**Data Analysis Project (DAP): Instructions and Guidelines.**

Pol 501: Introductory Statistics

# Overview and Objective

Objective and Goal:

This project is designed to apply the statistical concepts and techniques learned throughout the course to a real-world problem relevant to public policy, public administration, or political science. By the end of the project, students will have demonstrated their ability to analyze and interpret data to draw meaningful conclusions using sound principles from statistical inference.

# Project Stages, Timeline, and Deliverables

## Stage 1: (Optional) Form Your Team (Week 3-4)

Students can complete the DAP individually or in teams of up to two. Grading and instructions are the same for everyone. By Week 4, teams must be formed, and one team member should email the instructor with the names of the group members.

**Stage 1 Task**: Send an email to the instructor with the names of the group members.

**Due**: By the end of week 4 of the course schedule.

## Stage 2: Project Proposal (Week 7)

Each student/team must submit a one-page (or two-page maximum) “***Project Proposal*”** that includes:

* **A clear explanation of the research problem** or question, relevant to a policy issue, public administration problem, or political science question, which can be analyzed using data and statistical inference.
* **Justification of the problem’s relevance**, connecting it to broader political, public administration, or policy contexts.
* A **preliminary data plan**, specifying data sources and whether the team will collect original data or use existing datasets.

A “*research problem or question*” is the specific issue, topic, or question that a project aims to address or investigate. It defines the focus of the research and guides the project’s objectives. The problem or question should identify a gap in knowledge or an area where more information is needed, often related to real-world issues in a particular field. In research, the problem or question serves as the foundation for developing hypotheses, methodologies, and analyses.

A strong research problem or question is clear, focused, and feasible to explore within the project’s scope. The main goal of the DAP is to *use data and statistical inference to answer a research question or provide insights relevant to a specific policy issue*.

There are two main possible **approaches for the** **data plan**:

1. *Collecting Original Data*:

* If the team decides to collect their own data, they must define how they plan to do so, typically by designing a survey.
* The survey must be structured to gather the necessary data to analyze and perform statistical inference (test hypotheses) related to their problem.
* The proposal must describe the survey plan in detail, outlining how the data collection will be done, the target population, the sample, and what kind of questions will be included to gather relevant data for statistical analysis.

1. *Using Existing Data*:

* If the team chooses to use existing datasets, they must locate and download the relevant data, ensuring it’s appropriate for answering their research question.
* Teams can use public surveys or available data from reliable sources (e.g., Pew Research, ANES, World Bank datasets). The proposal must provide details about the dataset and how they plan to use it.
* In this case, teams should ensure they are doing new analysis and not merely replicating existing studies. They must explain how their analysis will be different from previous analyses done with this data and what new insights they aim to provide.

**Stage 2 Task**: Upload to Brightspace a Word or PDF document with the **Project Proposal**. Specific instructions:

1. Include the *Project’s Title* and the *Author(s) Names* at the beginning of the document.
2. Start with a section titled “*Research Problem*.”
3. Follow with a section titled “*Relevance*.”
4. Finish the document with a section titled “*Preliminary Data Plan*.”
5. The document should be written in Times New Roman, 11pt Font Size, and Single Space.
6. The document should have a maximum of two pages.

**Due**: By the end of ***week 7*** of the course schedule.

## Stage 3: Study Hypothesis and Data Analysis Plan (Week 11)

Each student/team must submit a three-page maximum “**Hypothesis Description and Data Analysis Plan”** that includes:

* **Hypothesis**: First, students will clearly define a hypothesis or a set of hypotheses that they will test in their project. The hypothesis should relate directly to the problem or question they have outlined in their proposal. Students’ must number each of the hypothesis to be tested (e.g.: *H*1: “…”, *H*2: “…”).
* **Dataset Description**: In this section, students must provide a detailed overview of the dataset they plan to use for their analysis. This description should include the source of the dataset, such as a publicly available database, survey, or any other data collection method. Additionally, students should describe key attributes of the dataset, such as the number of observations (sample size), the time period it covers, and the types of variables it contains (categorical, continuous, etc.), and other data collection methods (e.g., if it is an original survey how the survey was publicized and through what channels).
* **Data Preparation:** Students must create a table describing the variables they will use in the analysis. The table should be titled “***Variables’ Dictionary***.” The table should include the variable name, a conceptual description, the variable type, range or scale, and any transformations or creation of new variables. This process is known as the “operationalization” of the conceptual variables of interest.
* **Statistical Tests:** After defining the hypothesis and operationalizing the relevant variables, students must describe the statistical methods they will use to test their hypothesis. This section should directly link to the hypothesis and specify the statistical tests or techniques planned for analysis.

Notes:

1. *Project Hypothesis*.

Even for exploratory projects (i.e., investigating general patterns or insights), students are required to frame specific hypotheses. A hypothesis provides structure to the analysis, helping identify relationships, features, or differences to explore. It should be a clear, testable statement about a potential relationship between variables or a specific property (fact) of one variable. The hypothesis must directly connect to the project’s data, and it should make a claim that is testable using statistical methods. Students can develop multiple hypotheses as long as they relate to the core question of the project.

For example, in projects studying public opinion on specific policy issues, students could propose:

* “*H*1*: Individuals who identify as conservative are more likely to support less federal regulation of highways than those who identify as liberal*.”
* “*H*1*: Older individuals are more likely to support increased funding for schools compared to younger individuals*.”

Consider these other examples that posit hypothesis related to one specific variable (measured as a proportion):

* " *H*1*: The majority of the population supports raising the federal minimum wage to $15 per hour*."
* *" H*1*: A minority of the population is unaware of recent federal climate change policies."*

1. *Data Preparation*.

Creating a "Variables Dictionary" is essential for organizing and preparing data for analysis, as it ensures that the variables used to test hypotheses are clearly defined and understood. This process, known as operationalization, involves specifying how each conceptual variable will be expressed in numeric terms, allowing for statistical computation and hypothesis testing. By detailing the variable name, description, type, scale, and any transformations, students can ensure consistency and transparency in their analysis. It also clarifies which variables will be used to test specific hypotheses, providing a structured foundation for the investigation.

Here’s an example table using variables from a hypothetical Pew Research dataset, such as one focused on internet usage and social media habits:

**Table. Example Variables’ Dictionary.**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Variable Name** | **Conceptual Description** | **Variable Type** | **Range/Scale** | **Transformations/New Variables** |
| Age | Respondent's age in years | Continuous | 18-99 years | None |
| Internet\_Use | Frequency of internet use | Categorical Ordinal | Never, Daily, Weekly, Monthly | Recode to binary (1 = Uses internet, 0 = No) |
| Social\_Media\_Use | Whether the respondent uses social media | Binary (Categorical Nominal) | 1 = Yes, 0 = No | None |
| Education\_Level | Highest level of education attained | Categorical Ordinal | 1 = Less than high school, 2 = High school, 3 = Some college, 4 = Bachelor's or higher | Create a new variable that assigns the numbers to each categorical textual option. |
| Time\_On\_Social\_Media | Average hours spent on social media per day | Continuous | 0-24 hours | None |

1. *Statistical Tests*.
   * **Restate the Hypothesis**: The first step is to clearly restate the hypothesis in statistical terms. For example, if the hypothesis involves comparing the average opinions of two political groups on highway regulation, it can be framed as: "There is a significant difference in mean opinions on highway regulation between political party A and political party B."
   * **Selection of Statistical Tests**: Once the hypothesis is restated, students must identify the appropriate statistical tests for their data. The choice depends on the type of variables and the hypothesis. For instance:
     + *T-tests*: Suitable for comparing the means of a continuous variable between two groups.
     + *Chi-square tests*: Used for assessing the relationship between categorical variables.
     + *Correlation tests (Pearson or Spearman)*: Appropriate for evaluating the relationship between continuous variables, such as analyzing the correlation between age and income.
     + *ANOVA*: Useful for comparing means across more than two groups, for example, if opinions are being compared across multiple regions or age groups.
   * **Rationale for the Choice of Test and Test’s Assumptions**: For each test selected, students must provide a brief explanation of why that particular test is appropriate for the data and hypothesis. Each statistical test comes with assumptions. For example, a *t-test* assumes the data is normally distributed and that the variance is unknown. A *z-test*, in the case of a small sample, assumes the data is normally distributed and the population variance is known. If the data does not meet these conditions, adjustments may be necessary, or an alternative test might be more appropriate.

**Stage 3 Task**: Upload to Brightspace a Word or PDF document with the **Hypothesis Description and Data Analysis Plan**. Specific instructions:

1. Include the *Project’s Title* and the *Author(s) Names* at the beginning of the document.
2. Start with a section titled “*Hypothesis*.”
3. Follow with a section titled “*Dataset Description*.”
4. Follow with a section titled “*Data Preparation*.”
5. Finish the document with a section titled “*Statistical Tests*.”
6. The document should be written in Times New Roman, 11pt Font Size, and Single Space.
7. The document should have a maximum of three pages (without counting the appendix).
8. If you are carrying out an original survey, you must include the Distribution Strategy Plan as an appendix (see the subsection “*Guidelines for Collecting Original Data Using Surveys*”)
9. If you are carrying out an original survey, you must include the Survey’s questionnaire as an appendix of your document.

**Due**: By the end of ***week 11*** of the course schedule.

## Stage 4: Submission of the Final Written Report (Week 15).

Sections to be included in the Final Report:

* **Introduction**: Rewrite the submitted Project Proposal as an introduction to the Final Report (omit the Preliminary Data Plan section).
* **Hypothesis Description and Data Analysis Plan**: Include the submitted content from stage 3’s deliverable.
* **Description of Results:** Write a section that uses tables and graphs to report the results of the statistical methods/tests you carried to test your project’s hypothesis.
* **Discussion**: Write a section that interprets your results and discusses how these contribute to the research problem addressed by your project.

Notes:

1. *Description of Results*.

The “**Description of Results**” section should begin by clearly presenting the key findings of your statistical analyses, linking them to the hypotheses you tested. Start by specifying the statistical tests you used (e.g., t-tests, ANOVA, regression) and explain why these were appropriate for evaluating your hypotheses. For each hypothesis, describe the relevant estimates (means, regression coefficients, odds ratios), and provide exact values, such as means or proportions, to give a detailed understanding of the results. Be sure to connect these estimates directly to the hypothesis, indicating whether the results support or refute your initial expectations.

Next, report the significance of your findings by including *p-values* for each test, emphasizing whether the results were statistically significant (e.g., p < 0.05). Clearly state if the statistical evidence supports or rejects each hypothesis. Mention both significant and non-significant results, ensuring the reader understands how these outcomes align with your research hypotheses. Also, if applicable, include confidence intervals and computed estimates to provide further context on the magnitude and reliability of your findings.

Lastly, complement your results with visual aids such as graphs, tables, or figures to make your data more accessible. Use visualizations like bar charts or scatterplots to show trends or group differences and tables to summarize key statistics. Make sure that each graph or table is directly tied to a hypothesis and is clearly referenced in the text, reinforcing how your statistical tests relate to the research questions you aimed to answer.

1. *Discussion of Findings*.

In the “**Discussion of Findings**” section, start by interpreting your results in the context of the research questions and hypotheses you initially posed. Explain how your findings connect to the specific policy or political problem you are addressing. Simply put, you must explain to the reader how your findings matter and contribute to shed light to the research problem your project addressed.

Next, explore the broader implications of your results for politics, policy or administrative decision-making. Discuss how your findings might influence future political problems, policy formulation, implementation, or reform. Address the potential real-world impact, such as how your research could inform politics, policy, and public administration problems.

Finally, identify the limitations of your study. Acknowledge any methodological limitations, such as sample size, data quality, confounder variables, or other biases, and explain how they might affect the generalizability of your findings.

**Stage 4 Task**: Upload to Brightspace a Word or PDF document with the **Final Written Report**. Specific instructions:

1. Include the *Project’s Title* and the *Author(s) Names* at the beginning of the document.
2. Start with a section titled “*Introduction*.”
3. Follow with a section titled “*Hypothesis*.”
4. Follow with a section titled “*Data Analysis Plan.” (This section includes the following subsections: “Dataset Description,” “Data Preparation,” and “Statistical Tests”).*
5. Follow with a section titled *“Description of Results.”*
6. Finish the document with a section titled *“Discussion of Findings.”*
7. The document should be written in Times New Roman, 11pt Font Size, and Single Space.
8. The Final Written Report document should have a minimum of 2,000 and a maximum of 5,000 words.

**Due**: By the end of week 15 of the course schedule.

## Final Stage: Revisions and Final PowerPoint Presentation (Week 16-17).

In this stage, students start by receiving detailed feedback on their final written report. Instead of submitting a revised written report, students are expected to address and incorporate the feedback directly into their final PowerPoint presentation. They have two weeks to carefully revise their analysis, clarify any issues, and enhance their explanation of the research process, using the feedback provided.

During the final week of the semester, students will submit a 10-minute recorded presentation via Zoom. This presentation must describe the research problem, the statistical analysis, the results, and offer a discussion of the key insights revealed by the analysis. The presentation should directly address the feedback given on the final report, ensuring that all areas of concern or improvement are effectively incorporated. Students will use a slide deck to support their explanation, making sure the slides are clear, well-organized, and visually engaging.

The presentation quality will be assessed on students’ ability to explain the overall research process, articulate the hypothesis, and justify the specific statistical tests used. They should demonstrate a clear understanding of their results, explaining the statistical methods and tests, their appropriateness and assumptions, and their broader implications. A critical component of the assessment will be how well students integrate and respond to the feedback provided on their final written report, as well as their ability to clearly communicate statistical findings within the 10-minute time frame.

**Final Stage Task**: Upload to Brightspace two deliverables:

1. A link with the recorded zoom presentation.
2. The slide deck used in the presentation (as a PowerPoint or PDF file).

**Due**: By the end of week 17 of the course schedule.

# Final Thoughts: The Value of Data Analysis as a Learning Exercise

This assignment is designed as a learning exercise to help you develop practical data analysis skills. Whether you collect your own data through a survey or use publicly available datasets, the primary goal is to engage with the process of preparing and analyzing data. This is a valuable skill that many of you will likely need at some point in your professional careers, particularly in the field of public policy.

By working through the steps of this project, you will gain hands-on experience in:

* Defining a research question or identifying a policy problem that interests you.
* Collecting data—either by creating a survey or sourcing existing datasets.
* Preparing data for analysis, including cleaning and organizing it.
* Performing basic statistical analysis to draw meaningful insights.

This project is not about creating a perfect survey or working with an ideal dataset. Moreover, the project is not about writing a scientific article or producing publishable research. Instead, it is about learning how to apply the logic of data analysis. You will start with a question or problem you want to explore, gather or access quantifiable information to investigate it, and then use that data to perform simple statistical tests. This will help you address the problem and gain insights from a data-driven perspective.

One of the key challenges of this assignment is finding an interesting policy problem or research question that you care about. Once you have identified your topic, the rest of the task is about figuring out how to collect and analyze relevant data. The statistical techniques you will use will be straightforward and align with the material covered in this course. Remember, this is an introduction to statistics, so the analyses you conduct do not need to be complex or perfect—what matters most is that you go through the process and understand how to use data to explore real-world questions.

# Appendix

## Guidelines for Collecting Original Data Using Surveys

For students choosing to collect their own data, the use of surveys is a practical and instructive option. This section outlines the expectations and guidelines for conducting a survey and using the data for your project analysis.

**Survey Platform:**

Students should use either *Google Forms* or *Qualtrics* to create and distribute their survey. Both platforms are available through your Stony Brook University account, providing you with free and easy-to-use tools for survey creation.

**Target Audience and Distribution Plan:**

Your survey must aim to gather responses from a wide range of people beyond close friends and family. Although this will be a **convenience sample**, the goal is to achieve a diverse set of responses. Aim to reach at least 100 participants for your survey, though there is no strict requirement. The key is to demonstrate effort in distributing the survey widely.

**Distribution Strategy:**

To expand your survey’s reach, you are encouraged to:

* Share the survey on social media platforms (e.g., Facebook, Twitter, Reddit, Instagram).
* Post the survey in relevant forums or groups (e.g., Reddit communities or Facebook groups that align with your topic).
* Ask your friends, family, and classmates to share the survey with their networks (*a form of snowball sampling)*.
* Use email or messaging apps to encourage participation but make it clear that participation is voluntary.

Although there is no penalty for not reaching 100 responses, you must have a clear and well-defined plan for engaging as many people as possible. **You must include this plan in the Stage 3’s Deliverable (Study Hypothesis and Data Analysis Plan)**. The distribution strategy should be carefully considered and aim for the maximum *feasible* reach

**Survey Content and Target Population:**

It is crucial that your survey targets the right population. If your research question relates to the United States, for example, you should ensure that the majority of your responses come from individuals residing in the U.S. Include relevant questions such as “What is your country of residence?” to help control for this in your analysis. Thus, depending on your target population you must include at the beginning of the survey a set of “screening questions.” Screening questions are preliminary questions used at the start of a survey to ensure respondents meet specific criteria relevant to the study, such as their location, age, or occupation. For example, in Google Forms, you can implement screening questions by using multiple-choice or drop-down options, then applying the "*Go to section based on answer*" feature to redirect or exclude respondents who don't fit your target population.

You must design a set of questions for your survey, using benchmarks from Pew Research studies or other reputable institutions. **You must include the survey’s questionnaire in the Stage 3’s Deliverable (Study Hypothesis and Data Analysis Plan).** Ensure that your survey does not collect personally identifiable information (PII), such as names, email addresses, or other sensitive data. However, you may ask basic demographic questions (e.g., age, gender, location, political affiliation, education level) for subgroup analysis. Inform participants that their responses will remain anonymous and that the data will be analyzed only in aggregate form

**Survey Statement of Purpose:**

At the beginning of your survey, include a clear statement to participants explaining the nature of the survey. This will help set expectations and ensure transparency:

“*This survey is part of a course assignment for the Introductory Statistics course at Stony Brook University. The data collected will not be used for any research projects, publications, or commercial purposes. Participation is voluntary and anonymous. No personal information such as your name or contact details is requested through this poll. The responses will only be analyzed in aggregate, and the results will not be made publicly available. This survey’s goal is for educational purposes only and it is not part of any research activities conducted at Stony Brook University.”*

This statement ensures that respondents understand the purpose and scope of your survey and that their participation is for educational purposes only.

**Expectations for Survey Design:**

Your survey should be designed in a way that aligns with your project goals and hypotheses. While this is a non-scientific poll, it is still important to carefully craft questions that will allow for meaningful statistical analysis. Consider the following when creating your survey:

* Clear and concise questions: Avoid ambiguity, and make sure the wording of your questions is easily understandable.
* Relevance to your hypothesis: Each question should be designed to gather information that will help you test the hypotheses you’ve formulated.
* Consider how each question will be measured or turned into a numeric quantity (such as using rating scales or using binary variables for multiple-choice options) to ensure that statistical analysis is feasible later.
* Use of demographic questions: Ask basic demographic questions to allow for subgroup analysis later in your project. However, ensure that these questions do not collect sensitive or identifying information.[[1]](#footnote-1)

**Survey Ethics and Privacy:**

In line with ethical standards, your survey must:

* Not collect personal identifiers (e.g., names, emails, phone numbers, zip-code).
* Ensure anonymity of all participants.

**Survey Data Collection Goal:**

Your goal is to gather responses that allow for statistical analysis relevant to your project. While the data does not need to actually be representative of any larger population, you should still aim to distribute the survey widely and thoughtfully. A target of at least 100 responses is suggested, but this is not a strict requirement. The key is that your distribution efforts are clear, and that you demonstrate an intentional plan to reach a broad audience.

## Suggested Sources of Publicly Available Data.

* [Kaggle](https://www.kaggle.com/)
* [Roper Center Public Opinion Archives](http://proxy.library.stonybrook.edu/login?url=https://ropercenter.cornell.edu/ipoll)
* [The World Bank Open Data](https://data.worldbank.org/)
* [Google Dataset Search](https://datasetsearch.research.google.com/)
* [IPUMS](https://ipums.org/)
* [Datasets - Pew Research Center](https://www.pewresearch.org/datasets/)
* [American National Election Studies](https://electionstudies.org/data-center/)
* [General Social Survey](https://gss.norc.org/us/en/gss/get-the-data.html)
* [World Values Survey](https://www.worldvaluessurvey.org/WVSContents.jsp)
* [Gallup Public Dataset Collection](https://www.gallup.com/analytics/318923/world-poll-public-datasets.aspx)
* [Data.gov](https://www.data.gov)
* [European Social Survey](https://ess.sikt.no/en/?tab=overview)
* [Princeton’s Politics Research Guide](https://libguides.princeton.edu/politics/data) (highly recommended).

1. See the following link to understand what is considered to be Personal Identifiable Information (PII): <https://www.ibm.com/topics/pii>. [↑](#footnote-ref-1)