Fall 2021 GEOL2300–Homework 1
Based on Chapter 13 of the Notes (Stats)

1 Prejudice and Bayes

In class, we mentioned the logical fallacy: “which is more likely: a Brown student is an engineer, or a Brown student with glasses is an engineer?” This is commonly taken to be an example of improper statistical thinking. Use the following statistics to make some guesses: at Brown the number of engineering concentrators is about 100 per year out of about 1700 undergraduate students per year. 64% of people wear glasses. Use Venn diagrams and Bayes’s theorem to examine the following cases:

- a) what’s $p(E)$: the likelihood that a Brown student is an engineer?
- b) What’s $p(g)$, the likelihood that someone wears glasses?
- c) What’s $p_g(E)$: the likelihood that a Brown student with glasses is an engineer if wearing glasses and being an Engineer are statistically independent?
- d) The fallacy exposed: What can you say about the number of engineering students who wear glasses $N(Eg)$, versus the number of engineering students $N(E)$? What does this say about the probability that any student at Brown is one or the other of these $(p(Eg), p(E)$?
- e) What is the likelihood that a Brown student with glasses is an engineer $(p(Eg))$ if the rate of wearing glasses among engineers $(p_E(g))$ is 100%?
- Finally, here’s what the fallacy seems to suggest, but actually doesn’t: how does $p_E(g)$ compare to $p(g)$ and $p_g(E)$ compare to $p(E)$ given $N(Eg)/N(E) > N(g)/N$?

a) $p(E) = 100/1700 \approx 0.059$.
b) $p(g) = 0.64$
c) $p_g(E) = p(Eg)/p(g) = p(E)p(g)/P(g) = p(E) = 100/1700$.
d) $N(Eg) \leq N(E)$, thus $p(Eg) \leq p(E)$.
e) $p(Eg) = p_E(g)p(E) = 1.0p(E) = 100/1700$.
f) $\frac{p_E(g)}{p(g)} = \frac{p(Eg)}{p(E)p(g)} = \frac{N(Eg)N}{N(E)N(g)} > 1$, $\frac{p_g(E)}{p(E)} = \frac{p(Eg)}{p(g)p(E)} = \frac{N(Eg)N}{N(g)N(E)} > 1$, and indeed $\frac{p_E(g)}{p(g)} = \frac{p_g(E)}{p(E)}$.

2 Data Manipulate

Make up a small dataset of 10 or so data. a) Calculate the mean, variance, and standard error of the mean. b) Describe what statistics of the data are expected to have a distribution where the variance describes the spread and what statistics are expected to have a distribution where the standard error describes the spread of the statistic. c) Describe how jackknife estimation and bootstrap estimation can be used to produce a histogram categorizing the uncertainty in the mean. d) (optional) You may carry out the bootstrap and jackknife estimates computationally for extra credit.

dataset: 7.0605 0.3183 2.7692 0.4617 0.9713 8.2346 6.9483 3.1710 9.5022 0.3445.
a) mean= 3.9782, variance= 13.0080, standard error = $\sqrt{13.0080}/\sqrt{10} = 1.1405$.
b) The mean and variance of this data (generated from a random uniform distribution in matlab) describe the spread of the data points themselves, and would describe the histogram of more points drawn from the same distribution. The mean and standard error give the statistics of the *sample mean* of 10 points averaged together. This distribution is more normal (Gaussian) than the original distribution and also more narrow (smaller standard deviation by $\sqrt{(}10))$.
c) In this case, there are 10 jackknife estimates of the mean, which are: 3.6357 4.3848 4.1125 4.3689 4.3123 3.5052 3.6481 4.0678 3.3644 4.3819. Each one is the average of 9 of the original data leaving one value out. A histogram can be made of these data, and it is clear that they cluster roughly around the mean of the original distribution with roughly the right standard error (mean of jackknife estimates is the same as sample mean, 3.9782, but in this case the variance of the jackknifes is much smaller than predicted by central limit theory 13.0080/10 0=1.30. Jackknife variance is 0.1606). To form a distribution using bootstrapping, we randomly choose sets of 10 data from the original data with replacement (i.e., repeated values are OK). Doing this, I found a mean of 3.9684, quite near the sample mean and a variances between 1.14 and 1.18.
d) I did the manipulations using matlab functions jackknife and bootstrp: e.g., mean(jackknife(@mean,A)) var(jackknife(@mean,A)) mean(bootstrp(10000,@mean,A)) var(bootstrp(10000,@mean,A)).

3 Mostly False

Use Table 13.4 to consider the following scenario.

- a) Suppose an unbiased single-investigator performs a study where ten independent possible linkages between evolution of angiosperms and plate tectonics are tested. The different possibilities are mutually exclusive and estimated to be equally likely, but only one is true. This investigator works hard on the method, and she estimates only a 5% risk of a false positive and a 5% risk of a false negative, but there is only the right kind of data to test 4 of the possible linkages (each with equal power of detection). i) What are the odds that her experiment will result in a yes relationship with the correct true linkage? ii) What are the odds that her experiment will result in a no relationship with the true correct linkage? iii) What are the odds that her experiment will result in a yes relationship with a false linkage? iv) What are the odds that her experiment will result in a no relationship with a false linkage?

- b) A different investigator works on finding relationships performs a study where ten independent possible linkages between water in the mantle and melting are tested. The different possibilities are mutually exclusive and estimated to be equally likely, but only one is true. This investigator works hard on the method, and he estimates only a 5% risk of a false positive and a 5% risk of a false negative, but there is only the right kind of data to test 6 of the possible linkages (each with equal power of detection). Unlike the investigator in a), this investigator is going up for tenure, and so really wants to publish a significant result. Thus, he *does not report all 6 possible linkages tested, instead only reports 4*, and so when writing up the paper is drawn toward reporting primarily the linkages that were detected as "true" in his study. This "file-drawering" of negative results and "over-hyping" of positive results can be modeled with a $u = 0.5$ bias, meaning that he is 50% more likely to report a positive result than a negative one. i) What are the odds that his experiment will result in a yes relationship with the correct true linkage? ii) What are the odds that his experiment will result in a no relationship with the true correct linkage? iii) What are the odds that his experiment will result in a yes relationship with a false linkage? iv) What are the odds that his experiment will result in a no relationship with a false linkage?

- c) Worldwide, the urgency of climate change has driven ten groups to consider possible linkages between temperature and carbon dioxide. Each of the 10 groups is able to reach the same experimental accuracies as the investigator looking into evolution and plate tectonics, and they are all unbiased (because they have built elaborate double-blind studies from fear of climate-gate like investigations by skeptics!) and independent from one another (i.e., they don't share data or methods until after the experiments and analysis are done). i) What are the odds that one group will report a yes relationship when testing the correct true mechanism? ii) What are the odds that one group will report a no relationship when testing the true correct mechanism? iii) What are the odds that one group experiment will report a yes relationship when testing a false mechanism? iv) What are the odds that one group will report in a no relationship when testing a false mechanism?

- d) Do your results make you concerned about the state of accuracy in the scientific literature?

a) $R = 1/9, \alpha = \beta = 0.05, c = 4$. i) research yes for true yes, $c(1-\beta)R/(R+1)/c = 0.095$. ii) research no for true yes, $c\beta R/(R+1)/c = 0.005$. iii) research yes for true no, $c\alpha/(R+1)/c = 0.045$. iv) research no for true no, $c(1-\alpha)/(R+1)/c = 0.855$.
b) $R = 1/9, \alpha = \beta = 0.05, c = 4, u = 0.5$. i) research yes for true yes, $c((1-\beta)+u\beta)R/(R+1)/c = 0.0975$. ii) research no for true yes, $c(1-u)\beta R/(R+1)/c = 0.0025$. iii) research yes for true no, $c(\alpha+u(1-\alpha))R/(R+1)/c = 0.4725$. iv) research no for true no, $c(1-\alpha)(1-u)/(R+1)/c = 0.4275$. This researcher is thus about 50 times more likely to report a false positive than a true positive!
c) $R = 1/9, \alpha = \beta = 0.05, c = 4, u = 0.0, n = 10$. i) research yes for true yes, $c((1-\beta^n))R/(R+1)/c = 0.1$. ii) research no for true yes, $c\beta^n R/(R+1)/c = 1 \cdot 10^{-14}$. iii) research yes for true no, $c(1-(1-\alpha)^n)/(R+1)/c = 0.361$. iv) research no for true no, $c(1-\alpha)^n/(R+1)/c = 0.539$. Positive results are 3.6 times more likely to be false positives than true positives!
d) Absolutely!

4 Uniform Stats

Find the zeroth, first, and second moments (not normalized or centralized) of the continuous uniform distribution. Use these to derive the mean and variance of the continuous uniform distribution given in Table 13.2.

The continuous uniform probability density function is constant over the interval of possible values (here taken to be $a \leq x \leq b$) as in Table ??. Thus,

$$\begin{aligned}\rho(x;a,b) &= \frac{1}{b-a},\\ \langle x^0\rangle &= \int_a^b x^0 \frac{1}{b-a}\,\mathrm{d}x = \int_a^b \frac{1}{b-a}\,\mathrm{d}x = 1,\\ \langle x^1\rangle &= \int_a^b x^1 \frac{1}{b-a}\,\mathrm{d}x = \int_a^b \frac{x}{b-a}\,\mathrm{d}x = \frac{b^2-a^2}{2(b-a)} = \frac{(b-a)(b+a)}{2(b-a)} = \frac{b+a}{2},\\ \langle x^2\rangle &= \int_a^b x^2 \frac{1}{b-a}\,\mathrm{d}x = \int_a^b \frac{x^2}{b-a}\,\mathrm{d}x = \frac{b^3-a^3}{3(b-a)},\end{aligned}$$

Then the mean and variance are

$$\begin{aligned}\langle x\rangle = \langle x^1\rangle &= \frac{b+a}{2},\\ \langle x^2\rangle - \langle x\rangle^2 &= \frac{b^3-a^3}{3(b-a)} - \frac{(b+a)^2}{4} = \frac{a^2+ab+b^2}{3} - \frac{a^2+2ab+b^2}{4} = \frac{(a-b)^2}{12}.\end{aligned}$$

5 CreatiVenn

Make a Venn diagram that describes an aspect of your life or work. Does it reflect independence or mutual exclusivity?

I'm pretty happy with the ones in Fig. 13.3. Fig. 13.3b is a nice example of mutual exclusivity.

6 Bootstrapping Exercise

This portion of the homework is to get you thinking about probability distribution functions, averages, bootstrapping, and the Central Limit Theorem.

6.1 A

First, grab this dataset: http://fox-kemper.com/data/Reid-Mantyla/Reid-Mantyla.mat Open it up in matlab[1]. First, we'll do some routine checks just to understand what's in the file. Nothing needs to be turned in for this portion A.

In the matlab file version of the data, the casts are broken up so that each location, each depth, each time counts at a new row in every variable. Locate one "cast" that contains the 10,000th row of all of the data points, but find it at all depths from the following code snippet,

```
>> latitude(10000)


ans =
```

[1] if you want to use python, there's an netcdf version here: http://fox-kemper.com/data/Reid-Mantyla/data_from_Reid-Mantyla.nc

```
   36.2533

>> longitude(10000)

ans =

  309.1120

>> day(10000)

ans =

    23

>> depth(latitude==36.2533&longitude==309.1120&day==23)

ans =

              0
             10
             20
             30
             40
             50
             ...
```

6.2 B

Open the Reid-Mantyla dataset above, and find all temperature measurements in the latitude range from 30N to 35N and longitude range from 150 to 155 and shallower than 500m depth. For example,

```
hist(temperature(find(latitude<35&latitude>30&longitude>150&longitude<155)))
```

Make histograms of the depths and temperature values of these data at these locations. (Matlab functions hist and histc).

6.3 C

Make a histogram representing the distribution of "average temperature of all points" from this selected region's data, also known as the standard error for the temperature data in the selected region's average. You should do this by bootstrapping (see matlab function bootstrp) and by calculation from the mean and standard deviation of the original data points and using the central limit theorem assuming all measurements are independent (you should get nearly the same result). You can use the inclass.m matlab script to help with this part of the process.

The mean of the whole dataset is 17.4961. If we subsampled 100 values from it, we'd expect to find nearly the same mean, but it would vary depending on which 100 we chose. The central limit theorem here tells us that the uncertainty in that should be 17.4961 ± 0.2704, or respecting the conventions of significant digits, 17.5 ± 0.3. Similarly, we expect the uncertainty in the whole 485 values to be 17.5 ± 0.1, if they are interpreted as a sample drawn from a much larger set.

6.4 D

Now, find a way to change the data being used (synthetic data is OK, or a manipulation of this dataset), or change the statistic being examined from mean to another function for bootstrp, or other manipulation so as to *make the bootstrap estimate clearly erroneous in some way.* We'll discuss these in class!