# Statistics

## Lecture 05: |Independence – Types of Data

Prof. Dr. Mostafa Elhosseini

https://youtube.com/drmelhosseini

# Independence

- The **items in a sample are said to be independent** if knowing the values of some of them does not help to predict the values of the others.
- With a **finite**, **tangible population**, the items in a simple random sample **are not strictly independent**, because as each **item is drawn**, the **population changes**.
- This change can be **substantial** when the **population is small**.
- However, when the **population is very large**, this **change is negligible,** and the **items can be treated as if they were independent**.

# Independence

- To illustrate this idea, imagine that we **draw a simple random sample of 2 items** from the population
- **For the first draw**, the numbers **0 and 1 are equally likely**.
- But the **value of the second item** is clearly **influenced** by the first; if the first is 0, the second is more likely to be 1, and vice versa. Thus the **sampled items are dependent**

# Independence

- Again on the **first draw**, the **numbers 0 and 1 are equally likely**. But unlike the previous example, these **two values remain almost equally likely the second draw as well**, no matter what happens on the first draw.
- With the **large population**, the sample items are for all practical purposes **independent**.
- It is **reasonable to wonder how large a population must be** in order that the **items** in a simple random sample may be treated as **independent**.
- **A rule of thumb** is that **when sampling from a finite population**, the items may be treated as **independent** so long as the sample **contains 5% or less of the population**

# Independence

- Interestingly, it is **possible to make a population behave** as though it were **infinitely large**, by **replacing** each item after it is sampled.
- This method is called **sampling with replacement**. With this method, the **population is exactly the same on every draw** and the sampled items are **truly independent**

# Independence

- With a **conceptual population**, we require that the sample items be produced **under identical experimental conditions**.
- In particular, then, **no sample value may influence the conditions** under which the others are produced. Therefore, the items in a **simple random sample** from a **conceptual population** may be treated as **independent**.
- We may **think** of a **conceptual population** as being **infinite**, or equivalently, that the items **are sampled with replacement**.

# To Conclude…

- The items in a sample are independent if knowing the values of some of the items does not help to predict the values of the others.
- Items in a simple random sample may be treated as independent in many cases encountered in practice.
- The exception occurs when the population is finite and the sample consists of a substantial fraction (more than 5%) of the population.

# Other Sampling Methods

- In addition to **simple random sampling** there are other sampling methods that are useful in various situations.
- In **weighted sampling**, **some items are given a greater chance** of being **selected** than others, **like a lottery** in which some people **have more tickets than others**
- In **stratified random sampling**, the population is divided up into **subpopulations**, called **strata**, and a simple random sample is drawn from each stratum. In **cluster sampling**, items are drawn from the population in groups, or clusters.

# Other Sampling Methods

- **Cluster sampling** is useful when the population is too large and spread out for simple random sampling to be feasible
- many U.S. government agencies use cluster sampling to sample the U.S. population to measure sociological factors such as income and unemployment
- Simple random sampling is not the only valid method of random sampling. But it is the most fundamental, and we will focus most of our attention on this method.
- From now on, unless otherwise stated, the terms "**sample**" and "**random sample**" will be taken to mean "**simple random sample**."

# Types of Experiments

- There are many types of experiments that can be used to generate data
- In a one-sample experiment, there is only one population of interest, and a single sample is drawn from it.
- One-sample experiments can be used to determine whether a process meets a certain standard
- In a multisample experiment, there are **two or more populations** of interest, and a **sample is drawn from each population**
- The usual purpose of multisample experiments is to **make comparisons among populations**

# Types of Experiments

- In many multisample experiments, the populations are distinguished from one another by the varying of one or more factors that may affect the outcome.
- Such experiments are called **factorial experiments**
- In a **factorial experiment**, each combination of the factors for which data are collected defines a population, and a simple random sample is drawn from each population.
- The **purpose of a factorial experiment** is to determine **how varying the levels of the factors affects the outcome being measured**.

# Types of Data

- When a **numerical quantity** designating how much or how many is assigned to each item in a sample, the resulting set of values is called **numerical or quantitative**.
- In some cases, sample items are placed into **categories**, and category names are assigned to the sample items. Then the data are **categorical or qualitative**

# Numerical / categorical?

- The **torques**, in the middle column, are **numerical** data.
- The **failure locations**, in the rightmost column, are **categorical** data.

| Specimen | Torque (kN · m) | Failure Location |
|---|---|---|
| 1 | 165 | Weld |
| 2 | 237 | Beam |
| 3 | 222 | Beam |
| 4 | 255 | Beam |
| 5 | 194 | Weld |

# Controlled Experiments and Observational Studies

- Many scientific **experiments are designed** to determine the **effect of changing one or more factors on the value of a response**.
  - For example, suppose that a chemical engineer wants to determine how the concentrations of reagent and catalyst affect the yield of a process.
  - The engineer can run the process several times, changing the concentrations each time, and compare the yields that result.
- This sort of experiment is called a **controlled experiment**, because the values of the factors, in this case the concentrations of reagent and catalyst, are **under the control of the experimenter**

# Controlled Experiments and Observational Studies

- There are many situations in which scientists **cannot control the levels of the factors**.
  - To determine the effect of cigarette smoking on the risk of lung cancer
- The **experimenters cannot control who smokes and who doesn't**; people cannot be required to smoke just to make a statistician's job easier.
- This kind of study is called an **observational study**, because the experimenter **simply observes the levels** of the factor as they are, **without having any control** over them