5 Averages, Law of Large Numbers, and Central Limit Theorem

Expectation, standard deviation and correlation.

Expectation is the most widely used notion of average in statistics, because of its intuitive interpretations and convenient properties.

Linearity is the most important property of expectation.

The law of large numbers (LLN) and central limit theorem (CLT) are powerful results about the sample mean of a large number of IID RVs.

The LLN says that the sample mean is likely to be close to the theoretical expectation. The CLT says that the sample mean will be approximately Normal.

5.1 Expectation

Arithmetic mean of $x_1, x_2, \ldots, x_n$:

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j.$$

More generally, weighted mean of $x_1, x_2, \ldots, x_n$:

$$weighted\_mean(x) = \sum_{j=1}^{n} x_j p_j.$$

Where the weights $p_1, p_2, \ldots, p_n$ are pre-specified nonnegative numbers with $\sum_{j=1}^{n} p_i = 1$ (so for unweighted mean $p_j = 1/n \; \forall \; i$).

The definition of *expectation of DRV* is inspired by weighted mean, weights are probabilities:

The **expected value** (or **expected mean)** of DRV X whose possible values are $x_1, x_2, \ldots$ is:

$$E(X) = \sum_{j=1}^{\infty} x_j P(X = x_j).$$

If support is finite,

$$E(X) = \sum_x xP(X = x)$$

Where x is value and $P(X = x)$ is PMF at x.

The expectation is undefined if $\sum_{j=1}^{\infty} |x_j|P(X = x_j)$ diverges, since then the series for $E(X)$ diverges or depends on the order in which the x_j are listed.

Warning!

For any DRV X, $E(X)$ is a number (if exists). Common mistake: to replace an RV by its expectation which is wrong because X is a function, $E(X)$ is a constant, and ignores the variability of X.

Notation:

$E(X)$ is abbreviating to EX, similarly EX^2 is $E(X^2)$ not $(E(X))^2$!

The order of operations here is very important!

5.2 Linearity of expectation

The most important property of expectation is *linearity*: expected sum of RVs is the sum of expectations: $\forall X, Y$ $\forall$ constant c:

$$E(X + Y) = E(X) + E(Y),$$
$$E(cX) = cE(X).$$

Averages can be calculated in two ways, *ungrouped* or *grouped*, is all that is needed to prove linearity!

It allows us to work with the distribution X directly without returning to the sample space.

But we can't use the same super-pebbles for another RV Y on the same sample space.

We can take a weighted average of the values of individual pebbles. If $X(s)$ is the value assigns to pebble s:

$$E(X) = \sum_s X(s)P(\{s\}),$$

where $P(s)$ is the weight of pebble s. This corresponds to the ungrouped way of taking averages! It breaks down the sample space into the smallest possible units, so we are now using the *same* weights $P(\{s\})$ for every random variable:

$$E(Y) = \sum_s Y(s)P(\{s\}).$$

Now we can combine $E(X)$ and $E(Y)$:

$$E(X) + E(Y) = \sum_s X(s)P(\{s\}) + \sum_s Y(s)P(\{s\}) =$$
$$\sum_s (X+Y)(s)P(\{s\}) = E(X+Y)$$

Using this property, we can calculate expectations for *Binomial* and *Hypergeometric* distributions!

Binomial expectation: for $X \sim Bin(n,p)$:

$$E(X) = \sum_{k=0}^{n} kP(X=k) = \sum_{k=0}^{n} k\binom{n}{k}p^k q^{n-k}.$$

Linearity of expectation provides a much shorter path to the same result: RV X is the sum of n independent $Bern(p)$ RVs:

$$X = I_1 + \ldots + I_n,$$

each I_j has $E(I_j) = 1p + 0q = p$. By linearity,

$$E(X) = E(I_1) + \ldots E(I_n) = np.$$

Hypergeometric expectation: for $X \sim HGeom(w,b,n)$:

We can write X as a sum of Bernoulli RVs,

$$X = I_1 + \ldots + I_n,$$

Where I_j equals 1 if jth ball is white and 0 otherwise. By symmetry, $I_j \sim Bern(p)$, where $p = w/(w+b)$.

These I_j aren't independent, since balls aren't replacing. However, linearity still holds for dependent RVs. Thus,

$$E(X) = nw/(w+b).$$

5.3 Geometric and Negative Binomial

Geometric distribution:

A sequence of independent Bernoulli trials, each with the same success probability $p \in (0,1)$, with trials performed until a success occurs. RV X is the number of failures before the first successful trial. X has the Geometric distribution with parameter p: $X \sim Geom(p)$.

Geometric PDF: If $X \sim Geom(p)$, then PMF of X is:

$$P(X = k) = q^k p$$

for $k = 0, 1, 2, \ldots$ where $q = 1 - p$.

This is a valid PMF because

$$\sum_{k=0}^{\infty} q^k p = p \sum_{k=0}^{\infty} q^k = p \frac{1}{1-q} = 1.$$

Warning: In our convention, the *Geometric* distribution **excludes** the success, and the *First Success* distribution **includes** the success $Y \sim FS(p)$.

If $Y \sim FS(p)$, then $Y - 1 \sim Geom(p)$

Geometric expectation:

By definition,

$$E(X) = \sum_{k=0}^{\infty} k q^k p,$$

where $q = 1 - p$. Each term looks similar to kq^{k-1}.

$$\sum_{k=0}^{\infty} q^k = \frac{1}{1-q}.$$

This series converges because $0 < q < 1$. Differentiating both sides,

$$\sum_{k=0}^{\infty} k q^{k-1} = \frac{1}{(1-q)^2}.$$

Multiply both sides by pq:

$$E(X) = \sum_{k=0}^{\infty} k q^k p = pq \sum_{k=0}^{\infty} k q^{k-1} = pq \frac{1}{(1-q)^2} = \frac{q}{p}$$

First Success expectation:

$$E(Y) = E(X+1) = \frac{q}{p} + 1 = \frac{1}{p}.$$

Negative Binomial distribution

Sequence of independent Bernoulli trials which success probability p, X is the number of failures before the rth success, $X \sim NBin$.

Binomial counts the number of *successes* in a fixed number of *trials*, while the *Negative Binomial* counts the number of *failures* until a fixed number of *successes*.

Negative Binomial PDF: If $X \sim NBin(r,p)$, then PMF of X is:

$$P(X=n) = \binom{n+r-1}{r-1} p^r q^n$$

for $n = 0, 1, 2, \ldots$ where $q = 1-p$.

Theorem:

If $X \sim NBin(r,p)$, We can write $X = X_1 + \ldots X_r$ were X_i are IIDs $\sim Geom(p)$.

Negative Binomial expectation:

By previous theorem and linearity,

$$E(X) = E(X_1) + \ldots + E(X_r) = \frac{rq}{p}.$$

5.4 Indicator RVs and the fundamental bridge