Data Sampling and Probability

How to sample effectively, and how to quantify the samples we collect. (Continued Discussion)

Announcement for enrollment

Recap: Generalization of binomial probabilities

If we are drawing at random with replacement **n** times, from a population in which a proportion **p** of the individuals are called “successes” (and the remaining **1 - p** are “failures”), then the probability of **k successes** (and hence, **n - k failures**) is

$$P(k \text{ successes}) = \binom{n}{k} p^k (1-p)^{n-k}$$

Multinomial probabilities

Suppose we again sample at random with replacement 7 times from a bag of marbles, but this time, 60% of marbles are **blue**, 30% are **green**, and 10% are **red**.

- What is P(**bgbbbgr**)?
  - Following the same steps as before:

$$\text{P(bgbbbgr)} = = 0.6 \times 0.3 \times 0.6 \times 0.6 \times 0.6 \times 0.3 \times 0.1 = (0.6)^4(0.3)^2(0.1)^1$$

- What is P(4 blue, 2 green, 1 red)?
  - As we saw before, we multiply the above probability by the total number of ways to draw 4 blue, 2 green, and 1 red marbles. This gives

$$\text{P(4 blue, 2 green, 1 red)} = \frac{7!}{4!2!1!}(0.6)^4(0.3)^2(0.1)^1$$

Generalization of multinomial probabilities

If we are drawing at random with replacement **n** times, from a population broken into three separate categories (where $p_1 + p_2 + p_3 = 1$):

- Category 1, with proportion $\mathbf{p_1}$ of the individuals.
- Category 2, with proportion $\mathbf{p_2}$ of the individuals.
- Category 3, with proportion $\mathbf{p_3}$ of the individuals.

Then, the probability of drawing $\mathbf{k_1}$ individuals from Category 1, $\mathbf{k_2}$ individuals from Category 2, and $\mathbf{k_3}$ individuals from Category 3 (where $k_1 + k_2 + k_3 = n$) is

$$\frac{n!}{k_1!k_2!k_3!}p_1^{k_1}p_2^{k_2}p_3^{k_3}$$

At no point in this class will you be forced to memorize this! In practice, we use np.random.multinomial to compute these quantities.

Summary

- Formalized various ideas about sampling
  - Why we need to sample
  - What it means for the sample to biased
  - How to prevent these biases in the samples

- Compute probabilities from samples
  - Binomial and multinomial probabilities

Random Variables

Random Variables, Distributions, and Expectations

What is a random variable?

A ***random variable***, is a numerical outcome of a random phenomenon.
A random variable is a **numerical function** of **a random sample.**

- Another name for a numerical function of a sample is a “statistic.”

We typically denote random variables with uppercase letters (e.g. *X*, *Y*).

Why **random**? Because the sample on which it is a function was drawn at random.

Why **variable**? Because its value depends on how the sample came out.

Function of the sample

Population

- Let s be a sample of size 3.
- Let X be the number of blue people in our sample.
  - X, then, is a random variable!

X(s) = 1

X(s) = 2

X(s) = 0

Another Example: Let X be the outcome of the roll of a die. Then X is a random variable. Its possible values are 1, 2, 3, 4, 5, and 6.

The word "random" doesn't necessarily imply that the outcome is completely random in the sense that are values are equally likely; "random" simply means that the value is uncertain.

Distribution

Terminology and notation (possible values, probabilities)

For now, assume our random variables have a finite number of **possible values**.

$$P(X = x)$$

This is the **probability that random variable X takes on the value x**.

- For instance, P(X = 20) is the chance that X has the value 20.
- The probabilities of each possible value must each be non-negative, and must sum to 1:

$$\sum_{\text{all } x} P(X = x) = 1$$

The probabilities assigned to the possible values of a random variable are its **distribution**. A distribution completely describes a random variable

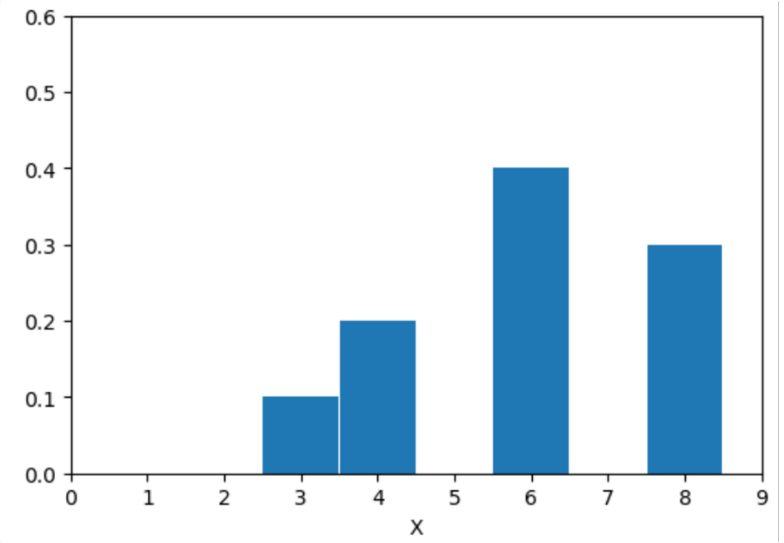
Example distribution

Consider a random variable X with the following **distribution table**:

values

x	$P(X = x)$
3	0.1
4	0.2
6	0.4
8	0.3

probabilities of those values



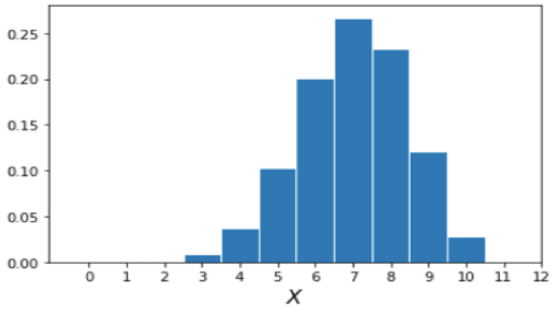
To compute related probabilities, we add up the probabilities belonging to that event.

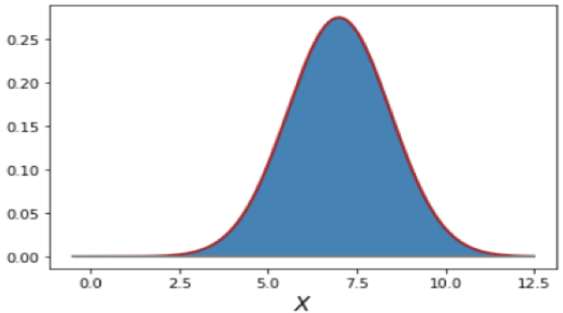
(For instance, X < 6 happens if X = 3 or X = 4).

Types of distributions

Probability distributions largely fall into two main categories.

- **Discrete**. (probability mass function)
  - The set of possible values that X can take on is either finite or countably infinite.
  - Values are separated by some fixed amount.
  - For instance, X = 1, 2, 3, 4, …
- **Continuous**. (probability density function)
  - The set of possible values that X can take on is uncountable.
  - Typically, X can be any real number in **some interval** (not just our counting numbers).
  - Probability is represented by the area under a curve.





Common distributions

Discrete

- **Bernoulli** (p).
  - Takes on the value 1 with probability p, and 0 with probability 1-p.
- **Binomial** (n, p).
  - Number of 1s in n independent Bernoulli (p) trials.
  - Probabilities given by the binomial formula.
- **Uniform on a finite set.**
  - Probability of each value is 1 / (size of set). For example, a standard die.

Continuous

- **Uniform on the unit interval.**
  - U could be any real number in the range [0, 1].
- Normal(μ, σ^2).

Expectation

Definition of expectation

The **expectation** of a random variable X is the weighted average of the values of X, where the weights are the probabilities of the values.

It is the long run average of the random variable, if you simulate the variable many times.

The most common formulation applies the weights one possible value at a time:

$$\mathbb{E}(X) = \sum_{\substack{\text{all possible} \\ x}} x\mathbb{P}(X = x)$$

Example: The expected value of the roll of a die is

$$1(\frac{1}{6}) + 2(\frac{1}{6}) + 3(\frac{1}{6}) + \cdots + 6(\frac{1}{6}) = 21/6 = 3.5.$$

Notice that the expected value is not one of the possible outcomes: you can't roll a 3.5. However, if you average the outcomes of a large number of rolls, the result approaches 3.5.

Examples

Consider the random variable X we defined earlier:

x	$P(X = x)$
3	0.1
4	0.2
6	0.4
8	0.3

$$
\begin{aligned}
E(X) &= \sum_x x \cdot P(X = x) \\
&= 3 \cdot 0.1 + 4 \cdot 0.2 + 6 \cdot 0.4 + 8 \cdot 0.3 \\
&= 0.3 + 0.8 + 2.4 + 2.4 \\
&= 5.9
\end{aligned}
$$

Note, 5.9 is not one of the possible values that X can take on!

Expectation of functions of random variables

More generally, if X is a random variable and g is any function (not necessarily linear), we have

$$\mathbb{E}(g(X)) = \sum_x g(x)P(X = x)$$

For example, if X is uniform on {1, 2, 3, 4, 5, 6}, we have

$$\begin{aligned}\mathbb{E}(X^2) &= \sum_x x^2 P(X = x) \\ &= 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + 3^2 \cdot \frac{1}{6} + 4^2 \cdot \frac{1}{6} + 5^2 \cdot \frac{1}{6} + 6^2 \cdot \frac{1}{6} \\ &= \frac{91}{6}\end{aligned}$$

$E(X^2) = 91/6 = 15.166\ldots$
$E(X)^2 = 3.5^2 = 12.25$

$E(X^2)$ and $E(X)^2$ are different!

Note: The property that held on the last slide for linear functions **g** does not hold in general!

$$\mathbb{E}(g(X)) \neq g(\mathbb{E}(X))$$

Variance

- The variance of a random variable X is denoted by either Var(X) or σ_X^2
  - $\sigma_X^2 = E[(X - E(X))^2]$
  - Expected value of the square difference between X and its mean
- For a discrete distribution, we can write the variance as
  - $\sigma_X^2 = \sum_x \big(x - E(X)\big)^2 P(X = x) = \mathrm{E}(X^2) - \ E(X)^2$

Find the variance and standard deviation for the roll of one die. Solution: We use the formula $Var[X] = E[X^2] - (E[X])^2$. We found previously that $E[X] = 3.5$, so now we need to find $E[X^2]$. This is given by

$$E[X^2] = \sum_{x=1}^{6} x^2 P_X(x) = 1^2(\frac{1}{6}) + 2^2(\frac{1}{6}) + \cdots + 6^2(\frac{1}{6}) = 15.167.$$

Thus,

$$\sigma_X^2 = Var[X] = E[X^2] - (E[X])^2 = 15.167 - (3.5)^2 = 2.917$$

and $\sigma = \sqrt{2.917} = 1.708$.

Transformations

Let X be a random variable, and a and b be constants.

- We call aX + b a **linear transformation** of X.
- The expectation of a linear transform of X is equal to the linear transform applied to the expectation of X.
- Put less cryptically:

$$\mathbb{E}(aX + b) = a\mathbb{E}(X) + b$$

Note, this means that E[c] = c, where c is a constant.

Why does this matter?

- We will often manipulate the sample mean of several random variables. This is a linear transformation of the sample sum.
- Many unit conversions are also linear transformations (e.g. ºF = 9/5 * ºC + 32).

Additivity

For **any** two random variables X and Y (regardless of their relationship):

$$\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$

We call this property the **additivity of expectation**. We will not prove this, but you can do so yourself using the second definition of expectation (summing over all samples).

For example: Consider the "sample sum" $S_n = \sum_{i=1}^{n} X_i$ where $\mathbb{E}(X_i) = \mu$ for each i. Then,

$$\begin{aligned}\mathbb{E}(S_n) &= \sum_{i=1}^{n} \mathbb{E}(X_i) \\ &= n\mu\end{aligned}$$

Linearity

Two of the properties we just established were

- Linear transformations apply to expectations.
- Expectation is additive.

Combining these gives us a single property, which is sometimes referred to as the **linearity of expectation**. For any random variables X, Y and constants a, b:

$$E(aX + bY) = aE(X) + bE(Y)$$

Summary

Summary

- Random variables are functions of random sample.
- The expectation of a random variable is the weighted average of its possible values, weighted by the probabilities of those values.
    - Expectation behaves nicely with linear transformations of random variables.
    - Expectation is also additive.