W4

The data set consists of information on 300 mothers with regards to:

agefirst	Age of mother when she had her first child in year 2017
mother30	30 years old or older at the time of the baby's birth
lone	Whether the mother is a lone (single) parent
momleave	Number of days of maternity leave

four three two seven one no six five

Use the appropriate descriptive indices to identify potential typos in the data and if so, clean the data set. Use the space below to keep a record of the typos you found.

The variable **agefirst** had one ✓ dubious entry/entries.

The variable **mother30** had three ✓ dubious entry/entries.

The variable **lone** had two ✓ dubious entry/entries.

The variable **momleave** had no ✓ dubious entry/entries.

111 | 62.30 | 24.80 | 28.53 | 61.70 | 116 | 2.14 | 19.12 | 22.28 | 33.88 | 24.70 | 28.56

Fill in the blanks below to appropriately describe the data (use the clean data set).

In our sample, the mothers reported having their first child between the ages [22.28] ✓ and [33.88] ✓, (mean=[28.56] ✓, SD=[2.14] ✓). Among the 300 mothers in our sample, [24.80] ✗ *[24.70]* % were single (lone) parents. Among the women whose data are known, [62.30] ✓ % of the mothers were 30 years old or older at the time of the current birth (in 2017). On average, the women used [111] ✓ days of annual leave (SD=[19.12] ✓).

Your answer is partially correct.

0.724 | 1.000 | -0.353 | was | smaller | 2.14 | larger | χ2= | 298 | was not | 28.56 | chi-square | t | 28.6 | ≠ | t= | =

In 2016, "the average age of first-time mothers was 28.6 years". Use the appropriate test to see if this is the case for 2017, based on our data.

We need to test if the mean age of the 'first-time mothers' in 2017 is different than the mean age in 2016. Therefore the test value is [28.56] ✗ *[28.6]*. The hypotheses are:

H_0: the average age of first-time mothers in 2017 [was] ✗ *[=]* 28.6 yo.

H_a: the average age of first-time mothers in 2017 [was not] ✗ *[≠]* 28.6 yo.

The correct test to use is the one sample [t] ✓ test. In our data, the mean age was [smaller] ✓ than the mean age in 2016. This difference [was not] ✓ statistically significant ([t=] ✓ [-0.353] ✓, df=[298] ✓, p=[0.724] ✓).

was | null | 185 | 148.5 | π | μ≠30 | 112 | =0.000 | <0.001 | 17.943 | =0.001 | π=0.5 | π≠0.5 | chi-square | alternative | μ=30 | t | μ | was not

In 2016, "over half (53%) of all live births in England and Wales were to mothers aged 30 and over". Use the appropriate test to see if in 2017 among the women who gave birth, the proportion of women who were 30 or older, is equal to the proportion of the women who were younger than 30 years old.

Let [π] be the proportion of 'women aged 30 and over when they gave birth' in the population. We test the hypotheses:

H_0: [π=0.5] versus H_a: [π≠0.5]

The observed number of women in our sample who aged 30 and over was [185]. The expected number of women to be aged 30 and over under the [null] hypothesis was [148.5]. According to the one sample [chi-square] test, this difference [was] statistically significant (χ^2=[17.943], df=1, p [<0.001]).

<0.05 | significantly higher | one-sample t-test | 120 | 299 | do not reject | reject | 300 | 111 | 7.152 | accept | =0.000 | significantly lower | paired sample t-test | <0.001 | 110 | -8.480

In 2016, "on average, the mothers used 120 days of maternity leave". Use the appropriate test to infer if in 2017 the women used on average the same or different number of maternity leave days.

According to the [one-sample t-test], the average number of maternity leave days in 2017 was [significantly lower] than the corresponding in 2016 (test value μ_0=[111] *[120]*). Therefore, we [reject] the null hypothesis (t=[-8.480], df=[299], p-value[<0.001]).

=1 | reject | is | <0.05 | do not reject | -0.5 | is not | 0.004 | accept | >0.05 | 224 | do not accept | =0.95 | =0.947

In 2016, "25% of the mothers were lone parents". Use the appropriate test to infer if in 2017, the proportion of women who were single mothers was also 0.25.

According to the one-sample chi square test (χ^2 = [0.004], df=1,p-value [=0.947]) the proportion of women who were lone parents in 2017 [is not] significantly different than the one in 2016. Based on our data, we [do not reject] the null hypothesis.

W5

sex is a categorical ✓ variable (binary ✓ in particular)

age is a numerical ✓ variable (continuous ✓ in particular)

ethnicity is a categorical ✓ variable (nominal ✓ in particular)

likA is a categorical ✓ variable (ordinal ✓ in particular)

likB is a categorical ✓ variable (ordinal ✓ in particular)

anxA is a numerical ✓ variable (continuous ✓ in particular)

anxB is a numerical ✓ variable (continuous ✓ in particular)

fearA is a categorical ✓ variable (ordinal ✓ in particular)

fearB is a categorical ✓ variable (ordinal ✓ in particular)

The programme facilitator reports that the CBT programme they had run in Tanzania had equal proportions of males and females. Run the appropriate test to see if the proportions of males and females in the UK population are statistically different.

To decide on which test to use, we consider the following:

the design is a one sample design ✓ , and
gender is a categorical ✓ variable.

Hence we can use either a test from the family of the chi-square tests ✓ or, if the assumptions do not hold, a test from the family of the exact tests ✓ .

The appropriate test to use here is the one sample chi-square test ✓ .

According to the test, the proportion of males was not significantly ✓ different than that of females (chi square=2.482 ✓ , df=1 ✓ , p=0.115 ✓).

The programme facilitator reported that the average age of people who attended their programme in the USA was 45 years old. Does this compare to the UK results? Interpret the results.

To decide on which test to use, we consider the following:

the design is a one sample design ✓ , and
age is a numerical ✓ variable.

Hence we can either use a test from the family of the parametric tests ✓ or, if the assumptions do not hold, a test from the family of the non-parametric tests ✓ .

The appropriate test to use is the one sample t-test ✓ .

According to our data, the mean ✓ age of the sample is 41.68 ✗ *[significantly]* different than the test value of μ_0= 45 ✓ years old (t=-25.710 ✓ , df= 1497 ✓ , p<0.001 ✓).

The programme facilitator wants to understand if ethnicity distribution was different between the two sex groups. Run the appropriate test to see if this is the case.

To decide on which test to use, we consider the following:

the design is a [two independent samples t-test] ✗ *[two independent samples design]*, and
ethnicity is a [categorical] ✓ variable.

Hence we can either use a test from the family of the [chi-square tests] ✓ or, if the assumptions do not hold, a test from the family of the [exact tests] ✓ .

The appropriate test to use here is [Fisher's exact test] ✓ .

According to the test, the proportion of each ethnicity group [did not differ] ✓ significantly between two sex groups (p-value=[0.425] ✓). We [do not reject] ✓ the null hypothesis and we infer that there [was not] ✓ sufficient data to indicate association between the two variables.

The programme facilitator wants to test if males were more willing than females to volunteer to give a talk, before their CBT treatment. Use the appropriate test and infer on the results.

To decide on which test to use, we consider the following:

the design is a [two independent samples design] ✓ , and
likA is a [categorical] ✓ [ordinal] ✓ variable
but because the variable has more than [5] ✓ points on a response scale, we
decide to treat it as a [numerical] ✗ *[continuous]* variable for our analysis.

Hence we can either use a test from the family of the [parametric tests] ✓ or, if the assumptions do not hold, a test from the family of the [non-parametric tests] ✓ .

The appropriate test to use here is the [two independent samples Mann-Whitney test] ✓ because the likA variable has a [skewed] ✓ distribution for each sex.

According to our data, males [were not] significantly more willing than females to give a talk before their CBT treatment ([U=281621.5] ✓ , [p=0.843] ✓).

The programme facilitator want to test if males experienced more anxiety than females before their CBT treatment. Use the appropriate test and infer on the results.

To decide on which test to use, we consider the following:

the design is a [two independent samples design], and
anxA is a [numerical] ✓ variable.

Hence we can either use a test from the family of the [parametric tests] ✓ or, if the assumptions do not hold, a test from the family of the [non-parametric tests] ✓ .

The appropriate test to use here is the [two independent samples t-test] ✓ because anxA is a [symmetrical] ✓ variable within each group.

According to our data, males [experienced] ✓ significantly more anxiety than females ([t=-24.275] ✓ , [df=1497] ✗ *[df=1172.128]* , [p<0.001] ✓).

W6 & 7

Then the type of variable: **'Nominal, Ordinal, Discrete, Continuous'**

height is a Numerical ✓ Continuous ✓ variable

weight is a Numerical ✓ Continuous ✓ variable

goals is a Numerical ✓ Continuous ✓ variable

throws is a Numerical ✓ Continuous ✓ variable

points is a Numerical ✓ Continuous ✓ variable

team is a Categorical ✓ Nominal ✓ variable

Investigate the linear relationship between **'goals'** and **'points'** and fill the gaps.

There is a linear correlation × relationship between the two variables. The higher ✓ the average points scored per game, the higher ✓ the percent of successful field goals. Variable **'goals'** is normally ✓ distributed. We can see in the histogram ✓ that it is bell-shaped ✓ , and symmetrical ✓ around the mean. Variable **'points'** is not-normally ✓ distributed. We can see in the histogram ✓ that it is not bell-shaped ✓ , and not-symmetrical ✓ around the mean. Hence, we can estimate a Spearman ✓ correlation coefficient. It takes a value of 0.347 ✓ that can be interpreted as a weak ✓ positive ✓ linear relationship. The correlation can ✓ be inferred to the whole population of players because p value=0.010 ✓ . Therefore, we reject ✓ the null hypothesis of no linear correlation ✓ between the two variables and we infer that the **'goals'** are linearly ✓ associated with the **'points'**, in the whole population of players.

Our journalist knows the player scored 6 points in average. Help him estimate the percentage of successful field goals of the player, based on his points. Use the appropriate SPSS command to build a simple linear regression model that can be used to predict the **'goals'** knowing the **'points'** of a player and fill the gaps.

y=0.411+0.003x+e ✓ is the equation of the linear regression model. In context:

y represents the field goals ✓

b_0 represents the extrapolated ✓ field goals for a player that got no points ✓

b_1 represents that a 1 point difference ✓ in points is associated with a 0.003 difference ✓ in field goals.

e is the residual ✓ value between the predicted value on the regression line and the observed value.

y=0.411+0.003x+e	y=-4.080+35.338x+e	the field goals	the extrapolated	no points
a 1 point difference	0.003 difference	residual	a 2 points difference	no goals
the points	35.338 difference			

Check the inference assumptions for the linear model derived in Q6 and fill the gaps.

1) We already know from Q3 that there is a [linear] ✓ relationship between 'goals' and 'points'.

2) We plot the [histogram] ✓ of the errors and see they are [normally] ✓ distributed.

3) The error terms have [different variance] ✗ irrespective of the values of x. We can inspect this by plotting a [scatterplot] ✓ of the [standardised predicted versus standardised residual] ✓ values. The plot showed [no pattern] ✓ .

In summary conditions were [met] ✓ , which [allows] ✓ us to make inferences from the model.

Provided the four assumptions to carry out statistical inference from the model derived in Q6 were met, write the null and the alternative hypotheses for the regression coefficient b_1 and report on the results.

H_0 = [b1 coefficient equals 0] ✓

H_1 = [b1 coefficient is different than 0] ✓

Use the model derived in Q6 to help the journalist and predict the percent of successful field goals the player scored, having had a total of 6 points. Please fill the gap in his report for tonight's news:

"Kevin White, player of the Chicago Bears, played an amazing season, being a key role member for his team. The player scored 6 points in average per play, and he had a [43.03] ✓ percent of successful field goals. Kevin White, informing for ITV news, Chicago"

Half an hour before going on life, the journalist's colleague tell him the next: "Oh, I am so sorry mate, but I forgot to tell you that the percentage of successful field goals also depends on the size of the players, it seems that the weight is important." What would you recommend to the journalist?

As weight might be a confounder, please consider a multiple linear regression model, including both 'points' and 'weight' as independent variables and "goals' as the dependent variable. Fill the gaps:

The equation of the multiple linear regression model is Y= [0.207] ✓ + [0.003] ✓ x+ [0.001] ✓ z . The two variables explain a [38.5] ✓ percent of outcome variance.

Use the adjusted R2 values to compare the simple linear regression model derived in Q6 and the multiple linear regression model derived in Q10. Fill the gaps:

The model including 'points' and 'weight' shows a [higher] ✗ fit for the data, given that its adjusted R2 is [0.36] ✓ , which is [higher] ✓ than the adjusted R2 for the model just including 'points' as independent variable, that was [0.098]

If necessary help the journalist to correct his report. Use the best linear regression model to predict the field goals, given that the player got 6 total points in average per play, and his weight is 225. Please also give a 95% confidence interval for the prediction.

"Kevin White, player of the Chicago Bears, played an amazing season, being a key role member for his team. The player scored 6 points in average per play, and he had a [44.47] ✓ percent of successful field goals. Kevin White, informing for ITV news, Chicago"

The 95% confidence interval for the predicted value is [[35.20] ✓ , [53.73] ✓]

W8

Indicate whether 'Categorical' or 'Numerical' first.

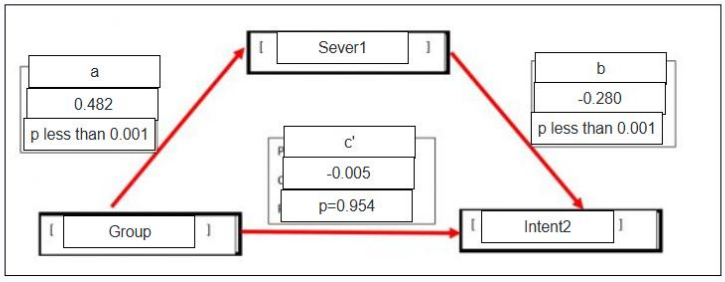
Then, the type of variable: 'Nominal, Ordinal, Discrete, Continuous'

Group is a categorical ✓ nominal ✓ variable

Intent 0, 1, 2 are numerical ✓ continuous ✓ variables

Sever 0, 1, 2 are numerical ✓ continuous ✓ variables

NB: Dragged items are bigger than the labels, so stack in the order of path, coefficient, p-value and place variable names in the boxes.



Path c' is the direct ✓ effect of treatment on the outcome.

Path c' ✓ (effect of treatment on Intent to use anabolic steroids controlling for the mediator) is equal to -0.005 ✓ , p =0.954 ✓ , with a 95% confidence interval of -0.177 to 0.166 ✓ .

Controlling for the mediator reduces ✓ the effect of the covariate (|c'| < ✓ |c|)

The direct effect is significantly not different ✓ from 0 and therefore we conclude there is complete ✓ mediation.

Indirect effect (ab ✓) can be calculated as the addition × ***[product]*** of paths a ✓ and b ✓ , which gives:

ab ✓ = (0.482 ✓) x (-0.280 ✓) = -0.135 ✓ .

The total effect c ✓ is therefore -0.135 × ***[-0.140]***

Sobel test of indirect effect is statistically significant ✓ (test statistic= -4.3197 ✓ , p <0.001 ✓).
This leads to the conclusion that perceived severity of anabolic steroid use is ✓ a mediator of the effect of treatment ✓ on Intent to use anabolic steroids ✓

W9

Then, the type of variable: 'Nominal, Ordinal, Discrete, Continuous'

BDI is a numerical ✓ continuous ✓ variable

epiNeur is a numerical ✓ continuous ✓ variable

stateanx is a numerical ✓ continuous ✓ variable

What does the regression coefficient of stateanx_X_epiNeur tells you?

Regression coefficient of stateanx_X_epiNeur represents the interaction effect ✓ between stateanx and epiNeur. The p-value 0.001 ✓ suggests that the interaction effect ✓ is statistically significant ✓ . This implies that both variables jointly affect depression ✓ , but their effects are not independent ✓ of each other. Effect of stateanx depends ✓ on epiNeur and vice-versa.

The coefficients of the variables stateanx and epiNeur do not ✓ carry their usual interpretations because of the presence of an interaction (cross-product) term involving these variables. For the interaction model, the coefficient of stateanx can be interpreted as the effect of stateanx on bdi when epiNeur=0 ✓ . The estimated coefficient 0.038 ✓ implies in people with 0 ✓ neuroticism symptoms, one unit increase in stateanx leads to 0.038 ✓ units increase in bdi. Similarly, the coefficient of epiNeur represents the effect of epiNeur on bdi when stateanx=0 ✓ . Both coefficients are not significantly different from 0 ✓ as the pvalue for the test for the stateanx's beta coefficient is 0.537 ✓ and for epiNeurs' coefficient is 0.435 ✓ .

'The estimated linear effect of epiNeur on bdi for a person with stateanx of 30 is 0.302 ✓ .'

'stateanx presents 0 ✓ influential value(s) because the absolute standardised DFBETA and DFFIT are lower than 1 ✓ . epiNeur presents 0 ✓ influential value(s) because the absolute standardised DFBETA and DFFIT are lower than 1 ✓ '

W10

The risk of having non adherance to hypertension treatment when patient received low social support is
2 ✓ times that of the patient receiving high social support

Calculate the odds of having non adherance to hypertension treatment for each group of social support and complete the interpretation below

The odds of having non adherance to hypertension treatment when the patient received low social support is about
3.857 ✓ times higher × *[larger]* than the odds of patients receiving high social support

Of those patients who did not adhere to hypertension treatment there was a higher ✓ proportion who received low social support compared to those who received high social support (65.0% ✓ versus 32.5% ✓). This difference was statistically significant ✓ according to Pearson's Chi Squared ✓ test (X2 = 5.735 ✓ , df=1, p = 0.017 ✓)

The investigator decides to further evaluate the association between social support and adherence among hypertension patients. Run the appropriate test and complete the inference paragraph below. With low social support being the reference category

The analysis results show that model was statistically significant ✓ , χ2(1) = 5.763 × *[χ2(1) = 5.763]* , p = 0.016 ✓ . The model explained (Nagelkerke R-squared) × *[12.3%]* 12.3% × *[(Nagelkerke R-squared)]* of variance in adherence. The correct classification rate has increased by 10 ✓ % to 66.7 ✓ %

Hypertension patients with high ✓ social support were 3.857 ✓ times more likely to indicate being adherent to treatment compared to those receiving low ✓ social support. This was a statistically significant result (Wald = 5.460 ✓ , df = 1 ✓ , p = 0.19 ✓ , 95% CI 1.243 - 11.968 ✓).

The investigator decides to estimate the effect of social support on adherence after controlling for age, disease years, depression symptoms, and self-compassion. Run the appropriate test and complete the inference paragraph below.

The model fit for the adjusted analysis was good ✓ , indicated by a Hosmer and Lemeshow ✓ test result that was lower × *[higher]* than the 0.05 level of significance.

Hypertension patients with high social support were 7.491 ✓ times more likely to indicate being adherent after adjusting for age, disease years, depression symptoms, and self-compassion compared to patients with low social support. The effect of social support on adherence was not ✓ statistically significant in the adjusted model (Wald = 3.796 ✓ , df= 1 ✓ , p = 0.051 ✓ , 95% CI 0.988 - 56.795 ✓).

Model shows that an increase in age by one year is associated with an increased ✓ likelihood of adherence.

The probability of a patient being adherant to hypertension treatment when they have high social support and are 58 years old, have a total score of 1 for depressive symptoms, 10 years of hypertension and a total score of 12 for self compassion is 99.8 ✓ %