Comparing whale guesstimates

Group assignment

Using the whale count data, compare the differences between first and second abundance guesstimates using R and excel and report the the following

1. state the null and alternative hypotheses being tested
2. the reason for choosing the statistical test you used
3. a summary of the results:
    - degrees of freedom, test statistic, p-values (at 5% level)?
    - did you accept or reject the null hypothesis?
    - is there a difference?
4. conduct the analysis in `R` and `excel` and submit:
    - a written report of points 1, 2 and 3 as PDF
    - an excel workbook showing your results
    - a saved R file showing results

So, the objective is to compare two samples. These samples actually represent *pairs* of values: each person provided an initial and an updated guesstimate of whale abundance. First we need to read in the data. While we are at it, let's look at the first 6 rows using the `head()` function:

```
whales <- read.csv("whales.csv")
head(whales)
```

```
##   species group recorder      date estimate estimate2 location
## 1   whale     1  Drummey 9/20/2016      150       130      MPA
## 2   whale     1  Drummey 9/20/2016      130        90      MPA
## 3   whale     1  Drummey 9/20/2016      118        97      MPA
## 4   whale     1  Drummey 9/20/2016      115        85      MPA
## 5   whale     1  Drummey 9/20/2016      112       125      MPA
## 6   whale     1  Drummey 9/20/2016       90       120      MPA
```

1. Hypothesis

- Hypotheses:
  - the null hypothesis is that there is *no difference* between the initial and updated guesstimates
  - the alternative hypothesis is that there *is a difference* between the initial and updated guesstimates

2. Which test and why

Next, we need to know the nature of the data in order to pick the correct statistical model. Here our interest is in comparing two samples so we need to look at two things:

1. are the samples independent?
2. are the data normally distributed?

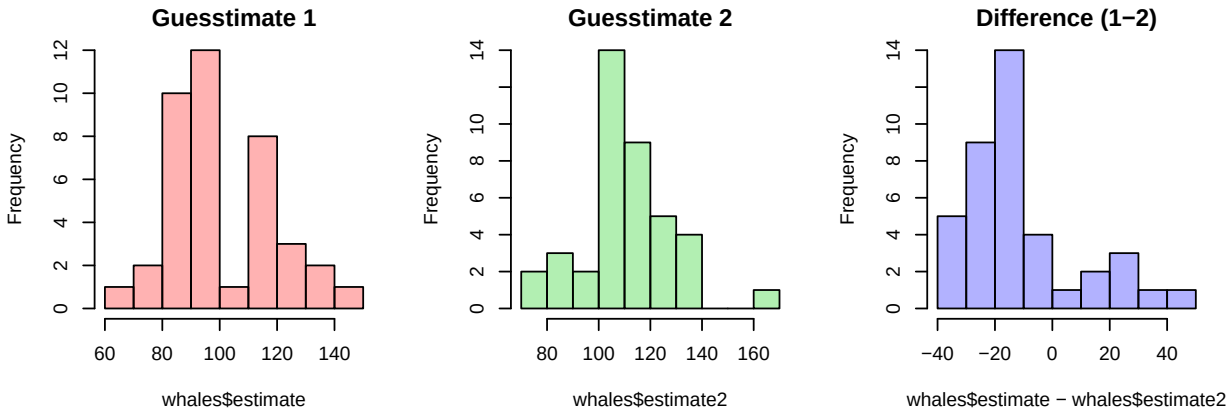
Independence

- data are pairs of values from the same individual
- data are not independent
- need a *paired* test

Normality

There are a couple of tools we can use to assess normality. First, symmetrical histograms are a good indicator of normality:

```
par(mfrow=c(1,3),oma=c(0,0,0,0),mar=c(4,4,2,2))
hist(whales$estimate)                  # first sample
hist(whales$estimate2)                 # second sample
hist(whales$estimate-whales$estimate2) # differences
```



We can also formally evaluate using the `shapiro.test()`. We are looking for a p-value > 0.05, i.e., not significantly different from normally distributed:

```
shapiro.test(whales$estimate)                    # first sample
```

```
##
##  Shapiro-Wilk normality test
##
## data:  whales$estimate
## W = 0.96042, p-value = 0.1731
```

```
shapiro.test(whales$estimate2)                   # second sample
```

```
##
##  Shapiro-Wilk normality test
##
## data:  whales$estimate2
## W = 0.9674, p-value = 0.2969
```

```
shapiro.test(whales$estimate-whales$estimate2) # differences
```

```
##
##  Shapiro-Wilk normality test
##
## data:  whales$estimate - whales$estimate2
## W = 0.85925, p-value = 0.0001521
```

It is important to consider the data we are analyzing

- `estimate` is normally distributed
- `estimate2` is normally distributed
- but the *differnce* is not

So, the data (the *differences*) are paired and *not* normally distributed which means we would choose the *Wilcoxon matched-pairs for skewed data*. For this assessment, however, I am willing to accept the *paired t-test*.

3. Summary of results

3a. Vital statistics

Wilcoxon matched-pairs test

```
wilcox <- wilcox.test(whales$estimate,whales$estimate2,paired=TRUE)
wilcox
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  whales$estimate and whales$estimate2
## V = 206, p-value = 0.01038
## alternative hypothesis: true location shift is not equal to 0
```

- Degrees of freedom: not required (but using df = number of pairs - 1 we get: df = 39)
- Test statistic: V = 206
- *p*-value: $p = 0.0103803$

Paired t-test

```r
ttest <- t.test(whales$estimate,whales$estimate2,paired=TRUE)
ttest
```

```
##
##  Paired t-test
##
## data:  whales$estimate and whales$estimate2
## t = -2.9321, df = 39, p-value = 0.005608
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -16.053467  -2.946533
## sample estimates:
## mean of the differences
##                    -9.5
```

- Degrees of freedom: df = 39)
- Test statistic: t = -2.9321217
- p-value: $p = 0.0056082$

3b. Hypothesis testing and differences

Both tests have a p-value below the 5% significance value value which mean that we are able to reject the null hypothesis that there is no differences between the first and second estimates. Based on our statistical test, we can concluse that the estimates are in fact differnt.

3. Documentation

You should have provided an `Excel` spreadsheet showing your working out, specifically using the appropriate test functions and/or the functions in the Analysis Toolpack. You should have also submitted an `R` script similar to the code outlined above, showing how you implemented the various analytical steps, and how you got the answers you presented.