Estimation Theory: The Big Picture

A **population** is a group of people who share a set of particular characteristics.

In research, we define the **population** as the group of people about whom we would like to make scientific inference.

A **random sample** is a subset of a population.

It is often infeasible to collect data on the entire population of interest.

In research studies, we collect data from a **random sample** of the population of interest in the hopes of understanding the entire population.

We assume that individuals in a random sample are **independent**.

Researchers are often interested in understanding the nature of a particular **event** that occurs in each member of a population.

In health research, examples of **events** of interest include:

- A genetic mutation observed in some members of the population but not in others.
- The age of onset of a particular mental health condition, which varies from person to person.
- The unique change in blood pressure that occurs, in each member of a population, after administration of an experimental blood pressure medication.
- The diagnosis of cancer, or lack thereof, in a population of individuals who carry a particular genetic mutation.

As scientists, we collect data from a subset of the population in order to gain insight into the range of events that occur in the entire population.

Statisticians use specific terminology to describe study designs, hypotheses, and data analyses.

Specifically, study designs, hypotheses, and data analyses can be described with precision by using the following statistical terms:

- Random variable
- Parameter
- Probability density function
- Likelihood function
- Distribution
- Estimator
- Estimate.

A **random variable** symbolically stores information about all of the possible **observable events** that might occur, at random, if we were able to study the entire population. X

A **parameter** is an **unobservable, defining characteristic** of the **events** that occur in the entire population.

A **probability density function** is an algebraic function of a particular, recognizable form.

Common forms of probability density functions include the Bernoulli, the Binomial, the Normal, the Poisson, the Beta, and the Gamma.

Example: The general form of the Bernoulli probability density function is:

Probability Density Function	**Interpretation**
$f_X(x\|p) = p^x(1-p)^{1-x}.$	**Summary:** This is a Bernoulli probability density function for as of yet unobserved data, with an unknown population proportion. Hypothesized distribution: $Bernoulli(p)$ Population parameter for the proportion: p "True" value of the population parameter: Unknown Observed values of the random variable: To be determined

Specific cases of the Bernoulli probability density function could be any of the following:

Probability Density Function for Observed Data	**Interpretation**
$f_X(x = 1\|p = 0.4) = 0.4^1(1-0.4)^{1-1}$	Random variable: X Number of observations: 1 Population parameter: p Hypothesized value of the population parameter: p=0.4 Observed value of the random variable: x=1 Calculated value of the pdf: $f_X(x = 1\|p = 0.4) = 0.4$
$f_Z(z = 0\|p = 0.4) = 0.4^0(1-0.4)^{1-0}$	Random variable: Z Number of observations: 1 Population parameter: p Hypothesized value of the population parameter: p=0.4 Observed value of the random variable: z=1 Calculated value of the pdf: $f_Z(z = 0\|p = 0.4) = 0.6$
$f_Y(y = 1\|p = 0.9) = 0.9^1(1-0.9)^{1-1}$	Random variable: Y Number of observations: 1 Population parameter: p Hypothesized value of the population parameter: p=0.9 Observed value of the random variable: y=1 Calculated value of the pdf: $f_Y(y = 1\|p = 0.9) = 0.9$

Example: The general form of the Normal probability density function is:

Probability Density Function	**Interpretation**
$f_X(x\|\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$	**Summary:** This is a Normal probability density function for as of yet unobserved data, with an unknown population mean and unknown population variance. Hypothesized distribution: $Normal(\mu,\sigma^2)$ Population parameter for the mean: μ Population parameter for the variance: σ^2 "True" value of the population parameter for the mean: Unknown "True" value of the population parameter for the variance: Unknown Observed values of the random variable: To be determined

Specific cases of the Normal probability density function could be any of the following:

Probability Density Function for Observed Data (i.i.d. observations) (n=3)	**Interpretation**
Marginal distribution for participant #1. $f_X(x_1 = 3\|\mu = 5,\sigma^2 = 1) = \frac{1}{1\sqrt{2\pi}} e^{-\frac{1}{2*1}(3-5)^2}$	Marginal or joint pdf? Marginal Random variable: X Number of observations: 1 Observation number: 1 Hypothesized value of the population parameter for the mean: $\mu = 5.$ Hypothesized value of the population parameter for the variance: $\sigma^2 = 1.$ Observed value of the random variable: $x_1 = 3$ Calculated value of the pdf: $f_X(x = 1\|\mu = 5,\sigma^2 = 1) =$
Marginal distribution for participant #2. $f_X(x_2 = 8\|\mu = 5,\sigma^2 = 1) = \frac{1}{1\sqrt{2\pi}} e^{-\frac{1}{2*1}(8-5)^2}$	Marginal or joint pdf? Marginal Random variable: X Number of observations: 1 Observation number: 2 Hypothesized value of the population parameter for the mean: $\mu = 5.$ Hypothesized value of the population parameter for the variance: $\sigma^2 = 1.$

	Observed value of the random variable: $x_2 = 8$ Calculated value of the pdf: $f_X(x = 8 \mid \mu = 5, \sigma^2 = 1) =$
Marginal distribution for participant #3. $f_X(x_3 = 5 \mid \mu = 5, \sigma^2 = 1) = \frac{1}{1\sqrt{2\pi}} e^{-\frac{1}{2*1}(5-5)^2}$	Marginal or joint pdf? Marginal Random variable: X Number of observations: 1 Observation number: 2 Hypothesized value of the population parameter for the mean: $\mu = 5.$ Hypothesized value of the population parameter for the variance: $\sigma^2 = 1.$ Observed value of the random variable: x=8 Calculated value of the pdf: $f_X(x = 5 \mid \mu = 5, \sigma^2 = 1) =$
JOINT distribution for participants 1, 2, and 3 **= the LIKELIHOOD function for participants 1, 2, and 3** $f_X(\boldsymbol{x} = 3, 8, 5 \mid \mu = 5, \sigma^2 = 1) = \frac{1}{1\sqrt{2\pi}} e^{-\frac{1}{2*1}(3-5)^2} * \frac{1}{1\sqrt{2\pi}} e^{-\frac{1}{2*1}(8-5)^2} * \frac{1}{1\sqrt{2\pi}} e^{-\frac{1}{2*1}(5-5)^2}$	Marginal or joint pdf? Joint Random variable: X Number of observations: 3 Observation number(s): 1, 2, 3 Hypothesized value of the population parameter for the mean: $\mu = 5.$ Hypothesized value of the population parameter for the variance: $\sigma^2 = 1.$ Observed value of the random variable: x=8 Calculated value of the joint pdf: $f_X(\boldsymbol{x} = 3, 8, 5 \mid \mu = 5, \sigma^2 = 1) =$

The right-hand side of a probability density function contains "inputs."

The "inputs" of the probability density function are **random variables** and **parameters**.

Each combination of random variables and parameters produces a specific numerical value of $f_X(x|\theta)$, the pdf.

The left-hand side of a probability density function, the "output," are points along the probability density function.

Combining many points produces a distribution.

We collect sample data to estimate population parameters.

We conduct research studies on a sample of the population, we assume that the scope of events that could conceivably occur in the entire population is the **same** as the scope of events that we will observe in our sample.

Once you have collected data, you can test hypotheses about the population parameter(s).

We conduct hypothesis tests to assess if our assumptions about the population parameter(s) are correct.

We use **estimatORS** to generate **estimATES**.

An **estimator** is a general algebraic equation composed of observations of the random variable.

When we collect data, we plug the values of our observations into the equation for our **estimator** to calculate an **estimate**.

In any hypothesis test, we evaluate whether the hypothesized value of the population **parameter** matches the value of our estimate.

You need a good estimator to conduct a good hypothesis test.

Examples of common estimators are:

Estimator of the population mean, $\mu = \hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Estimator of population proportion: $\hat{p} = \frac{x}{n}$