Exercise 1

The most important issue here is to use a descriptive measure that matches the type of data (numerical vs. categorical). For numerical data, some measure of central tendency (mean/median) and some measure of variation should be included, while for categorical data, frequencies and percentage is the natural choice. An example could be as follows:

Gender (n (%))	
Male	86 (57%)
Female	65 (43%)
Age (mean (SD))	54.7 (7.3)
Height (mean (SD))	171.9 (9.1)
Weight (mean (SD))	76.9 (14.3)
Marital status (n (%))	
Married/cohabitant	121 (81.7%)
Separated/divorced	10 (6.8%)
Widow/widower	9 (6.1%)
Single	8 (5.4%)
Work status (n (%))	
Yes, full time	70 (46.4%)
Yes, part time	17 (11.3%)
No	64 (42.4%)
Education (n (%))	
Primary education	30 (19.9%)
Upper secondary school	60 (39.7%)
University (1-4 years)	47 (31.1%)
University (> 4 years)	14 (9.3%)

Exercise 2

a)

Testing for associations between education and the different paint sites is most naturally done by chi-square tests. The p-values are as follows:

Education – headache: $p = 0.53$

Education – low back pain: $p = 0.76$

Education – knee pain: $p = 0.51$

Based on this, there is no association between education and any of the three pain sites. The assumptions regarding expected frequencies are also met (one cell with < 5 in all tables).

Trying to summarize, based on the observed frequencies, the proportion with headache is higher among the higher educated group. With regard to low back pain it doesn't seem to be much of a difference, while for knee pain, the proportion is higher among the lower educated group.

In the chi-square tests above, we are using all educational groups. We realize that the exercise could be read in direction of using only two groups of education also when testing. If this is done, the p-values are different (0.26, 0.71, 0.19, respectively), but we will also accept this. However, it should be noted that in this case, the assumptions are not met (still one cell < 5). Fisher's exact test is the alternative.

b)

The prevalence among women is 34/86, while among men it is 13/65. Thus, the prevalence ratio is

$\frac{34/86}{13/65} = 1.98 \approx 2$, suggesting that the prevalence of neck pain is twice as high among women as among men.

c)

The confidence interval for the odds ratio does not include the value '1'. Thus, we will conclude that the difference is statistically significant (on 5% level).

Exercise 3

In this exercise, we are looking for arguments for and against the different study designs, ideally in relation to how you decide to measure exposure and outcome.

For outcome, it might be relevant to count new cases of heart disease over time (incidence). An alternative could be to measure cholesterol level.

For exposure, it is more difficult. For the observational designs, this is not measurable and we will have to rely on questions (orally or by questionnaires) about intake of different types of coffee (or coffee prepared in different ways).

For the two types of observational study, one could use the "hard endpoint", heart disease, which is an advantage, but the disadvantage is that one has to rely on questionnaire information about intake of coffee. Since this is often a mix of different types, and it might change over time, the exposure information becomes imprecise for both cohort studies and case-control studies. In case-control studies, we also have the extra aspect of having to memorize coffee habits years back in time. In addition, there will always be confounding issues in observational studies. In a randomized study, people can be randomized to drink specific types of coffee for a given (relatively short) period, but then the natural outcome is change in cholesterol level rather than heart disease.

Exercise 4

a)

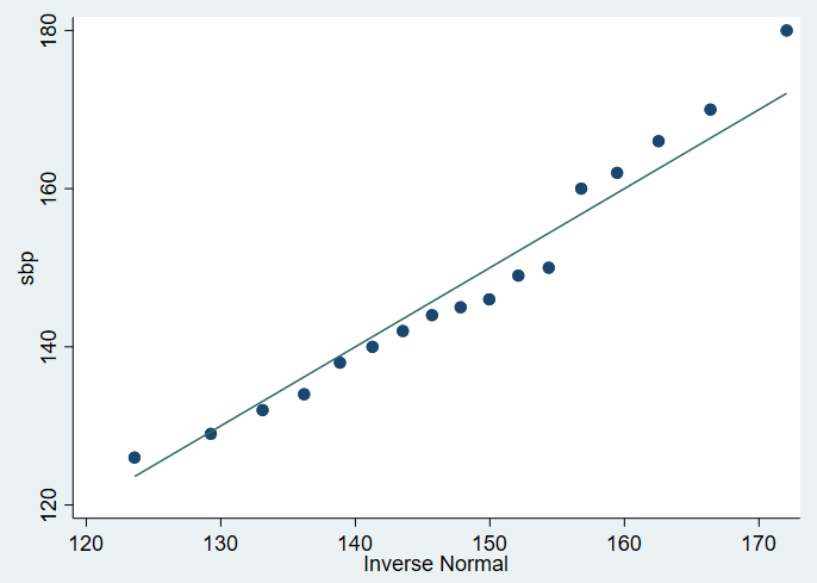
Systolic blood pressure is a numerical variable, and we are to compare two independent groups (smokers / non-smokers). This is a setup for a two-sample t-test.

The t-test gives a p-value = 0.17, hence no significant difference.

A relevant measure of association is the difference in mean values (with 95% confidence interval), given by Stata as -7.0 (-17.3, 3.2). The -7.0 indicates that the mean systolic blood pressure among non-smokers is 7 mmHg lower than among smokers.

b)

The assumptions are about normality (within each group), which can be checked by normality plots. A plot for smokers is given below, and the assumption seems okay.



c)

Stata output:

sbp	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
age	1.6045	.2387159	6.72	0.000	1.116977	2.092023
_cons	59.09162	12.81626	4.61	0.000	32.91733	85.26592

The estimated coefficient for age is 1.6 (1.1, 2.1), which suggests that for each one year increase in age, systolic blood pressure will on average increase by 1.6 mmHg. The association is highly significant ($p<0.001$).

d)

Stata output:

sbp	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
age	1.70916	.2017587	8.47	0.000	1.296517	2.121803
smk	10.29439	2.768107	3.72	0.001	4.632978	15.95581
_cons	48.0496	11.12956	4.32	0.000	25.2871	70.81211

There is still a highly significant association between age and systolic blood pressure. The estimated coefficient (1.7) hasn't changed dramatically, indicating that the influence of smoking is minor, although the estimated association has become slightly stronger (1.7 > 1.6).