Random samples as i.i.d. random variables

Random samples

Often it's of interest to estimate some property of a population by taking a random sample.

Example: Poll 20 "randomly chosen" voters. Estimate the proportion of all voters voting for Trump by the proportion of the 20 voting for Trump.

Example: Survey 10 "randomly chosen" MATH 131 students to estimate the average number of pairs of shoes a student owns, similarly.

We would like to know how reliable our estimate is. To make sense of this, we need a mathematical framework.

Independent random variables

Recall, events A, B in a sample space are independent if

$$\text{Prob}(A \cap B) = \text{Prob}(A) \cdot \text{Prob}(B) \ .$$

(Knowing whether an experimental outcome is in the set B gives you no information about whether the experimental outcome is in A.) The idea of indepence for random variables is essentially the same:

Discrete random variables X and Y are independent if for all numbers s and t,

$$\text{Prob}(X = s \text{ and } Y = t) = \text{Prob}(X = s) \cdot \text{Prob}(Y = t) \ .$$

Continuous random variables X and Y are independent if for all numbers intervals (a, b) and (c, d) in $\mathbb{R}$,

$$\text{Prob}(a < X < b \text{ and } c < Y < d) = \text{Prob}(a < X < b) \cdot \text{Prob}(c < Y < d) \ .$$

(Knowing the outputs of one of the random variables gives you no information about the outputs of the other random variable.) Here, "and" means $\cap$.

The definition for independence of n random variables is similar, using a product of n probabilities.

Independent identically distributed (i.i.d.) random variables

Random variables are identically distributed if the have the same probability law. They are i.i.d. if they are also independent.

I.i.d. random variables $X_1, \dots X_n$ give a mathematical framework for "random sample".

Example. For $1 \le k \le n$, let X_k be the random variable which is 1 with probability p and zero otherwise, and suppose these r.v.s are independent. This is a model for flipping a coin n times, where the coin lands heads with probability p, and X_k records the number of heads (zero or one) seen on the kth flip.

Example. "Randomly" choose n voters and ask each if he/she would vote for Trump. Let $X_k = 1$ if the answer is yes, and $X_k = 0$ otherwise. Here we think of each X_k as an independent sample of the underlying population distribution.

Now, to understand the reliability of an estimate from a random sample, we want to say something mathematical about the sums and averages,

$$\begin{aligned} S_n &= X_1 + \cdots + X_n \\ \overline{X}_n &= \frac{X_1 + \cdots + X_n}{n} = \frac{S_n}{n} . \end{aligned}$$

What we'll "say" are the Law of Large Numbers and the Central Limit Theorem. For this we'll need

Some mathematical facts about i.i.d. random variables:

If X and Y are independent, then $E(XY) = E(X)E(Y)$
and $\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$.
If $X_1, X_2, \dots X_n$ are i.i.d., each with mean μ and variance σ^2, then

$$\begin{aligned} E(X_1 + \cdots + X_n) &= n\,\mu \\ \mathrm{Variance}(X_1 + \cdots + X_n) &= n\,\sigma^2 \\ \text{st. dev.}(X_1 + \cdots + X_n) &= \sqrt{n}\,\sigma . \end{aligned}$$

These facts aren't hard to prove, but we'll skip proof.
That square root will be very important! It reflects that in the sum, there is some cancellation of variations above and below the mean.