Lecture 1 – Introduction to Statistical Inference

Last Modified: September 1, 2026

Outline:

1.1 Motivation

How do we learn from data? How do we quantify the uncertainty in our conclusions?

To appreciate the problem, consider the paper *Adult Persistence of Head-Turning Asymmetry* (*Nature*, 2003). The author wanted to know whether kissing couples have a preference to turn their heads to the right. He observed $n = 124$ couples kissing and found that

$$80 \text{ turned right}, \qquad 44 \text{ turned left}.$$

Summarizing the data

The first question we will look at is how to efficiently summarize the data. Clearly, the following way is inefficient:

$$\begin{array}{cc} \text{Jack and Annie} & R, \\ \text{Bob and Ricky} & R, \\ \text{Mandy and Morris} & L, \\ \vdots & \vdots \end{array}$$

Intuitively (and we will make this rigorous!), all the relevant information is captured in the observed proportion who turned right:

$$\frac{80}{124} \approx 0.645,$$

or approximately 64.5%.

Two inferential questions

We can answer questions about the *population* using our *sample* of 124 couples. For example, we might ask:

Question 1. Do humans prefer to turn their heads to the right when they kiss?

This question asks for a yes-or-no conclusion. A simple rule might be to answer "yes" whenever the observed proportion is greater than 50%. For the present sample,

$$\frac{80}{124} \approx 0.645 > 0.5,$$

so we conclude "yes." If instead we had observed two right turns among only three couples, then

$$\frac{2}{3} \approx 0.667 > 0.5,$$

and we would again conclude "yes."

Question 2. What proportion p of all humans turn right when kissing?

This question asks for a numerical value. Our best guess based only on the observed sample is

$$\widehat{p} = \frac{80}{124} \approx 0.645.$$

If we had observed two right turns among three couples, our analogous guess would be $\widehat{p} = 2/3$.

Just giving a yes/no answer, or a single best guess for a numerical value is easy. In this course, we will go further and additionally specify our uncertainty: how reliable is our yes-or-no conclusion, and how much uncertainty is associated with the numerical estimate. Clearly, our answers based on 124 couples are more precise than based on 3, and we will make this quantitative.

Statistics provides the formal machinery to do this.

1.2 Modeling Principles

Principle #1: the data are realizations of a random process.
We assume the observations are independent and identically distributed (i.i.d.) random variables:

$$X_i \overset{\text{iid}}{\sim} f, \qquad i = 1, \ldots, n,$$

where f is a probability mass function (pmf) in the discrete case or a probability density function (pdf) in the continuous case.

Definition 1 (Random sample)**.** A *random sample* from f is the random vector

$$\underline{X} = (X_1, \ldots, X_n)$$

whose components are i.i.d. from f. Its joint pmf or pdf is therefore

$$f_{\underline{X}}(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i). \tag{1.1}$$

The vector

$$\underline{x} = (x_1, \ldots, x_n)$$

that we actually observe is called a *realization* of the random sample.

In the kissing-couples example, define

$$X_i = \begin{cases} 1, & \text{if couple } i \text{ turns right,} \\ 0, & \text{if couple } i \text{ turns left.} \end{cases}$$

Then an observed data set could be

$$\underline{x} = (1, 1, 0, 0, 1, 1, 0, 1, 0, 1, \ldots).$$

Each X_i represents the outcome for a randomly drawn couple, and x_i is the outcome that was actually observed.

Remark 1. The i.i.d. assumption is a modeling choice that should not be taken for granted. It says both that the couples are independent and that they follow a common distribution. Whether this is a reasonable approximation depends on how the couples were sampled and on the population to which we want to generalize.

Principle #2: the distribution f is unknown but belongs to a specified family.

Assuming f lies in a specified family simplifies the inference problem and is usually necessary to draw reliable conclusions from the data. Sometimes, the family is automatically specified based on the type of data we have. In the kissing-couples example, $X_i \in \{0, 1\}$, so in this case, the only choice for the distribution of the X_i is Bernoulli(p). The distribution f being unknown means that p is unknown; it is some value in $[0, 1]$. We write the family of possible f's as

$$\big\{\text{Bernoulli}(p) : p \in [0, 1]\big\}.$$

The following definition formalizes this concept.

Definition 2 (Statistical model)**.** A *statistical model* is a collection of distributions indexed by a parameter θ in a parameter space Θ. It is commonly represented as

$$\big\{f_\theta(x) : \theta \in \Theta\big\},$$

where each f_θ is a pmf or pdf.

Example 1.

- For the above Bernoulli example, we have $\theta = p$ and $\Theta = [0, 1]$. The f_θ's are the pmf's of Bernoulli(p) for each p.
- For the statistical model
$$\big\{\mathcal{N}(\mu, \sigma^2) \,:\, \mu \in \mathbb{R},\ \sigma^2 \in (0, \infty)\big\},$$
we have $\theta = (\mu, \sigma^2)$ and $\Theta = \mathbb{R} \times (0, \infty)$.
- Now suppose σ^2 is some known value, and consider the model
$$\big\{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}\big\}.$$
Then $\theta = \mu$ and $\Theta = \mathbb{R}$.
- For
$$\big\{\text{Poisson}(\lambda) : \lambda \in (0, \infty)\big\},$$
the model is $\theta = \lambda$ and $\Theta = (0, \infty)$.

Definition 3 (Parametric statistical model)**.** If the parameter space Θ is finite-dimensional, the model is called *parametric*. Here "finite-dimensional" means that each $\theta \in \Theta$ can be represented by a finite number of numbers. The dimension is 2 in the $\theta = (\mu, \sigma^2)$ Gaussian example, and the dimension is 1 in the other examples.

The pdf or pmf of a random sample $\underline{X}$ from f_θ is just like in (1.1), but now with $f = f_\theta$. We denote it with a subscript θ:

$$f_{\underline{X},\theta}(\underline{x}) = \prod_{i=1}^{n} f_\theta(X_i).$$

Putting the two modeling principles together, we can state the general problem as follows:

$$\text{Given a realization } \underline{x} \text{ of a random sample } \underline{X} \text{ from } f_\theta,$$
$$\text{make inference on (draw a conclusion about) the unknown } \theta.$$

We will take the *frequentist*, or *classical* viewpoint: the value θ is unknown but fixed, meaning there is one true value of θ. We do *not* think of θ as random.

1.3 Statistical Inference

The three goals of statistical inference:

1. **Point estimation.** Give a single best guess for θ. In the kissing-couples example, this addresses Question 2: estimating the population proportion p.
2. **Hypothesis testing.** Make a decision between competing claims about a parameter. For example, for the kissing couples, Question 1 asks: is $p \leq 1/2$ (preference to the left) or $p > 1/2$ (preference to the right)?
3. **Confidence intervals.** Find an interval containing θ with high probability.

For each of these goals, we will draw our conclusions from the data – specifically, from some *summary statistic* of the data.

Definition 4 (Statistic)**.** A *statistic* is a function of the random sample and does not depend on the unknown parameter θ:

$$T = T(\underline{X}).$$

Because T is a function of random variables, it is itself a random variable. Once the data are observed, $T(\underline{X})$ is its realized numerical value.

For the Bernoulli model

$$\big\{\text{Bernoulli}(\theta) : \theta \in [0,1]\big\},$$

both

$$T(\underline{X}) = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and} \qquad T(\underline{X}) = X_1$$

are statistics. In contrast,

$$T(\underline{X}) = \overline{X}_n - \theta$$

is not a statistic because it depends on the unknown parameter θ.

1.4 Point Estimation

We may be interested in estimating θ itself or only "part" of it. For example, if $\theta = (\mu, \sigma^2)$, then μ may be the parameter of interest, while σ^2 is a nuisance parameter. In general, the *parameter of interest* is the value $g(\theta)$ we want to estimate. For example, we could have $g(\mu, \sigma^2) = \mu$.

Definition 5 (Estimator)**.** A statistic $T(\underline{X})$ that is used as a point estimate of a parameter $g(\theta)$ is called an *estimator* of $g(\theta)$. Its observed value $T(\underline{X})$ is called an *estimate*.

Thus every estimator is a statistic, but a statistic becomes an estimator only when it is assigned a particular quantity to estimate. In the kissing-couples example,

$$\widehat{p} = \overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

is an estimator of the population proportion p.
The next questions are:

- How do we find an estimator?
- How do we determine whether an estimator is good?

These will be discussed in subsequent lectures.