Forensic Sampling Course

Excel Lab

1 Sampling Approaches

In this exercise, we will consider how various sampling schemes affect estimates for the population mean. Open the `glass.xlsx` data set containing elemental concentrations of the isotope Manganese-55 in fragments sampled from panes of float glass manufactured by two different companies. We will behave as if the entire data set is our population of interest (meaning $N = 1147$) from which we draw samples. Note that this is a simplified version of a data set accessible here.

There are three tabs in this `.xlsx` file. The first tab, `glass`, shows the original data. We will use the other two tabs, `SRS` and `Stratified`, later in the exercise. The column descriptions in the `glass` tab are as follows:

1. `manufacturer`: letter identifying the manufacturer of the pane of glass (either A or B).
2. `pane`: alphabetic code identifying a pane of glass from which samples were drawn. The first letter corresponds to the manufacturer (A or B) and the rest of the code uniquely identifies the pane (i.e., pane AA is different from pane AB).
3. `fragment`: number identifying the fragment sampled from the pane (24 fragments were sampled per pane).
4. `Mn55`: concentration (in ppm) of the isotope Manganese-55 in the fragment.

1. What is the population mean? To calculate this quickly, select a blank cell and type `=AVERAGE(`, hover your cursor over column `D` header (above the `MN55` column title) until it turns into a black arrow pointing downwards, and left-click. This should select the entire column. Press `)` followed by `enter` to complete the formula.

   The true population average concentration is 56.994.

2. For our first sampling scheme, we will consider a non-random, "convenience" sample. Suppose that it was easy for us to obtain 24 fragments from pane AA but not from any other panes. What is the estimated population mean using only the pane AA data? How does it compare to the true population mean calculated in question 1?

   The estimated population mean using only the pane AA data is about 18.356. This is pretty far below the true population mean.

3. Scroll down to rows 743/744 in the data set. This marks the end of the manufacturer A data and the beginning of the manufacturer B data, respectively. How do the Mn55 concentration values for manufacturer A compare to those of manufacturer B? Why might this explain what you observed in question 2?

   The Mn55 values in manufacturer B panes are distinctly larger than those in manufacturer A panes. Large values tend to "pull" an average up which explains why the population average is larger than the average calculated from a single manufacturer A pane.

Now to perform Simple Random Sampling. Click on the `SRS` table. You will notice that the Mn55 column from the `glass` tab has been copied here. To the right of this column is a `random` column that is currently empty. The contents of columns `D` and `E` are intended to represent results from 5 samples drawn from the overall population (for the sake of illustration, we'll show 5 samples when in practice you'll commonly only work with one). Complete the following steps to draw samples from the data set. If you have problems executing these steps, please refer to this article.

(a) In cell B2, enter the formula =RAND(). This will generate a random number.

(b) Left-click on cell B2 and note the small square in the bottom-right corner of the cell. Your cursor should turn into a black + symbol when you hover over this square. Double-click the square and the remaining rows in the table should be filled in with random numbers. (Note: this is a very useful Excel trick for repeating formulas)

(c) Now select column B (left-click on the column B header) and copy the contents (either Ctrl+C or right-click and select Copy). Click the arrow below Paste in the Home tab at the top of the spreadsheet. Select the first option under Paste Values. This will ensure that Excel treats the cells' contents as numbers rather than as a formula (allowing us to sort by these numbers).

(d) Now select columns A and B by holding left-click on the A column header and dragging your cursor to the B column header. The 2 columns should be highlighted.

(e) With the 2 columns selected, click on the Sort button under the Data tab at the top of the spreadsheet. In the menu that pops up, click the arrow next to Sort by and select Column B. Click OK.

(f) We will behave as if every 50 data rows constitutes a sample drawn from this population. The values in column E should have updated. The text in cell E2 is meant to show how we would calculate the estimated mean using the first 50 rows of observations (which, because they were randomly shuffled due to the random number generation + sorting, will just be 50 random values in the population). Left-click on cell E2 and add an = sign before AVERAGE to convert this into an Excel formula. Press enter to evaluate this formula.

7. How do the sample averages calculated in column E relate to the population average, specifically compared to the convenience sample average you calculated in question 2?

   These sample average values are now much closer to the true population mean than the convenience sample average calculated before. However, the values are still (probably) fairly variable around the true population mean.

8. Note that there are $N_1 = 742$ observations from manufacturer A panes and $N_2 = 405$ observations from manufacturer B panes. Given this and your previous answers, why might it be appropriate to employ stratified sampling for this data set?

   Stratified sampling is used when there are different types of items (e.g., panes from different manufacturers) and some types are more abundant than others. Since we identified in question 3 that there is a systematic difference in the Mn55 concentrations between manufacturer A and B panes *and* that there are almost twice as many observations from manufacturer A panes, stratified sampling seems appropriate.

Now to perform Stratified Sampling. Click on the Stratified tab. You will again notice that some data have been copied into this tab from the glass tab. For your convenience, the data have already been stratified by manufacturer so that the manufacturer A observations occupy columns A and B while the manufacturer B observations occupy columns E and F. The contents of columns I, J, and K are again intended to represent results from 5 samples drawn from the overall population. Recall from the course slides that we can draw a stratified sample from the overall population by drawing simple random samples from each stratum. Repeat the same SRS process outlined above for each stratum (two random columns have already been added for you) and note how the sample values in column K update after sorting.

9. Suppose that our desired stratified sample size will be $n_h = \sqrt{N_h}$ for $h = 1, 2$ (rounding n_h up to the nearest whole number if-needed). Calculate n_1 and n_2 using this formula.

   $\sqrt{742} = 27.240$ **meaning** $n_1 = 28$. **Similarly,** $\sqrt{405} = 20.125$ **meaning** $n_2 = 21$.

10. The sample sizes you calculated in question 9 have been pre-programmed into the formulas in column K. Explain in words what the un-evaluated formulas in cells K2, K3, and K4 represent.

    Cells K2 and K3 are the estimated Mn55 concentration for manufacturers A and B, respectively. Cell K4 is the estimate of the overall population mean (see the formula on slide 78 of the course slides).

11. Add = signs to cells `K2`, `K3`, and `K4` to turn them into evaluated formulas. Then, calculate the true mean Mn55 concentration for each manufacturer (i.e., in a blank cell type `=AVERAGE(B:B)` and do the same for column `F`). How do the estimated average concentrations for each manufacturer compare to the true average concentrations for each manufacturer?

    The true "within-manufacturer" average concentrations are about 18.075 and 128.297 for manufacturers A and B, respectively. The 5 sample average concentrations are quite close to these true values.

12. Now compare the 5 estimated population averages to the true population average (calculated in question 1). How do these values compare to the convenience and SRS estimated values from before?

    Similar to the SRS estimates, these stratified estimates are much closer to the true average concentration than the convenience sample estimate. However, these stratified samples are also much less *variable* around the true average concentration than the SRS estimates. This highlights the benefit of using a stratified approach (when it is needed).

    As an ending thought, note that this exercise is obviously not reflective of how we would actually approach sampling panes of glass in that we were "given" the entire population up-front. This made it easy for us to determine that, for example, the concentration of Mn55 varied considerably between the two manufacturers. This wouldn't necessarily be so obvious in-practice. However, the exercise is meant to demonstrate a few points. Firstly, not treating the sample design with care can lead to extreme cases such as the convenience sample average seen in this exercise. In-practice, we wouldn't actually *know* that the convenience sample average is so far away from the population average. Secondly, simple random sampling is certainly an improvement as we will be close to the true population mean *on average*. Finally, stratified sampling plans can be employed to reduce estimate variability if we can identify ahead-of-time some factor (e.g., the manufacturer) that could systematically influence the variable that we're interested in (e.g., the Mn55 concentration). This article contains more details about the sampling procedure used to obtain the elemental concentrations presented in this data set. Finally, if the random number generation + sorting process used above to draw samples seems tedious, then you may be interested in learning how to use the R programming language that has many built-in functions to perform sampling and many other statistical processes.

2 Sample Size Calculation

The mandatory minimum sentence for trafficking one gram of lysergic acid diethylamide (LSD) in the US is 5 years in prison for first-time drug offenders (Source). A common way to take LSD is by dissolving it into a solution, spraying this solution onto a sheet of blotting paper, cutting the paper into one-dose squares or "tabs", and ingesting one of these tabs (Source). Assume that one gram of LSD is equivalent to about 200 tabs.

Suppose that 300 suspected tabs of LSD were seized. It is to be established beyond a reasonable doubt whether at least 200 of these 300 tabs actually contain LSD. Due to time or budget constraints, it may be unreasonable to actually test all of the tabs. If we are willing to tolerate a little uncertainty (e.g., a "reasonable doubt"), then we may be able to test considerably fewer tabs. The following exercises explore how to calculate a sample size using the European Network of Forensic Science Institute (ENFSI) Drugs Working Group's Calculator for Qualitative Sampling of Seized Drugs (CQSSD). More information about the calculator can be found in the UN Guidelines for Representative Drug Sampling.

2.1 Frequentist Hypergeometric Approach

For a Frequentist approach, we follow the guidelines presented in ASTM Standard E2548-16 and the UN Guidelines for Representative Drug Sampling. We will use the `Hypg_Number` tab in the CQSSD for this exercise. Note that the `Hypg_Proportion` tab can also be used to obtain equivalent results, but it requires data to be entered in a slightly different form.

2.1.1 Sample size calculation in Excel

We will now determine the sample size using the `Hypg_Number` tab in the CQSSD. Navigate to this tab. You will need to enter data into some of these cells. The calculated sample size can be seen in cell `B14`. The steps below detail how to enter data into the calculator.

Step 1. Enter the population size ($N = 300$) into this cell.

Step 2. Enter the number of positive LSD tabs that we would like to show in this sample (which is $K = 200$ tabss).

Step 3. We will assume for now that the expected number of negatives is 0. Enter 0 into this cell.

Step 4. For a confidence level $p \in [0, 1]$, we might interpret $1 - p$ as our level of doubt. What we consider to be a "reasonable doubt" will likely depend on context, but in-general we should aim for doubt to be small (somewhere between 0 and 0.05). For now, enter 0.99 into this cell (which corresponds to a "doubt level" of 0.01).

You'll note that Step 7 (cell `C14`) shows the actual confidence levels associated with the calculated sample size in cell `B15`. This slightly different from the confidence level entered into cell `B14` because of rounding error.

Report the calculated sample size below.

The calculated sample size is 11 tabs.

2.1.2 Changing case facts

We will now analyze how the calculated sample size changes by changing the data entered into the calculator.

1. Suppose we anticipate one of the sampled tabs to be negative. Change the value in the Step 3 cell to 1. How does the sample size change? Why does this make sense?

   The calculated sample size increases. Observing one negative tab in the sample means that there may be more than one negative tab in the overall population. Thus, we'll need to sample additional tabs to be confident that there are at least 200 positive tabs in the population.

2. Change the value in the Step 3 cell back to 0. Suppose now that what we consider to be a "reasonable doubt" is less restrictive. Change the confidence level of 0.99 in the Step 4 cell to 0.95. How does the sample size change? Why does this make sense?

   The calculated sample size decreases. We do not need to be as confident that at least 200 tabs are positive for LSD, so it makes sense that we do not need to sample as many tabs.

2.2 Bayesian Beta-Binomial Approach

For a Bayesian approach, we follow the guidelines presented in ASTM Standard E2548-16 and the UN Guidelines for Representative Drug Sampling. We will use the `Bayesian N >= 50` tab in the CQSSD for this exercise.

2.2.1 Prior distribution

The first step in a Bayesian approach is to choose a prior probability distribution that the tabs are positive for LSD, $P(\text{pos})$. The `Bayesian N >= 50` tab in the CQSSD uses the Beta distribution as a prior. There are two parameters of the Beta distribution, a and b, that affect the shape of the probability density function. Changing the shape of the probability density function can be interpreted as changing one's opinion of where the true probability that the tabs are positive is believed to lie (between 0 and 1). For example, $P(\text{pos}) = 1$ means that we are certain that the tabs contain LSD while $P(\text{pos}) = 0$ means that we certain that the tabs do *not* contain LSD.

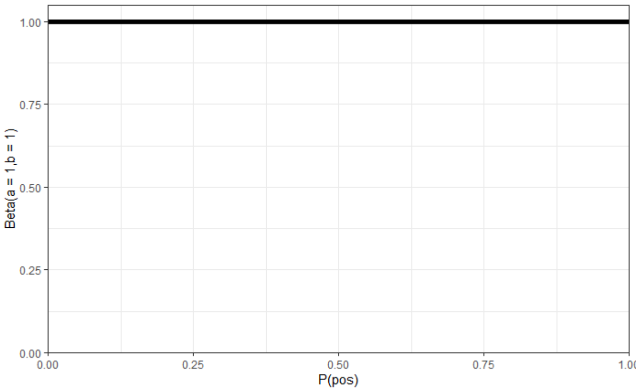
The Beta distribution parameters a and b are to be chosen by the examiner based on visual inspection of the tabs, other facts about the case, and the examiner's prior experience.[1] Consider, for example, that LSD blotting papers commonly bear an artistic design. One such popular design is called the "Man on Bicycle." Suppose that the 300 seized tabs depict the "Man on Bicycle" design. This information can be incorporated into your calculations by choosing a Beta distribution for $P(\text{pos})$ that places more probability on values closer to 1.[2] The following graphs

[1] You may feel uncomfortable with choosing values for these probabilities. It's important to note however that we rely on subjective assumptions frequently when analyzing data. The Bayesian paradigm allows us to make these assumptions explicit by incorporating them into our calculations.

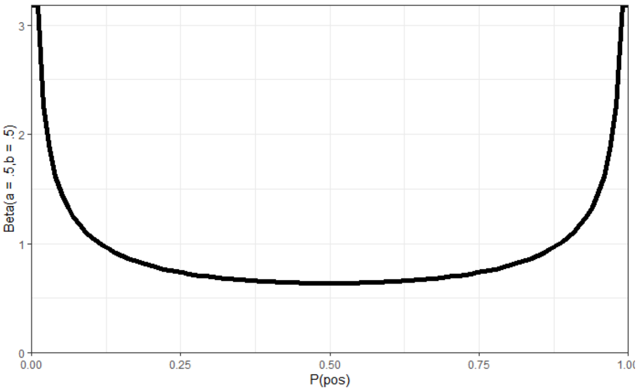
[2] Said another way, the fact that the seized tabs depict a popular design used on LSD paraphernalia can increase your belief that the tabs contain LSD even before the tabs are tested.

show Beta distributions for four different choices of a and b. Select amongst these four the most appropriate prior distribution in your opinion

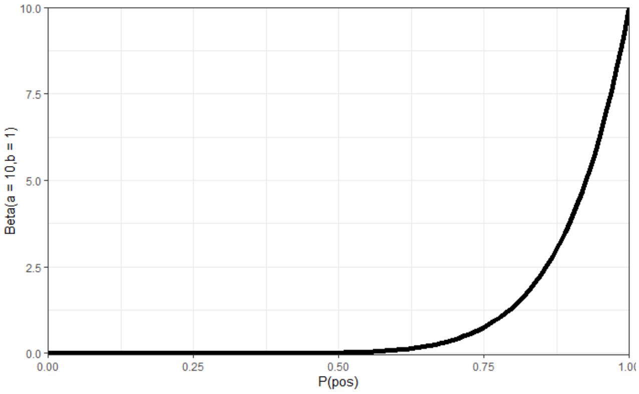
(Note: there isn't one "correct" answer to this exercise. The whole point of choosing a prior is that it's based on *your* educated opinion, but certain facts about the case may make one of these priors a reasonable choice over others).



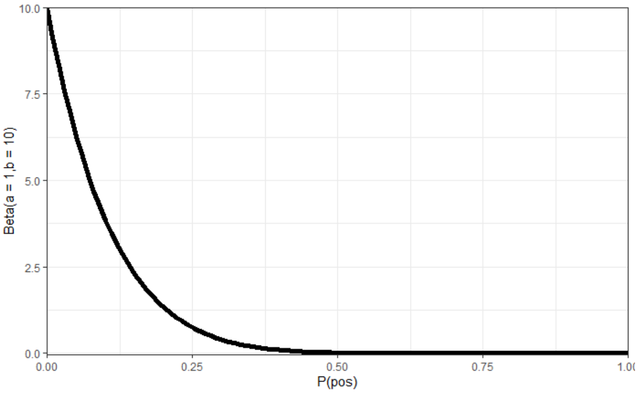
(a) Beta(a = 1,b = 1) probability density function



(b) Beta(a = 0.5,b = 0.5) probability density function



(c) Beta(a = 10,b = 1) probability density function



(d) Beta(a = 1,b = 10) probability density function

Explain your choice:

Answers will vary depending on your opinion. The following provides an interpretation of these different priors and how they represent differing opinions about the true value of $P(\text{pos})$. Note that none of these four may be the "most" appropriate prior distribution for this situation, so experimentation with different priors is encouraged.

(a) **The Beta(a = 1,b = 1) distribution is more commonly called the *Uniform* distribution on the interval [0,1]. This represents a completely impartial opinion on the true value of $P(\text{pos})$. That is, our belief that all of the tabs are positive for LSD is the same as our belief that none of the tabs are positive for LSD which is the same as our belief that any number of the tabs are positive for LSD. In this situation, it seems odd to assume that only *some* of the tabs contain LSD, so this is perhaps an inappropriate prior to assume.**

(b) **The Beta(a = 0.5,b = 0.5) distribution can be interpreted as putting belief near the endpoints of [0,1]. That is, we believe either that all of the tabs contain LSD or none of them do. This seems like a more reasonable assumption than the Uniform distribution on [0,1], but still doesn't incorporate the facts of the case (e.g., the Man on Bicycle design).**

(c) **The Beta(a = 10,b = 1) distribution can be interpreted as placing most belief near 1, which if you recall corresponds to all of the tabs containing LSD. An examiner who has seen many instances of the Man on Bicycle design on LSD blotting paper might be more inclined to believe that these 400 seized tabs all contain LSD. As such, this seems like a reasonable choice for a prior.**

(d) **The Beta(a = 1,b = 10) distribution can be interpreted as placing more belief near 0, which if you recall corresponds to none of the tabs containing LSD. Given the facts of the case, this doesn't seem to be as reasonable as some of the other choices.**

2.2.2 Sample size calculation in Excel

We will now determine the sample size using the `Bayesian N >= 50` tab in the CQSSD. Navigate to this tab. You will need to enter data into some of these cells. The calculated sample size can be seen in cell `B15`. The steps below detail how to enter data into the calculator.

Step 1. Enter the population size ($N = 300$) into this cell.

Step 2. Enter the proportion of positive LSD tabs that we would like to show in this sample (which is $k = \frac{200}{300} = 0.66$ approximately).

Step 3. We will assume for now that the expected number of negatives is 0. Enter 0 into this cell.

Step 4. For a confidence level $p \in [0, 1]$, we might interpret $1 - p$ as our level of doubt. What we consider to be a "reasonable doubt" will likely depend on context, but in-general we should aim for doubt to be small (somewhere between 0 and 0.05). For now, enter 0.99 into this cell (which corresponds to a "doubt level" of 0.01).

Step 5. Enter into this cell the value for a you've chosen based on your answer to the previous question.

Step 6. Enter into this cell the value for b you've chosen based on your answer to the previous question.

Report the calculated sample size below. How does this compare to the calculated sample size using the `Hypg_Number` tab? Why does this make sense?

Depending on your choice of prior, the sample size will be (a) 11, (b) 8, (c) 2, or (d) 40. The sample sizes associated with priors (a), (b), and (c) are less than or equal to the sample size calculated using the `Hypg_Number` tab. This makes sense because we've incorporated additional case information into the sample size calculation by way of the prior distribution. Choosing prior (d) leads to a larger sample size because we have placed more of our prior belief on the claim that none of the tabs contain LSD. A Frequentist approach does not allow us to incorporate such information in the CQSSD (as some say, Frequentist calculations happen "in a vacuum").

2.2.3 Changing case facts

We will now analyze how the calculated sample size changes by changing the data entered into the calculator.

1. Suppose that, instead of depicting the Man on Bicycle, the blotting paper is a blank sheet. Use this new information to choose a prior distribution among the four previously presented. Enter the new values of a and b into the calculator. How does the sample size change? Why does this make sense?

   Answers will vary, but this new information should push us to choose a prior distribution that places less probability near $P(\text{pos}) = 1$ than the previously-chosen prior. This will cause the calculated sample size to increase. For example, assuming that prior (c) was chosen previously, priors (b) or (d) now seem like a more reasonable choice. The fact that the design on the blotting paper isn't obviously associated with LSD may reduce our prior belief that the tabs are positive. The choice between priors (b) and (d) would likely depend on other facts about the case (e.g., if the blotting paper was seized from a cosmetician, then it's entirely possible that it could have been used to absorb excess sebum oil in which case we may not want to place much belief near 1).