STA 131 B: Mathematical Statistics **Spring 2024**

Lecture 1: Introduction to Statistics

Lecturer: Hang Zhou

What is statistics?

Statistics is the art and science of gathering, modeling and making inference from data.

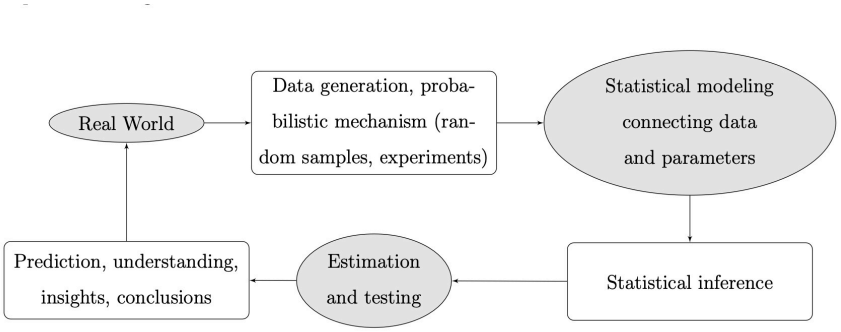

Figure 1.1: what is statistics?

Example 1.1 (coin toss) *A coin has two sides: heads (H) and tails (T). If you toss a coin n times and record each result as $X_i = 0$ for tails (T) and $X_i = 1$ for heads (H), can these recordings $X_1, \ldots, X_n$ tell us whether the coin is fair?*

In this example, we typically assume that the outcome of one toss does not affect the outcome of another; that is, each toss is considered statistically independent of the others. Then, we consider a generic version of X_i, denoted by X. X can be regarded as a Bernoulli random variable with a success probability of θ, that is, $P(X = 1) = \theta$. Thus, our question becomes: Is $\theta = 0.5$?

In this course, we study cases where a (probability) distribution is known except for a parameter θ. Such statistical models are called parametric statistical models. When we talk about a statistical model in this course, it always refers to the parametric statistical model. Note that the parameter θ can be multi-dimensional. An important example is the two-dimensional parameter vector $\theta = (\mu, \sigma)$, where the distribution is $\mathcal{N}(\mu, \sigma^2)$, also referred to as the normal or Gaussian distribution, characterized by its mean μ and variance σ^2.

Following is the formal (yet not overly formal) definition of a statistical model:

Definition 1.1 *A statistical model is a specification of the distribution of observed data, which depends on a parameter $\theta \in \Theta$, where $\Theta \subseteq \mathbb{R}^p$. Here Θ is the parameter space, the set of all possible values of the parameter. We write $\{f(x|\theta), \theta \in \Theta\}$ for the model, where f is the pmf/pdf with known form, and θ is an unknown parameter.*

In this course, our default assumption is that θ is fixed and non-random (frequentist statistics). Sometimes

θ will be considered as random (Bayesian statistics); whenever we adopt the Bayesian point of view, we will make it very clear that we deal with random parameters.

Example 1.2 (coin toss) *In the coin toss example, we have the random samples* $X_1, \ldots, X_n$ *i.i.d. the* Bernoulli(p)*, where* Bernoulli(p) *is the same as* binomial$(1, p)$*, which we abbreviate as* $\mathcal{B}(1, p)$ *in the following. Here the parameter is* $\theta = p$ *and the sample space* $\Theta = [0, 1]$*.*

What is statistics inference?

Statistical inference is the process of using data analysis to infer properties of an underlying distribution of probability. In the case of a parametric statistical model, this is equivalent to inferring the parameter θ.

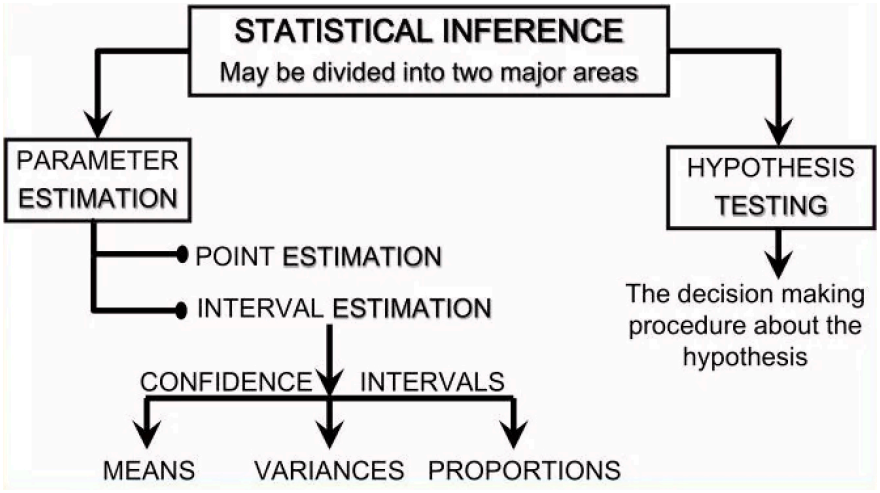

Figure 1.2: what is statistical inference?

Example 1.3 (point estimation) *Suppose we want to estimate the average length of newborn babies in a city. We take a random sample of 50 babies and measure their lengths. The sample mean length is found to be 49.5 cm. In this case, the sample mean of 49.5 cm serves as a point estimate for the population mean length of newborn babies in the city.*

Example 1.4 (confidence interval) *Continuing with the length estimation, it's often insufficient to rely solely on a point estimate of the mean length. Instead, we seek to determine an interval within which the true mean length is likely to lie, with a specified level of confidencecommonly 95%. This interval is known as a confidence interval. It is constructed so that, if we were to repeat the sample and calculation numerous times, we expect the true mean to fall within this interval in 95% of the cases.*

Example 1.5 (hypothesis testing) *A beverage company claims that its new sports drink contains, on average, 20 grams of sugar per bottle. A nutrition agency believes this is not accurate and decides to test this claim. They collect a random sample of 40 bottles and find that the mean sugar content is 21.5 grams with a standard deviation of 1.5 grams. To test the company's claim, they set up the following hypotheses:*

- *Null hypothesis* (H_0)*: The mean sugar content is 20 grams* ($\mu = 20$)*.*
- *Alternative hypothesis* (H_A)*: The mean sugar content is not 20 grams* ($\mu \neq 20$)*.*

Hypothesis testing is all about figuring out whether there's strong enough evidence from the data to reject the null hypothesis (H_0) and accept the alternative hypothesis (H_1). (think about why reject H_0, not accept?) If the data is very unlikely under the assumption that H0 is true, we conclude that we have enough evidence to reject H0 and accept H1. However, if the data is what we would expect if H0 is true, we do not reject H0. It's important to note that not rejecting H0 does not prove it's true, it just means we haven't found strong evidence against it.

Population, Sample and Statistic:

The population is the entire group from which you want to draw conclusions. It encompasses all possible observations or outcomes of interest.

A sample is a subset of the population chosen for the actual study. The objective of sampling is to infer properties of the population through the data gathered from the sample.

In this course, the terms 'sample' and 'population' are defined as follows:

Definition 1.2 *Let $X_1, \ldots, X_n \sim_{i.i.d.} f(x \mid \theta), \theta \in \Theta$. Then $\{X_1, \ldots, X_n\}$ is called a random sample (or just sample for notation simplicity) from $f(x \mid \theta)$. The population is the generic version of X_i represented by X. Here $\sim$ is shorthand for "distributed as" and i.i.d. for "independently and identically distributed".*

Statistic is a function mapping from the sample space to $\mathbb{R}^1$.

Definition 1.3 *Let $X_1, \ldots, X_n$ be n random samples and $r : \mathbb{R}^n \to \mathbb{R}^1$ a function. Then a r.v. $T = r(X_1, \ldots, X_n)$ is called a statistic. Here, r.v. stands for random variable, i.e., a manifestation of a random mechanism that assigns a value to an outcome randomly.*

Example 1.6 (Examples of Statistic) *$X_1, \ldots, X_n$ are random samples, the following are all statistic*

- $T = r(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^n X_i = \bar{X}$ *(sample mean)*
- $T = r(X_1, \ldots, X_n) = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2 = \bar{X}$ *(sample variance)*
- *Reorder the sample $X_1, \ldots, X_n$ from the smallest to the largest value, you get an ordered sample: $X_{(1)} \leq X_{(2)} \leq \ldots \leq X_{(n)}$, where $X_{(1)}$ is the smallest value (minimum), $X_{(n)}$ is the largest value (maximum), and $X_{(k)}$ represents the k-th smallest value in the sample. $T = X_{(k)}$ is the order statistic.*
- $T = r(X_1, \ldots, X_n) = \arg\min_{x \in \mathbb{R}} \sum_{i=1}^n |X_i - x|$ *(sample medium)*