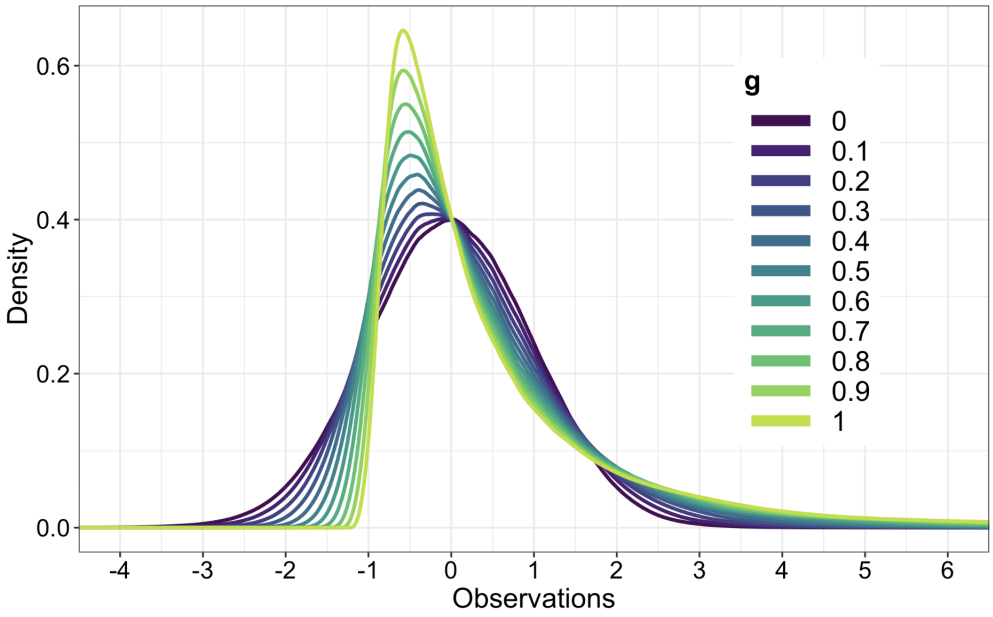
Lab 4: Simulation Experiments



In this lab we are going to experiment with implications of sampling from a probability distribution. First we will import, define, evaluate, and plot a single normal (Gaussian) distribution. We will then sample the distribution and look for relationships and trends of sampling related to sample size. After we write code to generate n-samples from a distribution, we will then look at effects of stochastic sampling which relate to the potential for type I errors (false positives) and type II errors (false negatives). Finally, we will use bootstrapping to estimate a confidence interval for the median of a data set.

Before anything else in this lab, import numpy and set a random seed of 42 using .random.seed(42). This will ensure that your results are replicable and consistent.

1 Create and plot a normal distribution

1. Import scipy and instantiate a normal distribution with $\mu = 0$ and $\sigma = 1$.
2. Plot the pdf of this distribution (you may find .arange or .linspace useful).
3. Plot the cdf of this distribution (you may find .arange or .linspace useful).
4. Using the inverse cdf and a unifrom distribution in the range of $[0, 1)$, sample the distribution 1000 times. Plot the histogram of these 1000 observations (hint: lookup the probability point function).
5. Answer the following in your writeup:

(a) What is the mean and standard deviation of the samples? Is it the same as the normal distribution that you instantiated?

(b) Can samples generated in this way do a good job of approximating the true distribution? What parameter can you change to get closer to the actual means and standard deviations?

2 Method for Sampling

1. Using the results from problem 1, create a function that accepts 2 parameters - a scipy class distribution object and an integer n. Here, n is the number of samples to take from the distribution. This method should return a list of length n elements, where each element is the value of a single sample.

2. Use your function to generate the following data and make histograms for each:

   - Normal distribution - 10 samples
   - Normal distribution - 20 samples
   - Normal distribution - 40 samples
   - Normal distribution - 80 samples
   - Normal distribution - 160 samples

3. For each one of these histograms, directly plot the normal pdf over the histogram (Hint: seaborns .histplot with stat="probability" may be useful).

4. Answer the following in your writeup:

   (a) What do you observe about the histograms as the number of samples gets larger? What property of sampling discussed in lecture accounts for this?

3 Type I Errors

1. Using the function from the previous problem, draw two sample groups using the same distribution and the same n. E.g. the first group should be 25 samples drawn from a normal distribution with $\mu = 0$ and $\sigma = 1$, the second group should be 25 different samples drawn from a normal distribution with $\mu = 0$ and $\sigma = 1$.

2. Calculate the estimated effect size between these two groups. (the absolute value of the difference between group 1 sample mean and group 2 sample mean).

3. Create a function that accepts two distribution objects, two values of n for number of samples to draw from the distributions, and an integer k which is the number of times to loop the sample generations. This method should draw from each distribution n times and repeat this process for k iterations. The method should return three lists that are k elements long. The elements in the first list are the estimated effect sizes

at iteration k between the drawn samples of the two distributions. Each element in the second list is a list of the drawn samples from distribution 1 at iteration k. Each element in the third list is a list of the drawn samples from distribution 2 at iteration k.

4. Use the defined method to generate two sample groups, each with 25 observations, for 1000 iterations. Plot a histogram of the effect sizes.

5. Find the index of the largest observed effect size. Use the index to plot the histograms of the two associated sample groups.

6. Answer the following in your writeup:

   (a) Describe in your own words a type I error.

   (b) What is the estimated effect size between the two groups in step 3.2.? What should the expected effect size be and how did you calculate it?

   (c) What is the maximum observed effect size in step 3.4.?

   (d) In step 3.5. what do you see in the histograms you created? If this happened in a single trial (you found this in an experimental data set) - what conclusion would you draw? Would this conclusion be correct or mistaken?

4 Type II Errors

1. Instantiate another normal distribution object with parameters $\mu = 1$ and $\sigma = 1$.

2. Using the function developed in problem 3.3., repeat the experiment in part 3.4. using two different distributions. This time the first sample group should be 1000 iterations of 25 samples drawn from a normal distribution with $\mu = 0$ and $\sigma = 1$ and the second group should be 1000 iterations of 25 different samples drawn from a normal distribution with $\mu = 1$ and $\sigma = 1$. Plot the histogram of the estimated effect sizes. Find the index of the minimum effect size and plot the two sample groups associated with this index using a histogram.

3. Answer the following in your writeup:

   (a) Describe in your own words a type II error.

   (b) What is the minimum observed effect size in step 4.2.? What should the expected effect size be and how did you calculate it?

   (c) In step 4.2. what do you see in the histograms you created? If this happened in a single trial (you found this in an experimental data set) - what conclusion would you draw? Would this conclusion be correct or mistaken?

5 Bootstrapping

In this section we will implement a simple bootstrapping algorithm from scratch to estimate a confidence interval around the median of a data set (this is essentially the only way to do this for the median!). A confidence interval is a measurement of how much you should expect the median to vary – in other words it is a measurement of the error of your estimated median.

1. Load in the data set from the assignment on canvas called some_data.csv. There should be three variables in this data set. We will produce bootstrapped estimates of the confidence interval around the median for each of the three variables.

2. For each of the three variables perform the following. For 1000 iterations: create a new bootstrap data set of 500 obsverations by sampling with replacement from the 500 observations generated in the data set; for each bootstrap data set compute the median and record it; sort the recorded values you have estimated for the median; take the 25th and 975th values as the bounds of your confidence interval.

3. Answer the following in your writeup:

   (a) How do you interpret the confidence interval you generated? What does it mean?

   (b) Are the medians of the three variables 'different' from one another? How can you tell?

6 Submission Instructions

- Write up the answers to the questions in a short word document; aim for around 2 pages of text and include all graphics generated. Add footnotes identifying which sentence addresses which questions. Write in complete sentences organized into paragraphs – your goal is to explain what you've done and what you've learned to your audience (me!). Accordingly, you may seek to emulate some of the sections of the whale paper describing their data. Include the appropriate plots you've generated as mentioned above. Convert this to pdf and submit it. Submit your .ipynb file as well.

- The grading rubric for this assignment will be available in Canvas.

- NO OTHER SUBMISSION TYPES WILL BE ACCEPTED.

- **Late policy**: 5% of total points deducted per day for three days – after that no submissions allowed.