5.1 Data Management and Exploration

a)

```
. use preg.dta , clear

. tab smoke

    Smoking |
     status |      Freq.     Percent        Cum.
------------+-----------------------------------
     Smoker |         14       48.28       48.28
 Non-smoker |         15       51.72      100.00
------------+-----------------------------------
      Total |         29      100.00

. summ bwt

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
         bwt |         29    3.291724    .4653575       2.54       4.29

. codebook smoke

-------------------------------------------------------------------------------
smoke                                                           Smoking status
-------------------------------------------------------------------------------
                  Type: Numeric (double)
                 Label: smokelab

                 Range: [1,2]                         Units: 1
         Unique values: 2                         Missing .: 0/29

            Tabulation: Freq.   Numeric  Label
                           14         1  Smoker
                           15         2  Non-smoker
```

There are 29 observations and no missing values. The variable smoke takes values 1 for Smokers and 2 for Non-smokers.

5.2 Comparison of population means

a) Obtain summary statistics for weight by smoking group using tabstat command:

```
. bysort smoke:summ bwt

-------------------------------------------------------------------------------
-> smoke = Smoker

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
         bwt |         14    3.071429    .1791371       2.74       3.35

-------------------------------------------------------------------------------
-> smoke = Non-smoker

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
         bwt |         15    3.497333    .5563204       2.54       4.29
```

Alternatively using `tabstat`:

```
. tabstat bwt , by(smoke) stat(n mean var)

Summary for variables: bwt
Group variable: smoke (Smoking status)

     smoke |         N      Mean  Variance
-----------+------------------------------
    Smoker |        14  3.071429  .0320901
Non-smoker |        15  3.497333  .3094924
-----------+------------------------------
     Total |        29  3.291724  .2165576
-------------------------------------------
```

	Smoking group (i = 1)	Non-smoking group (i = 2)
Sample size, n_i	14	15
Sample mean, $\bar{x}_i$	**3.071**	**3.497**
Sample variance, $\hat{\sigma}_i^2$	0.0321	0.3095
Pooled variance $\hat{\sigma}^2$	0.1759	

Pooled variance $\hat{\sigma}^2 = \dfrac{(n_1-1)\hat{\sigma}_1^2 + (n_2-1)\hat{\sigma}_2^2}{n_1+n_2-2} = \dfrac{13\times 0.0321 + 14\times 0.3095}{14+15-2} = 0.1759$

Note: this is not the same as the overall variance i.e. if we were to combine the two groups.

b) Testing H_0: $(\mu_2 - \mu_1) = 0$ against the alternative hypothesis H_1: $(\mu_2 - \mu_1) \neq 0$.

Since the pooled variance σ^2 is unknown, we need to use the following *T*-statistic:

$$T = \frac{(\bar{Y}_2 - \bar{Y}_1) - (\mu_2 - \mu_1)}{\hat{\sigma}\sqrt{((1/n_2) + (1/n_1))}} = \frac{(3.497 - 3.071) - 0}{\sqrt{0.1759}\times\sqrt{((1/15) + (1/14))}} = 2.73 .$$

Under H_0, *T* follows a *t*-distribution with $(n_1 + n_2 - 2)$ degrees of freedom. From statistical tables (or use commands such as **display invttail(27,0.025)** in Stata), $t_{27,0.975} = 2.05$, $t_{27,0.99} = 2.47$ whilst $t_{27,0.995} = 2.77$. Since 2.73 > 2.47 but < 2.77 the 2-sided *p*-value can be reported as being < 0.02.

Formal interpretation of the *p*-value: The probability of observing a difference in sample means as large or larger than 0.426 kg (the observed difference) in magnitude if in truth there is no difference in the population means is < 0.02.

A 95% CI for $(\mu_2 - \mu_1)$ is: $(\bar{Y}_2 - \bar{Y}_1) \pm t_{(n_1+n_2-2,\, 1-\alpha/2)}\sqrt{\hat{\sigma}^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$

$= 0.426 \pm 2.0518\times(0.1559) = (0.11 \text{ to } 0.75)$ kg.

Formal interpretation of the confidence interval: If we repeatedly sample from the population, each time constructing a 95% confidence interval, then there is a 95% probability that a particular confidence interval will include the difference in mean birth weights between children born to non-smoking and smoking mothers in the population.

Alternatively we could say that the observed difference (0.43kg) is consistent (in the sense that a hypothesis test would yield $p>0.05$) with all population differences (mean birth weight in non-smokers – mean birth weight in smokers) in the range 0.11 to 0.75 kg and inconsistent (in the sense that a hypothesis test would yield $p < 0.05$) with all population differences outside this range.

To perform the unpaired t-test in Stata use the following command:

```
. ttest bwt, by(smoke)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       1 |      14    3.071429    .0478764    .1791371    2.967998    3.174859
       2 |      15    3.497333    .1436413    .5563204    3.189253    3.805413
---------+--------------------------------------------------------------------
combined |      29    3.291724    .0864147    .4653575    3.114712    3.468737
---------+--------------------------------------------------------------------
    diff |           -.4259047    .1558681               -.7457197   -.1060898
------------------------------------------------------------------------------
    diff = mean(1) - mean(2)                                      t =  -2.7325
Ho: diff = 0                                     degrees of freedom =       27

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0055         Pr(|T| > |t|) = 0.0109          Pr(T > t) = 0.9945
```

From the Stata output, we can see that we should reject the null hypothesis with 2-sided p-value = 0.0109 which agrees with our result in b) above (2-sided p-value <0.02).

Further (apart from the trivial change of sign) the 95% confidence interval in the Stata output agrees with that calculated by hand above.

(c) The assumptions made for the t-test are:

(1) The observations in each group follow a normal distribution.

(2) All observations are independent.

(3) The two population variances σ_1^2 and σ_2^2 are equal.

d) Since the two sample variances are very different we suspect that the third assumption may be violated. To test the assumption that $\sigma_1^2=\sigma_2^2=\sigma^2$ use the variance ratio test.

By hand calculate F by dividing larger variance by smaller variance and comparing to F-distribution on n1-1, n2-1 degrees of freedom.

Using Stata:

```
. sdtest bwt, by(smoke)

Variance ratio test
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       1 |      14    3.071429    .0478764    .1791371    2.967998    3.174859
       2 |      15    3.497333    .1436413    .5563204    3.189253    3.805413
---------+--------------------------------------------------------------------
combined |      29    3.291724    .0864147    .4653575    3.114712    3.468737
------------------------------------------------------------------------------
    ratio = sd(1) / sd(2)                                         f =   0.1037
Ho: ratio = 1                                 degrees of freedom =   13, 14

    Ha: ratio < 1               Ha: ratio != 1                 Ha: ratio > 1
  Pr(F < f) = 0.0001         2*Pr(F < f) = 0.0002          Pr(F > f) = 0.9999
```

The output demonstrates that there is statistically significant evidence that the variances are not equal.

e) Relaxing the assumption of equal variance the test statistic becomes:

$$\mathrm{T} = \frac{(\bar{Y}_2 - \bar{Y}_1) - (\bar{\mu}_2 - \bar{\mu}_1)}{\sqrt{\hat{\sigma}_2^2/n_2 + \hat{\sigma}_1^2/n_1}} = \frac{(3.497 - 30.71) - (0)}{\sqrt{\frac{0.3095}{15} + \frac{0.0321}{14}}} = 2.81$$

Under H_0, T follows a t-distribution with n^* degrees of freedom, where

$$n^* = \frac{\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right)^2}{\left[\left(\frac{\hat{\sigma}_1^2}{n_1}\right)^2 \Big/ (n_1 - 1)\right] + \left[\left(\frac{\hat{\sigma}_2^2}{n_2}\right)^2 \Big/ (n_2 - 1)\right]} = 17.06 \quad .$$

From statistical tables (or Stata), $t_{17,0.975} = 2.11$, $t_{17,0.99} = 2.57$ whilst $t_{17,0.995} = 2.90$. Since 2.81 > 2.57 but < 2.90, we reject the null hypothesis that there is no difference between the mean birth weights of babies in the two groups, with 2-sided p-value < 0.02.

In Stata this test can be performed as follows.

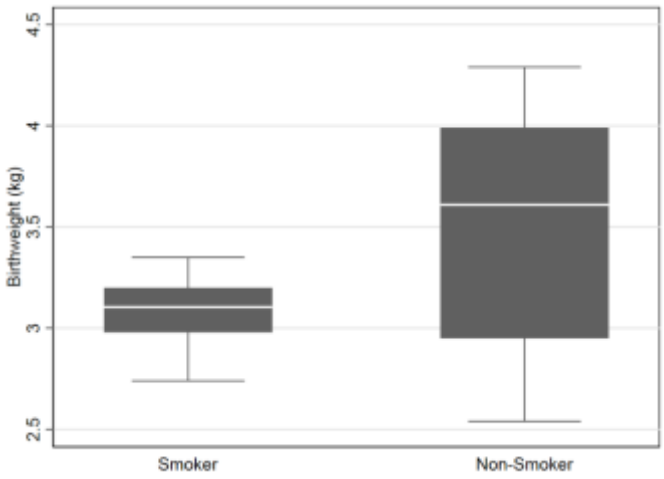
```
. ttest bwt, by(smoke) unequal

Two-sample t test with unequal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       1 |      14    3.071429    .0478764    .1791371    2.967998    3.174859
       2 |      15    3.497333    .1436413    .5563204    3.189253    3.805413
---------+--------------------------------------------------------------------
combined |      29    3.291724    .0864147    .4653575    3.114712    3.468737
---------+--------------------------------------------------------------------
    diff |           -.4259047      .15141                -.745271   -.1065385
------------------------------------------------------------------------------
    diff = mean(1) - mean(2)                                      t =  -2.8129
Ho: diff = 0                     Satterthwaite's degrees of freedom =  17.0567

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0060         Pr(|T| > |t|) = 0.0119          Pr(T > t) = 0.9940
```

f) Create a box-plot of babies' birthweight by mother's smoking status

```
. graph box bwt, over(smoke)
```



The box plot suggests that both the mean and variance of birth weight are greater in the babies of non-smoking mothers than smoking mothers. This visual impression is supported by the hypothesis tests in c) and d).

g) For both variants of the t-test, as the value taken by the most extreme observation increases the statistical significance of the difference between the groups decreases, even though the difference in the means increases. This is because extreme values have a greater impact on the estimated variance (common or group specific) than on the estimated difference in the means.

Table of results from the two sample t-test with equal variances

Max value	Mean diff	P-value	SE diff
5.29	0.49	0.0166	0.193
6.29	0.56	0.0300	0.244
7.29	0.63	0.0483	0.303
8.29	0.69	0.0684	0.365
9.29	0.76	0.0882	0.429
10.29	0.83	0.1066	0.495
11.29	0.89	0.1234	0.561
12.29	0.96	0.1384	0.628
13.29	1.03	0.1519	0.696
14.29	1.09	0.1638	0.763

5.3 Comparison of population proportions

(a) Create a binary indicator variable for low birthweight.

```
. gen lbw=bwt<3
. lab var lbw "Low birthweight"
. lab def lbwlab 1 "<3kg" 0 "3+ kg"
. lab val lbw lbwlab

. tab lbw

        Low |
birthweight |      Freq.     Percent        Cum.
------------+-----------------------------------
      3+ kg |         21       72.41       72.41
       <3kg |          8       27.59      100.00
------------+-----------------------------------
      Total |         29      100.00
```

(b) Obtain two-way table to obtain the proportion of low-birthweight babies in each group.

```
. tab lbw smoke , col

       Low |   Mother's smoking
birthweigh |        status
         t |    Smoker  Non-Smoke |     Total
-----------+----------------------+----------
     3+ kg |        10         11 |        21
           |     71.43      73.33 |     72.41
-----------+----------------------+----------
      <3kg |         4          4 |         8
           |     28.57      26.67 |     27.59
-----------+----------------------+----------
     Total |        14         15 |        29
           |    100.00     100.00 |    100.00
```

Percentage low birth weight is 28.6% and 26.7% in the smoker and non-smoker groups respectively.

(c) Carry out test. Fisher's exact test more appropriate here given small numbers (expected values less than 5 in two cells).

```
. tab lbw smoke , col  exact

       Low |
birthweigh |        smoke
         t |         1          2 |     Total
-----------+----------------------+----------
         0 |        10         11 |        21
           |     71.43      73.33 |     72.41
-----------+----------------------+----------
         1 |         4          4 |         8
           |     28.57      26.67 |     27.59
-----------+----------------------+----------
     Total |        14         15 |        29
           |    100.00     100.00 |    100.00

          Fisher's exact =                 1.000
   1-sided Fisher's exact =                 0.617
```

Interpretation of the *p*-value: The probability of observing a difference in sample means as large, or larger, than the observed difference if in truth there is no difference in the population means is 1.0. Here the *p*-value equals 1 because, conditional on the marginal means, there is no other arrangement of the data that is less extreme than that observed.

In reporting the result we could simply state that there is no evidence (p = 1.0) of a difference in the proportion of babies with a birth weight below 3kg in the two groups.

(d) By hand calculate relative risk (95% CI) for LBW comparing smoking to non-smoking mothers. Denote the observed probability of a low birth weight baby in smoking mothers by P_1 (R_1/n_1) and that in non-smoking mothers by P_2 (R_2/n_2). Denote the respective population parameters by π_1 and π_2.

Observed relative risk $\frac{P_1}{P_2} = \frac{4/14}{4/15}$ = 1.07. So $\log\left(\frac{P_1}{P_2}\right)$ = 0.06899

Standard error of $\log\left(\frac{P_1}{P_2}\right) = \sqrt{\frac{(1-P_1)}{R_1} + \frac{(1-P_2)}{R_2}} = \sqrt{\frac{1-(4/14)}{4} + \frac{1-(4/15)}{4}} = 0.6016$

95% CI for $\log(\pi_1/\pi_2)$ = 0.069 ± 1.96*0.6016 = (-1.110, 1.248)

Approximate 95% CI for (π_1/π_2) = (exp(-1.110) , exp(0.1535)) = (0.33, 3.48)

As anticipated from the result of the hypothesis test, the 95% CI does include the null value of one. The size of the confidence interval shows that the data are compatible with a wide range of population values for the relative risk, ranging from a three-fold relative risk in smoking mothers (compared with non-smoking mothers) to a three-fold relative risk in non-smoking mothers (compared with smoking mothers). The wide confidence interval reflects the small size of the study.

(e)

```
. cs 4 4 10 11
                 | smoker                  |
                 |   Exposed   Unexposed   |      Total
-----------------+-------------------------+-----------
           Cases |         4           4   |          8
        Noncases |        10          11   |         21
-----------------+-------------------------+-----------
           Total |        14          15   |         29
                 |                         |
            Risk |  .2857143    .2666667   |   .2758621
                 |                         |
                 |      Point estimate     |    [95% Conf. Interval]
                 |-------------------------+------------------------
 Risk difference |         .0190476        |   -.3066504    .3447456
      Risk ratio |         1.071429        |    .3295285    3.483642
 Attr. frac. ex. |         .0666667        |   -2.034639    .7129441
 Attr. frac. pop |         .0333333        |
                 +-------------------------------------------------
                              chi2(1) =     0.01  Pr>chi2 = 0.9087
```

The risk ratio and 95% CI is 1.07 (0.33, 3.48).