USP 634 Spring 2016
Assignment 1: Descriptive Statistics; Probability and Distribution

ANSWER KEY

1. (*Sample to be posted*)

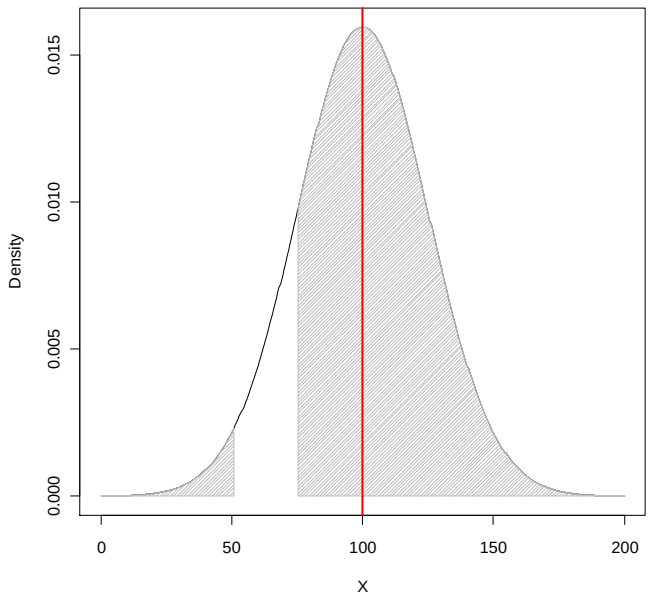
2. What is the probability of rolling two 5s with two fair dice? What is the probability of rolling snake eyes (two ones) twice in a row, followed by a four and a six, followed by a score adding to 10? (10 pts)

The probability of rolling two 5s with two fair dice:
Pr(5,5)=1/6 * 1/6 = 1/36

The probability of rolling snake eyes (two ones) twice in a row, followed by a four and a six, followed by a score adding to 10:
Pr = Pr(rolling snake eyes twice in a row) * Pr(a four and a six) * Pr(a score adding to 10)
= Pr(rolling snake eyes once) * Pr(rolling snake eyes once) * Pr(a four and a six) * Pr(a score adding to 10)
= Pr(1,1)*Pr(1,1)*(Pr(4,6)+Pr(4,6))*(Pr(4,6) + Pr(6,4) + Pr(5,5))
= (1/6 * 1/6) * (1/6 * 1/6) * (1/6*1/6 + 1/6*1/6) * (1/6*1/6 + 1/6*1/6 + 1/6*1/6)
= 1/36*1/36*2/36*3/36

Note that there are two possible ways to roll a 4 and a 6, and three possible ways to roll a score adding to 10 (4-6, 6-4, or 5-5).

3. In 2005, the average annual ozone levels in Smogsville were normally distributed with a daily mean of 100 ppb (parts per billion) and a standard deviation of 25 ppb. How many days in 2005 were smog levels either above 75 ppb (their air quality standard) or below 50 ppb? (10 pts)



You could convert this to z-scores and look up the result in the "Area Under the Normal Curve" table in the

appendix of the textbook. Better yet, note that 75ppb is one standard deviation from the mean and 50 ppb is 2 standard deviations from the mean. For the first you can add 50% to 34.13% (area between mean and Z = 1) = 84.13% * 365 = 307. For the number of days below 50 ppb you can get the area from Area Beyond Z = 2, which is 2.28% * 365 = 8. So in total 307+8 = 315 days with ozone levels **either** above 75 ppb **or** below 50ppb.

You can also use R:
Pr(X < 50 OR X>75) = Pr(X < 50) + Pr(X>75) = Pr(X<50) + 1 – Pr(X<75)

```
p = pnorm(50, mean=100, sd=25) + 1 - pnorm(75, mean=100, sd=25)
days = p * 365
p
[1] 0.8640949
days
[1] 315.3946
```

4. (15pts) Inferring the direction and existence of causal relationships from observational data is plagued by omitted-variables bias, selection bias, and reverse causality (simultaneous determination of dependent and independent variables). The following empirical patterns have been cited in press reports as potential evidence of causal relationships.

- Oakland is considering a Fresh Food Financing program that incentivizes grocery stores to locate in East Oakland. This program is based on studies showing that residents of neighborhoods without stores selling fresh foods have an unhealthy diet.
- Two percent of residents in Fresno, CA bike to work while eight percent bike in Berkeley. Berkeley has 50 more miles of bike lanes on their roads than Fresno. Therefore, if Fresno were to add more bike lanes its bike ridership would increase.
- A recent study in Minneapolis found that people who live in neighborhoods where the majority of houses have porches are more likely to talk to their neighbors at least once a week in comparison with people who live in neighborhoods where there are few porches. To encourage social cohesion in neighborhoods, Minneapolis is therefore considering a new grant program to help people add porches to their houses.

All three empirical patterns are seen in observational (non-experimental) data. Can you apply any of the criticisms of non-experimental empirical results to these three examples? If these criticisms were true, how do they alter interpretation of these patterns?

Full credit was given for plausible explanations for how each of these could be seen as an example of omitted variable bias, selection bias, or reverse causality, and in so doing, explained *which* variables might be missing, *how* selection might be happening, and *what* the reverse causal process might be.

For the grocery store example, there is a *sample selection* issue which is that the study may only be looking at the diets of people in neighborhoods without fresh food stores, and not looking at diets of those living in neighborhoods *with* such stores. There may also be self-selection bias, in which the apparent "treatment" is actually chosen by the population: that is, people choose to live in places without fresh food because they don't like fresh food and it's cheaper to live there. Note that sample selection is a different issue than (self-) selection bias: sample selection bias draws conclusion from

an un-representative sample, while self-selection bias produces biased estimate of the effectiveness of the treatment.

Also on the grocery store example, there may be *omitted variables bias*. In this case, household income is a plausible omitted variable that might cause both poor diets and fewer stores of any kind. Fresh food is often more expensive than packaged food. But note: omitted variables bias only exists if there are omitted variables that are correlated with both the independent and dependent (outcome) variable. Other omitted variables that are **not** correlated with independent do not cause omitted variable bias.

Arguably there may be *reverse causality*, as it may be the case where fresh foods retailers strategically placed stores in neighborhoods where residents have preference for and/or can afford healthy diet: lack of stores selling fresh foods may be an outcome of residents' leaning of an unhealthy diet.

For the bike lanes example, we have two cities that are very different in many ways, and *omitted variables* like weather and urban density may be more important than bike lanes but correlated with them in this very small sample of two cities. We could also be seeing *reverse causality*: that is, when the outcome causes the treatment, rather than the reverse. In this case the "treatment" is bike lanes and the "outcome" is bike commuting, but what if more bike enthusiasts demand more bike lanes and get them?

Finally, when it comes to porches, there could be *self-selection bias*: people may choose to live in homes with porches because they like to chat with neighbors. This could also be described as *reverse causality*: more social people are more likely to build porches, rather than the other way around. And there may be *omitted variables bias*: perhaps homes with porches happen to be in places with older populations, who are more culturally oriented to neighborhood relationships, or in traditional neighborhoods where side-walk and other urban form encourage interaction among neighbors.