QUANTITATIVE RESEARCH

WILLIAM KWESI DONKOH
(UD95680AC104902

Lesson 1: Parametric Tests

1. Definition and Foundational Characteristics

Parametric tests are conventional statistical methods that assume a given data set has been drawn from a population with a specific, known probability distribution—specifically, a normal distribution. As stated in the lesson notes, "parametric tests make certain assumptions about a data set; namely, that the data are drawn from a population with a specific (normal) distribution." This stands in contrast to non-parametric tests, which make fewer assumptions about the data set's underlying distribution.

The term "parametric" was coined by statistician Jacob Wolfowitz in 1942 to define the case where "the distribution functions of the various stochastic variables which enter into their problems are assumed to be of known functional form, and the theories of estimation and of testing hypotheses are theories of estimation of and of testing hypotheses about, one or more parameters... the knowledge of which would completely determine the various distribution functions involved." Parametric tests, therefore, operate within a framework where knowing the parameters (such as mean and standard deviation) completely defines the distribution.

2. Statistical Power and the Trade-Off of Assumptions

Parametric tests generally possess higher statistical power compared to non-parametric alternatives. The lesson notes explicitly state that "parametric tests generally have higher statistical power." This means parametric methods are more likely to detect a true effect when one exists, and they can produce more accurate and precise estimates.

However, this power comes at a cost: reliance on assumptions. "Generally speaking parametric methods make more assumptions than nonparametric methods. If those extra assumptions are correct, parametric methods can produce more accurate and precise estimates." Conversely, the notes warn that "if assumptions are incorrect, parametric methods can be very misleading." For this reason, parametric methods are often not considered robust. Their simplicity (formulae are easier to write down and faster to compute) sometimes compensates for their non-robustness, "especially if care is taken to examine diagnostic statistics."

3. The Hypothesis Testing Framework

Parametric tests operate within a formal hypothesis testing structure. A researcher sets up a null hypothesis against an alternative hypothesis. The lesson notes describe this as "testing, for instance, whether or not the population mean is equal to a certain value, and then using an appropriate statistic to calculate the probability that the null hypothesis is true. You then reject or accept the null hypothesis based on this calculated probability."

Null Hypothesis (H_0): The proposition that nothing happened—no differences, no cause and effect, no effect. As the notes explain, "when you test a statistical hypothesis, you are trying to see if something happened and are comparing against the possibility that nothing happened."

Alternative Hypothesis (H_1): The proposition that something did happen—that groups differ, that a treatment has an effect, or that one variable predicts another.

The logic is confirmatory. You attempt to disprove the null hypothesis. "If you disprove that nothing happened, then you can conclude that something happened." If you reject the null hypothesis, you claim the result is statistically significant and did not occur by chance, thereby proving the alternative hypothesis. If you fail to reject the null hypothesis, you conclude that you did not find an effect or difference.

4. Relationship to Inferential Statistics

Parametric tests are a subset of inferential statistics. The lesson notes define inferential statistics as methods used "to try to infer from the sample data what the population might think" and "to make judgments of the probability that an observed difference between groups is a dependable one or one that might have happened by chance."

Parametric tests fall within a broader family known as the General Linear Model. This family includes the t-test, Analysis of Variance (ANOVA), Analysis of Covariance (ANCOVA), regression analysis, and multivariate methods such as factor analysis, multidimensional scaling, cluster analysis, and discriminant function analysis. The lesson notes emphasize that "given the importance of the General Linear Model, it's a good idea for any serious social researcher to become familiar with its workings."

5. Developing a Parametric Model: Seven Sequential Steps

The lesson notes outline a structured process for developing a parametric estimating model, which consists of seven distinct stages.

Step 1: Cost Model Scope Determination. The first step involves establishing the model's end use, physical characteristics, cost basis, and critical components and cost drivers.

Step 2: Data Collection. This requires significant effort. "The quality of the resulting parametric model can be no better than the quality of the data it is based upon." Both cost and scope information must be identified and collected, often using standardized data input forms.

Step 3: Data Normalization. Before analysis, data must be normalized—adjusted to account for differences between each project's actual basis and a desired standard basis. Normalization includes adjustments for escalation, general location, site conditions, system specifications, estimate title information, total digital and analog I/O (addressed by controller hardware, purchased, and installed), control valves purchased, hardware type (distributed, programmable logic controller), redundant hardware/spare capacity information, and the type of process being controlled.

Step 4: Data Analysis. This step involves performing regression analysis of costs versus selected design parameters to determine key cost drivers. Most spreadsheet applications now provide regression analysis and simulation functions.

Step 5: Data Application. This stage establishes the user interface and presentation form. Using algorithms developed in the data analysis stage, an interface is developed to allow straightforward user input.

Step 6: Testing. Testing the model's accuracy and validity is critical. A key indicator is R^2 (the coefficient of determination), which measures how well the resulting algorithm predicts calculated costs. "An R^2 value of 1 indicates a perfect fit, while an R^2 value of .89 indicates an 89 percent confidence that the regression equation explains the variability in cost." However, the notes caution that "a high R^2 value by itself does not imply that the relationships between the data inputs and the resulting cost are statistically significant."

Step 7: Documentation. The resulting model must be thoroughly documented, including a user manual, clear input descriptions, data normalization discussions, regression data sets, test results, assumptions, allowances, exclusions, applicable input ranges, and model limitations.

6. Practical Example from the Notes

The lesson notes provide a concrete example illustrating parametric logic. Suppose a sample of 99 test scores has a mean of 100 and a standard deviation of 1. If we assume

all 99 test scores are random samples from a normal distribution, we predict a 1% chance that the 100th test score will be higher than 102.365 (mean plus 2.365 standard deviations). This calculation is possible because "the normal family of distributions all have the same shape and are parameterized by mean and standard deviation. That means if you know the mean and standard deviation, and that the distribution is normal, you know the probability of any future observation."

A non-parametric estimate of the same phenomenon would be simpler: there is a 1% chance that the 100th score is higher than any of the 99 preceding scores, requiring no assumption about the distribution's shape. This contrast highlights the core distinction between parametric and non-parametric approaches.

7. Key Inferential Tests Within the Parametric Family

The notes identify several specific parametric tests. The t-test is used "whenever you wish to compare the average performance between two groups." It is appropriate for determining whether eighth-grade boys and girls differ in math test scores or whether a program group differs from a control group on an outcome measure. ANOVA extends this logic to comparisons among more than two groups. ANCOVA incorporates covariates to adjust for pre-existing differences. Regression analysis examines relationships between variables and predicts outcomes.

All of these methods share the parametric assumption of normality. They are appropriate when the data set meets the assumptions of the specific distribution. When those assumptions cannot be made, non-parametric tests serve as alternatives, though they generally have lower statistical power.

Lesson 2: Types- Scientific Method

1. Definition and Core Characteristics

The scientific method is a body of techniques for investigating phenomena, acquiring new knowledge, or correcting and integrating previous knowledge. As defined in the Oxford English Dictionary and cited in the lesson notes, the scientific method is "a method or procedure that has characterized natural science since the 17th century, consisting in systematic observation, measurement, and experiment, and the formulation, testing, and modification of hypotheses."

For a method of inquiry to be termed scientific, it must be based on empirical and measurable evidence subject to specific principles of reasoning. This requirement distinguishes the scientific method from other methods of acquiring knowledge. The chief characteristic distinguishing the scientific method from all other methods is that "scientists seek to let reality speak for itself, supporting a theory when a theory's predictions are confirmed and challenging a theory when its predictions prove false."

Scientific inquiry is intended to be as objective as possible in order to minimize bias. Another basic expectation is the documentation, archiving, and sharing of all data collected or produced and of the methodologies used, so they may be available for careful scrutiny and attempts by other scientists to reproduce and verify them. This practice, known as full disclosure, also means that statistical measures of reliability may be made. The steps of the scientific method must be repeatable to guard against mistake or confusion in any particular experimenter.

2. The Historical Development of the Scientific Method

The development of the scientific method is inseparable from the history of science itself. Ancient Egyptian documents describe empirical methods in astronomy, mathematics, and medicine. In the 7th century BCE, Daniel, a Jewish captive of the Babylonian king Nebuchadnezzar, conducted a scientific experiment complete with a hypothesis, a control group, a treatment group, and a conclusion. The control group partook of the king's delicacies and wine, whereas Daniel's test group limited themselves to vegetables and water. At the end of the test, Daniel's hypothesis was proven true.

The ancient Greek philosopher Thales in the 6th century BCE refused to accept supernatural, religious, or mythological explanations for natural phenomena, proclaiming that every event had a natural cause. The development of deductive

reasoning by Plato was an important step towards the scientific method. Empiricism seems to have been formalized by Aristotle, who believed that universal truths could be reached via induction. According to David Lindberg, Aristotle wrote about the scientific method even if he and his followers did not actually follow what he said. Lindberg also notes that Ptolemy (2nd century CE) and Ibn al-Haytham (11th century CE) are among the early examples of people who carried out scientific experiments. John Losee writes that "Aristotle viewed scientific inquiry as a progression from observations to general principles and back to observations."

However, in order for true scientific method to develop, Aristotle could not be taken at face value. Errors in his "On the Heavens" and "Physics" had to be realized and corrected. Moreover, the pagan view common in the world during that era followed two concepts that prevented progress toward a functional scientific method: (1) an organismic view of nature, where nature and created objects are divine or are themselves without beginning or end, and (2) circular reasoning as opposed to linear reasoning. According to Haffner, cultures debilitated by these concepts included Chinese, Hindu, Meso-American, Egyptian, Babylonian, Greek, and Arabic.

3. Types of Scientific Methods

The lesson notes identify nine distinct types of scientific methods.

3.1 Empirical Method. Empirical research is a way of gaining knowledge by means of direct and indirect observation or experience. Empirical evidence, defined as the record of one's direct observations or experiences, can be analyzed quantitatively or qualitatively. Through quantifying the evidence or making sense of it in qualitative form, a researcher can answer empirical questions, which should be clearly defined and answerable with the evidence collected, usually called data. Research design varies by field and by the question being investigated. Many researchers combine qualitative and quantitative forms of analysis to better answer questions which cannot be studied in laboratory settings, particularly in the social sciences and in education.

3.2 Experimental Method. The experimental method involves manipulating one variable to determine if changes in one variable cause changes in another variable. This method relies on controlled methods, random assignment, and the manipulation of variables to test a hypothesis.

3.3 Hypothetico-Deductive Method. The hypothetico-deductive model or method is a proposed description of scientific method. According to it, scientific inquiry proceeds by formulating a hypothesis in a form that could conceivably be falsified by a test on observable data. A test that could and does run contrary to predictions of the

hypothesis is taken as a falsification of the hypothesis. A test that could but does not run contrary to the hypothesis corroborates the theory. It is then proposed to compare the explanatory value of competing hypotheses by testing how stringently they are corroborated by their predictions.

Concisely, the method involves the traditional steps of observing the subject, in order to elaborate upon an area of study. This allows the researcher to generate a testable and realistic hypothesis. The hypothesis must be falsifiable by recognized scientific methods but can never be fully confirmed, because refined research methods may disprove it at a later date. From the hypothesis, the researcher must generate some initial predictions, which can be proved or disproved by the experimental process. These predictions must be inherently testable for the hypothetico-deductive method to be a valid process.

3.4 Method of Scientific Observation. Scientific observation is the central element of scientific method or process. The core skill of a scientist is to make observation. Observation consists of receiving knowledge of the outside world through our senses, or recording information using scientific tools and instruments. Any data recorded during an experiment can be called an observation.

3.5 Method of Measurement. The technique or process used to obtain data describing the factors of a process or the quality of the output of the process is called the method of measurement. Measurement methods must be documented as part of a Six Sigma project or other process improvement initiative, in order to ensure that measurements of improvements to a process are accurate.

3.6 Dialectic Method. Dialectic, also called dialectics and the dialectical method, is a method of argument for resolving disagreement that has been central to European and Indian philosophy since antiquity. The word dialectic originated in ancient Greece and was made popular by Plato in the Socratic dialogues. The dialectical method is discourse between two or more people holding different points of view about a subject, who wish to establish the truth of the matter guided by reasoned arguments.

The term dialectics is not synonymous with the term debate. While in theory debaters are not necessarily emotionally invested in their point of view, in practice debaters frequently display an emotional commitment that may cloud rational judgment. Debates are won through a combination of persuading the opponent, proving one's argument correct, or proving the opponent's argument incorrect. The term dialectics is also not synonymous with the term rhetoric, a method or art of discourse that seeks to persuade, inform, or motivate an audience. Concepts like "logos" (rational appeal), "pathos" (emotional appeal), and "ethos" (ethical appeal) are intentionally used by rhetoricians to persuade an audience. The Sophists taught arete (quality, excellence) as the highest

value and the determinant of one's actions in life, teaching artistic quality in oratory as a manner of demonstrating one's arete.

3.7 Phenomenological Method. Phenomenology, from Greek "phainomenon" (that which appears) and "logos" (study), is the philosophical study of the structures of experience and consciousness. As a philosophical movement, it was founded in the early years of the 20th century by Edmund Husserl and was later expanded upon by a circle of his followers at the universities of Göttingen and Munich in Germany. Phenomenology, in Husserl's conception, is primarily concerned with the systematic reflection on and study of the structures of consciousness and the phenomena that appear in acts of consciousness. This ontology, or study of reality, can be clearly differentiated from the Cartesian method of analysis which sees the world as objects, sets of objects, and objects acting and reacting upon one another.

The object of phenomenological research is to draw from other people's experiences. Phenomenological researchers figuratively live through their subjects so they can better understand the meaning of their experiences. Phenomenological research poses inherent challenges, as lived experience descriptions are never identical to lived experience itself.

3.8 Historical Method. Historical method comprises the techniques and guidelines by which historians use primary sources and other evidence, including the evidence of archaeology, to research and then to write histories in the form of accounts of the past. The question of the nature, and even the possibility, of a sound historical method is raised in the philosophy of history as a question of epistemology. The study of historical method and of different ways of writing history is known as historiography.

3.9 Inductive Logical Method. An inductive logic is a system of evidential support that extends deductive logic to less-than-certain inferences. For valid deductive arguments, the premises logically entail the conclusion, where the entailment means that the truth of the premises provides a guarantee of the truth of the conclusion. In a good inductive argument, the premises should provide some degree of support for the conclusion, where such support means that the truth of the premises indicates with some degree of strength that the conclusion is true.

The lesson notes specify a Criterion of Adequacy (CoA) for inductive logic: "As evidence accumulates, the degree to which the collection of true evidence statements comes to support a hypothesis, as measured by the logic, should tend to indicate that false hypotheses are probably false and that true hypotheses are probably true."

4. The Relationship Between Theory, Hypothesis, and Experiment

In some fields, quantitative research may begin with a research question. The lesson notes provide the following example: "Does listening to vocal music during the learning of a word list have an effect on later memory for these words?" which is tested through experimentation in a laboratory. Usually, a researcher has a certain theory regarding the topic under investigation. Based on this theory, some statements called hypotheses will be proposed. For example: "Listening to vocal music has a negative effect on learning a word list."

From these hypotheses, predictions about specific events are derived. For example: "People who study a word list while listening to vocal music will remember fewer words on a later memory test than people who study a word list in silence." These predictions can then be tested with a suitable experiment. Depending on the outcomes of the experiment, the theory on which the hypotheses and predictions were based will be supported or not.

5. Summary of the Scientific Method's Essential Features

The scientific method, as synthesized from the lesson notes, rests upon several essential features. It requires empirical and measurable evidence. It demands systematic observation, measurement, and experimentation. It proceeds through hypothesis formulation, prediction generation, and experimental testing. It insists on repeatability to guard against experimenter error. It mandates full disclosure of data and methodologies to enable verification by other scientists. It seeks objectivity to minimize bias. It operates on the principle that reality speaks for itself: theories are supported when predictions are confirmed and challenged when predictions prove false. It recognizes that hypotheses must be falsifiable, and that scientific knowledge is always provisional, subject to revision or rejection as new evidence accumulates.

The scientific method encompasses multiple specific approaches: the empirical method (knowledge through observation), the experimental method (manipulation to determine causation), the hypothetico-deductive method (falsifiable hypotheses and testable predictions), the method of scientific observation (sensory or instrument-based data collection), the method of measurement (documented techniques for obtaining accurate data), the dialectic method (reasoned discourse to resolve disagreement), the phenomenological method (study of consciousness and experience), the historical method (analysis of primary sources to write history), and the inductive logical method (evidential support for less-than-certain inferences).

Lesson 3: Hypothetico-Deductive Method

The Hypothetico-Deductive Method

1. Definition and Basic Characterization

The hypothetico-deductive model or method, often abbreviated as the HD method, is a proposed description of scientific method. According to the lesson notes, "scientific inquiry proceeds by formulating a hypothesis in a form that could conceivably be falsified by a test on observable data." The HD method is described as "a very important method for testing theories or hypotheses" and is characterized as "one of the most basic methods common to all scientific disciplines including biology, physics, and chemistry."

The core logic of the HD method rests on two symmetrical outcomes. A test that could and does run contrary to predictions of the hypothesis is taken as a falsification of the hypothesis. A test that could but does not run contrary to the hypothesis corroborates the theory. The method then proposes "to compare the explanatory value of competing hypotheses by testing how stringently they are corroborated by their predictions."

The lesson notes make a critical epistemological limitation explicit: "This method can never absolutely verify (prove the truth of) 2. It can only falsify 2." This principle is reinforced by a direct quotation from Albert Einstein: "No amount of experimentation can ever prove me right; a single experiment can prove me wrong."

2. The Five Stages of HD Reasoning

The lesson notes divide the application of the hypothetico-deductive method into five distinct stages:

Stage 1: Form many hypotheses and evaluate each hypothesis. Before any testing occurs, the researcher generates multiple possible explanations for the phenomenon under investigation. This stage requires breadth of thinking rather than depth of analysis.

Stage 2: Select a hypothesis to be tested. From the set of formulated hypotheses, the researcher chooses one specific hypothesis for empirical examination. This selection may be based on plausibility, prior evidence, theoretical coherence, or other criteria.

Stage 3: Generate predictions from the hypothesis. The researcher deduces specific, observable consequences that would follow if the hypothesis were true. As the notes

state, "you have to deduce from a hypothesis and make predictions which can be tested through experiments."

Stage 4: Use experiments to check whether predictions are correct. The researcher designs and conducts experiments to determine whether the predicted outcomes actually occur under controlled conditions.

Stage 5: Draw conclusions based on experimental results. If the predictions are correct, the hypothesis is confirmed. If the predictions are incorrect, the hypothesis is disconfirmed.

The notes emphasize a crucial logical constraint: "Many hypotheses can't be tested directly; you have to deduce from a hypothesis and make predictions which can be tested through experiments."

3. The Four Operational Phases

The lesson notes also present a four-phase sequence for implementing the hypothetico-deductive method:

Phase 1: Use your experience. The researcher considers the problem and tries to make sense of it. This involves gathering data and looking for previous explanations. If the problem is new to the researcher, the process moves to Phase 2.

Phase 2: Form a conjecture (hypothesis). When nothing else is yet known, the researcher attempts to state an explanation, either to someone else or in a notebook.

Phase 3: Deduce predictions from the hypothesis. The researcher asks: if the hypothesis is assumed true, what consequences follow?

Phase 4: Test (or experiment). The researcher looks for evidence (observations) that conflict with these predictions in order to disprove the hypothesis. The notes provide a critical warning: "It is a logical error to seek 3 directly as proof of 2. This formal fallacy is called affirming the consequent."

The notes describe the iterative nature of the method: "If the outcome of 4 holds, and 3 is not yet disproven, you may continue with 3, 4, 1, and so forth; but if the outcome of 4 shows 3 to be false, you will have to go back to 2 and try to invent a new 2, deduce a new 3, look for 4, and so forth."

4. The Role of Observations and Data Collection

The lesson notes specify that before a hypothesis can be formulated or an experiment conducted, "a scientist needs to collect data and make some observations on the subject being studied." A hypothesis needs to be an educated guess, so "the observation and data collection step gives some background to be able to formulate a good hypothesis." The notes further state that "the more observations a scientist makes and research he does, the better the hypothesis will be and the more credibility his own research will get."

5. Criteria for a Proper Hypothesis

The lesson notes establish specific requirements for hypothesis formulation. First, "the hypothesis needs to be able to be proven true or false through the experiment with scientific data and proof to back it up." Second, "the hypothesis needs to be specific and not leave room for interpretation, because different interpretations cannot be backed up with scientific proof." The notes provide a concrete warning: "using words such as 'good' or 'bad' should not be used in a hypothesis because what the scientist constitutes as good may be different from what others see as good."

6. Experimental Requirements

Once the hypothesis has been formulated, "the experiment can be conducted." The notes specify that "the experiment should have a control group and a real group so the scientist can compare the differences between the two groups." Additionally, "the experiment should contain some way to measure changes in both groups, so that data can be collected and research shown."

7. The Importance of HD Reasoning in Learning and Society

The lesson notes identify two domains where HD reasoning holds particular importance.

In learning and concept construction: The notes state that "students typically do not come to the learning situation as blank slates. Rather, they come with alternative conceptions (i.e., hypotheses) that must be modified or replaced by scientific conceptions." Through HD reasoning and experimentation, "students can test their preconceptions against scientific concepts and find out which match experimental results. This promotes conceptual change."

In everyday life: The notes provide a concrete example. Suppose your portable music player fails to switch on. You hypothesize that the batteries are dead. From this

hypothesis, you predict that the music player should work properly if you replace the batteries with new ones. You replace the batteries (the experiment). If the player works again, the hypothesis is confirmed. If the player still does not work, "the prediction was false, and the hypothesis is disconfirmed. You might reject your original hypothesis and come up with an alternative one to test, such as the batteries are fine but your music player is broken."

8. Piaget and Developmental Levels of HD Reasoning

According to Piaget's theory of intellectual development, cited in the notes, "HD reasoning appears in the formal operational stage." However, the notes also reference Lawson et al. (2000), who claim "that there are two general developmentally-based levels of hypothesis-testing skill." The first level involves skills associated with testing hypotheses about observable causal agents. The second level involves testing hypotheses about unobservable entities. The notes state that "the ability to test alternative explanations involving unseen theoretical entities is a fifth stage of intellectual development that goes beyond Piaget's four stages."

9. Simplified Example: Red Blood Cells and Salt Water

The lesson notes provide a detailed worked example to illustrate HD reasoning.

Observation: A student looked at red blood cells under a microscope and saw little round balls. After adding salt water, the cells appeared smaller.

Question: Why do the red blood cells appear smaller?

Two competing hypotheses:

- **Hypothesis I:** Salt ions (Na+ and Cl-) push on the cell membranes and make the cells appear smaller.
- **Hypothesis II:** Water molecules are attracted to the salt ions so the water molecules move out of the cells and leave the cells smaller.

Experiment: The student used salt water, a very accurate weighing device, and water-filled plastic bags (assuming the plastic behaves like red blood cell membranes). The student weighed a water-filled bag in a salt solution for ten minutes and then reweighed the bag.

Predictions and logical deduction:

- If Hypothesis I is correct (salt ions pushing), then no molecules or ions are coming into or going out of the bag. Therefore, the weight of the bag will not change.
- If Hypothesis II is correct (water molecules moving out), then the bag will lose weight because water molecules exit the bag.

Testing which hypothesis is probably wrong:

- What result would show Hypothesis I is probably wrong? Answer A: The bag loses weight. (If the bag loses weight, Hypothesis I's prediction of no weight change is false.)
- What result would show Hypothesis II is probably wrong? Answer B: The bag weighs the same. (If the bag weighs the same, Hypothesis II's prediction of weight loss is false.)

This example demonstrates the core HD logic: different hypotheses generate different predictions. The experiment produces evidence that can falsify one hypothesis while corroborating the other.

Experimentum Crucis (Crucial or Critical Experiment)

1. Definition and Origin

In the sciences, an experimentum crucis (English: crucial experiment or critical experiment) is defined in the lesson notes as "an experiment capable of decisively determining whether or not a particular hypothesis orthodoxy is superior to all other hypotheses or theories whose acceptance is currently widespread in the scientific community."

The notes specify the decisive characteristic of such an experiment: "such an experiment must typically be able to produce a result that rules out all other hypotheses or theories if true, thereby demonstrating that under the conditions of the experiment (i.e., under the same external circumstances and for the same 'input variables' within the experiment), those hypotheses and theories are proven false but the experimenter's hypothesis is not ruled out."

Historical origins of the concept:

- Francis Bacon first described the concept in his *Novum Organum*, using the name "instantia crucis" (the crucial instance) to denote a situation in which one theory but not others would hold true.
- The phrase *experimentum crucis*, denoting the deliberate creation of such a situation for the purpose of testing rival theories, was later coined by Robert Hooke.
- The term was famously used by Isaac Newton.

The notes state that "the production of such an experiment is considered necessary for a particular hypothesis or theory to be considered an established part of the body of scientific knowledge." However, the notes also acknowledge that "it is not unusual in the history of science for theories to be developed fully before producing a critical experiment. A given theory which is in accordance with known experiment but which has not yet produced a critical experiment is typically considered worthy of exploration in order to discover such an experimental test."

2. Historical Examples of Experimentum Crucis

Example 1: Isaac Newton (17th Century). In his *Philosophiae Naturalis Principia Mathematica* (1687), Newton presented a disproof of Descartes' vortex theory of the motion of the planets. This represented a crucial experiment that ruled out the prevailing alternative theory. Additionally, in his *Opticks*, Newton described an optical *experimentum crucis* in the First Book, Part I, Proposition II, Theorem II, Experiment 6, "to prove that sunlight consists of rays that differ in their index of refraction."

Example 2: Eddington's 1919 Solar Eclipse Expedition. The notes describe this as "a famous example in the 20th century of an experimentum crucis." The expedition, led by Arthur Eddington to Principe Island in Africa in 1919, recorded the positions of stars around the sun during a solar eclipse. "The observation of star positions confirmed predictions of gravitational lensing made by Albert Einstein in the general theory of relativity published in 1915." The notes conclude that "Eddington's observations were considered to be the first solid evidence in favor of Einstein's theory."

Example 3: Planck's Quantum Hypothesis (1900). The notes provide a nuanced example. Max Planck proposed the quantum hypothesis to account for the observed black-body spectrum, "an experimental result which the existing classical Rayleigh-Jeans law could not predict." However, the notes caution that "such cases are not considered strong enough to fully establish a new theory." In the case of quantum mechanics, "it took the confirmation of the theory through new predictions for the theory to gain full acceptance." This example illustrates that explaining existing anomalous results, while

important, does not constitute a true experimentum crucis unless the new theory also generates novel predictions that are subsequently confirmed.

3. Distinction Between Simple Confirmation and Crucial Experiment

A standard experiment can confirm or disconfirm a single hypothesis. An experimentum crucis, by contrast, must discriminate decisively between competing hypotheses or theories. The notes emphasize that a crucial experiment "rules out all other hypotheses or theories if true." This is a higher standard than simple hypothesis testing. It requires that the experimental design be capable of producing results that are logically incompatible with the predictions of every alternative theory except the one being advanced.

Summary Table: Key Differences Between HD Method and Experimentum Crucis

Feature	Hypothetico-Deductive Method	Experimentum Crucis
Primary purpose	Test a single hypothesis or theory	Decisively determine superiority between competing theories
Outcome	Confirmation or disconfirmation of the tested hypothesis	Ruling out all alternative hypotheses
Logical structure	If predictions correct → confirmed; if incorrect → disconfirmed	Produces result that is incompatible with all rival theories
Historical origin	Formalized by Popper and others in 20th century	Bacon (*instantia crucis*), Hooke (coined term), Newton (famously used)
Epistemological status	Can never absolutely verify; can only falsify	Considered necessary for a theory to be established as scientific knowledge

Lesson 4: Scientific Theory

1. Definition and Core Characteristics

A scientific theory is defined in the lesson notes as "a well-substantiated explanation of some aspect of the natural world that is acquired through the scientific method, and repeatedly confirmed through observation and experimentation." This definition establishes three essential criteria that distinguish scientific theories from other forms of knowledge. First, the explanation must be well-substantiated, meaning it rests upon substantial empirical evidence rather than speculation. Second, it must be acquired through the scientific method, which ensures systematic, repeatable, and objective inquiry. Third, it must be repeatedly confirmed through observation and experimentation, meaning the same results can be obtained by multiple investigators under similar conditions.

The notes further characterize scientific theories as "inductive in nature and aim for predictive power and explanatory force." Inductive reasoning moves from specific observations to general principles. A scientific theory does not merely describe what has been observed; it provides a framework for predicting what will be observed under conditions not yet tested. Explanatory force refers to the theory's ability to account for why phenomena occur, not just that they occur.

The lesson notes draw a sharp distinction between scientific usage of the term "theory" and common usage. "This is significantly different from the common usage of the word 'theory', which implies that something is a guess (i.e., unsubstantiated and speculative)." In everyday language, people say "I have a theory" to mean a hunch or an untested idea. In science, a theory represents the highest, most reliable, most rigorous, and most comprehensive form of scientific knowledge.

2. The Strength of a Scientific Theory

The strength of a scientific theory is determined by three factors enumerated in the notes: "the diversity of phenomena it can explain, and to its elegance and simplicity (Occam's razor)."

Diversity of phenomena: A theory that explains a narrow range of observations is weaker than a theory that explains a wide range of seemingly unrelated observations. For example, the theory of evolution explains not only fossil records but also comparative anatomy, embryology, genetics, biogeography, and observed speciation events. This broad explanatory scope strengthens the theory.

Elegance and simplicity (Occam's razor): Occam's razor is the principle that, among competing explanations that account for the same phenomena, the simplest explanation requiring the fewest assumptions is preferred. A theory that explains complex phenomena with parsimonious assumptions is stronger than a theory that requires many ad hoc adjustments or auxiliary hypotheses.

The notes provide a concrete example of how a theory can remain useful even when not perfectly accurate: "Newton's laws of motion as an approximation to special relativity at velocities which are small relative to the speed of light." Newtonian mechanics is known to be inaccurate at velocities approaching the speed of light. However, for everyday human experience involving comparatively low velocities, Newton's laws are "almost exactly correct" and remain useful approximations. A less-accurate theory can still be treated as a theory "if it is useful (due to its sheer simplicity) as an approximation under specific conditions."

3. Testability and Falsifiability

Scientific theories, according to the notes, "are testable and make falsifiable predictions." Two criteria are embedded in this statement.

Testability means that the theory generates specific predictions about observable phenomena that can be examined through empirical investigation. A theory that makes no testable predictions cannot be evaluated scientifically. It remains outside the domain of science.

Falsifiability means that the predictions are stated in such a way that they could be proven false by experimental evidence. A theory that can accommodate any possible outcome is not falsifiable and therefore not scientific. The notes emphasize that "the predictions made by classical mechanics are known to be inaccurate in the relativistic realm," but this inaccuracy was discovered precisely because the predictions were specific enough to be proven wrong.

The notes also state that theories "describe the causal elements responsible for a particular natural phenomenon." This causal requirement distinguishes scientific theories from mere correlational descriptions. A theory must specify not only that two variables are associated but also the mechanism through which one produces changes in the other.

4. The Formation Process of Scientific Theories

The lesson notes describe a sequential process through which hypotheses become theories.

Step 1: Proposal and testing of hypotheses. The scientific method involves "the proposal and testing of hypotheses, by deriving predictions from the hypotheses about the results of future experiments, then performing those experiments to see whether the predictions are valid." This provides evidence either for or against each hypothesis.

Step 2: Accumulation of experimental results. "When enough experimental results have been gathered in a particular area of inquiry, scientists may propose an explanatory framework that accounts for as many of these as possible." This explanatory framework is a candidate theory.

Step 3: Testing the explanatory framework. The proposed explanation itself is tested. "If it fulfills the necessary criteria, then the explanation becomes a theory." The notes emphasize that "this can take many years, as it can be difficult or complicated to gather sufficient evidence."

Step 4: Acceptance by the scientific community. "Once all of the criteria have been met, it will be widely accepted by scientists as the best available explanation of at least some phenomena." The theory must have "made predictions of phenomena that previous theories could not explain or could not predict accurately, and it will have resisted attempts at falsification."

Step 5: Replication of critical experiments. "The strength of the evidence is evaluated by the scientific community, and the most important experiments will have been replicated by multiple independent groups." Replication serves as a safeguard against experimenter bias, methodological error, or fraud. A finding that cannot be replicated does not provide reliable support for a theory.

5. The Provisional Nature of Scientific Theories

The lesson notes make explicit that no scientific theory is ever absolutely certain. "Like all knowledge in science, no theory can ever be completely certain, since it is possible that future experiments might conflict with the theory's predictions." This provisionality is not a weakness of scientific knowledge. It is a defining feature. Science progresses by subjecting its most cherished theories to continuous testing and potential falsification.

However, the notes also establish that some theories have such overwhelming supporting evidence that they are treated as certain for practical purposes. "Theories supported by the scientific consensus have the highest level of certainty of any scientific knowledge; for example, that all objects are subject to gravity or that life on Earth evolved from a common ancestor."

The notes also clarify that a theory can be accepted without all of its predictions being tested. "Acceptance of a theory does not require that all of its major predictions be tested, if it is already supported by sufficiently strong evidence. For example, certain tests may be unfeasible or technically difficult." In such cases, the predicted results are described informally as "theoretical." These predictions can be tested at a later time, and if they are proven incorrect, this may lead to revision or rejection of the theory.

6. Modification and Improvement of Theories

The lesson notes describe a systematic process for responding to experimental results that contradict a theory's predictions.

Step 1: Evaluate experimental design. "Scientists first evaluate whether the experimental design was sound." This step checks for methodological flaws that could produce spurious results.

Step 2: Confirm by independent replication. "If so they confirm the results by independent replication." A single anomalous result does not refute a theory; the anomaly must be reproducible.

Step 3: Search for improvements. "A search for potential improvements to the theory then begins. Solutions may require minor or major changes to the theory, or none at all if a satisfactory explanation is found within the theory's existing framework."

Step 4: Progressive improvement over time. "Over time, as successive modifications build on top of each other, theories consistently improve and greater predictive accuracy is achieved." The notes emphasize a critical constraint: "Since each new version of a theory (or a completely new theory) must have more predictive and explanatory power than the last, scientific knowledge consistently becomes more accurate over time."

Step 5: Replacement of theory. "If modifications to the theory or other explanations seem to be insufficient to account for the new results, then a new theory may be required." The notes note that "this occurs much less commonly than modification" because "scientific knowledge is usually durable."

Step 6: Retention of previous theory until replacement is available. "Until such a theory is proposed and accepted, the previous theory will be retained. This is because it is still the best available explanation for many other phenomena, as verified by its predictive power in other contexts."

The notes provide a concrete historical example: "It was known in 1859 that the observed perihelion precession of Mercury violated Newtonian mechanics, but the theory remained the best explanation available until relativity was supported by sufficient evidence." For over fifty years, Newtonian mechanics was known to be incorrect in one specific domain, yet it remained the accepted theory because no superior alternative existed. When Einstein's general theory of relativity provided both a more accurate explanation of Mercury's orbit and correct predictions of novel phenomena (such as gravitational lensing), Newtonian mechanics was replaced.

7. Unification of Theories

The lesson notes describe a process called unification, where "two or more theories may be replaced by a single theory which explains the previous theories as approximations or special cases." Unification is analogous to the way a theory provides a unifying explanation for many confirmed hypotheses.

The notes provide one clear example: "For example, electricity and magnetism are now known to be two aspects of the same phenomenon, referred to as electromagnetism." Prior to unification, separate theories existed for electrical phenomena and magnetic phenomena. Maxwell's theory of electromagnetism demonstrated that these were not independent domains but different manifestations of a single underlying phenomenon.

The notes also discuss the resolution of contradictory predictions between theories. "When the predictions of different theories appear to contradict each other, this is also resolved by either further evidence or unification." The notes provide an example from 19th-century physics: "Physical theories in the 19th century implied that the Sun could not have been burning long enough to allow certain geological changes as well as the evolution of life. This was resolved by the discovery of nuclear fusion, the main energy source of the Sun." Contradictions between theories can also be explained "as the result of theories approximating a more fundamental, non-contradictory explanation."

8. Summary: Scientific Theory vs. Common Usage

The following table synthesizes the key distinctions between scientific theory and common usage as presented in the lesson notes.

Feature	Scientific Theory	Common Usage of "Theory"
Epistemological status	Well-substantiated, repeatedly confirmed	Unsubstantiated, speculative
Evidence base	Empirical, replicable, gathered over many years	Little or no evidence
Predictive power	Makes testable, falsifiable predictions	Makes no specific predictions
Explanatory scope	Explains diverse phenomena across multiple domains	Explains single observation or anecdote
Community acceptance	Accepted by scientific consensus after extensive testing	Personal opinion or hunch
Provisionality	Subject to modification or replacement with new evidence	Resistant to counter-evidence
Example	Theory of evolution, theory of relativity	"I have a theory about why my keys are missing"

Lesson 5: Population Parameters

1. Definition and Purpose of Hypothesis Testing

Hypothesis testing is defined in the lesson notes as "a way of systematically quantifying how certain you are of the result of a statistical experiment." This definition establishes two essential features. First, the process is systematic, meaning it follows a predetermined, repeatable set of procedures rather than relying on intuition or subjective judgment. Second, it quantifies certainty, meaning it produces numerical probabilities that allow researchers to make decisions with known levels of confidence.

The lesson notes also describe hypothesis testing as "a process of evaluating a research question" that is "sometimes also referred to as significance testing." The term "significance" refers to whether an observed effect is likely to have occurred by chance alone or reflects a genuine phenomenon in the population being studied.

The fundamental purpose of hypothesis testing is to determine whether a research hypothesis extends beyond the specific individuals examined in a single study to the broader population from which the sample was drawn. The notes provide a concrete illustration: "If Sarah and Mike wanted to know which teaching method was the best, they could simply compare the performance achieved by the two groups of students... and conclude that the best method was the teaching method which resulted in the highest performance. However, this is generally of only limited appeal because the conclusions could only apply to students in this study." Hypothesis testing allows researchers to generalize findings from a sample to a population.

The notes provide another example: "taking a sample of 200 breast cancer sufferers in order to test a new drug that is designed to eradicate this type of cancer. As much as you are interested in helping these specific 200 cancer sufferers, your real goal is to establish that the drug works in the population (i.e., all breast cancer sufferers)."

2. Population Parameters

A population parameter is defined as "a term used in statistics" that "refers to a measurement of a population that is being studied." More precisely, "the parameter is a number that is used by scientists to describe a specific population or group." Before any hypothesis testing can occur, researchers must identify the population of interest and the specific parameters they intend to measure.

A population is defined as "any group of people or objects that a researcher wants to measure." Examples provided include "people in the United States, or trees in a forest." A population parameter must also have a unit that is being measured. For people in the United States, "possible units that could be measured include number, height and eye color."

The notes distinguish between small and large populations. "Small populations can be measured directly, but when researchers study large populations, they often take samples to represent the entire population." Statistics is defined as "a branch of mathematics that allows researchers to make guesses about population parameters based on data gathered from random samples of a population." Depending on the size of the sample, "researchers will use a margin of error to describe how close they think a sample statistic is to a population parameter."

The notes specify several families of probability distributions characterized by different parameters. The normal distribution has two parameters: the mean and the variance. If these are specified, "the distribution is known exactly." The chi-squared distribution has one parameter: "the number of degrees of freedom." The Poisson distribution, binomial distribution, and exponential distribution are also identified as parameterized families.

In statistical inference, parameters are "sometimes taken to be unobservable, and in this case the statistician's task is to infer what they can about the parameter based on observations of random variables." The notes identify specific types of parameters: location parameter, dispersion parameter or scale parameter, shape parameter, and concentration parameter. Regression coefficients are also identified as "statistical parameters in the above sense, since they index the family of conditional probability distributions that describe how the dependent variables are related to the independent variables."

3. The Null and Alternative Hypotheses

The lesson notes establish a clear distinction between two competing hypotheses.

Null hypothesis (H_0): The notes define the null hypothesis as "a statement about the world which can plausibly account for the data you observe." It is also described as "the hypothesis of 'no difference'" and "usually formulated for the purpose of being rejected." The notes caution: "Don't read anything into the fact that it's called the 'null' hypothesis — it's just the hypothesis we're trying to test."

An example is provided: "the coin is fair" is a null hypothesis, as is "the coin is biased." The critical requirement is that "the null hypothesis be able to be expressed in simple, mathematical terms."

For a two-tailed test, "the null hypothesis is typically that a parameter equals zero although there are exceptions. A typical null hypothesis is $\mu_1 - \mu_2 = 0$ which is equivalent to $\mu_1 = \mu_2$."

For a one-tailed test, "the null hypothesis is either that a parameter is greater than or equal to zero or that a parameter is less than or equal to zero." If the prediction is that μ_1 is larger than μ_2, "then the null hypothesis (the reverse of the prediction) is $\mu_2 - \mu_1 \geq 0$. This is equivalent to $\mu_1 \leq \mu_2$."

Alternative hypothesis (H_a or H_1): The notes define the alternative hypothesis as "the hypothesis that contradicts the null hypothesis." It is also described as "the operational statement of the experimenter's research hypothesis." If the null hypothesis is rejected, "the alternative is being supported."

The relationship between the two hypotheses is a logical complement. If evidence contradicts the null hypothesis sufficiently, researchers reject it and support the alternative. If evidence does not sufficiently contradict the null, researchers fail to reject it.

4. The Four Basic Steps of Hypothesis Testing

The lesson notes outline a four-step procedure for hypothesis testing.

Step 1: Specify the null hypothesis. As described above, this involves stating the null hypothesis in precise mathematical terms appropriate to a one-tailed or two-tailed test.

Step 2: Specify the α level (significance level). The alpha level is defined as "the significance level." The notes state that "typical values are 0.05 and 0.01." The alpha level represents the probability of committing a Type I error (rejecting a true null hypothesis) that the researcher is willing to accept.

Step 3: Compute the probability value (p-value). The p-value is defined as "the probability of obtaining a sample statistic as different or more different from the parameter specified in the null hypothesis given that the null hypothesis is true."

Step 4: Compare the p-value with the α level. "If the probability value is lower then you reject the null hypothesis." The notes clarify that "rejecting the null hypothesis is not an all-or-none decision. The lower the probability value, the more confidence you can have that the null hypothesis is false."

However, "if your probability value is higher than the conventional α level of 0.05, most scientists will consider your findings inconclusive." The notes emphasize a critical principle: "Failure to reject the null hypothesis does not constitute support for the null hypothesis. It just means you do not have sufficiently strong data to reject it."

5. Type I and Type II Errors

The lesson notes identify two types of errors that can be committed in hypothesis testing.

Type I error: Defined as "the situation where we incorrectly reject H_0 when in fact it is true." This is also called "a false positive result (as we incorrectly conclude that the research hypothesis is true when in fact it is not)."

When researchers run a test of hypothesis and decide to reject H_0, "then either we make a correct decision because the research hypothesis is true or we commit a Type I error." The notes state that "we will never know whether the null hypothesis is really true or false" in any given study. The alpha level set in Step 2 is the probability of committing a Type I error.

Type II error: While not explicitly defined in the provided pages, the concept is implied by the discussion of Type I error and the mention of two types of errors. In standard statistical terminology (consistent with the notes' framing), a Type II error is failing to reject a false null hypothesis (a false negative).

The notes present a summary table (referenced but not fully shown in the excerpt) that categorizes the four possible outcomes of hypothesis testing based on whether H_0 is actually true or false and whether the researcher rejects or fails to reject it.

6. The Sample-to-Population Generalization Problem

The lesson notes explain why hypothesis testing is necessary rather than simply comparing sample statistics directly.

"In statistics terminology, the students in the study are the sample and the larger group they represent (i.e., all statistics students on a graduate management degree) is called the population." The notes state that if the sample is representative of the population, "you can use hypothesis testing to understand whether any differences or effects discovered in the study exist in the population."

The notes provide a concrete illustration of the challenge: "If Sarah and Mike wanted to know which teaching method was the best, they could simply compare the performance achieved by the two groups of students... and conclude that the best method was the teaching method which resulted in the highest performance." However, "this is generally of only limited appeal because the conclusions could only apply to students in this study."

The problem arises because of sampling variation. The notes explain: "Because of the sampling variation, [samples] will not have the same exact value as the population parameter. Hence, differences between the sample information and the population under study might be due chance." The procedure of statistical testing provides the basis for deciding "whether differences between the sample observation and the hypothesized value could be due to sampling variation alone, or are so large enough as to make the proposed statement untenable."

7. The Nine-Step Hypothesis Testing Structure

The lesson notes provide a detailed nine-step structure for conducting hypothesis testing.

1. **Define the research hypothesis for the study.** This is the initial statement of what the researcher expects to find.
2. **Explain how you are going to operationalize what you are studying and set out the variables to be studied.** Operationalization means measuring or operationally defining the concepts of interest.
3. **Set out the null and alternative hypothesis (or more than one hypothesis; in other words, a number of hypotheses).** This step translates the research hypothesis into statistically testable form.
4. **Set the significance level.** This establishes the alpha threshold for rejecting the null hypothesis.
5. **Make a one- or two-tailed prediction.** This determines whether the test will examine differences in one direction (e.g., Group A > Group B) or both directions (Group A ≠ Group B).

6. **Determine whether the distribution that you are studying is normal.** The notes state that "this has implications for the types of statistical tests that you can run on your data." Normal distributions permit parametric tests; non-normal distributions may require non-parametric alternatives.
7. **Select an appropriate statistical test based on the variables you have defined and whether the distribution is normal or not.**
8. **Run the statistical tests on your data and interpret the output.**
9. **Reject or fail to reject the null hypothesis.** This final step produces the conclusion of the hypothesis test.
10.

8. Intuitive Understanding of Hypothesis Testing

The lesson notes provide an intuitive example to build understanding before introducing formal statistical concepts.

A coin is flipped. The null hypothesis is that the coin is fair. Flipping it 100 times produces 51 heads. The notes ask: "Do we know whether the coin is biased or not?" The expected number of heads is 50, and 51 is quite close. "But what if we flipped the coin 100,000 times and it came up heads 51,000 times? We see 51% heads both times, but in the second instance the coin is more likely to be biased."

The notes articulate a critical principle: "Lack of evidence to the contrary is not evidence that the null hypothesis is true. Rather, it means that we don't have sufficient evidence to conclude that the null hypothesis is false. The coin might actually have a 51% bias towards heads, after all."

If instead the coin produced 1 head in 100 flips, "intuitively we know that the chance of seeing this if the null hypothesis were true is so small that we would be comfortable rejecting the null hypothesis and declaring the coin to (probably) be biased."

9. Summary of Key Hypothesis Testing Concepts

The following table synthesizes the core concepts presented in the lesson notes.

Concept	Definition	Example
Population Parameter	A measurement of a population being studied	Mean height of all US adults
Null Hypothesis (H_0)	Statement of "no difference"; formulated to be rejected	The coin is fair; $\mu_1 = \mu_2$
Alternative Hypothesis (H_a)	Contradicts the null; operational statement of research hypothesis	The coin is biased; $\mu_1 \neq \mu_2$
Alpha Level (α)	Significance level; probability of Type I error accepted	0.05 or 0.01
P-value	Probability of obtaining sample statistic as extreme as observed, given H_0 true	Calculated from test statistic
Type I Error	Rejecting a true null hypothesis (false positive)	Concluding drug works when it does not
Type II Error	Failing to reject a false null hypothesis (false negative)	Concluding drug does not work when it does
Decision Rule	Reject H_0 if p-value < α; fail to reject if p-value ≥ α	$p = 0.03$, $\alpha = 0.05$ → reject H_0

Lesson 6: The Chi-Square Test

1. Definition and Foundational Characteristics

A chi-squared test, also referred to in the lesson notes as "chi-square test or χ^2 test," is defined as "any statistical hypothesis test in which the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true." This definition establishes the core condition: the test statistic follows a chi-square distribution under the null hypothesis.

The notes also identify a broader application: "a chi-squared test is a test in which this is asymptotically true, meaning that the sampling distribution (if the null hypothesis is true) can be made to approximate a chi-squared distribution as closely as desired by making the sample size large enough." Asymptotic means the approximation improves as sample size increases. For smaller samples, the chi-square distribution may be only approximately valid.

The chi-square test was first investigated by Karl Pearson in 1900. When the term "chi-squared test" is mentioned "without any modifiers or without other precluding context," the notes specify that "this test is usually meant" to refer to Pearson's chi-squared test.

2. Types of Chi-Square Tests

The lesson notes identify multiple specific chi-square tests, each suited to different research contexts.

Pearson's chi-squared test: Also known as "the chi-squared goodness-of-fit test or chi-squared test for independence." This is the most widely used chi-square test.

Yates's correction for continuity: Also referred to as "Yates' chi-squared test." This is a correction applied to improve the approximation for small samples.

Cochran-Mantel-Haenszel chi-squared test: A test used for stratified contingency tables.

McNemar's test: "Used in certain 2×2 tables with pairing." This test is appropriate when the same subjects are measured under two conditions (paired data).

Tukey's test of additivity: A test for interaction effects in two-way tables.

The portmanteau test in time-series analysis: "Testing for the presence of autocorrelation." This test examines whether correlations exist between a time series and its lagged values.

Likelihood-ratio tests: Used "in general statistical modelling, for testing whether there is evidence of the need to move from a simple model to a more complicated one (where the simple model is nested within the complicated one)."

The notes also mention one case "where the distribution of the test statistic is an exact chi-squared distribution." This is "the test that the variance of a normally distributed population has a given value based on a sample variance." However, the notes state that "such a test is uncommon in practice because values of variances to test against are seldom known exactly."

3. Chi-Square Test for Variance in a Normal Population

The lesson notes describe a specific application of the chi-square test for testing population variance.

If a sample of size n is taken from a population having a normal distribution, "then there is a result which allows a test to be made of whether the variance of the population has a pre-determined value." An example is provided: "a manufacturing process might have been in stable condition for a long period, allowing a value for the variance to be determined essentially without error." Then, "suppose that a variant of the process is being tested, giving rise to a small sample of n product items whose variation is to be tested."

The test statistic T "could be set to be the sum of squares about the sample mean, divided by the nominal value for the variance (i.e. the value to be tested as holding). Then T has a chi-squared distribution with n-1 degrees of freedom." A concrete example is given: "if the sample size is 21, the acceptance region for T for a significance level of 5% is the interval 9.59 to 34.17."

4. Minimum Chi-Square Estimation

The notes introduce a related concept: minimum chi-square estimation, defined as "a method of estimation of unobserved quantities based on observed data."

The logic proceeds as follows. "In certain chi-square tests, one rejects a null hypothesis about a population distribution if a specified test statistic is too large, when that statistic would have approximately a chi-square distribution if the null hypothesis is true." In

minimum chi-square estimation, "one finds the values of parameters that make that test statistic as small as possible."

Among the consequences of its use is that "the test statistic actually does have approximately a chi-square distribution when the sample size is large." The notes specify a rule for degrees of freedom: "Generally, one reduces by 1 the number of degrees of freedom for each parameter estimated by this method."

5. Pearson's Chi-Square Test: Core Properties

Pearson's chi-squared test (χ^2) is defined as "a statistical test applied to sets of categorical data to evaluate how likely it is that any observed difference between the sets arose by chance." Several properties are specified. It is "suitable for unpaired data from large samples." It is "the most widely used of many chi-squared tests."

The test evaluates a null hypothesis stating that "the frequency distribution of certain events observed in a sample is consistent with a particular theoretical distribution." The notes impose two critical conditions on the events considered: they "must be mutually exclusive and have total probability 1." A simple example provided is "the hypothesis that an ordinary six-sided die is 'fair', i.e., all six outcomes are equally likely to occur."

6. Two Types of Comparison: Goodness of Fit and Test of Independence

The lesson notes identify two distinct uses of Pearson's chi-square test.

Test of goodness of fit: This "establishes whether or not an observed frequency distribution differs from a theoretical distribution." The researcher has a specific expected distribution in mind and tests whether the observed data match that expectation.

Test of independence: This "assesses whether paired observations on two variables, expressed in a contingency table, are independent of each other." An example is provided: "polling responses from people of different nationalities to see if one's nationality is related to the response."

The notes elaborate on this distinction in Section 6.3: "With the goodness-of-fit test one is interested in determining whether a given distribution of data follows an expected pattern. For the test of association, however, one is interested in learning whether two (or more) categorical variables are related. Typically one will find two categorical variables depicted in a contingency table (a cross-tabulation of the frequencies for various combinations of the variables)."

7. The Three-Step Procedure

The lesson notes outline a three-step procedure for conducting a chi-square test.

Step 1: Calculate the chi-squared test statistic, χ^2. The statistic is described as "a normalized sum of squared deviations between observed and theoretical frequencies."

Step 2: Determine the degrees of freedom, df, of that statistic. The degrees of freedom is defined as "essentially the number of frequencies reduced by the number of parameters of the fitted distribution."

Step 3: Compare χ^2 to the critical value from the chi-squared distribution with df degrees of freedom. The notes state that "in many cases gives a good approximation of the distribution of χ^2."

8. Degrees of Freedom: Critical Distinctions

The lesson notes emphasize a critical distinction regarding degrees of freedom that is often misunderstood.

When testing whether observations are random variables "whose distribution belongs to a given family of distributions," the "theoretical frequencies" are calculated using a distribution from that family fitted in some standard way. "The reduction in the degrees of freedom is calculated as p = s + 1, where s is the number of covariates used in fitting the distribution."

Examples are provided. "When checking a three-covariate Weibull distribution, p = 4." "When checking a normal distribution (where the parameters are mean and standard deviation), p = 3." In other words, "there will be n - p degrees of freedom, where n is the number of categories."

The notes contain a crucial clarification: "It should be noted that the degrees of freedom are not based on the number of observations as with a Student's t or F-distribution." An example is given: "if testing for a fair, six-sided die, there would be five degrees of freedom because there are six categories/parameters (each number). The number of times the die is rolled will have absolutely no effect on the number of degrees of freedom."

9. Chi-Square as Proof of Association

The lesson notes discuss using chi-square analysis to determine "the statistical significance level of association rules." An association rule is defined as "a rule of the

form A → B (1) where A and B are item sets, that is, sets of items that appear in a database of transactions." This terminology originates from "market basket" analysis, where "each transaction item set represents the set of items that are purchased together in a single retail transaction."

The notes state: "We show that the chi-squared statistic of a rule may be computed directly from the values of confidence, support, and lift (interest) of the rule in question." This has three practical benefits: "facilitate pruning of rule sets obtained using standard association rule mining techniques, allow identification of statistically significant rules that may have been overlooked by the mining algorithm, and provide an analytical description of the relationship between confidence and support in terms of chi-squared and lift."

The approach involves viewing "the boolean product over each of these item sets as a single binary-valued random variable." This allows the researcher to work with "two-dimensional contingency tables regardless of the number of items that appear in a rule." The notes identify an advantage: "using lower-dimensional tables is that it becomes easier to achieve the minimum cell counts required for validity of chi-squared analysis."

10. Practical Example: Gender and Political Affiliation

The lesson notes provide a detailed worked example to illustrate chi-square calculation and interpretation.

Step 1: Begin with a hypothesis. "A common hypothesis in much research is that there is no correlation between the two variables of interest." The example examines data from "125 registered voters (65 women and 60 men)" and their political party affiliation (Democratic or Republican). "Suppose we know from previous research that 55 percent of voters identified themselves as Democrats. Our working hypothesis is that this 55 percent will be evenly distributed between men and women."

Step 2: Calculate expected values. "Based on 125 voters, we expect that 55 percent (69 voters) will identify themselves as Democrats." By gender: "we expect that 36 women and 33 men will express a preference for the Democratic Party, leaving 29 women and 27 men favoring the Republican Party." The data is organized in a 2×2 matrix: "party affiliation be the column variables and gender be the row variables."

Step 3: Compare actual values with expected values. "For this example, let's say that among the 65 women, 44 percent identified themselves as Democrats and 21 as Republicans, while 36 men claimed a Democratic affiliation and 24 preferred the Republican Party."

Step 4: Calculate the chi-square statistic. The statistic is "the sum of the squared differences between the observed and expected values (also known as the residuals), divided by the expected values." The researcher calculates this for "the four possible combinations of gender and political affiliation specified in your model." In the example, "the sum of squared differentials divided by expected values is 4.59."

Step 5: Determine statistical significance. To determine significance, the researcher needs "two things: the degrees of freedom and the significance level." Degrees of freedom is calculated as "the number of rows in your table minus one, times the number of columns minus one." For a 2×2 table, this equals (2-1) × (2-1) = 1 degree of freedom.

Significance level is defined as "the probability that the observed correlation could have occurred by chance alone." The notes state that "many researchers prefer a .05 significance level, meaning there is only a 5 percent likelihood that the observed relationship is pure chance."

The researcher then looks up "the chi-square value that corresponds to the significance level and degrees of freedom" in a statistics book appendix. "For our example, the chi-square value for 1 degree of freedom and .05 significance level is 3.84." The calculated value of 4.59 is greater than 3.84, meaning "there is a statistically significant relationship between gender and political party affiliation."

11. Summary of the Chi-Square Concept

The following table synthesizes the core concepts presented in the lesson notes.

Concept	Definition	Application
Chi-Square Test (χ^2)	Statistical hypothesis test where test statistic follows chi-square distribution under null hypothesis	Testing categorical data for goodness of fit or independence
Pearson's Chi-Square	Most widely used chi-square test; suitable for unpaired large samples	Goodness-of-fit; test of independence
Goodness-of-Fit	Determines whether observed frequency distribution differs from theoretical distribution	Testing if a die is fair; testing if population follows normal distribution
Test of Independence	Determines whether two categorical variables are related	Examining relationship between gender and political affiliation
Degrees of Freedom (df)	Number of frequencies reduced by number of parameters	For 2×2 table: df = 1; for six-sided die: df = 5 (NOT based on sample size)
Chi-Square Statistic	Sum of (observed - expected)^2 / expected across all cells	4.59 in gender-political affiliation example
Critical Value	Threshold from chi-square distribution at given df and α	3.84 for df = 1, α = 0.05
Decision Rule	If χ^2 > critical value → reject null hypothesis (relationship exists)	4.59 > 3.84 → reject null

Lesson 7: Non-Parametric Statistics

1. Definition and Foundational Context

Non-parametric statistics, as implied throughout the lesson notes, refers to statistical methods that do not assume a specific probability distribution for the population from which the sample is drawn. Unlike parametric tests that assume normality and require estimation of distribution parameters (mean, variance), non-parametric tests make fewer assumptions about the underlying data structure.

The lesson notes from Lesson 1 established the foundational distinction: "Parametric tests make certain assumptions about a data set; namely, that the data are drawn from a population with a specific (normal) distribution. Non-parametric tests make fewer assumptions about the data set." Lesson 7 provides the specific non-parametric tests that researchers use when the assumptions required for parametric testing cannot be satisfied.

Non-parametric methods are particularly valuable in three circumstances: (1) when the sample size is too small to assess normality reliably, (2) when the data are measured on ordinal rather than interval scales, and (3) when the population distribution is known to be non-normal (skewed, heavy-tailed, or multi-modal).

2. Anderson-Darling Test: Testing for Normality

The Anderson-Darling test is defined as "a statistical test of whether a given sample of data is drawn from a given probability distribution." In its basic form, "the test assumes that there are no parameters to be estimated in the distribution being tested, in which case the test and its set of critical values is distribution-free."

When applied to testing if a normal distribution adequately describes a set of data, "it is one of the most powerful statistical tools for detecting most departures from normality." The test is named after Theodore Wilbur Anderson (born 1918) and Donald A. Darling (born 1915), who invented it in 1952. Additionally, "K-sample Anderson-Darling tests are available for testing whether several collections of observations can be modeled as coming from a single population, where the distribution function does not have to be specified."

Interpretation of results: The test "rejects the hypothesis of normality when the p-value is less than or equal to 0.05." This means that "failing the normality test allows you to state with 95% confidence the data does not fit the normal distribution." Conversely, "passing the normality test only allows you to state no significant departure from

normality was found." The notes emphasize a critical distinction: passing the test does not prove normality; it only indicates insufficient evidence to reject it.

Serious flaw: The notes identify a significant limitation when applied to real-world data. "The Anderson-Darling test is severely affected by ties in the data due to poor precision. When a significant number of ties exist, the Anderson-Darling will frequently reject the data as nonnormal, regardless of how well the data fits the normal distribution." The notes provide an example: "data generated from the normal distribution but rounded to the nearest 0.5 to create ties." A tie is defined as "when identical values occurs more than once in the data set."

3. Cohen's Kappa Coefficient: Measuring Inter-Rater Agreement

Cohen's kappa coefficient is defined as "a statistical measure of inter-rater agreement or inter-annotator agreement for qualitative (categorical) items." It is "generally thought to be a more robust measure than simple percent agreement calculation since κ takes into account the agreement occurring by chance."

Critiques and limitations: The notes document several concerns. "Some researchers have expressed concern over κ's tendency to take the observed categories' frequencies as givens, which can have the effect of underestimating agreement for a category that is also commonly used; for this reason, κ is considered an overly conservative measure of agreement."

Others contest the assertion that kappa "takes into account" chance agreement. "To do this effectively would require an explicit model of how chance affects rater decisions. The so-called chance adjustment of kappa statistics supposes that, when not completely certain, raters simply guess—a very unrealistic scenario."

Factors affecting kappa magnitude: The notes identify multiple factors that influence kappa values beyond actual agreement. "Two important factors are prevalence (are the codes equiprobable or do their probabilities vary) and bias (are the marginal probabilities for the two observers similar or different)." Other things being equal, "kappas are higher when codes are equiprobable" but "kappas are higher when codes are distributed asymmetrically by the two observers." Another factor is "the number of codes. As number of codes increases, kappas become higher."

Based on a simulation study, Bakeman and colleagues concluded that "for fallible observers, values for kappa were lower when codes were fewer. And, in agreement with Sim & Wright's statement concerning prevalence, kappas were higher when codes were roughly equiprobable." The notes provide a concrete example: "given equiprobable

codes and observers who are 85% accurate, value of kappa are 0.49, 0.60, 0.66, and 0.69 when number of codes is 2, 3, 5, and 10, respectively."

Magnitude guidelines: The notes present two sets of guidelines, both characterized as arbitrary. Landis and Koch characterized values: "< 0 as indicating no agreement; 0-0.20 as slight; 0.21-0.40 as fair; 0.41-0.60 as moderate; 0.61-0.80 as substantial; and 0.81-1 as almost perfect agreement." However, "this set of guidelines is by no means universally accepted; Landis and Koch supplied no evidence to support it, basing it instead on personal opinion. It has been noted that these guidelines may be more harmful than helpful." Fleiss's guidelines characterize "kappas over 0.75 as excellent, 0.40 to 0.75 as fair to good, and below 0.40 as poor."

4. Friedman Test: Repeated Measures by Ranks

The Friedman test is defined as "a non-parametric statistical test developed by the U.S. economist Milton Friedman." It is "similar to the parametric repeated measures ANOVA" and "used to detect differences in treatments across multiple test attempts." The procedure involves "ranking each row (or block) together, then considering the values of ranks by columns." It is "applicable to complete block designs" and is "thus a special case of the Durbin test."

Classic examples of use:

- "n wine judges each rate k different wines. Are any wines ranked consistently higher or lower than the others?"
- "n wines are each rated by k different judges. Are the judges' ratings consistent with each other?"
- "n welders each use k welding torches, and the ensuing welds were rated on quality. Do any of the torches produce consistently better or worse welds?"

The Friedman test is described as "used for one-way repeated measures analysis of variance by ranks. In its use of ranks it is similar to the Kruskal-Wallis one-way analysis of variance by ranks." The test is "widely supported by many statistical software packages."

5. Kolmogorov-Smirnov Test: Comparing Distributions

The Kolmogorov-Smirnov test (K-S test) is defined as "a nonparametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one-sample K-S test), or to compare two samples (two-sample K-S test)."

Test statistic and interpretation: "The Kolmogorov-Smirnov statistic quantifies a distance between the empirical distribution function of the sample and the cumulative distribution function of the reference distribution, or between the empirical distribution functions of two samples." The null distribution is calculated "under the null hypothesis that the samples are drawn from the same distribution (in the two-sample case) or that the sample is drawn from the reference distribution (in the one-sample case)."

The notes emphasize that "the two-sample K-S test is one of the most useful and general nonparametric methods for comparing two samples, as it is sensitive to differences in both location and shape of the empirical cumulative distribution functions of the two samples."

Limitations for normality testing: "The Kolmogorov-Smirnov test can be modified to serve as a goodness of fit test. In the special case of testing for normality of the distribution, samples are standardized and compared with a standard normal distribution." However, "various studies have found that, even in this corrected form, the test is less powerful for testing normality than the Shapiro-Wilk test or Anderson-Darling test." The notes note that "other tests have their own disadvantages. For instance the Shapiro-Wilk test is known not to work well with many ties (many identical values)."

6. Kruskal-Wallis Test: Comparing Multiple Independent Samples

The Kruskal-Wallis one-way analysis of variance by ranks is defined as "a nonparametric method for testing whether samples originate from the same distribution." It is "used for comparing more than two samples that are independent, or not related." The parametric equivalent is "the one-way analysis of variance (ANOVA)."

Interpretation: "When the Kruskal-Wallis test leads to significant results, then at least one of the samples is different from the other samples." However, the notes emphasize a critical limitation: "The test does not identify where the differences occur or how many differences actually occur." It is described as "an extension of the Mann-Whitney U test to 3 or more groups." For identifying specific differences, "the Mann-Whitney would help analyze the specific sample pairs for significant differences."

Assumptions: "Since it is a non-parametric method, the Kruskal-Wallis test does not assume a normal distribution of the residuals, unlike the analogous one-way analysis of variance. However, the test does assume an identically shaped and scaled distribution for each group, except for any difference in medians."

7. Mann-Whitney U Test: Comparing Two Independent Samples

The Mann-Whitney U test (also called "the Mann-Whitney-Wilcoxon (MWW), Wilcoxon rank-sum test, or Wilcoxon-Mann-Whitney test") is defined as "a nonparametric test of the null hypothesis that two populations are the same against an alternative hypothesis, especially that a particular population tends to have larger values than the other."

Efficiency: The test "has greater efficiency than the t-test on non-normal distributions, such as a mixture of normal distributions, and it is nearly as efficient as the t-test on normal distributions." This means that when the normality assumption is violated, the Mann-Whitney U test performs better than the t-test. Even when normality holds, it performs almost as well.

Important distinction: The notes emphasize that "the Wilcoxon rank-sum test is not the same as the Wilcoxon signed-rank test, although both are nonparametric and involve summation of ranks."

8. Mood's Median Test: Testing Equality of Medians

Mood's median test is defined as "a special case of Pearson's chi-squared test." It is "a nonparametric test that tests the null hypothesis that the medians of the populations from which two or more samples are drawn are identical."

Procedure: "The data in each sample are assigned to two groups, one consisting of data whose values are higher than the median value in the two groups combined, and the other consisting of data whose values are at the median or below. A Pearson's chi-squared test is then used to determine whether the observed frequencies in each sample differ from expected frequencies derived from a distribution combining the two groups."

9. Spearman's Rank Correlation Coefficient: Monotonic Association

Spearman's rank correlation coefficient, "named after Charles Spearman and often denoted by the Greek letter ρ (rho) or as r_s," is defined as "a nonparametric measure of statistical dependence between two variables." It "assesses how well the relationship between two variables can be described using a monotonic function."

Interpretation: "If there are no repeated data values, a perfect Spearman correlation of +1 or -1 occurs when each of the variables is a perfect monotone function of the other." Unlike Pearson correlation which measures linear relationships, Spearman correlation

measures monotonic relationships (consistently increasing or consistently decreasing, not necessarily at a constant rate).

10. Wilcoxon Signed-Rank Test: Paired Differences

The Wilcoxon signed-rank test is defined as "a non-parametric statistical hypothesis test used when comparing two related samples, matched samples, or repeated measurements on a single sample to assess whether their population mean ranks differ (i.e. it is a paired difference test)."

Purpose: "It can be used as an alternative to the paired Student's t-test, t-test for matched pairs, or the t-test for dependent samples when the population cannot be assumed to be normally distributed."

Important distinction: The notes emphasize that "the Wilcoxon signed-rank test is not the same as the Wilcoxon rank-sum test, although both are nonparametric and involve summation of ranks." The test is "named for Frank Wilcoxon (1892-1965) who, in a single paper, proposed both it and the rank-sum test for two independent samples (Wilcoxon, 1945)." It was "popularized by Siegel (1956) in his influential text book on non-parametric statistics."

Assumptions: The test rests on three assumptions:

1. "Data are paired and come from the same population."
2. "Each pair is chosen randomly and independently."
3. "The data are measured at least on an ordinal scale, but need not be normal."

11. Summary Table of Non-Parametric Tests

The following table synthesizes the non-parametric tests presented in the lesson notes, their purposes, and their parametric equivalents.

Test	Purpose	Parametric Equivalent	Key Feature
Anderson-Darling	Test if data drawn from given distribution	Normality tests	Most powerful for detecting departures from normality; affected by ties

Test	Purpose	Parametric Equivalent	Key Feature
Cohen's Kappa	Measure inter-rater agreement for categorical items	None (unique)	Takes chance agreement into account; overly conservative
Friedman	Detect differences across multiple treatments (repeated measures)	Repeated measures ANOVA	Ranks rows together, considers column ranks
Kolmogorov-Smirnov	Compare sample distribution with reference or two samples	Goodness-of-fit tests	Sensitive to location and shape differences
Kruskal-Wallis	Compare three or more independent samples	One-way ANOVA	Does not identify where differences occur
Mann-Whitney U	Compare two independent samples	Independent t-test	Greater efficiency than t-test on non-normal distributions
Mood's Median	Test equality of medians across groups	One-way ANOVA (on medians)	Special case of Pearson's chi-square
Spearman's rho	Measure monotonic association between two variables	Pearson correlation	Uses ranks; detects monotonic (not just linear) relationships

Test	Purpose	Parametric Equivalent	Key Feature
Wilcoxon Signed-Rank	Compare two related/paired samples	Paired t-test	For non-normal paired data

Lesson 8: Shows Statistics

Sampling in Finite Populations

1.1 Definition of Population and Sample

In statistics, the lesson notes define a population as "a population or process to be studied." Populations can be diverse topics such as "all persons living in a country" or "every atom composing a crystal." The entire population, when compiled in its entirety, is called a census.

When a census is not feasible, "a chosen subset of the population called a sample is studied." The notes establish the fundamental reason for sampling: "Once a sample that is representative of the population is determined, data is collected for the sample members in an observational or experimental setting." The sample must be representative, meaning its characteristics should approximate those of the population from which it was drawn.

1.2 The Role of Descriptive and Inferential Statistics

The notes distinguish between two statistical methodologies applied to populations and samples. "Descriptive statistics can be used to summarize the population data. Numerical descriptors include mean and standard deviation for continuous data types (like income), while frequency and percentage are more useful in terms of describing categorical data (like race)."

However, the notes emphasize a critical point about samples: "the drawing of the sample has been subject to an element of randomness, hence the established numerical descriptors from the sample are also due to uncertainty. In order to still draw meaningful conclusions about the entire population, inferential statistics is needed. It uses patterns in the sample data to draw inferences about the population represented, accounting for randomness."

These inferences may take multiple forms: "answering yes/no questions about the data (hypothesis testing), estimating numerical characteristics of the data (estimation), describing associations within the data (correlation) and modeling relationships within the data (for example, using regression analysis)."

1.3 Parameter vs. Statistic

The lesson notes provide precise definitions. A parameter is defined as "a characteristic, feature, or measurable factor that can help in defining a particular system." More specifically, "a parameter is an important element to consider in evaluation or comprehension of an event, project, or situation."

In statistics and econometrics, "attention shifts to estimating the parameters of a distribution based on observed data, or testing hypotheses about them." The notes distinguish classical from Bayesian estimation: "In classical estimation these parameters are considered 'fixed but unknown', but in Bayesian estimation they are treated as random variables, and their uncertainty is described as a distribution."

A statistic is defined as "a numerical characteristic of a sample that can be used as an estimate of the corresponding parameter, the numerical characteristic of the population from which the sample was drawn." The notes provide a concrete example: "the sample mean (usually denoted $\bar{X}$) can be used as an estimate of the mean parameter (μ) of the population from which the sample was drawn."

1.4 Finite Populations vs. Infinite Populations

While the term "infinite population" does not explicitly appear in the provided pages, the concept is implicit in the distinction between population and process studies. The notes describe two types of studies. "With a population study, the analyst is interested in estimating or describing some characteristic of the population (inferential statistics)." This typically involves a finite population with a fixed size (e.g., all employees in a company, all loans processed in a year).

"With a process study, the analyst is interested in predicting a process characteristic or change over time." Process studies involve conceptually infinite populations because the process continues over time, generating an unlimited sequence of potential observations. The notes illustrate with the "I Love Lucy" television show's "Candy Factory" episode: "a population study, using samples, would seek to determine the average weight of the entire daily run of candies" (finite population with fixed daily production). "A process study would seek to know whether the weight was changing over the day" (infinite process continuing through time).

1.5 Non-Parametric vs. Parametric Approaches

The notes clarify when parametric assumptions about populations are unnecessary. "It is possible to make statistical inferences without assuming a particular parametric family of

probability distributions. In that case, one speaks of non-parametric statistics as opposed to the parametric statistics just described."

An example distinguishes the two approaches: "a test based on Spearman's rank correlation coefficient would be called non-parametric since the statistic is computed from the rank-order of the data disregarding their actual values (and thus regardless of the distribution they were sampled from), whereas those based on the Pearson product-moment correlation coefficient are parametric tests since it is computed directly from the data values and thus estimates the parameter known as the population correlation."

Sampling Error

2.1 Definition and Cause

Sampling error is defined as "incurred when the statistical characteristics of a population are estimated from a subset, or sample, of that population." The notes explain the fundamental cause: "Since the sample does not include all members of the population, statistics on the sample, such as means and quantiles, generally differ from parameters on the entire population."

The notes provide a concrete illustration: "if one measures the height of a thousand individuals from a country of one million, the average height of the thousand is typically not the same as the average height of all one million people in the country." The difference between the sample statistic and the population parameter is the sampling error.

2.2 The Measurement Problem

A critical limitation is noted: "Exact measurement of sampling error is generally not feasible since the true population values are unknown." If the true population parameter were known, there would be no need for sampling. Therefore, researchers face a paradox: sampling error cannot be directly measured because the benchmark against which it would be measured (the population parameter) is unknown.

However, the notes provide a solution: "sampling error can often be estimated by probabilistic modeling of the sample." Statistical theory provides methods to calculate standard errors and confidence intervals that quantify the expected magnitude of sampling error based on sample size and variability.

2.3 Relationship to Inferential Statistics

Sampling error is the reason inferential statistics exist. The notes state that "since the drawing of the sample has been subject to an element of randomness, hence the established numerical descriptors from the sample are also due to uncertainty. In order to still draw meaningful conclusions about the entire population, inferential statistics is needed."

The randomness inherent in sampling produces uncertainty. Inferential statistics provides the tools to quantify that uncertainty and to make probabilistic statements about the relationship between sample statistics and population parameters.

Sampling

3.1 Definition and Purpose

Sampling is defined as "the selection of a subset of individuals from within a statistical population to estimate characteristics of the whole population." The notes specify that "each observation measures one or more properties (such as weight, location, color) of observable bodies distinguished as independent objects or individuals."

Sampling is widely used across multiple disciplines. "In business and medical research, sampling is widely used for gathering information about a population." The notes also mention applications in "statistics, quality assurance, & survey methodology."

3.2 The Seven Stages of the Sampling Process

The lesson notes outline a comprehensive seven-stage sampling process:

1. **Defining the population of concern** — specifying exactly which group or process will be studied
2. **Specifying a sampling frame** — identifying "a set of items or events possible to measure"
3. **Specifying a sampling method** — selecting the approach for choosing items from the frame
4. **Determining the sample size** — deciding how many observations will be collected

5. **Implementing the sampling plan** — executing the predetermined selection procedure
6. **Sampling and data collecting** — gathering measurements from selected items
7. **Data which can be selected** — determining what data will be extracted from selected items

3.3 Probability vs. Non-Probability Sampling

The notes distinguish between two broad categories of sampling methods. Probability sampling, also called random sampling, is defined as "a sampling technique in which the probability of getting any particular sample may be calculated."

Nonprobability sampling is defined negatively: "does not meet this criterion and should be used with caution. Nonprobability sampling techniques cannot be used to infer from the sample to the general population."

The advantage of nonprobability sampling is noted: "its lower cost compared to probability sampling." However, the notes caution that "one can say much less on the basis of a nonprobability sample than on the basis of a probability sample." The notes raise a critical question about research practice: "many analysts draw generalizations (e.g., propose new theory, propose policy) from analyses of nonprobability sampled data. One must ask, however, whether those published works are publishable because tradition makes them so, or because there really are justifiable grounds for drawing generalizations from studies based on nonprobability samples."

Some researchers assert that "while probability methods are suitable for large-scale studies concerned with representativeness, non-probability approaches are more suitable for in-depth qualitative research in which the focus is often to understand complex social phenomena." The notes challenge this assertion: "how can one understand a complex social phenomenon by drawing only the most convenient expressions of that phenomenon into consideration?"

The notes conclude that there is "only one situation in which a non-probability sample can be appropriate—if one is interested only in the specific cases studied (for example, if one is interested in the Battle of Gettysburg), one does not need to draw a probability sample from similar cases."

Stratified Sampling Strategies

4.1 Definition and Basic Concept

Stratified sampling is defined as a method where "the population embraces a number of distinct categories, the frame can be organized by these categories into separate 'strata.' Each stratum is then sampled as an independent sub-population, out of which individual elements can be randomly selected."

The key insight is that dividing the population into homogeneous subgroups before sampling can improve efficiency and ensure representation of important subpopulations.

4.2 Four Potential Benefits of Stratified Sampling

The notes enumerate four distinct benefits.

First benefit: "dividing the population into distinct, independent strata can enable researchers to draw inferences about specific subgroups that may be lost in a more generalized random sample." Without stratification, a small but important subgroup might be underrepresented or entirely missed.

Second benefit: "utilizing a stratified sampling method can lead to more efficient statistical estimates (provided that strata are selected based upon relevance to the criterion in question, instead of availability of the samples)." Even if stratification does not increase efficiency, "such a tactic will not result in less efficiency than would simple random sampling, provided that each stratum is proportional to the group's size in the population."

Third benefit: "it is sometimes the case that data are more readily available for individual, pre-existing strata within a population than for the overall population; in such cases, using a stratified sampling approach may be more convenient than aggregating data across groups."

Fourth benefit: "since each stratum is treated as an independent population, different sampling approaches can be applied to different strata, potentially enabling researchers to use the approach best suited (or most cost-effective) for each identified subgroup within the population."

4.3 Potential Drawbacks

The notes identify three disadvantages. First, "identifying strata and implementing such an approach can increase the cost and complexity of sample selection, as well as leading to increased complexity of population estimates." Second, "when examining multiple criteria, stratifying variables may be related to some, but not to others, further complicating the design, and potentially reducing the utility of the strata." Third, "in some cases (such as designs with a large number of strata, or those with a specified minimum sample size per group), stratified sampling can potentially require a larger sample than would other methods (although in most cases, the required sample size would be no larger than would be required for simple random sampling)."

4.4 Three Conditions for Effectiveness

The notes specify three conditions under which stratified sampling is most effective:

1. "Variability within strata are minimized" — each stratum should be internally homogeneous
2. "Variability between strata are maximized" — strata should be distinctly different from each other
3. "The variables upon which the population is stratified are strongly correlated with the desired dependent variable" — the stratification factor must be relevant to what is being measured

4.5 Advantages Over Other Sampling Methods

The notes list four specific advantages:

1. "Focuses on important subpopulations and ignores irrelevant ones"
2. "Allows use of different sampling techniques for different subpopulations"
3. "Improves the accuracy/efficiency of estimation"
4. "Permits greater balancing of statistical power of tests of differences between strata by sampling equal numbers from strata varying widely in size"

4.6 Example of Stratified Random Sampling

The notes provide a concrete example. "The manager of a lending business wanted to estimate the average cycle time for a loan application process. She knows there are three types (strata) of loans (large, medium and small). Therefore, she wanted the sample to have the same proportion of large, medium and small loans as the population. She first separated the loan population data into three groups and then pulled a random sample from each group."

This example illustrates the key feature of stratified random sampling: the sample proportions mirror the population proportions, ensuring that each stratum is fairly represented.

4.7 Comparison with Other Sampling Strategies

The notes identify four primary sampling strategies. Stratified random sampling is distinct from simple random sampling (where each unit has equal probability without stratification), systematic sampling (taking samples according to a systematic rule such as every fourth unit), and rational subgrouping (grouping measurements produced under similar conditions to understand sources of variation).

Stratified random sampling shares with simple random sampling the benefit of randomness within strata but adds the efficiency gains from knowing the strata boundaries.

Summary Table: Sampling Strategies Comparison

Strategy	Primary Use	Key Feature	When Most Appropriate
Simple Random Sampling	Population study	Each unit has equal probability of selection	No prior knowledge about stratification factors; useful first step
Stratified Random Sampling	Population study	Population divided into homogeneous strata; independent random samples from each	Population has distinct subgroups that must be fairly represented
Systematic Sampling	Process study (real-time data collection)	Samples taken according to systematic rule (every fourth unit, etc.)	Data collected in real time during process operation; caution against bias from

Strategy	Primary Use	Key Feature	When Most Appropriate
			systematic rule matching underlying structure
Rational Subgrouping	Process study (real-time data collection)	Grouping measurements under similar conditions to understand short-term vs. long-term variation	Understanding sources of variation; minimizing special causes within subgroups

Lesson 9: The Determination of the Appropriate Size of a Sample Objective

1. Definition and Foundational Importance

Sample size determination is defined in the lesson notes as "the act of choosing the number of observations or replicates to include in a statistical sample." The sample size is described as "an important feature of any empirical study in which the goal is to make inferences about a population from a sample."

The notes specify two primary factors that determine the sample size used in a study: "the expense of data collection, and the need to have sufficient statistical power." These two factors often pull in opposite directions. Larger sample sizes increase statistical power but also increase data collection costs. Researchers must balance these competing demands.

The notes also clarify that sample size is not always a single number for an entire study. "In complicated studies there may be several different sample sizes involved in the study: for example, in a survey sampling involving stratified sampling there would be different sample sizes for each population." Similarly, "in experimental design, where a study may be divided into different treatment groups, there may be different sample sizes for each group."

In a census, the notes state that "data are collected on the entire population, hence the sample size is equal to the population size." However, censuses are often infeasible due to cost, time, or accessibility constraints, making sampling necessary.

2. Three Methods for Choosing Sample Sizes

The lesson notes identify three distinct approaches to selecting sample sizes.

Method 1: Expedience. This approach involves including "those items readily available or convenient to collect." While this method is practical, the notes issue a strong caution: "A choice of small sample sizes, though sometimes necessary, can result in wide confidence intervals or risks of errors in statistical hypothesis testing." Expedience should not be the primary driver of sample size decisions.

Method 2: Using a target variance for an estimate. The researcher determines the desired precision for the estimate that will be derived from the sample and selects a

sample size sufficient to achieve that precision. This approach focuses on the quality of the resulting estimate before data collection begins.

Method 3: Using a target for the power of a statistical test. The researcher determines the desired statistical power for a test to be applied once the sample is collected and selects a sample size sufficient to achieve that power. This approach is particularly relevant for hypothesis testing contexts.

3. The Effect of Sample Size on Precision and Accuracy

The notes establish a general principle: "Larger sample sizes generally lead to increased precision when estimating unknown parameters." A concrete example is provided: "if we wish to know the proportion of a certain species of fish that is infected with a pathogen, we would generally have a more accurate estimate of this proportion if we sampled and examined 200 rather than 100 fish."

Several fundamental mathematical principles explain this phenomenon: "the law of large numbers and the central limit theorem." The law of large numbers states that as sample size increases, the sample mean converges to the population mean. The central limit theorem states that the sampling distribution of the sample mean approaches normality as sample size increases, regardless of the shape of the population distribution.

However, the notes identify important exceptions where larger sample sizes do not improve accuracy. "This can result from the presence of systematic errors or strong dependence in the data, or if the data follow a heavy-tailed distribution." Systematic errors affect all measurements in the same direction regardless of sample size; increasing the number of observations does not eliminate bias. Strong dependence (e.g., autocorrelation) means that additional observations provide diminishing new information. Heavy-tailed distributions (e.g., power-law distributions) have infinite variance, violating the assumptions underlying standard sample size calculations.

4. Sample Size Judged by Quality of Resulting Estimates

Sample sizes are evaluated based on the quality of the estimates they produce. The notes provide two concrete examples.

Example for proportion estimation: "If a proportion is being estimated, one may wish to have the 95% confidence interval be less than 0.06 units wide." This specifies a desired precision before data collection. The researcher would calculate the sample size needed to achieve a margin of error of 0.03 (half of 0.06) given the desired confidence level.

Example for hypothesis testing: "If we are comparing the support for a certain political candidate among women with the support for that candidate among men, we may wish to have 80% power to detect a difference in the support levels of 0.04 units." This means the researcher wants an 80% probability of correctly detecting a true difference of 4 percentage points between the two groups.

5. Required Sample Sizes for Hypothesis Tests

The lesson notes describe a common problem faced by statisticians: "calculating the sample size required to yield a certain power for a test, given a predetermined Type I error rate α."

The notes identify three methods for estimating required sample sizes.

Method A: Using pre-determined tables. Tables can be used, for example, "in a two-sample t-test to estimate the sample sizes of an experimental group and a control group that are of equal size, that is, the total number of individuals in the trial is twice that of the number given, and the desired significance level is 0.05." The parameters used in such tables are:

- "The desired statistical power of the trial, shown in column to the left"
- "Cohen's d (=effect size), which is the expected difference between the means of the target values between the experimental group and the control group, divided by the expected standard deviation"

Cohen's d standardizes the effect size, making it comparable across studies with different measurement scales. A larger effect size requires a smaller sample size to achieve the same power; a smaller effect size requires a larger sample size.

Method B: Mead's resource equation. This method is "often used for estimating sample sizes of laboratory animals, as well as in many other laboratory experiments." The notes acknowledge its limitations: "It may not be as accurate as using other methods in estimating sample size, but gives a hint of what is the appropriate sample size where parameters such as expected standard deviations or expected differences in values between groups are unknown or very hard to estimate."

The notes specify an important detail: "All the parameters in the equation are in fact the degrees of freedom of the number of their concepts, and hence, their numbers are subtracted by 1 before insertion into the equation."

Method C: Using the cumulative distribution function. This more general approach requires understanding the distribution of the test statistic under both the null and alternative hypotheses.

6. Estimation Theory and Parameter Estimation

The notes introduce estimation theory as the broader context for understanding sample size. Estimation theory is defined as "a branch of statistics that deals with estimating the values of parameters based on measured/empirical data that has a random component." The parameters "describe an underlying physical setting in such a way that their value affects the distribution of the measured data." An estimator "attempts to approximate the unknown parameters using the measurements."

Two examples illustrate the concept. First, "it is desired to estimate the proportion of a population of voters who will vote for a particular candidate. That proportion is the parameter sought; the estimate is based on a small random sample of voters." Second, "in radar the goal is to estimate the range of objects (airplanes, boats, etc.) by analyzing the two-way transit timing of received echoes of transmitted pulses." Since the reflected pulses "are unavoidably embedded in electrical noise, their measured values are randomly distributed, so that the transit time must be estimated."

Two approaches to estimation are generally considered. "The probabilistic approach assumes that the measured data is random with probability distribution dependent on the parameters of interest." "The set-membership approach assumes that the measured data vector belongs to a set which depends on the parameter vector."

7. Estimate of a Proportion

The lesson notes provide a step-by-step procedure for estimating a proportion.

Definition: "The proportion of something is the number of observations that meet a certain criterion, divided by the total number of observations." For example, "the proportion of males in the population of Americans is the number of American males divided by the number of Americans." The population proportion "can rarely be calculated exactly, so it must be estimated."

Step 1: Get a random sample. "If your sample is not random, estimates of the proportion (and other quantities) may be biased." The example given: "if you want to estimate the proportion of boys in an elementary school, you could assign a number to each student, then randomly pick a sample by choosing random numbers." The notes emphasize that "the bigger your sample, the more accurate your estimate will be."

Step 2: Find the number of observations meeting the criterion. In the example, "we would find how many of the children in our sample were boys."

Step 3: Divide by the total number of observations. "This is the estimated proportion."

Step 4: Calculate the confidence interval. "The standard formula for a 95 percent confidence interval is p ± 1.96(pq/n)^0.5, where p is the proportion found in step 3, q = 1 - p, and n is the number of observations." This formula assumes the sample size is sufficiently large for the normal approximation to be valid (typically $np \geq 5$ and $n(1-p) \geq 5$).

8. Estimation of an Average: Point Estimates and Interval Estimates

The notes define estimation as "the process by which one makes inferences about a population, based on information obtained from a sample."

Two types of estimates are distinguished.

Point estimate: Defined as "a single value of a statistic." The notes provide examples: "the sample mean $\bar{x}$ is a point estimate of the population mean μ. Similarly, the sample proportion p is a point estimate of the population proportion P." Point estimates are simple to compute and communicate, but they provide no information about precision or uncertainty.

Interval estimate: Defined as "an interval estimate defined by two numbers, between which a population parameter is said to lie." For example, "$a < x < b$ is an interval estimate of the population mean μ. It indicates that the population mean is greater than a but less than b." Interval estimates convey both the best estimate and the range of plausible values.

9. Confidence Intervals

The lesson notes provide a comprehensive treatment of confidence intervals.

Definition and three parts: "Statisticians use a confidence interval to express the precision and uncertainty associated with a particular sampling method. A confidence interval consists of three parts: (1) A confidence level, (2) A statistic, (3) A margin of error."

Interpretation: "The confidence level describes the uncertainty of a sampling method. The statistic and the margin of error define an interval estimate that describes the

precision of the method. The interval estimate of a confidence interval is defined by the sample statistic ± margin of error."

Example: "Suppose we compute an interval estimate of a population parameter. We might describe this interval estimate as a 95% confidence interval. This means that if we used the same sampling method to select different samples and compute different interval estimates, the true population parameter would fall within a range defined by the sample statistic ± margin of error 95% of the time."

Advantage over point estimates: "Confidence intervals are preferred to point estimates, because confidence intervals indicate (a) the precision of the estimate and (b) the uncertainty of the estimate."

10. Confidence Level and Margin of Error

Confidence level: Defined as "the probability part of a confidence interval." It "describes the likelihood that a particular sampling method will produce a confidence interval that includes the true population parameter."

The notes provide a clear interpretation: "Suppose we collected all possible samples from a given population, and computed confidence intervals for each sample. Some confidence intervals would include the true population parameter; others would not. A 95% confidence level means that 95% of the intervals contain the true population parameter; a 90% confidence level means that 90% of the intervals contain the population parameter; and so on."

Margin of error: Defined as "the range of values above and below the sample statistic." The notes provide an example: "Suppose the local newspaper conducts an election survey and reports that the independent candidate will receive 30% of the vote. The newspaper states that the survey had a 5% margin of error and a confidence level of 95%. These findings result in the following confidence interval: We are 95% confident that the independent candidate will receive between 25% and 35% of the vote."

Critical caution: "Many public opinion surveys report interval estimates, but not confidence intervals. They provide the margin of error, but not the confidence level. To clearly interpret survey results you need to know both! We are much more likely to accept survey findings if the confidence level is high (say, 95%) than if it is low (say, 50%)." A 50% confidence interval would be narrower but would contain the true population parameter only half the time—essentially no better than a coin flip.

11. Comparison of Two Proportions

The notes briefly address the case where two proportions are being compared. This situation arises "when the variable is categorical (for example, smoker/nonsmoker, Democrat/Republican, support/oppose an opinion, and so on) and you're interested in the proportion of individuals with a certain characteristic — for example, the proportion of smokers."

The methodological requirement is specified: "two independent (separate) random samples need to be selected, one from each population." The null hypothesis H_0 is "that the two population proportions are the same; in other words, that their difference is equal to 0." The notation for the null hypothesis is "H_0: $p_1 = p_2$, where p_1 is the proportion from the first population, and p_2 is the proportion from the second population."

Summary Table: Sample Size Determination Concepts

Concept	Definition	Key Formula/Value	Notes
Sample Size (n)	Number of observations in sample	Varies by study	Determined by expense and power needs
Point Estimate	Single value of statistic estimating parameter	$\bar{x}$ for μ; p for P	No precision information
Interval Estimate	Range between two numbers where parameter is said to lie	$a < x < b$	Provides precision information
Confidence Level	Probability that sampling method produces interval containing true parameter	95% (standard), 90%, 99%	Higher level requires wider interval

Concept	Definition	Key Formula/Value	Notes
Margin of Error	Range above and below sample statistic	± 1.96(pq/n)^0.5 for proportion	Half the confidence interval width
Cohen's d	Standardized effect size	$(\mu_1 - \mu_2)/\sigma$	Larger effect → smaller needed n
Statistical Power	Probability of correctly rejecting false H_0	Typically 80%	Higher power requires larger n
Type I Error (α)	Probability of rejecting true H_0	Typically 0.05 or 0.01	Lower α requires larger n

Lesson 10: Sample Bias of Selection and Double-Blind

Sample Bias

1.1 Definition and Foundational Characteristics

Sampling bias, also referred to as sample bias, is defined in the lesson notes as "a bias in which a sample is collected in such a way that some members of the intended population are less likely to be included than others." This definition establishes the core problem: unequal probability of selection across population members.

The consequence of sampling bias is "a biased sample, a non-random sample of a population (or non-human factors) in which all individuals, or instances, were not equally likely to have been selected." The notes issue a critical warning: "If this is not accounted for, results can be erroneously attributed to the phenomenon under study rather than to the method of sampling." This means researchers may draw false conclusions about the real world when the real problem is simply how they selected their data.

Medical sources, the notes indicate, "sometimes refer to sampling bias as ascertainment bias." The notes state that "ascertainment bias has basically the same definition, but is still sometimes classified as a separate type of bias."

1.2 Types of Sampling Bias

The lesson notes identify multiple distinct types of sampling bias.

Selection from a specific real area. The notes provide an example: "a survey of high school students to measure teenage use of illegal drugs will be a biased sample because it does not include home-schooled students or dropouts." A sample is also biased "if certain members are underrepresented or overrepresented relative to others in the population." Another example: "a 'man on the street' interview which selects people who

walk by a certain location is going to have an overrepresentation of healthy individuals who are more likely to be out of the home than individuals with a chronic illness." The notes characterize this as "an extreme form of biased sampling, because certain members of the population are totally excluded from the sample (that is, they have zero probability of being selected)."

Self-selection bias. This occurs "whenever the group of people being studied has any form of control over whether to participate." The notes explain that "participants' decision to participate may be correlated with traits that affect the study, making the participants a non-representative sample." For example, "people who have strong opinions or substantial knowledge may be more willing to spend time answering a survey than those who do not." Another example is "online and phone-in polls, which are biased samples because the respondents are self-selected." The notes describe the mechanism: "Those individuals who are highly motivated to respond, typically individuals who have strong opinions, are overrepresented, and individuals that are indifferent or apathetic are less likely to respond. This often leads to a polarization of responses with extreme perspectives being given a disproportionate weight in the summary. As a result, these types of polls are regarded as unscientific."

Pre-screening of trial participants or advertising for volunteers within particular groups. The notes provide a pointed example: "a study to 'prove' that smoking does not affect fitness might recruit at the local fitness center, but advertise for smokers during the advanced aerobics class, and for non-smokers during the weight loss sessions." This design systematically biases the sample by recruiting smokers from a highly fit subgroup and non-smokers from a less fit subgroup.

Exclusion bias. This "results from exclusion of particular groups from the sample, e.g. exclusion of subjects who have recently migrated into the study area (this may occur when newcomers are not available in a register used to identify the source population)." The notes also note that "excluding subjects who move out of the study area during follow-up is rather equivalent of dropout or non response, a selection bias in that it rather affects the internal validity of the study."

Healthy user bias. This occurs when "the study population is likely healthier than the general population, e.g. workers (i.e. someone in ill-health is unlikely to have a job as manual laborer)."

Berkson's fallacy. This occurs when "the study population is selected from a hospital and so is less healthy than the general population." The notes explain the consequence: "This can result in a spurious negative correlation between diseases: a hospital patient

without diabetes is more likely to have another given disease such as cholecystis, since they must have had some reason to enter the hospital in the first place."

Overmatching. This involves "matching for an apparent confounder that actually is a result of the exposure. The control group becomes more similar to the cases in regard to exposure than the general population."

1.3 Symptom-Based Sampling

The notes discuss a specific form of bias in medical research. "The study of medical conditions begins with anecdotal reports. By their nature, such reports only include those referred for diagnosis and treatment." Examples are provided: "A child who can't function in school is more likely to be diagnosed with dyslexia than a child who struggles but passes. A child examined for one condition is more likely to be tested for and diagnosed with other conditions, skewing comorbidity statistics." Additionally, "as certain diagnoses become associated with behavior problems or intellectual disability, parents try to prevent their children from being stigmatized with those diagnoses, introducing further bias." The notes conclude that "studies carefully selected from whole populations are showing that many conditions are much more common and usually much milder than formerly believed."

1.4 The Caveman Effect: An Illustrative Example

The notes provide a memorable example of sampling bias called the "caveman effect." "Much of our understanding of prehistoric peoples comes from caves, such as cave paintings made nearly 40,000 years ago. If there had been contemporary paintings on trees, animal skins or hillsides, they would have been washed away long ago." Similarly, "evidence of fire pits, middens, burial sites, etc. are most likely to remain intact to the modern era in caves." The conclusion: "Prehistoric people are associated with caves because that is where the data still exists, not necessarily because most of them lived in caves for most of their lives." This example illustrates how the mechanism of preservation (not the actual behavior of prehistoric people) creates a biased record.

1.5 Problems Caused by Sampling Bias

The notes are explicit about the consequences: "A biased sample causes problems because any statistic computed from that sample has the potential to be consistently erroneous. The bias can lead to an over- or underrepresentation of the corresponding parameter in the population."

The notes acknowledge a practical reality: "Almost every sample in practice is biased because it is practically impossible to ensure a perfectly random sample." However, "if the degree of underrepresentation is small, the sample can be treated as a reasonable approximation to a random sample. Also, if the group that is underrepresented does not differ markedly from the other groups in the quantity being measured, then a random sample can still be a reasonable approximation."

The notes address the connotation of the word "bias." "The word bias has a strong negative connotation. Indeed, biases sometimes come from deliberate intent to mislead or other scientific fraud. In statistical usage, bias merely represents a mathematical property, no matter if it is deliberate or either unconscious or due to imperfections in the instruments used for observation. While some individuals might deliberately use a biased sample to produce misleading results, more often, a biased sample is just a reflection of the difficulty in obtaining a truly representative sample."

1.6 Statistical Corrections for a Biased Sample

The notes distinguish between uncorrectable and correctable bias. "If entire segments of the population are excluded from a sample, then there are no adjustments that can produce estimates that are representative of the entire population." However, "if some groups are underrepresented and the degree of underrepresentation can be quantified, then sample weights can correct the bias."

The notes provide an example: "The U.S. National Center for Health Statistics, for example, deliberately oversamples from minority populations in many of its nationwide surveys in order to gain sufficient precision for estimates within these groups. These surveys require the use of sample weights to produce proper estimates across all ethnic groups. Provided that certain conditions are met (chiefly that the sample is drawn randomly from the entire sample) these samples permit accurate estimation of population parameters."

1.7 Historical Examples

The lesson notes present two classic historical examples of sampling bias.

The 1936 Literary Digest poll. "In the early days of opinion polling, the American Literary Digest magazine collected over two million postal surveys and predicted that the Republican candidate in the U.S. presidential election, Alf Landon, would beat the incumbent president, Franklin Roosevelt, by a large margin. The result was the exact opposite." The explanation: "The Literary Digest survey represented a sample collected from readers of the magazine, supplemented by records of registered automobile

owners and telephone users. This sample included an over-representation of individuals who were rich, who, as a group, were more likely to vote for the Republican candidate." In contrast, "a poll of only 50 thousand citizens selected by George Gallup's organization successfully predicted the result, leading to the popularity of the Gallup poll."

The 1948 presidential election. "On election night, the Chicago Tribune printed the headline DEWEY DEFEATS TRUMAN, which turned out to be mistaken. In the morning the grinning president-elect, Harry S. Truman, was photographed holding a newspaper bearing this headline." The reason for the error: "their editor trusted the results of a phone survey. Survey research was then in its infancy, and few academics realized that a sample of telephone users was not representative of the general population. Telephones were not yet widespread, and those who had them tended to be prosperous and have stable addresses." Additionally, "the Gallup poll that the Tribune based its headline on was over two weeks old at the time of the printing."

Selection Bias

2.1 Definition and Core Concepts

Selection bias is defined as "a statistical bias in which there is an error in choosing the individuals or groups to take part in a scientific study." It is "sometimes referred to as the selection effect." The phrase "most often refers to the distortion of a statistical analysis, resulting from the method of collecting samples." The notes warn: "If the selection bias is not taken into account, then some conclusions of the study may not be accurate."

2.2 Distinction Between Sampling Bias and Selection Bias

The notes present a distinction, though they note it is "not a universally accepted one." Under this distinction: "sampling bias undermines the external validity of a test (the ability of its results to be generalized to the rest of the population), while selection bias mainly addresses internal validity for differences or similarities found in the sample at hand."

In this framework, "errors occurring in the process of gathering the sample or cohort cause sampling bias, while errors in any process thereafter cause selection bias." Examples of sampling bias include "self-selection, pre-screening of trial participants,

discounting trial subjects/tests that did not run to completion and migration bias by excluding subjects who have recently moved into or out of the study area."

2.3 Types of Selection Bias

The notes identify numerous specific types of selection bias.

Time interval biases.

- "Early termination of a trial at a time when its results support a desired conclusion."
- "A trial may be terminated early at an extreme value (often for ethical reasons), but the extreme value is likely to be reached by the variable with the largest variance, even if all variables have a similar mean."

Exposure biases.

- **Susceptibility bias:** "when one disease predisposes for a second disease, and the treatment for the first disease erroneously appears to predispose to the second disease." Example: "postmenopausal syndrome gives a higher likelihood of also developing endometrial cancer, so estrogens given for the postmenopausal syndrome may receive a higher than actual blame for causing endometrial cancer."
- **Protopathic bias:** "when a treatment for the first symptoms of a disease or other outcome appear to cause the outcome." This occurs "when there is a lag time from the first symptoms and start of treatment before actual diagnosis." It "can be mitigated by lagging, that is, exclusion of exposures that occurred in a certain time period before diagnosis."
- **Indication bias:** "a potential mix up between cause and effect when exposure is dependent on indication, e.g. a treatment is given to people in high risk of acquiring a disease, potentially causing a preponderance of treated people among those acquiring the disease. This may cause an erroneous appearance of the treatment being a cause of the disease."

Data biases.

- "Partitioning (dividing) data with knowledge of the contents of the partitions, and then analyzing them with tests designed for blindly chosen partitions."
- "Rejection of 'bad' data on arbitrary grounds, instead of according to previously stated or generally agreed criteria."
- "Rejection of 'outliers' on statistical grounds that fail to take into account important information that could be derived from 'wild' observations."

Study selection biases.

- "Selection of which studies to include in a meta-analysis."
- "Performing repeated experiments and reporting only the most favorable results, perhaps relabelling lab records of other experiments as 'calibration tests', 'instrumentation errors' or 'preliminary surveys'."
- "Presenting the most significant result of a data dredge as if it were a single experiment (which is logically the same as the previous item, but is seen as much less dishonest)."

Attrition bias. This "is a kind of selection bias caused by attrition (loss of participants), discounting trial subjects/tests that did not run to completion. It includes dropout, non response (lower response rate), withdrawal and protocol deviators." The bias occurs when attrition "is unequal in regard to exposure and/or outcome." Example: "in a test of a dieting program, the researcher may simply reject everyone who drops out of the trial, but most of those who drop out are those for whom it was not working. Different loss of subjects in intervention and comparison group may change the characteristics of these groups and outcomes irrespective of the studied intervention."

Observer selection. "Data is filtered not only by study design and measurement, but by the necessary precondition that there has to be someone doing a study. In situations where the existence of the observer or the study is correlated with the data observation selection effects occur, and anthropic reasoning is required."

The notes provide an example: "the past impact event record of Earth: if large impacts cause mass extinctions and ecological disruptions precluding the evolution of intelligent observers for long periods, no one will observe any evidence of large impacts in the recent past (since they would have prevented intelligent observers from evolving). Hence there is a potential bias in the impact record of Earth. Astronomical existential risks might similarly be underestimated due to selection bias, and an anthropic correction has to be introduced."

2.4 Avoidance and Correction of Selection Bias

The notes are sobering about the possibility of correction: "In the general case, selection biases cannot be overcome with statistical analysis of existing data alone, though Heckman correction may be used in special cases." An informal assessment "can be made by examining correlations between exogenous (background) variables and a treatment indicator."

The fundamental difficulty is explained: "in regression models, it is correlation between unobserved determinants of the outcome and unobserved determinants of selection

into the sample which bias estimates, and this correlation between unobservables cannot be directly assessed by the observed determinants of treatment."

2.5 Related Issues

The notes identify three issues closely related to selection bias.

Publication bias or reporting bias: "the distortion produced in community perception or meta-analyses by not publishing uninteresting (usually negative) results, or results which go against the experimenter's prejudices, a sponsor's interests, or community expectations."

Confirmation bias: "the distortion produced by experiments that are designed to seek confirmatory evidence instead of trying to disprove the hypothesis."

Exclusion bias: "results from applying different criteria to cases and controls in regards to participation eligibility for a study/different variables serving as basis for exclusion."

Double-Blind

3.1 Note on the Lesson Notes

The lesson notes provided for Lesson 10 do not contain a section on double-blind methodology. The file title includes the phrase "double-blind," and the document was presented with that title, but the actual content pages provided do not include any discussion of double-blind procedures. The notes cover sample bias (Section 10.1) and selection bias (Section 10.2) in detail, but the double-blind content appears to be either missing from the provided excerpt or referenced only in the title without corresponding text.

3.2 Standard Definition of Double-Blind (For Context)

Based on standard research methodology (not from the provided notes, as the notes do not contain this information), double-blind refers to an experimental procedure where neither the participants nor the experimenters know which participants belong to the control group and which belong to the treatment group. This design prevents bias from: (a) participant expectations affecting their responses (placebo effect), (b) experimenter

expectations affecting how they treat participants or interpret outcomes (experimenter bias), and (c) assessment bias in measuring outcomes.

Dimension	Sample Bias	Selection Bias
Primary focus	External validity (generalizability to population)	Internal validity (differences within sample)
When occurs	During process of gathering the sample or cohort	During any process after sample gathering
Core problem	Some population members less likely to be included	Error in choosing individuals or groups for study
Examples	Self-selection, pre-screening, exclusion, healthy user bias, Berkson's fallacy	Time interval bias, attrition bias, indication bias, protopathic bias, observer selection
Correctability	If underrepresentation quantified, sample weights can correct	Generally cannot be overcome with statistical analysis alone (except Heckman correction)

Since the lesson notes provided do not contain this material, the assignment cannot be completed on double-blind from the source alone. The notes cover sample bias and selection bias in extensive detail, as presented above, but the double-blind component is absent from the provided file content.

Summary Table: Sample Bias vs. Selection Bias