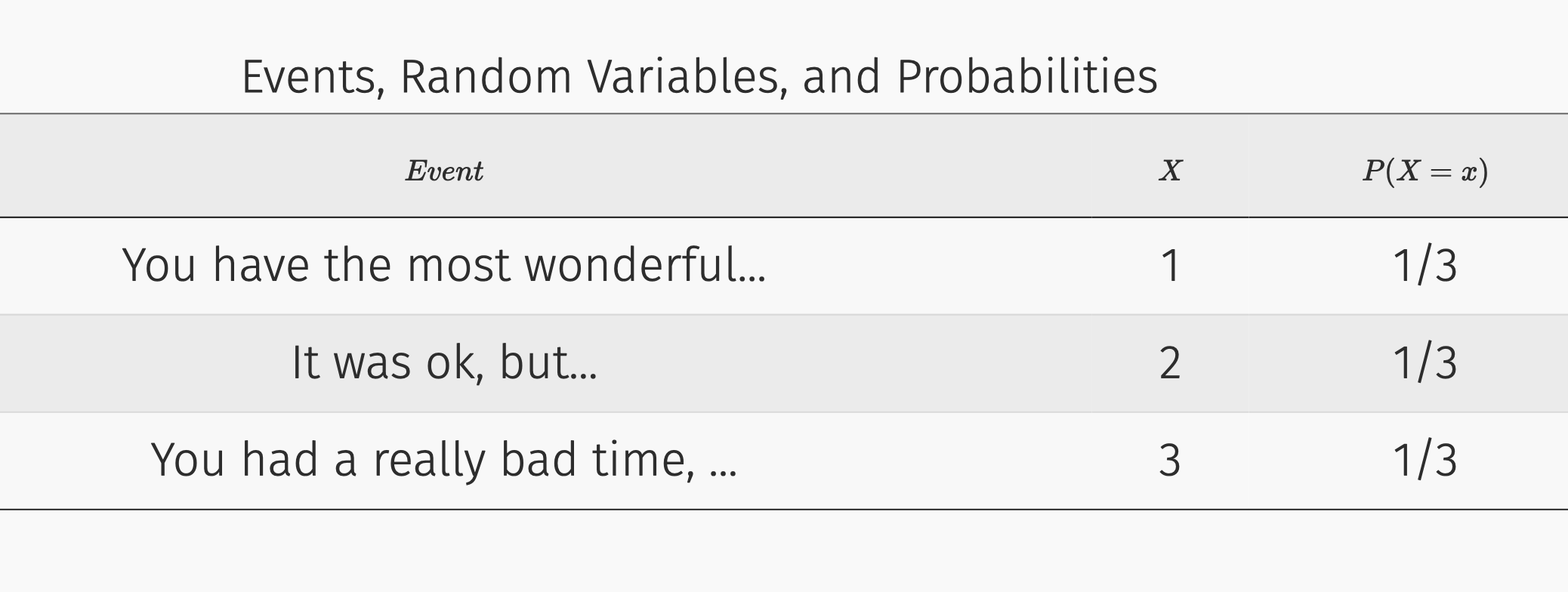
**Introduction**

1. Describe the two characteristics for the concept of “bullshit” as discussed in class.
2. The credibility revolution, has move economics away from X, and closer to Y. Where X and Y represent
   1. X = Engineering, Y = Sociology
   2. X =Medicine, Y = Sociology
   3. X =Physics, Y = Medicine
   4. X = Sociology, Y =Medicine

**Random variables and probabilities:**

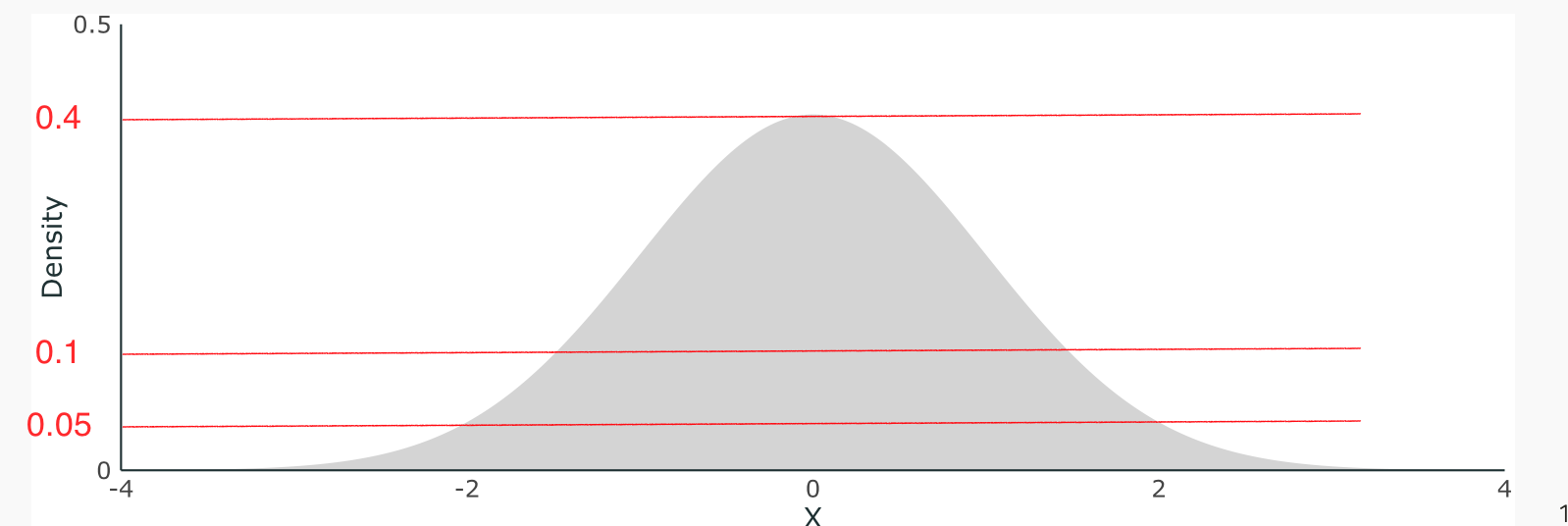
1. In what sense is a random variable similar to a variable? And in what sense is random variable similar to a function
2. What is the interpretation for the values of the y-axis in the in a density function? How can we compute the probability of an event?
3. Notation is only useful, as long as we understand what it represents. In class we discuss the following example of a random variable and its probability distribution:



(continuation of 3)

Now assume that instead of using the previous notation we would have represented the random variable but using the random variable K mapping into the numerical values k= {75, 23, -12}. Additionally, instead of using the previous notation now we want to use the letter Q to represent its probability distribution. Please rewrite the table above with the new corresponding notation.

1. For the case of a continues random variable: which distributions is a good representation of not knowing much about a phenomenon (represented by the random variable)? How about for a discrete random variable (assume 2 events for simplicity)?
2. Given the following continuous random variable X, with the following probability density functions (past any density), what is the probability that the X= -2



[from SW TB CH2]

1. The probability of an outcome
   1. is the number of times that the outcome occurs in the long run.
   2. equals *MN*, where *M* is the number of occurrences and *N* is the population size.
   3. is the proportion of times that the outcome occurs in the long run.
   4. equals the sample mean divided by the sample standard deviation.

State whether each of the following random variables is discrete or continuous: a) The number of defective tires on a car b) The body temperature of a hospital patient c) The number of pages in a book d) The number of draws (with replacement) from a deck of cards until a heart is selected e) The lifetime of a lightbulb

1. State whether each of the following random variables is discrete or continuous:

a) The number of defective tires on a car

b) The body temperature of a hospital patient

c) The number of pages in a book

d) The number of draws (with replacement) from a deck of cards until a heart is selected

e) The lifetime of a lightbulb

1. - Part 1: The probability that two random variables, X, Y take the values of x, **and** y can **always** be described as
   1. Pr(X=x) + (Y=y)
   2. Pr(X=x, Y=y).
   3. Pr(X=x) \* (Y=y)
   4. Pr(X=x) / (Y=y)

* Part 2: Using the concept of conditional probabilities, this same term (the probability that two random variables, X, Y take the values of x, **and** y) can be described as.

1. Pr(X=x|Y=y)\* Pr(X=x)
2. Pr(Y=y|X=x)\* Pr(Y=y)
3. Pr(X=x|Y=y)\* Pr(Y=y)
4. Pr(Y=y|X=x)
5. Define what is a random variable for the case of discrete events.
6. Two parts:
   1. Part1: define the concept of independence in plain Englis
   2. Part 2: Define the concept of independence in terms of conditional probabilities.
7. In which sense is the variance, an average. What is an average of?

**Expected Value, Variance, LLN and CLT**

1. The sample average \overline{Y} is

a. is a single number and as a result cannot have a distribution.

b. a random variable and has a probability distribution

c. has a probability distribution called the standard normal distribution.

d. has a probability distribution that is the same as for the  i.i.d. variables.

Answer: b

1. Compute the variance for a variable with values 1,2, and 3. Show your calculations.
2. CAn econometrics class has 80 students, and the mean student weight is 145 lb. A random sample of four students is selected from the class, and their average weight is calculated.

a) Will the average weight of the students in the sample equal 145 lb?

b) Explain why you answered yes or no to the above question

c) Is the sample average, bar{Y} , a random variable?

1. Based on what we discussed in class: what is the key characteristic that makes the data sets used in economics differ from the data sets used by an anthropologist conducting in-depth field interviews

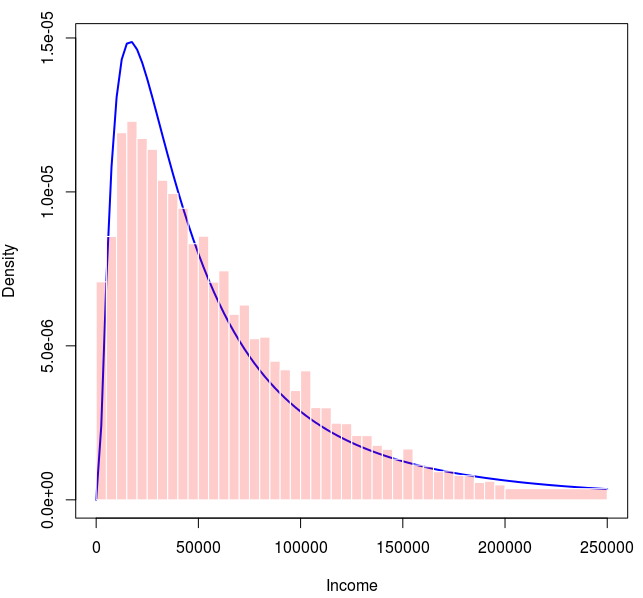
a. Economic data sets are stored in spreadsheets, while anthropologies are stored in notebooks.

b. Economic data sets are about objective indicators, while anthropologist data sets are about less tangible phenomena

c. Economic data sets are about economic activity while anthropology data sets are about cultural phenomenon

d. Economic data sets are structure data sets (typically in rectangular form) while anthropology data sets are unstructured.

1. Based on what we discussed in class: what is the key characteristic that makes the data sets used
2. In a rectangular data set: what is represented by rows and what by columns.
3. Assume that you have receive a data set with millions of observations on the income of all the American households. Which concept(s) discussed in class can you used to communicate some initial insights about these millions of observations?
4. What is the interpretation of the mean of a binary variable?
5. What is the connection between the mean of a r.v. and its expected value. In one (say which) with sum over proportions, in the sum over …?
   * Answer: E() is the population version of the mean. For the mean we some over proportions (or all obs divining by N), in the other we sum over probabilities.
6. Given a random variable X, please write out the formula for expected value of any generic function of this random variable g(X). E(g(X))=?
7. When comparing average ratings for two products, we learn that their average ratings are 3.5 and 4.1 respectively. Is this enough information two make an informed choice? What other concept from class could you used here? Propose hypothetical values for that concept (for each product), such that you are much more confident that the second product (the one with 4.1) is better.
8. What is the main reason we prefer to report (and read) standard deviations instead of variances?
9. Given a collection of random variables, Y1, Y2, …, Yn let \overline{Y} represent its sample mean. What happens with the its expected value as the number of random (n) variables increase. What happen to its variance?
10. Assuming independence and identically distributed rvs Y1, Y2, …, Yn. What is the expected value of \overline{Y}. What is its variance?
11. Given a random variable that represent the age of a population in years, choose which of the following could be a plausible value for its standard deviation, **and** explain why each of the other values are not plausible:
    * 1. 475
      2. 0.54
      3. -45
      4. 21
12. Given a data set with information on the age of each individual in the US (360 million), choose which of the following could be a plausible value for its standard deviation for the **sample mean,** and explain why.
    * 1. -32
      2. 32
      3. 320
      4. 0.32
13. Describe the law of large numbers in plain English.
14. The law of large numbers says something about the value of the sample mean as the sample size increases. But it also implies something about the variance of the sample means, what is this implication?
15. Describe the central limit theorem in plain English.
16. The income distribution is a highly asymmetrical distribution, where most of its mass (or the proportion of households) have values between 0 and $200,000. Resulting in a figure like that below. What does the central limit theorem tell us here:



* 1. That we cannot apply it because incomes are not independent from one another
  2. That as we increase the sample size distributions will become normal
  3. That as we increase the sample size, the distribution of its sample mean will become normal.
  4. That as we increase the number of observations the standard deviations will become larger and larger.

1. For the simulation of the use performed in class and sessions to explore the central limit theory:
   1. What happen with the distribution of the sample mean as we increase the sample size?
      1. Resemble more a normal distribution, and shrink its variance
   2. What happen with the distribution as we increase the number of simulations?

**Conditional probabilities:**

1. The conditional expectation of *Y* given *X*, , is calculated as follows:

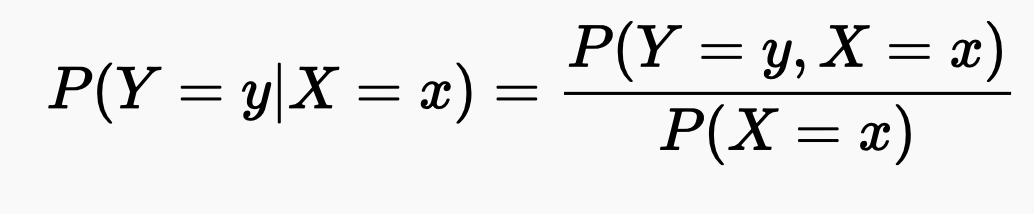
a.

b.

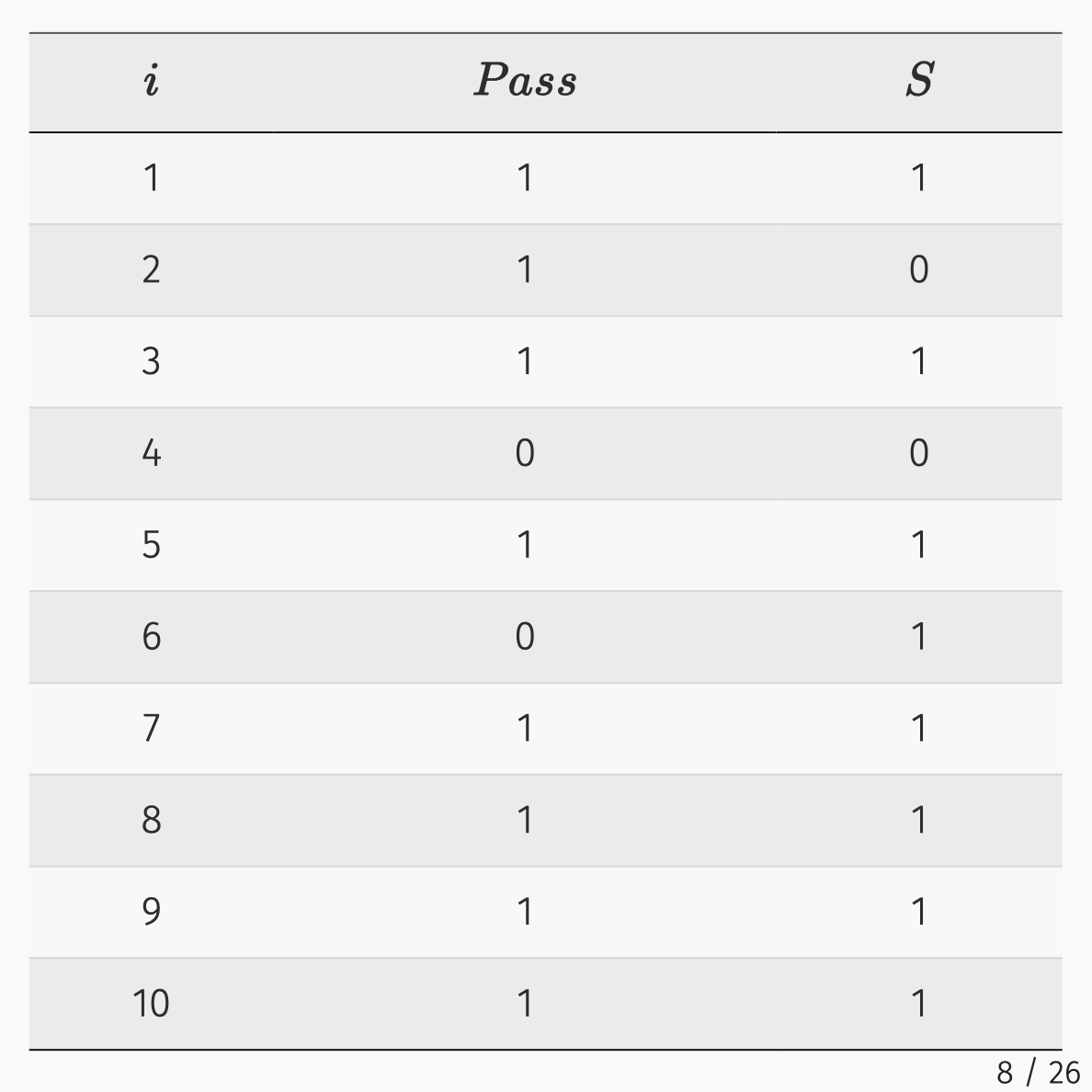
c. 

d. 

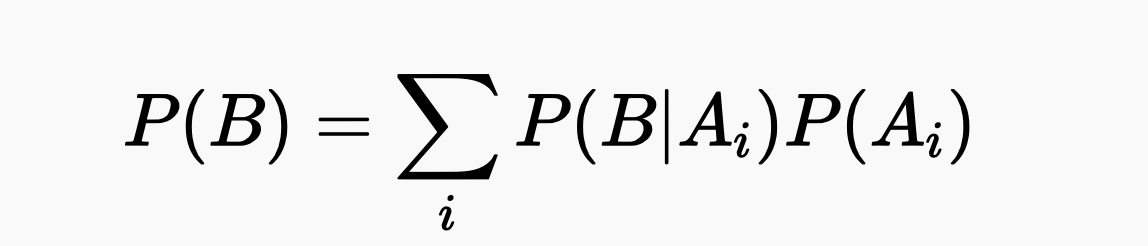
1. Remember the definition of conditional probability. Describe this equality in plain English (hint: remember the key is to “re-scale”)



1. Let’s look again at the table use to explain the intuition behind conditional probabilities. Add two new variables “Pass|S=1” and “Pass & S” and fill-in the corresponding value of each observation



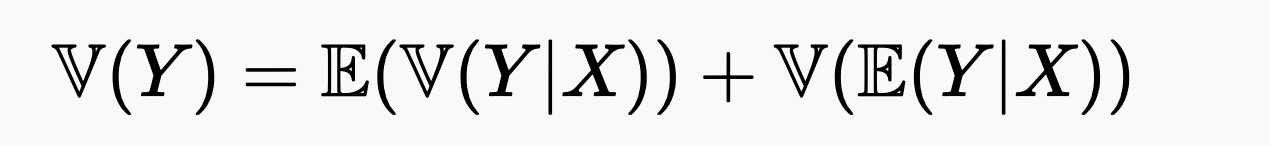
1. Given two random variable X, Y. Derive Bayes rule.
2. Explain the law of total probability (below) in plain English



1. A repeat of the [“conditionities” problem](https://www.youtube.com/watch?v=by3_weGwnMg), but with different values.

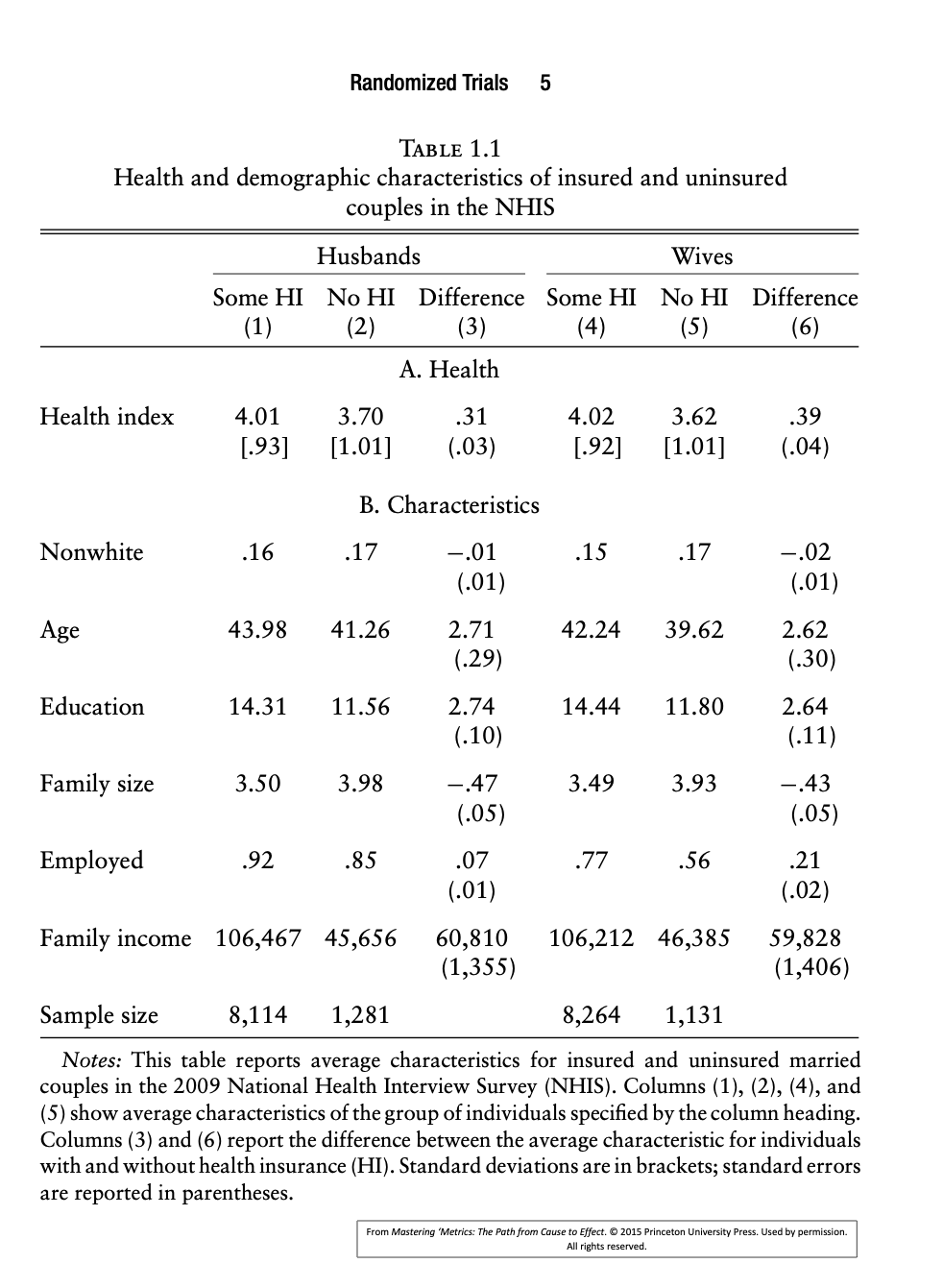
**Conditional Expectations**

1. Discuss a plausible explanation for the law of iterated expectations (Addams law). Use an example.
2. Describe Ev(v)es Law, the expression below, in plain English.

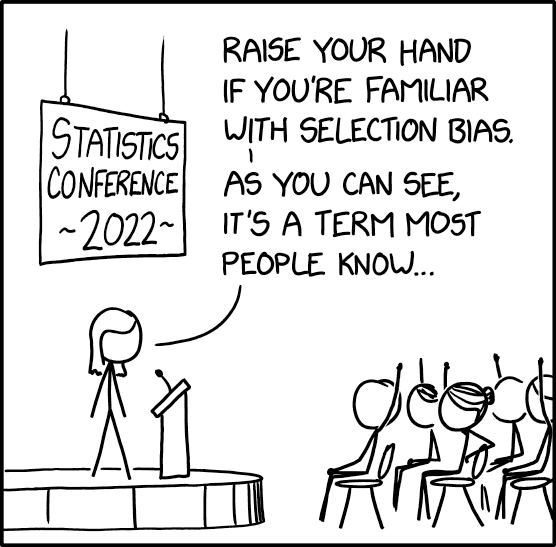


1. Ev(v)es Law tells us that the unconditional variance will always be
   1. Larger or equal than the conditional variance
   2. Smaller or equal than the conditional variance
   3. Larger when the Y is independent of X
   4. Smaller when Y is independent of X
2. [Bonus] Explain the solution to the Monty Python problem in plain english.
3. [Bonus] Derive the solution to the Monty Python problem using bayes rule

**Causal inference in the real world**

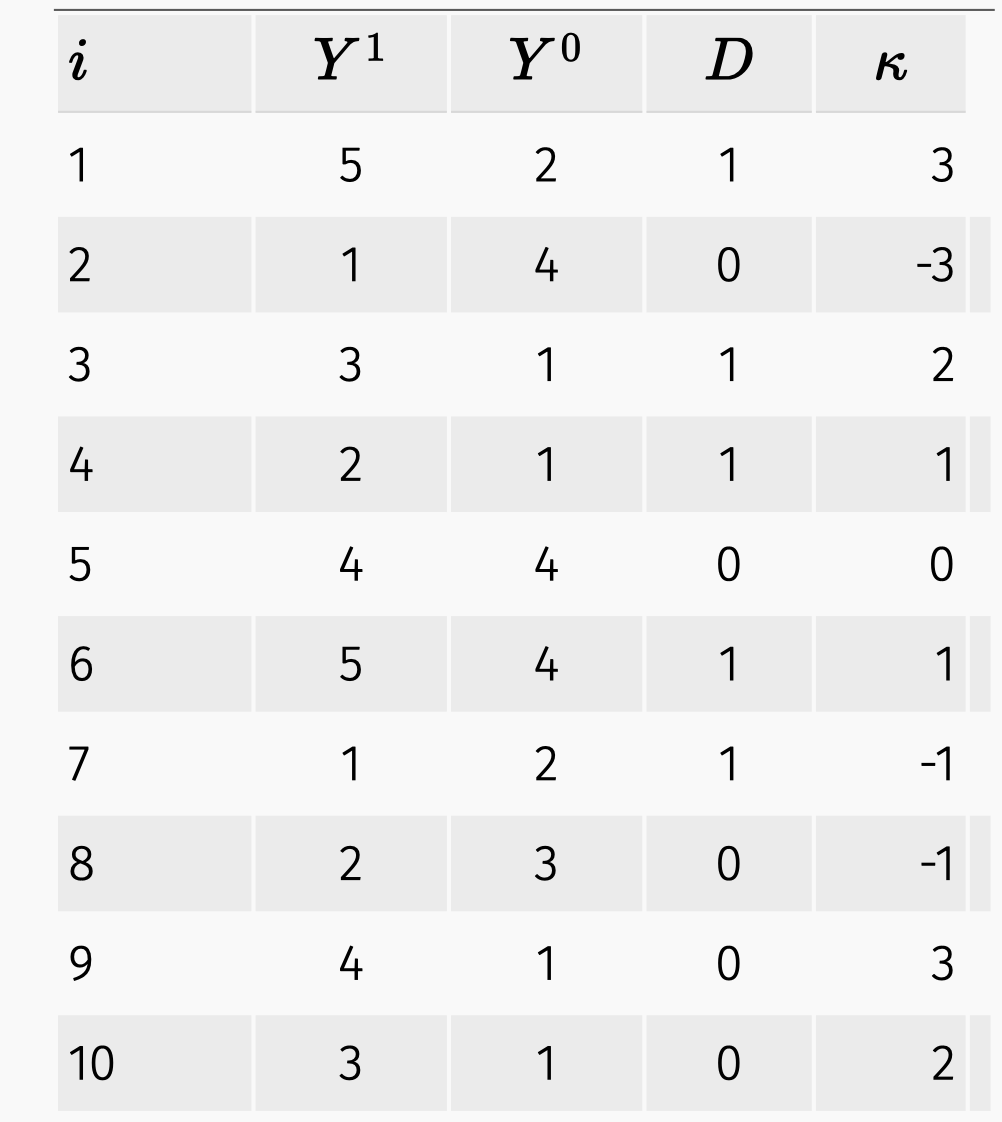
1. Ask to report the value of any given number of this table and interpret it.
2. Describe any given mean of this using the notation of conditional expectations
3. Describe what is the key assumption that is violated when we observe a spurious correlation.

**Selection Bias**

1. For the example of WWII airplanes and selection bias. What variable where the engineers conditioning on. Define DL as a random variable use to keep track of the damage in the plane due to bullets. Describe the conditional expectation that the engineers where looking at, and describe the conditional expectation that they should have been thinking about.
2. Describe the following comic by XQCD using conditional expectations (hint define two r.v.: one to keep track of knowledge of SB, and another to keep track of which group in the populations you are condition on
3. Provide a clear and short example of selection bias. Make sure to clearly describe the variable that you are interested in making inference about, and the variable that you are conditioning on.

**Potential Outcomes**

1. Describe the fundamental problem of causal inference.
2. Repeat example of Maria and Khuzdar, but with numbers changed (including treatment status). Describe what a simple comparison in outcomes would yield.
3. For the exercise of potential outcomes done in class with a fictional data of 10 individuals, cross with an x the potential outcomes that are missing from the data.



**RCTs [To be updated Tuesday Morning]**

1. Explain how randomization solves the fundamental problem of causal inference.
2. Explain any number of the tables for the RAND RCT and Oregon RCT.
3. Describe the intuition behind balance.
4. What is observational data? What is experimental data? Is the data analyzed in Table 1.1 observational or experimental? Is the data analyzed in Table 1.3 observational or experimental?
5. Describe why do we perform balance tests.

**Hypothesis testing [Not sure we will make it to here by Wednesday] RCTs [To be updated Tuesday Morning]**

1. Describe statistical significance.
2. Describe economic significance
3. Explain what is the p-value in plain English.
4. Explain the problem of p-hacking and potential solutions.

[From MRU]

Questions 60 and 61 below deal with the following study: Many college students struggle to finish their degrees. Only about half of American public college and university students finish in 6 years or less. In an effort to boost completion rates and reduce time to completion, many schools offer a variety of support services. One, called Accelerated Study in Associate Programs (ASAP), was piloted at the City University of New York (CUNY). ASAP provides additional support services like a dedicated adviser and financial aid. A recent randomized study evaluates ASAP effectiveness. Specifically, some portion of 896 eligible CUNY freshmen were assigned to receive ASAP services. Researchers estimated the causal effects of the opportunity to participate in ASAP.

1. 4. Many colleges and universities offer support services. Why do we need a randomized trial to study this? We can simply compare graduation rates for students who do and don’t use these services. What’s wrong with this reasoning?

a. The ASAP program has not been around long enough to adequately estimate its effect on graduation rates.

b. More determined students might enroll in the program, producing a selection bias.

c. No data is available for four-year colleges to compare.

d. If someone does not graduate, we do not have any information about them.

1. 5. Which of the following best describes a randomized research design? \*

a. Students chose to participate in ASAP (treatment group) or not to participate in ASAP (control group) and reported their choice to the researchers.

b. Researchers assigned students to ASAP (treatment group) based upon their need. Remaining students were placed in the control group.

c. Researchers randomly assigned students to the ASAP program (treatment group) or the control group.

d. Researchers randomly assigned students to the ASAP program (treatment group) or the control group initially, but students were allowed to switch groups after assignment to meet their needs.