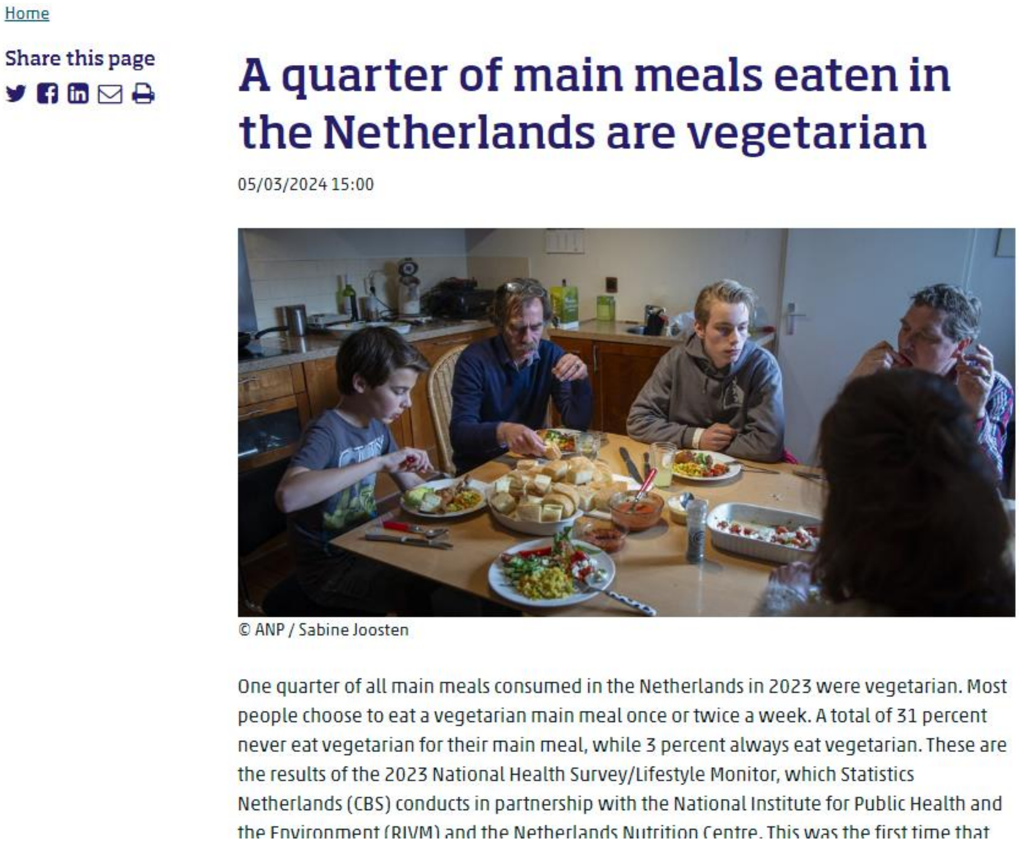
Practice exam Causal Inference 1

The article below was being posted by Statistics Netherlands



A quarter of main meals eaten in the Netherlands are vegetarian

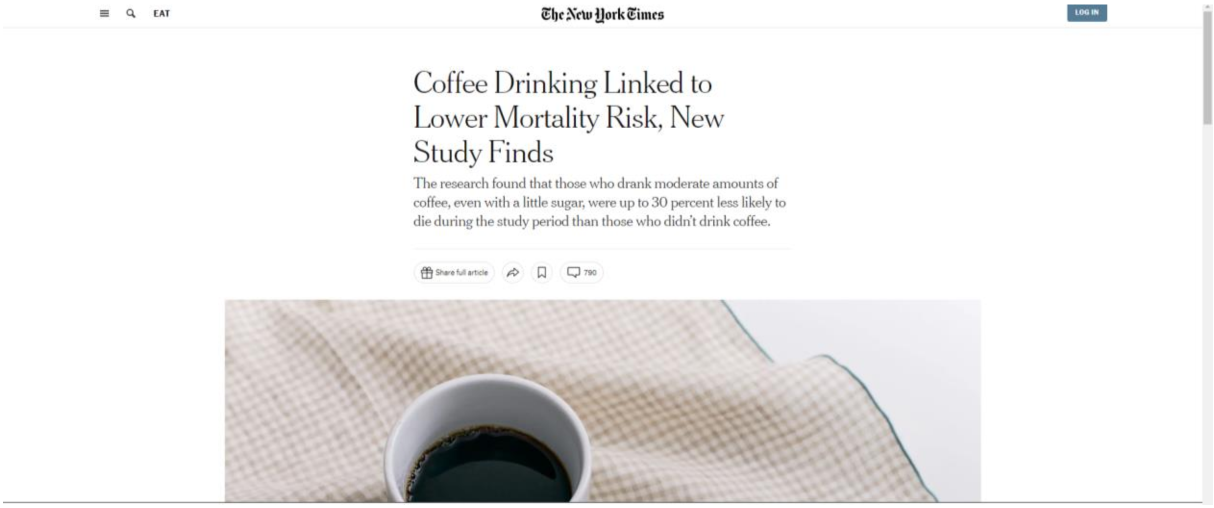
05/03/2024 15:00

© ANP / Sabine Joosten

One quarter of all main meals consumed in the Netherlands in 2023 were vegetarian. Most people choose to eat a vegetarian main meal once or twice a week. A total of 31 percent never eat vegetarian for their main meal, while 3 percent always eat vegetarian. These are the results of the 2023 National Health Survey/Lifestyle Monitor, which Statistics Netherlands (CBS) conducts in partnership with the National Institute for Public Health and the Environment (RIVM) and the Netherlands Nutrition Centre. This was the first time that

1. What type of question is being considered here? A causal question, a prediction question or a descriptive question. Descriptive

2. And what type of research question is touched below? Causal



Coffee Drinking Linked to Lower Mortality Risk, New Study Finds

The research found that those who drank moderate amounts of coffee, even with a little sugar, were up to 30 percent less likely to die during the study period than those who didn't drink coffee.

3. What are the three assumptions to identify a causal estimand? Exchangeability, positivity, consistency
4. Consider a study where the effect of following working groups on the grades of master students in Leiden is being studied. Let *Y* be the grade of a student and *A* be following working groups. Use potential outcomes notation to express the causal relative risk in the population.
   E(Y(1))/E(Y(0))

5. A randomized clinical trial compares two different forms of physical therapy for back pain. The standard form of physiotherapy is given individually, the new form is therapy in group sessions. Pain is measured with a questionnaire before and two months after the therapy has started. After randomization, 5 persons who are randomized for group sessions decline the group therapy, and receive standard therapy.

   What is the most valid way to handle the data of these 5 persons in the analysis, according to the intention to treat principle?

   Leave them in the group therapy group

6. In this randomized trial the treatment is not blinded. Which assumption could be violated because of this?

   Formulate the following key elements of a protocol for this study

   Eligibility criteria: People with backpain for whom therapy is being prescribed

   Exposure definition An offer for individual therapy vs group therapy

   Assignment procedures: Randomisation, before the start of therapy

   Follow-up period: Two months

   Outcome definition Pain measured with a questionnaire after two months

   Causal contrast of interest: E(Y(1))-E(Y(0))

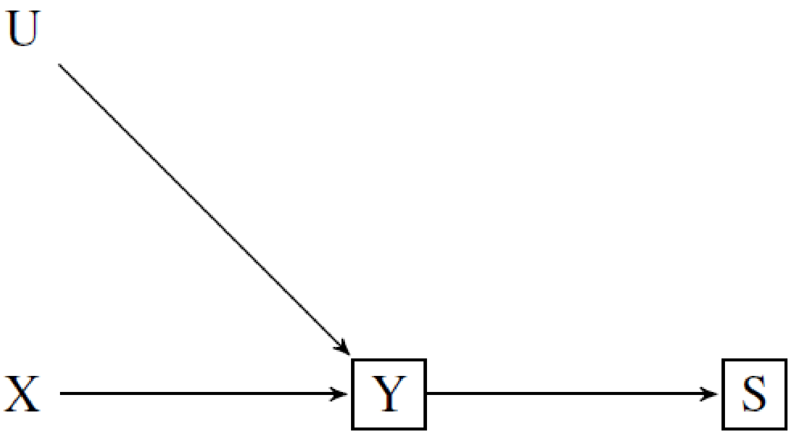
In a study, the relation between smoking of the mother during pregnancy (A) and death of infant after birth(Y) was studied. Therefore data of 100 infants who died within 28 day after birth and data of 100 infants who did not die were collected, and the mothers were asked whether they smoked during their pregnancy. The data are given below

Table: number of infants.

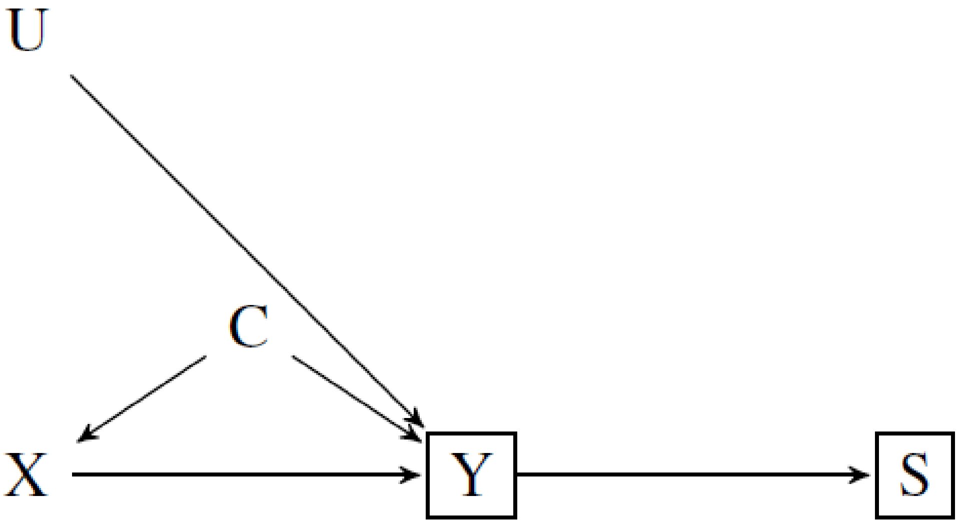
		Mortality	
		Yes	No
Smoking of mother	Yes	43	29
	No	57	71

7. What kind of study design is being used here? Case-control study (outcome dependent sampling)
8. Which measure of association would you use to summarize the difference in mortality between smokers and non smokers, and why? Odds ratio, you can show that there is selection bias in a case control study, case-control study (outcome dependent sampling)
9. Calculate this association measure. 1.85
10. Could there be confounding in this study? Draw a DAG to illustrate confounding
11. Could there be selection bias in this study? Draw a DAG to illustrate selection bias

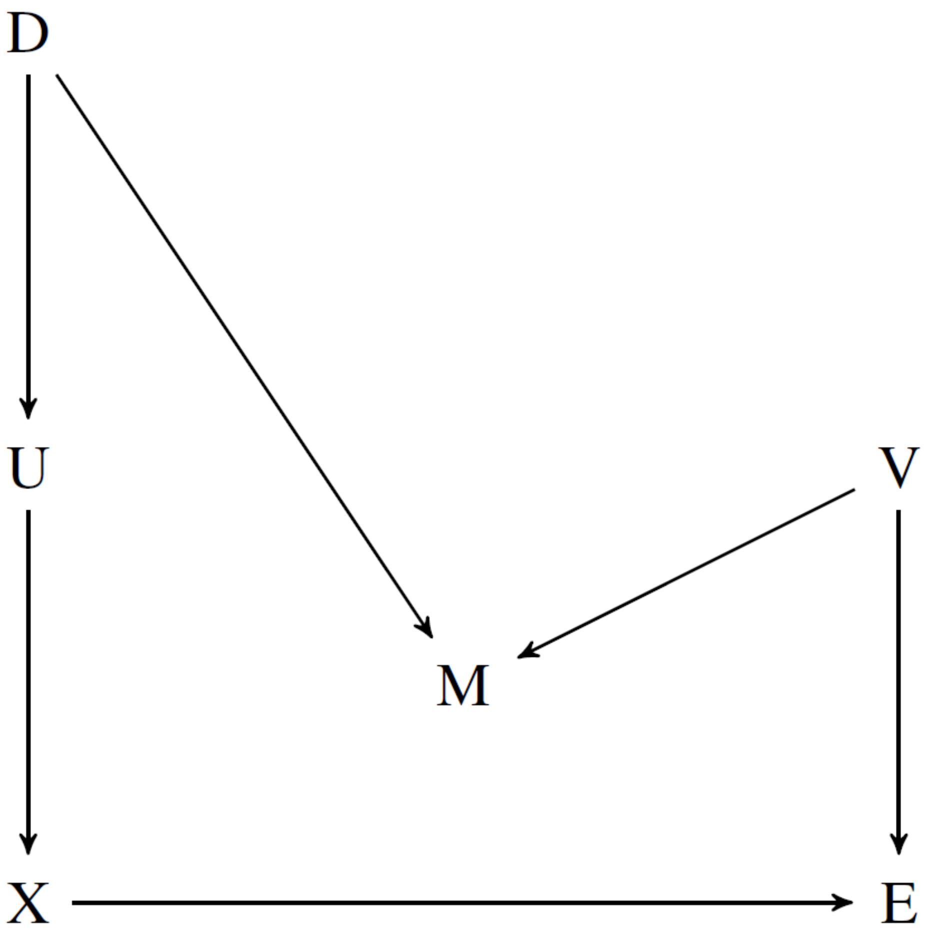
In a case control study, observations are selected based on Y. This results in a collider in Y, conditioning on Y opens a backdoor path from X to Y via U



There could also be confounding (for example alcohol use)



12. Consider the Directed Acyclic Graph (DAG) below. Consider the node X to be the treatment(exposure) of interest. In addition, consider E to be the outcome of interest. Please answer the following questions.



13. Name all the parents of M. D,Y
14. Who are the ancestors of X? E
15. Who are the children of D? U,M
16. Name all descendants of D. U,M,X,Y
17. Draw all directed paths from X to E.
18. Which set of variables d-separates X and V? empty set (they are independent)
19. What is the minimal adjustment set of variables that satisfy the backdoor criterion to determine the causal effect of X on E? ? empty set (they are independent)
20. Suppose in the analysis one is controlling for M. What is the minimal adjustment set of variables that satisfy the backdoor criterion to determine the causal effect of X on E? either U or D or V
21. Which of the listed conditional dependencies hold for the given DAG?

$D \perp V \mid M$

$U \perp M \mid D$

X ⊥ E | U, M
~~X ⊥ Z | T~~

22. Consider a study with a continuous outcome Y, an treatment X (0/1) and a confounder C (0/1). Assume that the assumptions of conditional exchangeability and consistency hold. Suppose that the following results are given.

    E(Y| C=0, X=0)= 2
    E(Y| C=0, X=1)= 3
    E(Y| C=1, X=0)= 4
    E(Y| C=1, X=1)=5
    P(C=1) = 0.4.
    P(X=1)=0.2

23. Estimate E(Y(1)) (the average potential outcome if X is set to 1 in the population), using outcome modelling with standardization (G computation) .
    $E\big(Y(1)\big)$= $E_C E(Y|C,\ X=1)$=
    $$E(Y\big|C=0,\ X=1)P(C=0)+E(Y|C=1,\ X=1)P(X=1)$$
    = 3 * 0.6 + 5* 0.4 = 3.8

In a small study, investigators examined the effect of regular alcohol use on physical exercise. Therefore they asked 8 individuals to fill in a questionnaire with questions on gender, alcohol use and daily exercise and sports. The results of this small study are given in the Table below.

Individual	Gender	More than 4 glasses of alcohol per week	Minutes of exercise per day
1	Male	Yes	30
2	Male	Yes	40
3	Male	Yes	50
4	Female	Yes	60
5	Male	No	50
6	Female	No	60
7	Female	No	70
8	Female	No	80

24. Calculate the difference in minutes of exercise per day for heavy drinkers compared to other group. Is this a causal estimate? Explain your answer. The difference is (30+40+50+60)/4-(50+60+70 +80)/4 = -20
25. Calculate the propensity score for the first individual (using gender as the only confounder)
    $P(X=1|C=1)$ = P(heavy drinker|male) = 3/4
26. Suppose that an inverse weighing analysis would be performed. What would be the weights for individual 3 and 4? Individual 3: Male, heavy drinker. P (heavy drinker|male) =3/4 → weight = 4/3

Individual 4. female, heavy drinker. P (heavy drinker|female) =1/4 → weight = 4

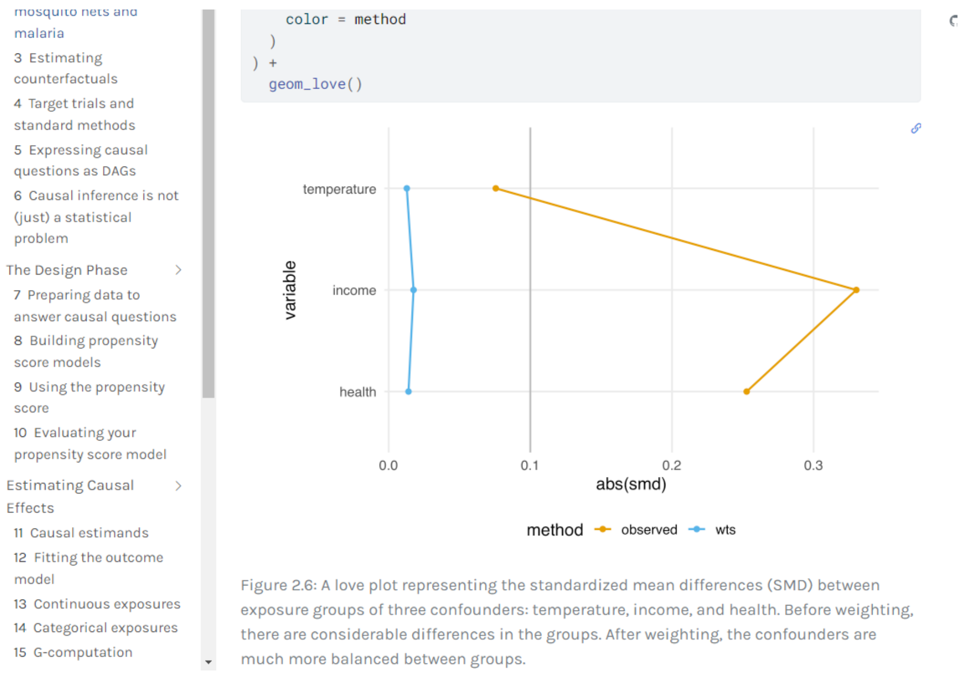
27. Suppose that the following weights would have been calculated

Individual	Gender	More than 4 glasses of alcohol per week	Minutes of exercise per day	Weights
1	Male	Yes	30	1.33
2	Male	Yes	40	1.33
3	Male	Yes	50	1.33
4	Female	Yes	60	4
5	Male	No	50	4
6	Female	No	60	1.33
7	Female	No	70	1.33
8	Female	No	80	1.33

Calculate what the mean minutes of exercise per day would be if the whole population would have been heavy drinkers

(30*1.33+40*1.33 +50*1.33 + 60*4)/(1.33+1.33+1.33+4) = 50

28. What would you conclude from a love plot which looks like this (blue is after weighting with propensity score methods, brown before)

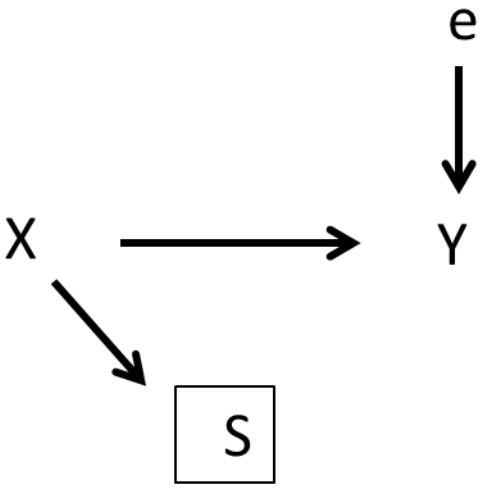


```
    color = method
  )
) +
  geom_love()
```

Figure 2.6: A love plot representing the standardized mean differences (SMD) between exposure groups of three confounders: temperature, income, and health. Before weighting, there are considerable differences in the groups. After weighting, the confounders are much more balanced between groups.

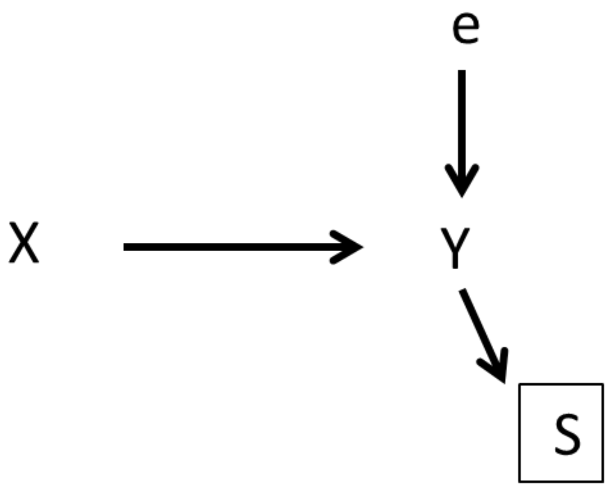
Before matching the SMDs of income and health were > 0.1 indicating that the distribution of these variables are not balanced over the groups. After matching all SMD's are small, indicating good balance and suggesting that the matched groups are exchangeable

29. Consider a study where people fill out online a questionnaire. Indicate for each of the following examples whether the missing is MCAR, MAR or MNAR.
    a. Some results have not been saved because of a power failure MCAR
    b. Older people tended to stop half way filling in the questions MAR
    c. Some people refused to fill in their weight MNAR
30. Consider a simulation study, where two variables A and Y are being generated. Interest is in the relation between A and Y. Suppose that Y is set missing if A > 3. What type of missing is this MCAR, MAR or MNAR.
31. For this simulation study, draw a DAG



32. For this example indicate whether the following approaches are consistent (i.e. asymptotically unbiased) to estimate the effect of A on Y, and whether the standard errors are estimated correctly
    a. Complete case analysis consistent, correct se
    b. Single impute missing values of Y by the mean of Y inconsistent
    c. Single impute missing values of Y by the mean of Y given A consistent, incorrect se
    d. Multiple imputation using linear regression with both Y and A consistent, correct se
33. Consider a simulation study, where two variables A and Y are being generated. Interest is in the relation between A and Y. Suppose that Y is set missing if Y > 3. What type of missing is this MCAR, MAR or MNAR.

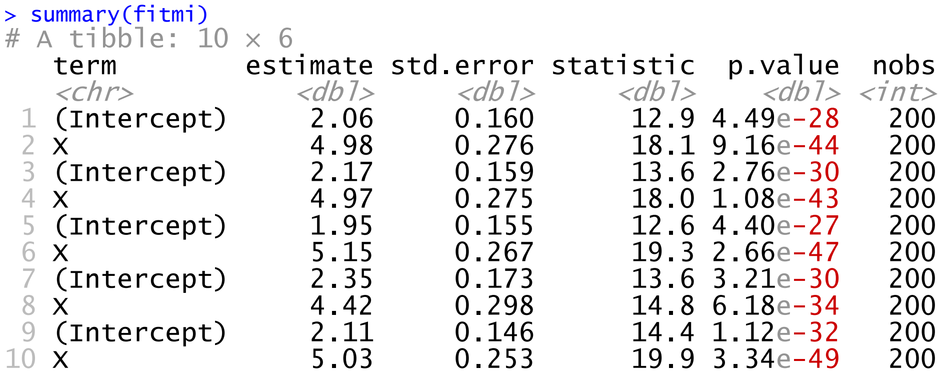
34. For this simulation study, draw a DAG



35. For this example indicate whether the following approaches are consistent (i.e. asymptotically unbiased) to estimate the effect of A on Y, and whether the standard errors are estimated correctly
    a. Complete case analysis
    b. Single impute missing values of Y by the mean of Y
    c. Single impute missing values of Y by the mean of Y given A
    d. Multiple imputation using linear regression with both Y and A

None of the methods will yield correct results

36. Below is the output of a linear model with X as only covariate, run on a 5 times imputed dataset.



```
> summary(fitmi)
# A tibble: 10 × 6
   term        estimate std.error statistic  p.value  nobs
   <chr>          <dbl>     <dbl>     <dbl>    <dbl> <int>
 1 (Intercept)     2.06     0.160      12.9 4.49e-28   200
 2 X               4.98     0.276      18.1 9.16e-44   200
 3 (Intercept)     2.17     0.159      13.6 2.76e-30   200
 4 X               4.97     0.275      18.0 1.08e-43   200
 5 (Intercept)     1.95     0.155      12.6 4.40e-27   200
 6 X               5.15     0.267      19.3 2.66e-47   200
 7 (Intercept)     2.35     0.173      13.6 3.21e-30   200
 8 X               4.42     0.298      14.8 6.18e-34   200
 9 (Intercept)     2.11     0.146      14.4 1.12e-32   200
10 X               5.03     0.253      19.9 3.34e-49   200
```

37. Use Rubin's rules to calculate the pooled estimate of the coefficient for X (use 3 decimals).

(4.98+4.97+5.15+4.42+5.03)/5=4.91

38. Calculate SD ($\hat{\beta}_i$)

$\sqrt{\frac{\Sigma(\widehat{\beta}\mathrm{i}-\widehat{\beta})^2}{n-1}}$=sqrt(((4.98-4.91)**2+(4.97-4.91)**2+(5.15-4.91)**2+(4.42-4.91)**2+(5.03-4.91)**2)/4)= 0.283

39. Suppose that SD ($\hat{\beta}_i$) = 0.283 Use Rubin's rules to calculate the pooled standard error of the coefficient for X (use 3 decimals).

$$se(\hat{\beta}) = \sqrt{\frac{1}{m}\sum_{i=1}^{m} se_i \;\; 2 + \left(1 + \frac{1}{m}\right) sd^2}$$

m=5, mean(se^2)= (0.276**2+0.275**2+0.267**2+0.298**2+0.253**2)/5 = 0.0751
Sd = 0.283

$$\sqrt{0.0751 + \left(1 + \frac{1}{5}\right) 0.283^2}$$