**Definitions and Terms to Know – Part 1: Random Sampling**

**Random** – An event is random if

1. we know what the possible outcomes of the event are and
2. we do not know which particular outcome will occur

**Distribution** – any random event has a distribution if

1. all possible outcomes of the event are known and
2. the probabilities associated with each of the outcomes is known (e.g. we know which outcomes (if any) are more likely than others)

**Population** – the entire group of individuals or instances about which we hope to learn

**Sample** – a representative subset of a population that is studied with the goal being to understand something about the population

**Sampling frame** – a list of individuals from whom the sample is drawn (e.g. phonebook)

**Census** – a sample that consists of the entire population

**(Population)** **parameter** – a numerically valued attribute of a model for a population; typically this is unknown but can be approximated with calculating an analogous value over a representative sample from the population

**(Sample)** **statistic** – any value calculated for sampled data; often times these values will correspond to a population parameter

|  |  |  |
| --- | --- | --- |
| **Name** | **Statistic** | **Parameter** |
| Mean |  | (mu, pronounced “mmm-you” |
| Standard deviation |  | (sigma) |
| Correlation |  | (rho, pronounced “row”) |
| Proportion |  | (pronounced “pee”) |

**Sample surveys**

Sample surveys are studies that ask questions of a sample drawn from some population with the goal being to understand something about the population

1. **Sample size** – often abbreviated as *n*, determines how well the sample represents the populations (note: this is different from the proportion of the size of the sample to the size of the population)
2. **Sampling bias** – any systematic failure of a sampling method to represent the population from which it is drawn
   1. **Undercoverage bias** – a sampling scheme that gives a part of the population less representation than it has in the population
   2. **Nonresponse bias** – when a large fraction of those sampled fail to respond then those who do respond are likely to not represent the entire sample
      1. **Voluntary response bias** – when individuals can choose on their own whether to participate in the sample; samples based on voluntary response are always invalid and cannot be recovered no matter how large the sample size. Is a form of nonresponse bias.
   3. **Response bias** – anything in a survey that influences responses (e.g. wording of questions)
   4. **Convenience sample** – a sample of only the individuals what are conveniently available; often these fail to be representative since every individual in the population is not equally convenient to sample
3. **Sampling methods** 
   1. **Simple random sample** – the gold standard of sampling from a population, a sample of size *n* is a simple random sample (SRS) from the population if every possible subset of *n* elements from the population had an equal chance of being selected
   2. **Stratified random sampling** – a sampling design in which the population is divided into several subpopulations (strata) and random samples are drawn from each of these subpopulations; want the samples to be homogeneous over each stratum
   3. **Cluster sample** – a sampling design in which entire groups (clusters) are chosen at random from a population; want each group to be sufficiently heterogeneous
   4. **Multistage sample** – sampling schemes that combine several sampling methods
   5. **Systematic sample** – a sample drawn by selecting individuals systematically from a sampling frame; can be representative of the population when there is no relationship between the order of the sampling frame and the variables of interest
4. **Sampling variability/error** – the natural tendency of randomly drawn samples to differ from one another. (the word “error” here is a bit of a misnomer, it is not a bad (or good) attribute, just the way data naturally varies)

|  |  |  |
| --- | --- | --- |
|  | **Random Assignment** | **No Random Assignment** |
| **Random Sampling** | Causal conclusion, generalized to the population | No causal conclusion (only correlation), generalized to the population |
| **No Random Sampling** | Causal conclusion for the sample only | No causal conclusion (only correlation), no generalization to the population |

**Observation studies and Experiments**

**Observational study** – a study based on data in which no manipulation of factors has been employed

1. **Retrospective study** – subjects are selected and then their previous conditions or behaviors are determined; not based on random samples; focus is on estimating differences between groups or associations between variables
2. **Prospective study** – subjects are followed to observe future outcomes; focus on estimating differences among groups that might appear as the groups are followed
3. **Lurking variable** – a variable that simultaneous affects two other variables of interest and can account for a strong correlation between these other two variables

**Experiment** – a study that manipulated different factor levels to create treatments, randomly assigns subjects to these different treatment levels, and then compares the responses of the subject groups across treatment levels

1. **Random assignment** – assigning experimental units to treatment groups at random
2. **Factor** – a variable whose levels are controlled by the experimenter; goal is to discover the effects that differences in factor levels may have on the response of the experimental units
3. **Response** – a variable whose values are compared across different treatments
4. **Experimental units** – individuals on whom an experiment is preformed (sometimes referred to as subjects or participants)
5. **Level** – the specific values that the experimenter chooses for a factor
   1. When levels of one factor are associated with levels of another factor making their effects inseparable, these two factors are said to be **confounded**
6. **Treatment** – the process, intervention, or other controlled circumstance applied to randomly assigned experimental units; are the different levels of a single factor or combinations of levels of two or more factors
   1. Any individual associated with an experiment who is not aware of how subjects have been allocated to treatment groups is said to be blinded
      1. **Single blind** – when either every individual who could influence the results or every individual who could evaluate the results of the experiment is blinded
      2. **Double blind** – when all individuals who could influence the results and/or evaluate the results of the experiment is blinded
7. Principles of experimental design
   1. **Control** aspects of the experiment that we know may have an effect on the response, but that are not the factors being studied
      1. **Control group** – the experimental units assigned to a baseline treatment level, typically either the default treatment, or a null, placebo treatment; their response provides the basis for comparison
      2. A treatment which is known to have no effect on the response variable of interest is called a **placebo**; the placebo effect is the tendency of many human subjects (often 20% or more of the experimental subjects) to show a response even when administered a placebo
   2. **Randomize** subjects to treatment to even out effects that we cannot control
   3. **Replicate** over as many subjects as possible; results for a single subject are just anecdotes; if sample is not representative of the population of interest, replicate the entire study with a different group of subjects from another part of the population
   4. **Block** to reduce the effects of identifiable attributes of the subjects that cannot be controlled
      1. Blocking sections off groups of similar experimental units; this helps isolate the variability attributable to the differences between the blocks to see the differences caused by the treatments more clearly
8. Experimental designs
   1. **Randomized block design** – the randomization occurs only within blocks
   2. **Completely randomized design** – all experimental units have an equal chance of receiving any treatment