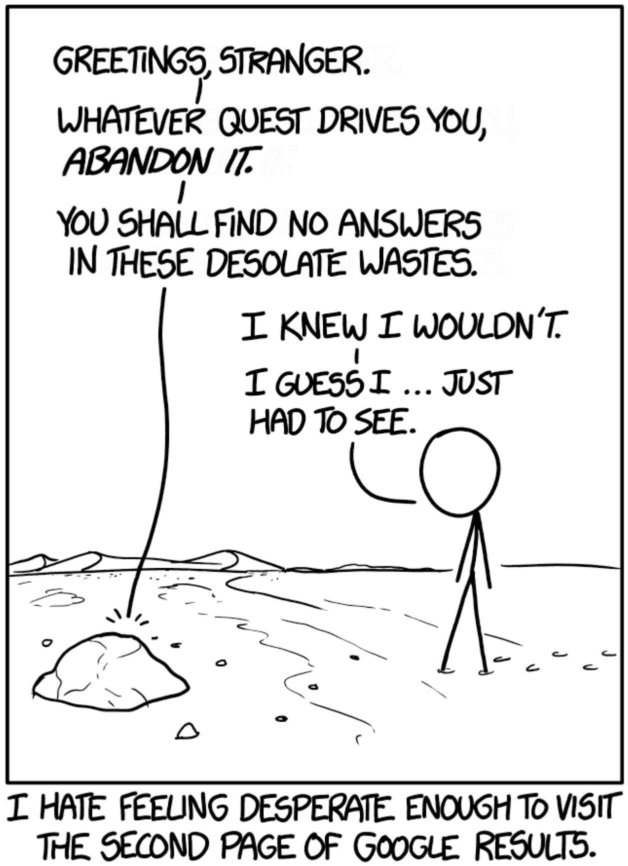
Stat 88: Probability & Mathematical Statistics in Data Science



https://xkcd.com/1334/

Lecture 14: 2/22/2021

Method of indicators

Sections 5.2, 5.3

Joint distributions

- Draw two tickets without replacement from a box that has 5 tickets marked 1, 2, 2, 3, 3. Let X_1 and X_2 represent the values of the tickets drawn on the first and second draws respectively.

	$X_2 = 1$	$X_2 = 2$	$X_2 = 3$	
$X_1 = 1$	0	2/20	2/20	4/20
$X_1 = 2$	2/20	2/20	4/20	8/20
$X_1 = 3$	2/20	4/20	2/20	8/20
	4/20	8/20	8/20	

- $S = X_1 + X\_2$, find $E(S)$

Warm up and review

- A joint distribution for two random variables, M and S, is given below. Find $E(M)$.
- Are M and S independent?

	$M = 2$	$M = 3$
$S = 2$	0	$1/3$
$S = 3$	$1/3$	$1/3$

Method of indicators

- Additivity of Expectation: This is a very useful property – no matter what the joint distribution of X and Y may be, we have:

$$E(X+Y) = E(X) + E(Y)$$

- Whether X and Y are dependent or independent, this holds, making it enormously useful.
- We also have linearity: $E(aX + bY) = aE(X) + bE(Y)$

- Recall that we talked about "classifying and counting" – so, we divide up the outcomes into those that we are interested in (successes), and everything else (failures), and then count the number of successes.
- We can represent these outcomes as 0 and 1, where 1 marks a success and 0 and failure, so if we model the trials as draws from a box, we can count the number of success by counting up the number of times we drew a 1.

- We can represent each draw as a Bernoulli trial, where $p = P(S)$

Using indicators and additivity

- For example, say we roll a die 10 times, and success is rolling a 1.
- Then p=1/6, and we can define a Bernoulli rv as $X = \begin{cases} 0, & w.p.\,5/6 \\ 1, & w.p.\,1/6 \end{cases}$
- We can also define an event A: let A be the event of rolling a 1 and define a rv I_A that takes the value 1 if A occurs and 0 otherwise.
- This is a Bernoulli rv, what is its expectation?
- Now let $X \sim Bin(10, \frac{1}{6})$, so X counts the number of successes in 10 rolls. Let's find $E(X)$ using additivity and indicators:

Using indicators

- Binomial

- Hypergeometric: Did we use the independence of the trials for the binomial? If not, we can use the same method to compute the expected value of a hypergeometric rv:

Exercise 5.7.6: A die is rolled 12 times. Find the expectation of:

a) the number of times the face with five spots appears

b) the number of times an odd number of spots appears

c) the number of faces that don't appear

d) the number of faces that do appear

Example

- Let X be the number of spades in 7 cards dealt **with replacement** from a well shuffled deck of 52 cards containing 13 spades. Find E(X).

1. Write down what X is

2. Define an indicator for the kth trial: I_k

3. Find $p = P(I_k = 1)$

4. Write X as a sum of indicators

5. Now compute $E(X)$ using additivity

- Do the same thing if we deal 7 cards **without replacement.**

Missing classes

- We can use indicators to compute the chance that something *doesn't* occur.

- For example, say we have a box with balls that are red, white, or blue, with 35% being red, 30% being white, and 35% blue. If we draw n times with replacement from this box, what is the expected number of colors that *don't* appear in the sample?

Examples

1. An instructor is trying to set up office hours during RRR week. On one day there are 8 available slots: 10-11, 11-noon, noon-1, 1-2, 2-3, 3-4, 4-5, and 5-6. There are 6 GSIs, each of whom picks one slot. Suppose the GSIs pick the slots at random, independently of each other. Find the expected number of slots that no GSI picks.

2. A building has 10 floors above the basement. If 12 people get into an elevator at the basement, and each chooses a floor at random to get out, independently of the others, at how many floors do you expect the elevator to make a stop to let out one or more of these 12 people?

5.4 Unbiased Estimators

- We showed the linearity of expectation earlier. That is,
$$E(aX + b) = aE(X) + b$$
- We often want to estimate a *population parameter*: some fixed number associated with the population
- A statistic is any number that is computed from the data sample. Usually we use a *random sample*.
- Note that the parameter is *constant* and the statistic is a *random variable*.
- We will use a *statistic* to *estimate* the parameter. It is called an *estimator* of the parameter.
- If the expectation of the statistic is the parameter that it is estimating, we call the statistic an unbiased estimator of the parameter.