Exercise 3.1 Exact test for a population proportion

A junior obstetrician working at one of London's Teaching hospitals decided to monitor his caesarean section rate to see whether it is comparable to the national average of 20%. Out of 20 patients he looked after during 3 subsequent on-call weekends, he found that he delivered 8 patients by caesarean section.

i) Write down the null and alternative hypotheses in this experiment.

ii) What values for the number of caesarean section deliveries out of 20 would be as or more extreme than that observed under the null hypothesis?

iii) Use the binomial tables in Neave's to calculate the exact *p*-value for this test.

iv) Comment on the wisdom of this experiment.

v) Check your results in Stata using the command:

```
. bitesti 20 8 0.2
```

Exercise 3.2: Approximate test and CI for a population proportion

In a survey of middle aged men in the Burnley area, 168 out of 286 men said they currently smoked cigarettes.

i) Is the proportion of middle aged men in the Burnley area who smoke significantly different from the national average of 0.3?

ii) Calculate an approximate 95% CI for the proportion of middle aged men in the Burnley who are smokers.

iii) What is the relationship between your answers to i) and ii)?

iv) Check the test result and 95% CI in Stata using the commands:

```
. bitesti 286 168 0.3
. cii proportions 286 168
```

Exercise 3.3: Paired t-test – test of means

High levels of vitamin E are thought by some to be protective against cancer. This hypothesis was investigated by measuring Vitamin E in stored blood samples from 271 men in a large cohort study initially aged 35-64 years who subsequently developed cancer. These values were compared with those from control men who had not, at that time, developed cancer. One control for each case was selected randomly from within the cohort study, subject to matching for age (within 5 years), duration of storage of the blood sample (within 3 months) and smoking status.

The data is in the Stata file **vit_e** which is available from the AT Moodle page. The variables are:

set: case-control unique identifier taking values from 1 to 271

case: 1=case, 2=control

vit_e: vitamin E, mg/dl

i) Launch Stata and load the **vit_e** data file. We start by sorting the data and then using browse to look at the data. Note that the data are in long format i.e. two rows per set.

```
. sort set case
. browse
```

ii) Produce a graph linking the vitamin E results for each of the matched sets.

```
. twoway connected vit_e case, c(L) xlab(1 2,value) lwidth(0.1pt)
      msize(1pt) lcol(gs2%30)
```

Note that **c(L)** means connect with a straight line in ascending order of x variable (case); **xlab(1 2, value)** means put tick marks at 1 and 2 and use the value labels as labels; **lwidth(0.1pt)** means make the line width 0.1 point; **msize(1pt)** means make the marker size 1 point in size; **lcol(gs2%30)** means use grey-scale 2 (black) for line colour with 30% transparency.

iii) Use Stata's **reshape** command to convert the data to wide format in order to carry out the paired t-test.

```
. reshape wide vit_e, i(set) j(case)
. list in 1/6
. label var vit_e1 Vitamin E Cases
. label var vit_e2 Vitamin E Controls
```

iv) Carry out a paired t-test to investigate whether or not vitamin E levels differ between cases and controls. Interpret your result.

```
. ttest vit_e1=vit_e2
```

v) What assumptions are made in carrying out this test? Use the **histogram** command to construct a graph to investigate whether one of these assumptions is justified. What do you conclude?