18.650: Fundamentals of Statistics	Fall 2023

Lecture 1 — September 6, 2023

Prof. Philippe Rigollet — *Scribe: Anya Katsevich*

1 Overview

Statistics is about **analyzing** and drawing **conclusions** from **data**.

What methods do we use to **analyze** data?

1. descriptive statistics: numbers that summarize the data (e.g. the mean) or visual representations (e.g. histograms)
2. estimation, confidence intervals, hypothesis testing, regression...
3. and many more! Most of which will be covered in this class.

What kinds of **conclusions** can we draw?

1. make predictions (how many goals will the US women's soccer team make at the World Cup?)
2. answer yes/no questions (can LLMs get an MIT degree? Does blue light make you age faster?)

What *is* the **data**?
Independent, identically distributed (i.i.d.) samples $X_1, X_2, \ldots, X_n \sim P$, *where* P *is unknown!* (See Section 3 for a concrete example).

The statistics pipeline:

$$\begin{array}{c}\text{i.i.d. data}\\ X_i \sim P\end{array} \longrightarrow \boxed{\begin{array}{l}\text{statistical}\\ \text{method}\end{array}} \longrightarrow \hat{P} \approx P \tag{1}$$

The fields of statistics and probability are opposites in the following sense:

1. **Probability: given P, what can we say about data from P?**
   Example: $P = \mathcal{N}(0,1)$. *Using probability, we can say a sample* $X \sim P$ *lies in the interval* $(-1,1)$ *with probability* 0.682.
2. **Statistics: given data from P, what can we say about P?**
   Example: $X = 100$. *Using statistics, we can say* X *is most likely* not *a sample from* $\mathcal{N}(0,1)$, *i.e.* $P \neq \mathcal{N}(0,1)$.

2 Key tools from probability

Though it is "opposite" of statistics, probability is the workhorse of statistical computations. So let's review some probability fundamentals.

2.1 Mean and variance of i.i.d. averages

Let $X_1, \ldots, X_n$ be i.i.d. with $\mathbb{E}[X_i] = \mu$ and $\mathbb{V}[X_i] = \sigma^2$ (we will use the notation $\mathbb{V}[X]$ for variance of X). Then the *sample mean* is

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i,$$

and

$$\mathbb{E}[\bar{X}_n] = \mu, \quad \mathbb{V}[\bar{X}_n] = \frac{\sigma^2}{n}.$$

This comes from the following calculations (make sure you understand each step):

$$\begin{aligned}
\mathbb{E}[\bar{X}_n] &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \mu, \\
\mathbb{V}[\bar{X}_n] &= \mathbb{V}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\mathbb{V}\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{V}[X_i] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.
\end{aligned} \tag{2}$$

The equality in blue expresses the important property that for independent variables, the *sum of the variances is the variance of the sum.*

2.2 The Law of Large Numbers (LLN)

Together, $\mathbb{E}[\bar{X}_n] = \mu$ and $\mathbb{V}[\bar{X}_n] = \sigma^2/n$ tell us that the fluctuations of $\bar{X}_n$ around μ get smaller and smaller as $n \to \infty$. This is expressed by the following law of large numbers (LLN):

$$\bar{X}_n \to \mu \quad \text{as} \quad n \to \infty.$$

2.3 The Central Limit Theorem (CLT)

We know that the fluctuations of $\bar{X}_n$ around μ are shrinking, but to do statistical inference, we need more fine-grained information about the distribution of $\bar{X}_n$. This is where the *Central Limit Theorem* (CLT) comes in.

To motivate the CLT, note that

$$\mathbb{E}\left[\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu)\right] = 0, \quad \mathbb{V}\left[\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu)\right] = 1.$$

(Make sure you can do this calculation). It turns out that this scaled, centered random variable $(\sqrt{n}/\sigma)(\bar{X}_n - \mu)$ converges to our favorite distribution which also has mean 0 and variance 1: the normal distribution $\mathcal{N}(0, 1)$.

Theorem 2.1: Central Limit Theorem

Let X_i, $i = 1, \ldots, n$ be i.i.d. with mean μ and variance σ^2. Then

$$\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \rightsquigarrow \mathcal{N}(0, 1),$$

where $\rightsquigarrow$ denotes convergence in distribution.

Convergence in distribution means that for all a, b we have

$$\mathbb{P}\left(a \leq \frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \leq b\right) \to \mathbb{P}(a \leq Z \leq b) \quad \text{as} \quad n \to \infty,$$

where $Z \sim \mathcal{N}(0, 1)$.

Remark.

> Note that $(\sqrt{n}/\sigma)(\bar{X}_n - \mu)$ itself need not be normally distributed. For example if X_i is Bernoulli, then it takes value 0 or 1, so $\bar{X}_n$ will take values in a discrete range: $\{0, 1/n, 2/n, \ldots, (n-1)/n, 1\}$, whereas the normal distribution has continuous range.

Remark.

> If $(\sqrt{n}/\sigma)(\bar{X}_n - \mu) \approx \mathcal{N}(0, 1)$ then
>
> $$\bar{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right). \tag{3}$$
>
> This is the form in which we'll typically use the CLT. As a rule of thumb, the approximation is reasonably accurate for $n \geq 30$.

3 An example

Do people prefer to turn their heads to the right when they kiss?

In an experiment in *Nature* [1], $n = 124$ couples were observed kissing. 80 of the couples turned their heads to the right when they kissed. That's a proportion of $80/124 = 0.645$, which is bigger than 0.5. Can we conclude for sure that humans have a preference to turn to the right? In other words — is 0.645 really "much bigger" than 0.5? Statistics will help us make this quantitative.

We model the n couples as $X_i \overset{\text{i.i.d.}}{\sim} \text{Ber}(p)$, $i = 1, \ldots, n$, where "Ber" stands for Bernoulli. Specifically, we let

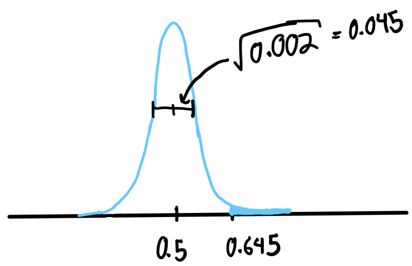
$$X_i = \begin{cases} 1 & \text{if turned right,} \\ 0 & \text{if turned left.} \end{cases}$$

Note that $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ is precisely the proportion of couples who turned their heads to the right. We have observed $\bar{X}_n = 0.645$.

We now want to know: what is the probability of observing $\bar{X}_n = 0.645$ given that $p = 1/2$? If this probability is sizeable, then it is reasonable that $p = 1/2$ is the true value of p which generated the data $X_i \sim \text{Ber}(p)$, and we can't conclude that there is a tendency to kiss turning your head to the right. But if the probability is very small, then we can confidently conclude that a right-turning preference does exist.

So let's do this computation: first we need the mean and variance of the X_i. The mean of $\text{Ber}(p)$ is p and the variance is $p(1-p)$, so if $p = 1/2$ then $\mu = \mathbb{E}[X_i] = 1/2$ and $\sigma^2 = \mathbb{V}[X_i] = 1/4$. By the CLT, we then have

$$\bar{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) = \mathcal{N}\left(1/2, \frac{1/4}{124}\right) = \mathcal{N}(0.5, 0.002).$$



We get $\mathbb{P}(\bar{X}_n \geq 0.645) \approx \mathbb{P}(\mathcal{N}(0.5, 0.002) \geq 0.645) \approx 0.003$. This is what's known as a p-value. Since it's tiny, we can be very confident that $1/2$ is *not* the right value, and that the true value of p is bigger than $1/2$ (meaning, there *is* a preference to turn your head to the right).

References

[1] Onur Güntürkün. *Adult persistence of head-turning asymmetry.* Nature, 2003