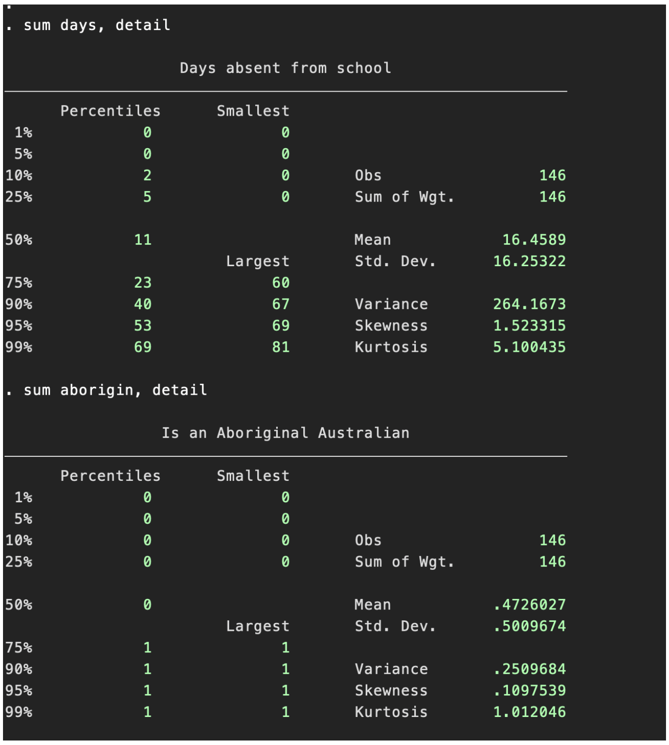
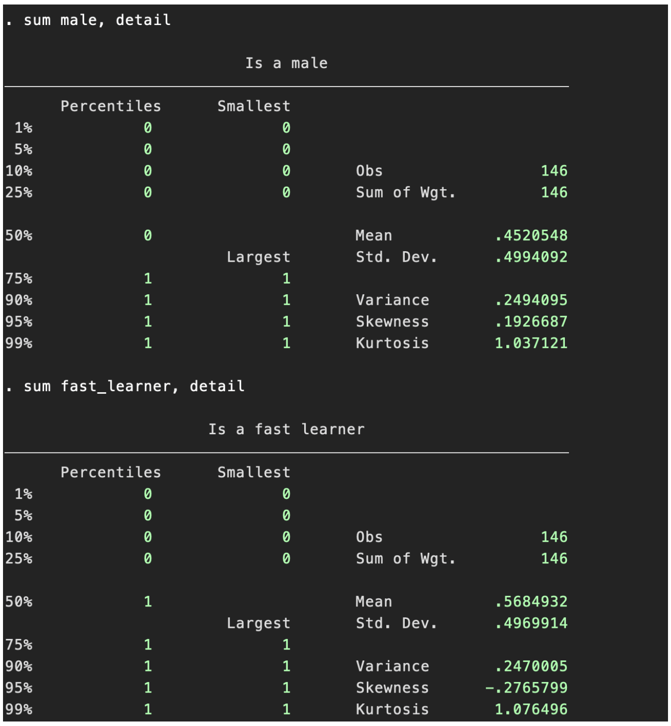
Exercise 1:

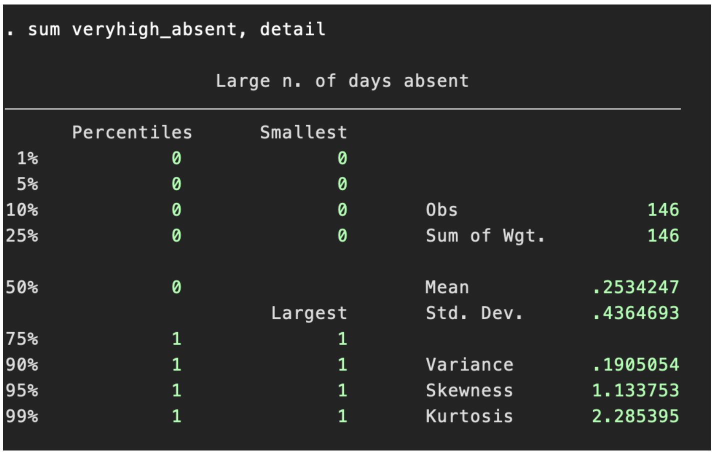
a)



```
.
. sum days, detail

                   Days absent from school
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            2              0       Obs                 146
25%            5              0       Sum of Wgt.         146

50%           11                      Mean            16.4589
                        Largest       Std. Dev.      16.25322
75%           23             60
90%           40             67       Variance       264.1673
95%           53             69       Skewness       1.523315
99%           69             81       Kurtosis       5.100435

. sum aborigin, detail

                 Is an Aboriginal Australian
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            0              0       Obs                 146
25%            0              0       Sum of Wgt.         146

50%            0                      Mean            .4726027
                        Largest       Std. Dev.       .5009674
75%            1              1
90%            1              1       Variance        .2509684
95%            1              1       Skewness        .1097539
99%            1              1       Kurtosis        1.012046
```



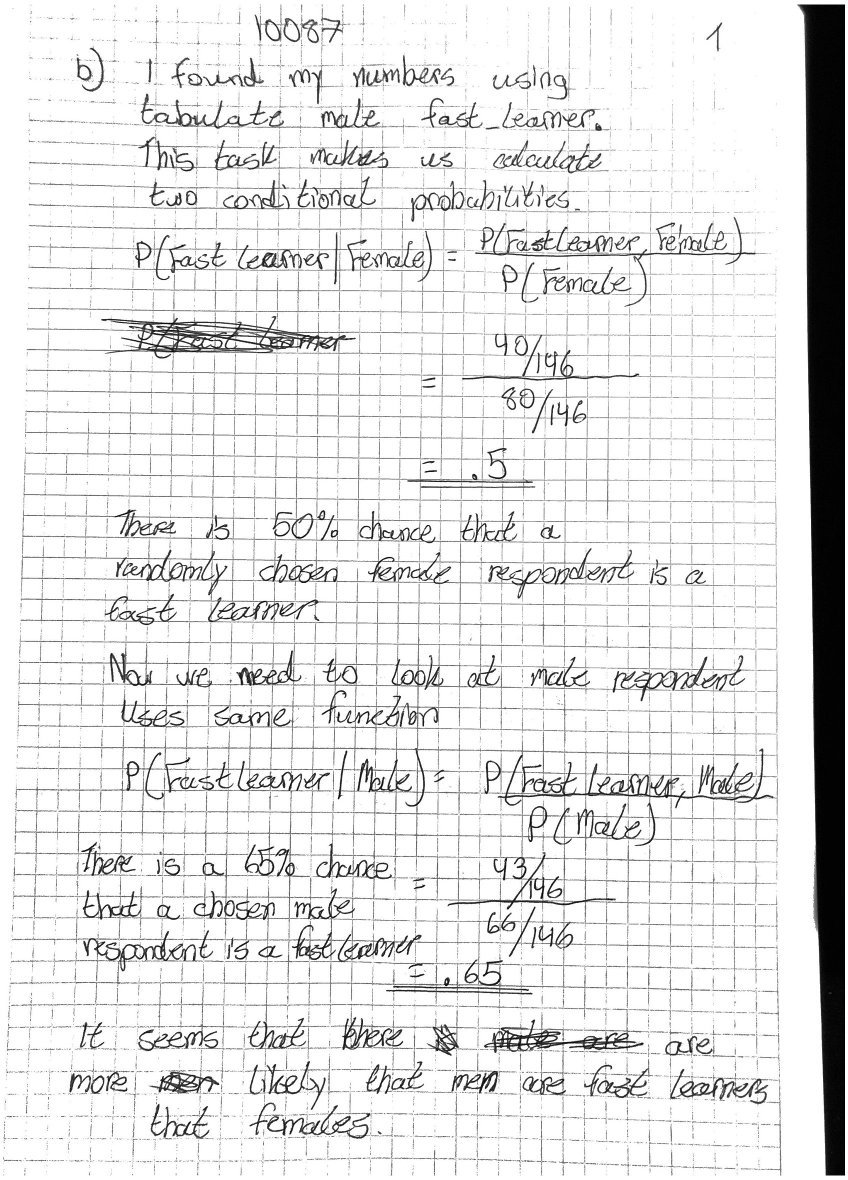
```
. sum male, detail

                          Is a male
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            0              0       Obs                 146
25%            0              0       Sum of Wgt.         146

50%            0                      Mean            .4520548
                        Largest       Std. Dev.       .4994092
75%            1              1
90%            1              1       Variance        .2494095
95%            1              1       Skewness        .1926687
99%            1              1       Kurtosis        1.037121

. sum fast_learner, detail

                      Is a fast learner
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            0              0       Obs                 146
25%            0              0       Sum of Wgt.         146

50%            1                      Mean            .5684932
                        Largest       Std. Dev.       .4969914
75%            1              1
90%            1              1       Variance        .2470005
95%            1              1       Skewness       -.2765799
99%            1              1       Kurtosis        1.076496
```



```
. sum veryhigh_absent, detail

                   Large n. of days absent
-------------------------------------------------------------
      Percentiles      Smallest
 1%            0              0
 5%            0              0
10%            0              0       Obs                 146
25%            0              0       Sum of Wgt.         146

50%            0                      Mean           .2534247
                        Largest       Std. Dev.      .4364693
75%            1              1
90%            1              1       Variance       .1905054
95%            1              1       Skewness       1.133753
99%            1              1       Kurtosis       2.285395
```

For days, we can see the average is 16.45, were the highest observation was 81 and lowest 0 with std.dev of 16.253. For aborigin, male, fast_learner and veryhigh_absent have the highest observation of 1 and lowest of 0. 1 means that a person is aborigin, male, fast learner and/or veryhigh_absent. For aborigin there is a mean of .47, so 47% is a aborigin of the sample. 45,2% are male. 56,84% are fast learner and 25% have very high absent. The standard deviation gives use the average spread amount our observations are from the mean.

b)



10087 1

b) I found my numbers using tabulate male fast_learner. This task makes us calculate two conditional probabilities.

$$P(\text{Fast learner} \mid \text{Female}) = \frac{P(\text{Fast Learner, Female})}{P(\text{Female})}$$

~~P(Fast learner~~

$$= \frac{40/146}{80/146}$$

$$\underline{\underline{= .5}}$$

There is 50% chance that a randomly chosen female respondent is a fast learner.

Now we need to look at male respondent Uses same function

$$P(\text{Fastlearner} \mid \text{Male}) = \frac{P(\text{Fast Learner, Male})}{P(\text{Male})}$$

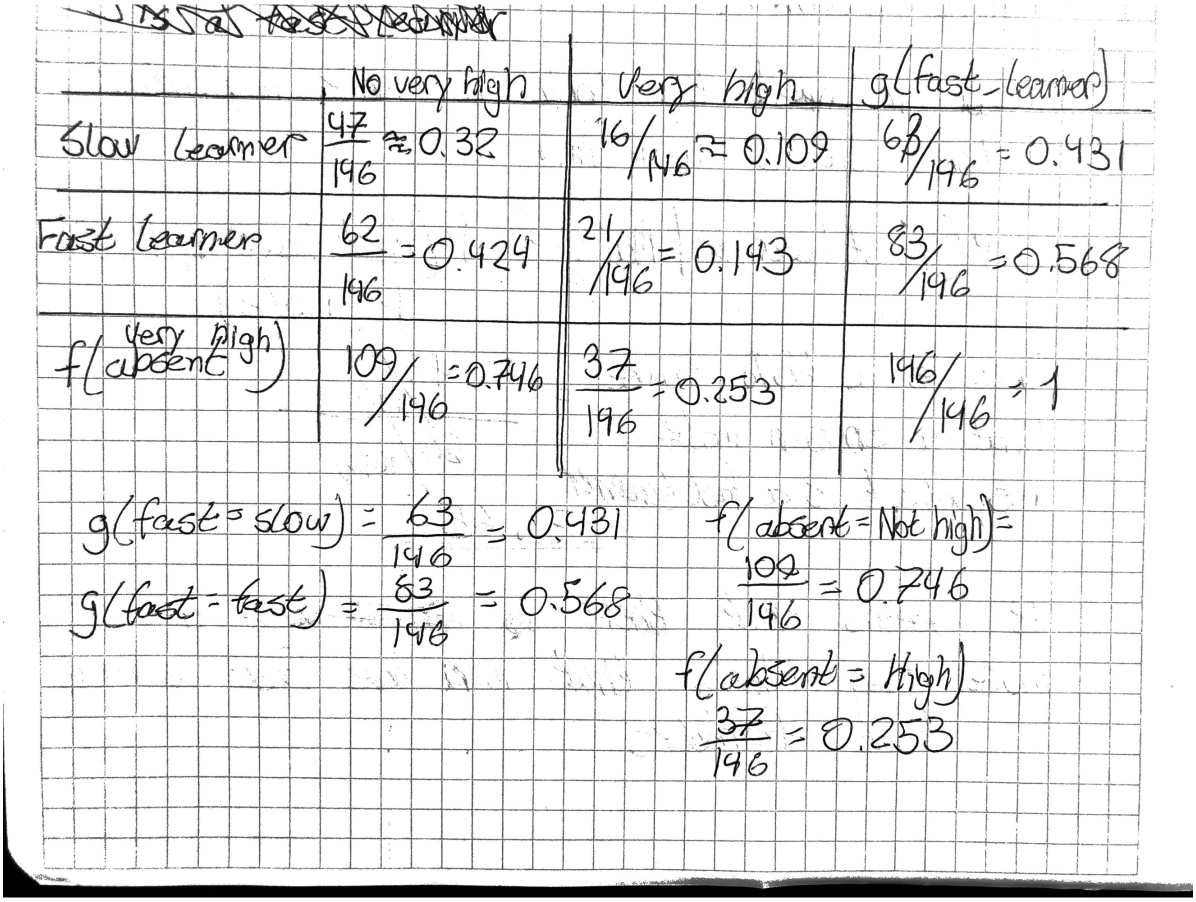
There is a 65% chance that a chosen male respondent is a fast learner

$$= \frac{43/146}{66/146}$$

$$\underline{\underline{= .65}}$$

It seems that there ~~male are~~ are more ~~men~~ likely that men are fast learners that females.

c)



	No very high	Very high	$g(\text{fast\_learner})$
Slow learner	$\frac{47}{196} \approx 0.32$	$16/146 \approx 0.109$	$63/196 = 0.431$
Fast learner	$\frac{62}{196} = 0.424$	$21/196 = 0.143$	$83/196 = 0.568$
$f(\text{absent very high})$	$109/196 = 0.746$	$\frac{37}{196} = 0.253$	$146/146 = 1$

$g(\text{fast} = \text{slow}) = \frac{63}{146} = 0.431$

$g(\text{fast} = \text{fast}) = \frac{83}{146} = 0.568$

$f(\text{absent} = \text{Not high}) = \frac{109}{146} = 0.746$

$f(\text{absent} = \text{High})$ $\frac{37}{146} = 0.253$

d)

I do not agree with this statement and its conclusion because independent is a stronger condition than that the covariance being zero. The covariance says something if there is a linear association between the fast_learner and veryhigh_absent. If on the other hand, we find that these variables are independent from each other, then there is no association at ALL (non-linear and linear) between the two variables. The covariance being zero, means that the correlation is also zero. Correlation being zero excludes as I have said earlier linear relationships between the two variables, but the relationship can be non-linear. That is why I disagree with this statement and its conclusion. The absence of days from school could have a non-linear effect on school performance, that is why it is important to try to have school open during these times.

e)

The first thing we need to consider is the sample size. Since the sample size is larger than 30, then the central limit theorem comes into play. The Central Limit Theorem says that if the sample size is big enough than the sample distribution has a normal distribution (could also use a Bernoulli distribution. We can use this to present a 99% confidence integral for the average number of days a student was absent from school.

$$\boldsymbol{Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}} \textbf{ has a N(0,1) distribution}$$

$$P(-c < Z < c) = 99\%$$

To get the c-value, we must look at the normal distribution table. Since the interval should be 99%, we need to divide 99/2. This is 49,5. The z-value that gives 0.495 is 2.58. c = 2.58, n = 146, x-bar = 16.45, sigma = 16.25. (Replacing sigma with s and saying that s is an unbiased estimator of sigma).

$$P\left(-c < \frac{\bar{X}-\mu}{\sqrt{\frac{\sigma^2}{n}}} < c\right)$$

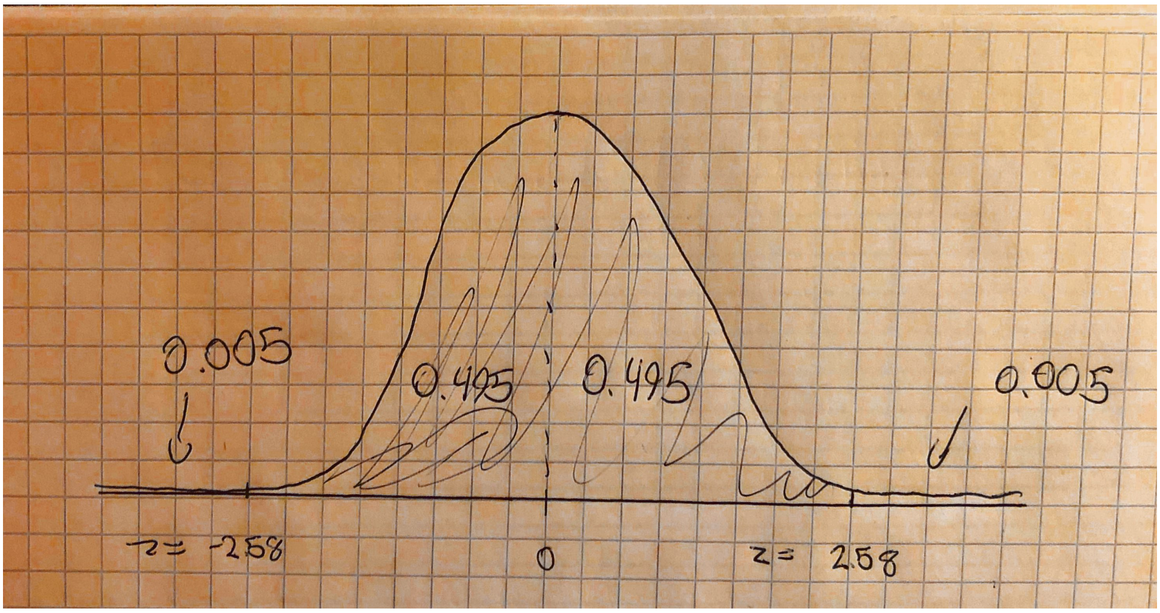
$$P\left(-c * \frac{\sigma}{\sqrt{n}} < \bar{X}-\mu < c * \frac{\sigma}{\sqrt{n}}\right)$$

$$P\left(\bar{X} - c * \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + c * \frac{\sigma}{\sqrt{n}}\right)$$

$$P\left(\bar{X} - c * \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + c * \frac{\sigma}{\sqrt{n}}\right) = 99\%$$

We can plot our values in stata. This gives us that:

P(12.98 < μ < 19.92) = 99%



This gives us a range that 99% of the times the population mean of days absent is between 12.98 and 19.92.

f)

Here we need to take into consideration what our variable veryhigh_absent says. 1 means that the person has more that 23 days absent, while 0 means that the person has less than 23 days absent. We will use this later in the task to say something about our integral.

To calculate the confidence interval, we use the same assumption as e), but change the variables. We could have used a Bernoulli normal distribution, but I´m using a normal distribution since stata from previous task has calculated the mean and standard deviation.

$$\boldsymbol{Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}} \textbf{ has a N}(\boldsymbol{\mu, \mu(1-\mu)/n}) \textbf{ distribution}$$

$$P(-c < Z < c) = 90\%$$

To get the c-value, we must look at the normal distribution table. Since the interval should be 90%, we need to divide 90/2. This is 45. The z-value that gives 0.45 is 1.64. c = 1.64, n = 146, x-bar = 0.25, sigma = 0.43. (Replacing sigma with s and saying that s is an unbiased estimator of sigma).

$$P\left(-c < \frac{\bar{X} - \mu}{\sqrt{\frac{\sigma^2}{n}}} < c\right)$$

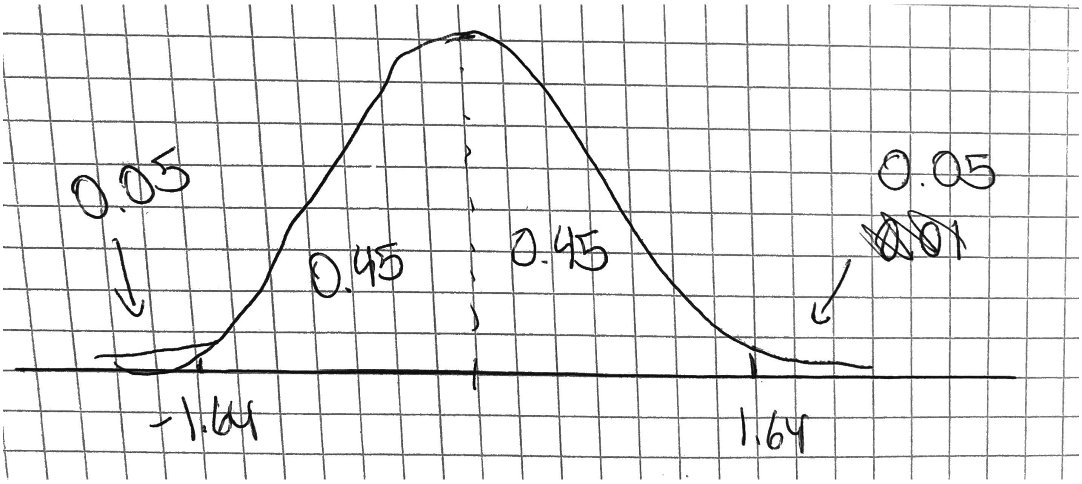
$$P\left(-c * \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < c * \frac{\sigma}{\sqrt{n}}\right)$$

$$P\left(\bar{X} - c * \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + c * \frac{\sigma}{\sqrt{n}}\right)$$

$$P\left(\bar{X} - c * \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + c * \frac{\sigma}{\sqrt{n}}\right) = 90\%$$

$$P\left(0.25 - 1.64 * \frac{0.43}{\sqrt{146}} < \mu < 0.25 + 1.64 * \frac{0.43}{\sqrt{146}}\right) = 90\%$$

$$P\left(0.25 - 1.64 * \frac{0.43}{\sqrt{146}} < \mu < 0.25 + 1.64 * \frac{0.43}{\sqrt{146}}\right) = 90\%$$



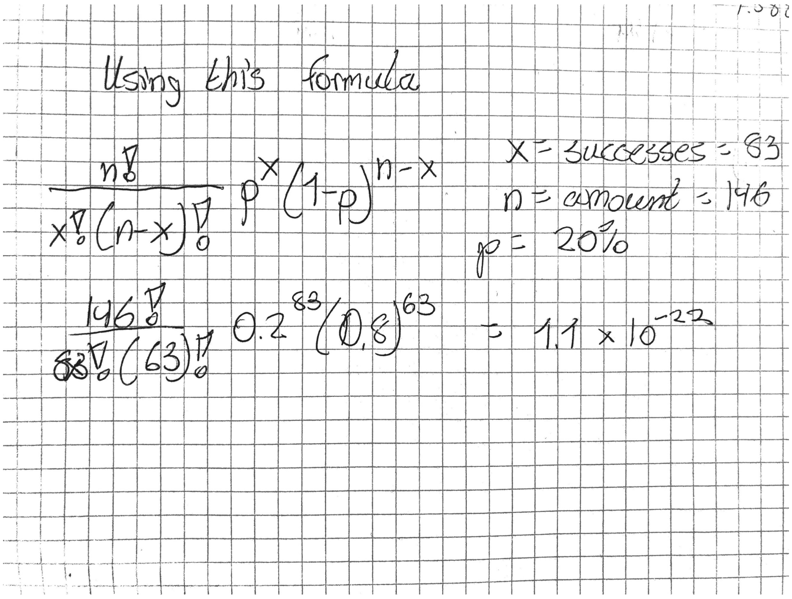
$$P(0.19 < \mu < 0.3) = 90\%$$

90% of the times our population mean of student absent over 23 days is between 0.19 and 0.3. I would say that there seems to be a low amount of student that are absent over 23 days, since the population means lays so low compared that 1 means absent over 23 days.

g)

I would have expected that the average share of fast learners in the population to be 20%, since the data from the Ministry says the share of fast learners was 20%.

h) Using Binomial distribution



Using this formula

$$\frac{n!}{x!(n-x)!} p^{x}(1-p)^{n-x}$$

x = successes = 83

n = amount = 146

p = 20%

$$\frac{146!}{83!(63)!} 0.2^{83}(0.8)^{63} = 1.1 \times 10^{-22}$$

The probability that exactly 83 students are fast learners in population of 146 students is $1.1*10^{-22}$. This is extremely low. This data suggests that there is a very small chance exactly 83 students are fast learners of the population.

i)

Four conditions need to hold:

1) Bivariate outcome, and each trial outcome can be classified as a success and a failure.

- In this setting it can be satisfied, since we can check if a student learns fast or not. But a student can also learn to be a fast learner.

2) Independent, the trails are independent.

- I would say that this isn't satisfied. They are students and students study together. So, fast learners can have influence on the slower ones. They can help the slower ones over time. So, it is difficult to say if it is independent.

3) The number of trails are fixed.

- Satisfied

4) The probability for success is the same for each trial.

- Satisfied

From our sample 56% (83/146) are fast learners.

j)

State null and alternative hypothesis.

State null = **H_0 : μ = 0.2** , which is the population mean.

Alternative hypothesis = **H_A : μ > 0.2**, increase because of the program. This means we are using a one-tail test.

Construct an appropriate test statistic.

The TS is when H_0 : μ = 0.2. Since we are saying something about the population, we should assume that H_0 is true, an allow Type 1 error (level of significance). We also assume the:

$$\boldsymbol{Z = \frac{\bar{p}-\mu}{\sqrt{p(1-p)}/\sqrt{n}}} \textbf{ has a N}(\boldsymbol{p, p(1-p)/n})\textbf{ distribution, where } \boldsymbol{\mu} \textbf{ is given}$$

Decide on the level of significance

The level of significance is 0.1.

Formulate a test criterion

The test criterion can be written as:

Reject H_0 if TS > 2.33: but reserve judgement if TS < 2.33

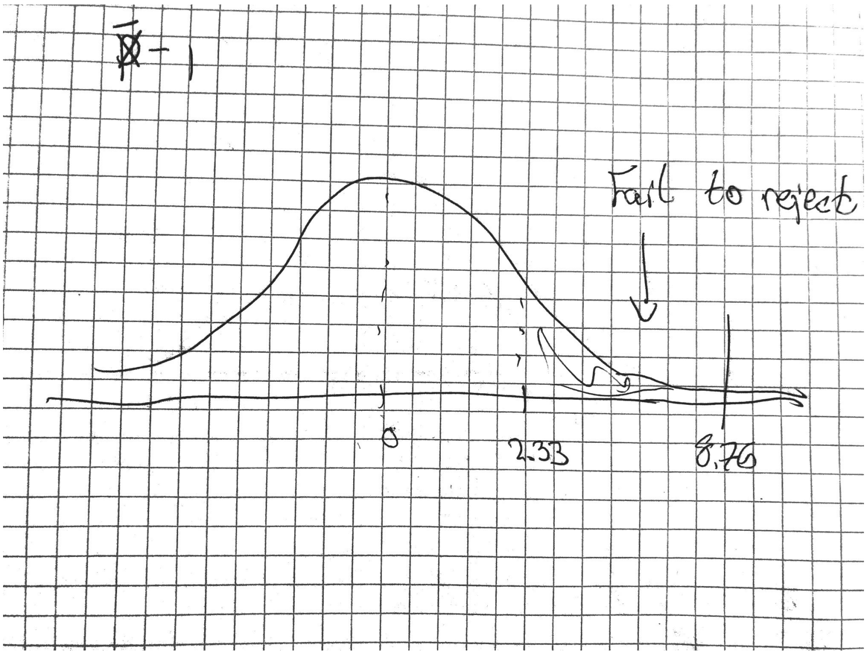
The value is taken from the table of Normal distribution table and area under right-hand tail equal 0.49. Since 0.5- 0.49 = 0.1

Take the sample and examine the sample information and compute the TS

Now we can plot in the data into our test statistic formula that we constructed. Assume that the s^2 is an unbiased estimator of variance. X-bar is 0.56, mean is 0.2, st.dev = sqrt(p(1-p)), n =146:

$$\boldsymbol{TS = \frac{0.56 - 0.2}{\sqrt{0.56 * 0.44}/\sqrt{146}}}$$

TS = 8.76



Wrote wrong

The |TS| value is 8.76. Using test criterion, we can say that:

We reject H_0 since 8.76 > 2.33, but we allow ourselves a level of significance of 0.1. It seems that the program was effective in increasing the share of fast learners.

k)

We can write the formula:

95% level gives critical value of 1.96 since we are looking at interval

$$1.96\sqrt{\frac{p(1-p)}{n}} = 0.03$$

$$1.96\sqrt{\frac{0.56(1-0.56)}{n}} = 0.03$$

$$\sqrt{n} = \frac{1.96 * \sqrt{0.56 * 0.44}}{0.03}$$

$$\sqrt{n} = \frac{1.96 * \sqrt{0.56 * 0.44}}{0.03}$$

$$\sqrt{n} = 32.43$$

$$n = 1051.74$$

The minimum sample size should be 1052 to estimate the share of fast learner in the population lies in an interval that is only 3 percentage points wide with a 95 % confidence integral.

i)

Now we are looking at the variance, not the population mean in the previous tasks, but we can still use the same steps with slight changes. We use the same variables from sum days.

State null and alternative hypothesis.

State null = $\mathbf{H_0 : \sigma^2 = 400.}$ The variance is the same. It has not changed.

Alternative hypothesis = $\mathbf{H_A : \sigma^2 <400}$. The variance has changed after the program was inroduced

Construct an appropriate test statistic.

The TS is when $H_0 : \mathbf{\sigma^2 = 2}$ is true. We are saying something about the population variance. The changes our distribution.

$\boldsymbol{Z = \frac{(n-1)s^2}{\sigma^2}}$ **has a χ^2 distribution with n – 1 degree freedom.**

Now we use chi-square distribution. χ^2 variable is the sum of all the squares of n outstanding standard normal Z variables.

Decide on the level of significance

The level of significance is 0.1.

Formulate a test criterion

The test criterion can be written as:

Reject H_0 if |TS| < 130: but reserve judgement if |TS| > 130

The value is taken from the table of χ^2 distribution with 145 degrees of freedom (146-1). Used stata command

Take the sample and examine the sample information and compute the TS

Now we can plot in the data into our test statistic formula that we constructed.

Assume that the s is an unbiased estimator of variance:

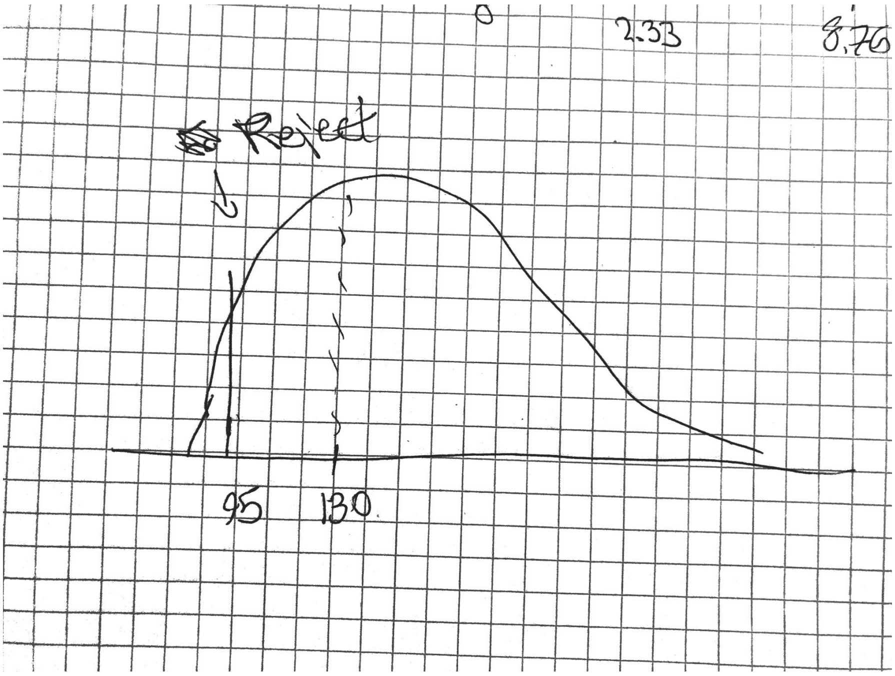
$$\boldsymbol{TS = \frac{(146-1)(16.23)^2}{(20)^2}}$$

$$\boldsymbol{TS = \frac{145 * 263,4}{400}}$$

TS = 95.48

The TS value is 95.48. Using test criterion, we can say that:

We reject H_0, since TS < 130, but we allow ourselves a failure of 0.1. It seems that the program was successful in reducing the variability in absenteeism.



Exercise 2:

a)

This is a normal distribution so we can use the table for normal distribution.

$P(-1 < X < 2) = P(-1 < X < 0) + P(0 < X < 2) = P(0 < X < 1) + P(0 < X < 2) = 0.3413 + 0.4772 = 0.8285$

b)

Rewrite:

Student t-distribution

Freedom of degree of 3

$P(X < -2.3) = P(X > 2.3)$ = around 0.05. Area 0.05 gives us the value 2.353

c)

Rewrite:

Chi-squared distribution

Freedom of degree of 9

Since the table give us the right-hand tail we can write:

$P(1.735 < X < 3.32) = P(X > 1.735) – P(X > 3.32) =$

1.735 gives us area of 0.995 while 3.32 gives us area of 0.95.

$P(1.735 < X < 3.32) = P(X > 1.735) – P(X > 3.32) = 0.995 – 0.95 = 0.045$

d)

F-distribution

1.degree = 12

2.defree = 3

Upper 5 % points

$P(X > c) = 0.05$

$P(X > 8.7446) = 0.05$. I get this from looking in the tables.

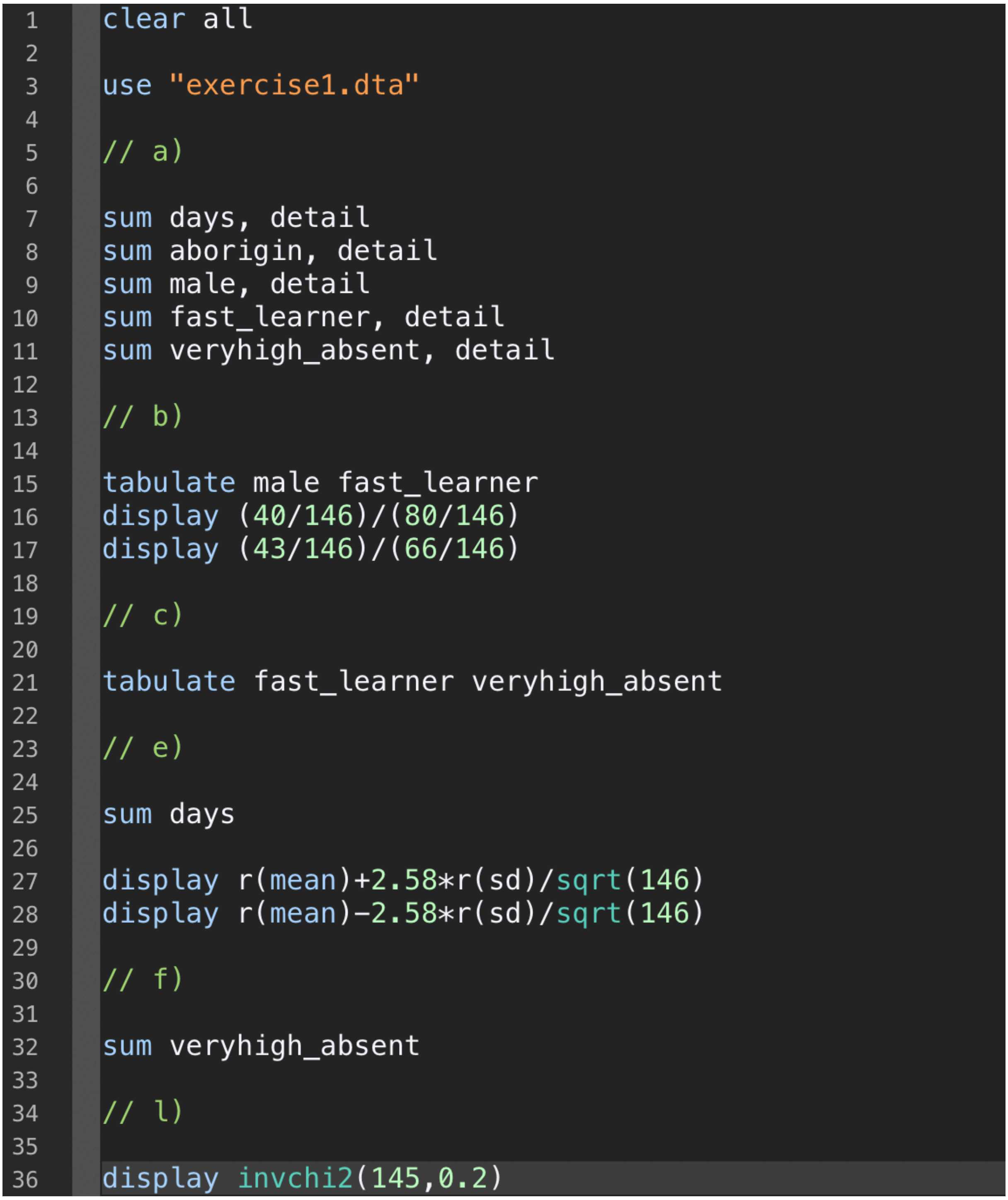
Exercise 3

a)

An estimator being unbiased means that if you take the average of your samples from a population multiple times your estimator will give you the parameter that you are interested in for example μ or σ. Say for example that you have an estimator Q (i.e sample mean). If you take the average of Q (E(Q)) multiple times, it will give eventually give you the parameter of interest (mean or variance for the population.) That is what it means for an estimator to be unbiased.

b)

To check whether an estimator is unbiased using Stata, I would draw random multiple samples of a certain size from the population that has a certain distribution (just like we did in one of our assignments). Then I would take the average of all the different estimator of each sample. If the estimator is close to the true parameters (i.e population mean) then I would say that my estimator is unbiased, since it on average delivers my parameter of interest.



```
clear all

use "exercise1.dta"

// a)

sum days, detail
sum aborigin, detail
sum male, detail
sum fast_learner, detail
sum veryhigh_absent, detail

// b)

tabulate male fast_learner
display (40/146)/(80/146)
display (43/146)/(66/146)

// c)

tabulate fast_learner veryhigh_absent

// e)

sum days

display r(mean)+2.58*r(sd)/sqrt(146)
display r(mean)-2.58*r(sd)/sqrt(146)

// f)

sum veryhigh_absent

// l)

display invchi2(145,0.2)
```