how many replicate birds would the ornithologist need to measure from each population (assume significance level of 0.05)? HINT

Q8-2. Often, it is difficult to obtain estimates of the likely population standard deviation. Similarly, it can be difficult to estimate the effect size (delta). Consequently, it is often more preferable to perform power calculations for a range of standard deviations or effect sizes and plot the relationships for each parameter set. To assist the ornithologist to determine sample sizes, estimate the following:

a. Relationship between power and sample size for a range of effect sizes (3, 4, 5 & 6). HINT. Note, you need to load the biology library
b. Relationship between power and sample size for a range of standard deviations (1.8,2,2.2). HINT. Note, you need to load the biology library

Q8-3. Alternatively, the ornithologist's sampling efforts may be constrained somewhat by either ethics or by difficulties in capturing birds, and thus the ornithologist may wish to estimate what the minimum detectable effect size would be for a given range of sample sizes. To assist the ornithologist to determine sample sizes, estimate the following:

a. Relationship between minimum detectable effect size and sample size for a range of standard deviations (1.8,2,2.2). HINT
b. Relationship between minimum detectable effect size and sample size for a range of power (0.7,0.8,0.9). HINT

Welcome to the end of Worksheet 2