Unit 2: Expectation and variance, Normal distribution and CLT

Ethan Levien

September 19, 2025

Contents

Introduction

In this section we introduce expectation, an operation which takes a random variable and produces a deterministic quantity. The expectation of a random variable can be approximated with sample averages and from them we can infer properties of the model (like parameters). Much of statistics relies on the fact that sample averages approximate expectations, and understanding how well these approximations work is a central goal of statistics. This will motivate us to study the probability distribution of sums of random variables, which leads to the CLT and the Normal distribution. If there is time, we will also learn about log normal distributions, which are a better model for many real world random variables, but can easily by mapped to normal random variables.

2.1 Expectation, variance and standard deviation

2.1.1 Sample averages and expectation

References: [Evans and Rosenthal, 2004, Ch. 3]

Usually it is difficult to obtain the full distribution of a random variable from data and it may not even be that relevant for the questions we are asking. Instead, we would like to summarize properties of a random variable by looking at averages. In other cases, we have a good idea what the type of distribution is, but there are unknown parameters which can be estimated by averages.

You are probably familiar with the sample mean, sample average, or empirical average. If $Y_1, Y_2, \dots, Y_n$ are iid samples of Y, the sample mean is defined as

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

Sometimes the notation $\langle\cdot\rangle$ is used. More generally, we might look at the average of some function of a random variable

$$\overline{g(Y)} = \frac{1}{n}\sum_{i=1}^{n} f(Y_i)$$

(Remember, a function of a random variable is just another random variable, so there is nothing too deep here). If we take the function to be

$$g(y) = 1_A(y) = \begin{cases} 1 & \text{if } y \in A \\ 0 & \text{if } y \in A \end{cases} \tag{1}$$

Then we can connect the idea of a sample average to estimates of probabilities via

$$\overline{g(Y)} = \frac{N(y \in A)}{n} \approx P(\{Y \in A\}) \tag{2}$$

Any quantity we compute from data is in some way a sample average, so a great deal of statistics is about understanding the behavior of sample averages.

Now we introduce the idea of expectation. I like to think of expectation as the mathematical idealization of a sample average, just as probabilities are mathematical idealizations of long-run frequencies[1]. Suppose each Y_i are iid random variables with each having sample space Y. If n is large, then of course the fraction of samples for which $Y_i = y$ will be $\approx P(\{Y_1 = y\})$. We can express the sample average in terms of the probabilities via

$$\overline{Y} = \frac{1}{n}\sum_i Y_i = \frac{1}{n}\sum_{y\in S} yN(Y=y) = \sum_{y\in S} y\frac{N(Y=y)}{n} \approx \sum_{y\in S} yP(Y=y).$$

The second equality just came from putting the samples in groups according to their value of y. The expression on the right is the definition of the mean, or expectation, and is denoted

$$\mathbb{E}[Y] = \sum_{y=1}^{m} yP(Y=y) \tag{3}$$

To summarize what we saw above (and should be intuitively clear)

$$E[Y] \approx \overline{Y} \tag{4}$$

Sometimes we use $\mathbb{E}$ instead of E to distinguish it from other variables names E, but I will try not to use E for other things. If we have a function $g : S \to S'$ from the sample space to some other space S', then $g(Y)$ is simply a new random variable with sample space S', but we don't usually need to find the distribution of $X = g(Y)$ to compute expectations, since this can be written

$$E[X] = E[g(Y)] = \sum_{y\in S} g(y)P(Y=y). \tag{5}$$

The nice thing here is that we can use the probability distribution for Y to compute the expectation of X, which in some cases may be simpler (this is the case for variance below).

It is important to understand that, just like probabilities, the expectation is an operation which takes a random variable to a deterministic number. The sample average is the approximate version of this. I like to think of expectations and sample averages as living in "math world" and "data world" respectively. After discussion variance, standard deviation and CV, Section 2.2 will take you on a deeper dive into this connection, which is made precise by the LLN and CLT.

[1] This is at least the frequentist interpretation. We can also interpret them as measures of belief. More on that in Unit 6 and 7

Example 1 (Expectation of a discrete random variable)**.** Let Y be a random variable taking values in $\{1, 2, 3\}$ with probabilities

$$P(Y = 1) = \tfrac{1}{2}, \quad P(Y = 2) = \tfrac{1}{3}, \quad P(Y = 3) = \tfrac{1}{6}.$$

Question: What is $E[Y]$?

Solution: By definition,

$$\begin{aligned} E[Y] &= \sum_{y=1}^{3} y P(Y = y) \\ &= 1 \cdot \tfrac{1}{2} + 2 \cdot \tfrac{1}{3} + 3 \cdot \tfrac{1}{6} \\ &= \tfrac{1}{2} + \tfrac{2}{3} + \tfrac{1}{2} \\ &= \tfrac{5}{3}. \end{aligned}$$

So the expected value of Y is $E[Y] = \frac{5}{3}$.
Python exercise: Write code that generates a dataframe with n simulated values of Y, computes the sample mean, and plots the sample mean as a function of n. Compare your results with the theoretical expectation $E[Y] = 5/3$.

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

# Define the probability distribution
values = [1, 2, 3]
probs = [1/2, 1/3, 1/6]

# Number of samples to generate
Nmax = 5000

# Draw samples from the distribution
samples = np.random.choice(values, size=Nmax, p=probs)

# Store in a dataframe
df = pd.DataFrame({"Y": samples})

# Compute running sample mean
sample_means = df["Y"].expanding().mean()

# Plot sample mean vs n
plt.figure(figsize=(6,4))
plt.plot(sample_means, label="Sample mean")
plt.axhline(5/3, color="red", linestyle="--", label="Theoretical $E[Y]=5/3$")
plt.xlabel("n (sample size)")
plt.ylabel("Sample mean")
plt.legend()
plt.title("Convergence of sample mean to expectation")
plt.show()
```

2.1.2 Measuring variation

One of the most important expectations is the variance, which measures the typical distance from the mean. This is defined as

$$\text{var}(Y) = E[(Y - E[Y])^2]$$

which you should comes from taking $f(Y) = (Y - E[Y])^2$ in Eq. 5. Another way to write this is

$$\text{var}(Y) = E[Y^2] - 2E[Y]^2 + E[Y]^2 = E[Y^2] - E[Y]^2$$

Example 2 (Mean and variance of Bernoulli random variable)**.** Let Y be a Bernoulli random variable with parameter q. We will use the convention that $Y = 1$ with probability q.

Question: What is $E[Y]$ and $\text{var}(Y)$?

Solution: Using the definitions above

$$E[Y] = P(Y = 0) \times 0 + P(Y = 1) \times 1 = q$$

similarly you should be able to see that $\text{var}(Y) = q(1 - q)$. Try testing this formula with Python (this is one of the exercises).

Based on this example, we can see that to estimate q we can use $\hat{q} = \bar{Y}$ (as expected). Here I'm defining $\hat{q}$ as shorthand for an estimator of q.

To measure "how much variation" there is in a random variable, we need to compare the variance to the mean. However, there is a problem with doing this directly: the variance has different units than the mean. For example, say we are looking at human height. If the mean height is about 170 cm, then the variance might be something like 100 cm^2. This is hard to interpret because the mean is measured in centimeters, while the variance is in squared centimeters. To make the comparison meaningful, we first take the square root of the variance to obtain the standard deviation, which brings the measure of spread back into the same units as the mean (in this case, centimeters). Now we can meaningfully say that the spread around the mean height is about 10 cm.

But even the standard deviation is not enough by itself when we want to compare variability across different contexts. Suppose we measure the weights of the same individuals, where the mean is around 70 kg and the standard deviation is 10 kg. The "10" here is not directly comparable to the "10" cm in height, because the scales of measurement are different. What matters is not the absolute size of the variation, but its size *relative to the mean*.

This leads us to the coefficient of variation (CV), defined as

$$\text{CV} = \frac{\sigma}{\mu}, \tag{6}$$

where σ is the standard deviation and μ is the mean. The CV is unitless and therefore allows for comparisons across variables measured in different units or across distributions with very different scales. For example, a CV of 0.06 in human height (10/170) indicates less relative variability than a CV of 0.14 in weight (10/70).

Thus, the CV is the correct measure of variation when our goal is to compare variability across different contexts, because it removes dependence on units and scale while still preserving the intuitive meaning of variation as "spread relative to the average."

2.1.3 Conditional expectation

References: [Evans and Rosenthal, 2004, Ch. 1 Sec. 2 and Ch. 2.1]

We define the conditional expectation [Evans and Rosenthal, 2004, Definition 3.5.1] as the expectation of the conditional variable; that is,

$$E[X|Y=y] = \sum_x xP(X|Y=y)$$

With samples

$$\{(x_1,y_1),\dots,(x_n,y_n)\}$$

we have

$$E[X|Y=y] \approx \frac{1}{N(Y=y)}\sum_{i=1}^{n} 1_{\{y_i=y\}}x_i$$

where $1_{\{y_i=y\}}$ is the indicator function

$$1_{\{y_i=y\}} = \begin{cases} 0 & \text{if } y_i \neq y \\ 1 & \text{if } y_i = y \end{cases}$$

Don't get too hung up on the notation, put simply: we compute the conditional expectation from a sample average by taking the sample average among samples satisfying a condition.

As we already noted, the conditional probabilities can tell us whether two variables are independent. That is, $P(X|Y) = P(X)$ if and only if X and Y are independent. If X and Y are independent, then $E[X|Y=y] = E[X]$ for all y but the converse is false: **it is possible that this is true but X and Y are not independent!** We will say about this later.

Example 3 (Computing conditional expectation)**.** Consider the pair of random variables (Y_A, Y_B) defined by the probability distribution we saw in week 1:

$$P(Y_A,Y_B) = \begin{cases} 1/2 & \text{if } Y_A = 0 \text{ and } Y_B = 0 \\ 1/8 & \text{if } Y_A = 0 \text{ and } Y_B = 1 \\ 1/8 & \text{if } Y_A = 1 \text{ and } Y_B = 0 \\ 1/4 & \text{if } Y_A = 1 \text{ and } Y_B = 1 \end{cases} \tag{7}$$

Question: Compute $E[Y_A|Y_B=1]$

Solution: We can obtain the conditional distribution of Y_A as

$$P(Y_A=1|Y_B=1) = \frac{P(Y_A=1,Y_B=1)}{P(Y_B=1)} = \frac{1/4}{3/8} = \frac{2}{3}$$

Note that his means $P(Y_A=0|Y_B=1) = 1/3$ and so the conditional distribution of Y_A is

$$Y_A|(Y_B=1) \sim \text{Bernoulli}(2/3)$$

which means

$$E[Y_A|Y_B=1] = \frac{2}{3}.$$

Example 4 (Mean of a Geometric random variable via simulation)**.** Let $X \sim \text{Geometric}(q)$ with support $\{1,2,\dots\}$, i.e.,

$$P(X=k) = q(1-q)^{k-1}, \quad k \in \{1,2,\dots\}.$$

Question: Use simulations to confirm that $E[X] = 1/q$.

Solution:

```python
import numpy as np
import matplotlib.pyplot as plt

rng = np.random.default_rng(42)

q = 0.3        # success probability
N = 100_000    # total number of samples

# NumPy's geometric uses support {1,2,...} with P(X=k)=q(1-q)^{k-1}
X = rng.geometric(p=q, size=N)

# Running sample mean
running_mean = np.cumsum(X) / np.arange(1, N+1)

print(f"Theoretical E[X] = {1/q:.6f}")
print(f"Estimated  E[X] = {running_mean[-1]:.6f}")
```

Analytically,

$$E[X] = \sum_{k=1}^{\infty} k\,q(1-q)^{k-1} = q \cdot \sum_{k=1}^{\infty} k r^{k-1}\Big|_{r=1-q} = q \cdot \frac{1}{(1-r)^2}\Big|_{r=1-q} = \frac{1}{q}.$$

but you don't need to be able to do this calculation.

Example 5 (Computing conditional expectation from data)**.** Consider the following data containing children's test scores and some other information.

```
# Here is some data on children's test scores
url = (
    "https://raw.githubusercontent.com/"
    "avehtari/ROS-Examples/"
    "master/KidIQ/data/kidiq.csv"
)
df = pd.read_csv(url)
df
```

Let Y be the test score and X be a binary variable representing whether the mother graduated high school.

Question: Compute $E[Y|X=0]$ and $E[Y]$. Do you think X and Y are independent?

Solution: See Python notebook.

2.1.4 Properties of expectation

Expectation has some important properties. These become particularly relevant when we work with linear regression models, which are defined in terms of conditional expectations.

1. **Linearity [Evans and Rosenthal, 2004, Theorem 3.1.2]:** For two random variables X and Y

$$E[X+Y] = E[X] + E[Y]$$

Proof. We do the proof when the sample spaces S_X and S_Y are discrete:

$$\begin{aligned}
E[X+Y] &= \sum_{y\in S_Y}\sum_{x\in S_X}(x+y)P(X=x,Y=y) \\
&= \sum_{x\in S_X}\sum_{y\in S_Y} xP(X=x,Y=y) + \sum_{x\in S_X}\sum_{y\in S_Y} yP(X=x,Y=y) \\
&= \sum_{x\in S_X} x\left(\sum_{y\in S_Y} P(X=x,Y=y)\right) + \sum_{y\in S_Y} y\left(\sum_{x\in S_X} P(X=x,Y=y)\right) \\
&= E[X]+E[Y]
\end{aligned}$$

□

2. **Multiplication by a constant [Evans and Rosenthal, 2004, Theorem 3.1.2]:** If a is a constant (meaning it is not random), then

$$E[aX] = aE[X]$$

Proof. Left as an exercise. □

3. **Factoring for independent variables [Evans and Rosenthal, 2004, Theorem 3.1.3]:** If X and Y are independent, the

$$E[XY] = E[X]E[Y]$$

Proof. Using independence, we have

$$\begin{aligned}
E[XY] &= \sum_{x\in S_X}\sum_{y\in S_Y} xyP(X=x,Y=y) \\
&= \sum_{x\in S_X}\sum_{y\in S_Y} xP(X=x)yP(Y=y) \\
&= \left(\sum_{x\in S_X} xP(X=x)\right)\left(\sum_{y\in S_Y} yP(X=x)\right) = E[X]E[Y]
\end{aligned}$$

□

4. **Tower property [Evans and Rosenthal, 2004, Theorem 3.5.2]:** Let X and Y be two random variables,

$$E[E[X|Y]] = E[X]$$

where by $E[X|Y]$ we mean the random variable constructed by taking the conditional expectation of X given a random value of Y. Another way to define this is to introduce the deterministic function $f(y) = E[X|Y=y]$ which outputs a number for every value $y \in Y$. Then we define the random variable $E[X|Y] = u(Y)$. Therefore

$$E[E[X|Y]] = E[f(Y)]$$

Proof. Left as an exercise. □

Example 6 (Calculating conditional expectations)**.** Consider the probability model for a variable X

$$P(X = x) = \begin{cases} \frac{1}{2}, & x = 1, \\ \frac{1}{8}, & x = 2, \\ \frac{1}{8}, & x = 3, \\ \frac{1}{4}, & x = 4, \end{cases} \tag{8}$$

and define

$$Y = X \cdot Z, \qquad Z \sim \text{Geometric}(X/4), \tag{9}$$

Question: Compute $E[Y]$ using the tower property and check your answer using simulations.

Solution: By the tower property,

$$\begin{aligned} E[Y] &= E[\,E[Y \mid X]\,] && (10) \\ &= E[\,E[XZ \mid X]\,] && (11) \\ &= \sum_{x \in S_X} P(X = x) E[xZ \mid X = x] && (12) \\ &= \sum_{x \in S_X} P(X = x) x E[Z|X = x] = \sum_{x \in S_X} P(X = x) x \frac{4}{x} && (13) \\ &= E\left[X \cdot \frac{4}{X}\right] && (14) \\ &= E[4] \ = \ 4. && (15) \end{aligned}$$

Hence, regardless of the distribution in (8), as long as $Z \sim \text{Geometric}(X/4)$ (with support starting at 1), we have

$$E[Y] = 4. \tag{16}$$

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

rng = np.random.default_rng(2025)

# Distribution of X
x_vals  = np.array([1, 2, 3, 4])
x_probs = np.array([1/2, 1/8, 1/8, 1/4])

def simulate(N):
    # Sample X
    X = rng.choice(x_vals, size=N, p=x_probs)
    # For each X, sample Z ~ Geometric(p=X/4)
    Z = np.array([rng.geometric(p=xi/4) for xi in X])
    Y = X * Z
    return pd.DataFrame({"X": X, "Z": Z, "Y": Y})

# Run simulation
N = 200_000
df = simulate(N)
```

```
# Estimate E[Y]
EY_hat = df["Y"].mean()

# Empirical E[E[Y|X]]
EY_given_X_hat = df.groupby("X")["Y"].mean().sort_index()
pX_hat         = df["X"].value_counts(normalize=True).sort_index()
EEY_given_X_hat = (EY_given_X_hat * pX_hat).sum()

print(f"Theoretical E[Y]     = 4.000000")
print(f"Estimated E[Y]       = {EY_hat:.6f}")
print(f"Estimated E[E[Y|X]] = {EEY_given_X_hat:.6f}")
print("\nConditional means E[Y|X=x] (each should be ~4):")
print(EY_given_X_hat)
```

Example 7 (Expectation of binomial)**.** Let Y be a binomial random variable.

Question: What are $E[Y]$ and $\text{var}(Y)$?

Solution:

$$E[Y] = \sum_{k=1}^{N} kP(Y=k) = \sum_{k=1}^{N} k\binom{N}{k} q^k (1-q)^{N-k} = \cdots .$$

A much easier way is to use the definition of a Binomial random variable and exceptions

$$\begin{aligned} E[Y] &= E\left[\sum_{j=1}^{N} X_i\right] \\ &\underset{(1)}{=} \sum_{j=1}^{N} E\left[X_i\right] = Nq \end{aligned}$$

where we are using the fact that averages are additive (property (1)). Similarly,

$$\begin{aligned} E[Y^2] = E\left[\left(\sum_{j=1}^{N} X_i\right)^2\right] &= E\left[\sum_{i=1}^{N}\sum_{j=1}^{N} X_i X_j\right] \\ &\underset{(1)}{=} \sum_{i=1}^{N}\sum_{j=1}^{N} E[X_i X_j] \underset{(3)}{=} \sum_{i=1}^{N}\sum_{j\neq i}^{N} q^2 + Nq(1-q) + Nq^2 \\ &= N(N-1)q^2 + Nq(1-q) + Nq^2 = Nq(1-q) + N^2q^2 \end{aligned}$$

Therefore

$$\text{var}(Y) = E[Y^2] - E[Y]^2 = Nq(1-q)$$

To summarize what we learned in Example 7

$$E[Y] = qN \qquad \text{var}(Y) = Nq(1-q). \tag{17}$$

The important observation that the mean grows much faster with N than the variance is also captured by the coefficient of variation:

$$\text{CV} = \frac{\sqrt{\text{var}(Y)}}{E[Y]} = \sqrt{\frac{(1-q)}{q}\frac{1}{N}}.$$

The idea is that we are measuring the variation *relative* to the average. This is relevant for many applications where we only care about the relative deviations.

Example 8 (Election modeling)**.** Consider a model of votes in an election involving two candidate. Let q be the fraction of people in the population who support candidate one and suppose N people vote at the election (you can assume N is much less than the total number of people in the population, as voter turnout is low). Then the number of people, M, who vote for the first candidate can be modeled as

$$M \sim \text{Binomial}(N, q)$$

Think about the assumption we are making when we use this model.

Question: Suppose there is a city in which a fraction $q = 0.51$ of people support a candidate for city council. If $N = 1000$ people turnout for the election, what is the chance that the actual vote share, $\phi = M/N$, differs from the actual fraction of support throughout the population by more than 1%?

Solution:

```
import numpy as np
import matplotlib.pyplot as plt

# Parameters
q = 0.51        # true population support
N = 1000        # number of voters
delta = 0.01    # tolerance for deviation in vote share
trials = 200_000   # Monte Carlo repetitions

rng = np.random.default_rng(123)

# --- Monte Carlo estimation ---
# Draw M ~ Binomial(N, q) many times, compute phi = M/N, estimate P(|
    phi - q| > delta)
M_samples = rng.binomial(n=N, p=q, size=trials)
phi = M_samples / N
prob_est = np.mean(np.abs(phi - q) > delta)

print(f"Monte Carlo estimate P(|phi - {q}| > {delta}) = {prob_est:.6f}
    "
      f"(N={N}, trials={trials})")
```

In the problem above the vote share is the same as the sample mean:

$$\overline{X_i} = \frac{M}{N} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

You should be able to see that $E[\phi] = q$. What about the variance?

$$\text{var}(\phi) = \text{var}(Y/N) = \frac{1}{N^2}\text{var}(Y) = \frac{q(1-q)}{N}$$

Notice that this will tend towards zero as $N \to \infty$. Meanwhile, $E[\phi]$ has no dependence on N. This is a consequence of the fact that the CV is decreasing with N and it allows us to determine q by approximating $E[\phi]$ with the sample mean.

2.2 Connecting "math world" and "data world": LLN and CLT

The goal of this section is to understand the distribution of a sum of iid random variables when N is large. This is obviously relevant if we want to be more precise about how accurate our estimates are. We begin with the law of large numbers, which simply makes Eq. 4 – the statement that the sample average approximates the expectation – precise.

2.2.1 LLN

The binomial distribution illustrates a very basically principle that we have already used a number of times: When we sum over a large number of independent random variables and divide by the total number, the result is close to the mean. This is the Law of Large Numbers (LLN).

Theorem 1 (Law of Large numbers). *Let X_i be independent and identically distributed and set*

$$S_N = \sum_{i=1}^{N} X_i.$$

If $E[X_i] < \infty$, then $S_N/N \to E[X_i]$.

This is not very precise, since we should really be specific about what it means for a random number to converge to something, but for our purposes it will suffice to think of this as saying that for large enough N, S_N/N will not differ from $E[X_i]$ very much. See [Evans and Rosenthal, 2004, Theorem 4.2.1] for a more technical statement. Another way to say this is that for iid random variables X_i, $i = 1, \dots, N$, the sample average $\overline{X}$ approach $E[X_i]$. The binomial distribution actually tell us more, it tell us that the variation around $E[X_i]$ is proportional to $1/\sqrt{N}$. It is natural to ask whether this is also true for other random variables. The key is that the dependence on N in Equations 17 does not depend on the distribution of X_i! So if X_i is the roll of a dice, or a geometric distribution, we expect the same thing to hold. The behavior of random sums is in-fact even more universal than this argument suggests. We can actually describe the distribution of any[2] random sum with a single distribution. In order to describe this distribution, we need to introduce the notion of continuous random variables.

2.2.2 Continuous probability distributions

To understand what happens to the distribution of sums and averages of random variables, we need to extend our framework to *continuous probability distributions*. The motivation comes from the sample average of many i.i.d. random variables,

$$\overline{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i.$$

Even if each Y_i only takes finitely many values (for instance, a Bernoulli random variable), the average can take increasingly many distinct values as n grows.

For example, starting with iid Bernoulli variables $X_1, \dots, X_n$, the sample average $\bar{X}$ (which is a Binomial random variable divided by n) has sample space

$$S_{\bar{X}} = \left\{0, \tfrac{1}{n}, \tfrac{2}{n}, \dots, 1\right\}.$$

[2] With the caveat that here we only deal with the case where $\mathrm{var}(X_i) < \infty$

As $n \to \infty$, this set of possible outcomes becomes dense in the interval $[0, 1]$. In the limit we naturally want to talk about a probability distribution supported on a continuum of values—but our previous discrete framework does not cover this situation.

Fortunately, much of the discrete theory still applies once we replace sums with integrals. You will not be asked to evaluate integrals in this course, but we need to set up the basic definitions.

The Uniform Distribution

A simple starting point is the *uniform distribution* on an interval:

$$Y \sim \text{Uniform}(a, b).$$

Here Y is equally likely to fall anywhere in the interval $[a, b]$ (with $a < b$). If we let $L = b - a$, then for any subinterval $y_1 < y_2$ inside $[a, b]$,

$$P(y_1 \leq Y \leq y_2) = \frac{y_2 - y_1}{L}.$$

This is shown in Figure 1. In words: the probability of landing in an interval is proportional to its length. This ensures normalization:

$$P(a \leq Y \leq b) = 1.$$

Notice that as $y_2 \to y_1$, the probability goes to zero. Thus $P(Y = y) = 0$ for any specific y. This reflects the fact that there are uncountably many possible outcomes in any interval, so no single point can carry positive probability.

Densities

This example motivates the general notion of a *probability density function* (pdf). A continuous random variable Y is characterized by a nonnegative function $f(y)$, called its density, such that for any $a < b$,

$$P(a < Y < b) = \int_a^b f(y)\, dy.$$

Geometrically, the probability is given by the *area under the curve* of $f(y)$ between a and b.

For small intervals,

$$P(y \leq Y \leq y + dy) \approx f(y)\, dy,$$

so $f(y)$ plays the role of "probability per unit length."

In the uniform case,

$$f(y) = \begin{cases} 1/L, & y \in [a, b], \\ 0, & \text{otherwise.} \end{cases}$$

Every pdf $f(y)$ must satisfy:

1. **Nonnegativity:** $f(y) \geq 0$ for all y.
2. **Normalization:** $\int_{-\infty}^{\infty} f(y)\, dy = 1$.

These are the continuous analogues of the conditions we imposed on discrete probability distributions. Note that $f(y)$ need NOT be less than 1, because $f(y)$ is not a probability, rather integrals $\int_a^b f(y) dy$ is a probability. For example, suppose $f(y)$ is a uniform distribution on $[0, L] = [0, 1/1000]$. Then $f(y) = 1/L = 1000$ for y between 0 and 1/1000, and zero otherwise. The fact that the density is concentrated in a very small region cancels with the large values it takes in this region so that the integrals are always ≤ 1.

For a density, the expectation is defined by replacing the sum with a integral:

$$E[g(Y)] = \int_{S_Y} g(y) f(y) dy \tag{18}$$

You won't have to calculate integrals in this class, but it's important to understand where the expectation comes from for a continuous distribution.

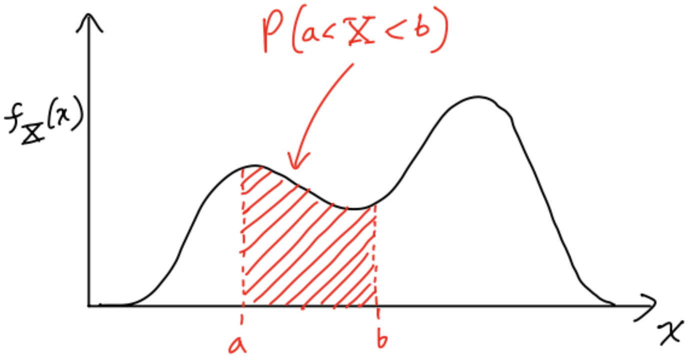

Figure 1: The density and its relationship to probabilities

Example 9 (Condition with continuous random variables)**.** If Y is uniform on $[0, 1]$.

Question: What is the density of $Y|(Y < 1/2)$? Check the answer with simulations.

Solution: We can start with the definition of density

$$P(y_1 < Y < y_2|Y < 1/2) = \frac{P(y_1 < Y < y_2, Y < 1/2)}{P(Y < 1/2)}$$

What is the think on the top? We will assume $y_1 > 0$ and $y_2 < 1/2$, then the numerator is $y_2 - y_1$, since $Y < 1/2$. The key here is that if $Y \in [y_1, y_2]$ $Y < 1/2$ is automatically true, so the chance that BOTH of these things are true in a sample is the chance that the more restrictive one is true.
The denominator is $P(Y < 1/2) = 1/2$. This means

$$P(y_1 < Y < y_2|Y < 1/2) = 2(y_2 - y_1)$$

This means the density is

$$f(y|Y < 1/2) = 2$$

Thus

$$Y|(Y < 1/2) \sim \text{Uniform}(0, 1/2)$$

```
import numpy as np
import matplotlib.pyplot as plt

rng = np.random.default_rng(123)

# Number of samples
N = 200_000

# Sample Y ~ Uniform(0,1)
Y = rng.uniform(0, 1, size=N)

# Condition on Y < 1/2
Y_cond = Y[Y < 0.5]
```

```
print(f"Proportion of samples kept (should be ~0.5): {len(Y_cond)/N:.3f
    }")
print(f"Sample mean of conditional distribution: {Y_cond.mean():.3f}")
print(f"Theoretical mean of Uniform(0,0.5): {0.25}")

# Plot histogram of conditional samples
plt.figure(figsize=(6,4))
plt.hist(Y_cond, bins=40, density=True, alpha=0.7, label="Simulated
    density")
plt.axhline(2.0, color="red", linestyle="--", label="Theoretical
    density f(y|Y<1/2)=2")
plt.xlabel("y")
plt.ylabel("Density")
plt.title("Conditional distribution of Y given Y < 1/2")
plt.legend()
plt.tight_layout()
plt.show()
```

2.2.3 The Gaussian curve

We know meet the most important probability model of all. This is the Normal distribution, which has a density

$$g(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{19}$$

Despite the simplicity of the density function, calculating probabilities for Normal random variable by computing the area under the curve (integrating) is difficult. Instead we can remember some rough estimates based on the following figure. You should also be able to justify (to yourself) the bell curve shape by looking at the function. Hint: near $x = \mu$, the tangent line to $(x-\mu)^2$ is horizontal, then it decays exponentially. There is an inflection point, where is it?

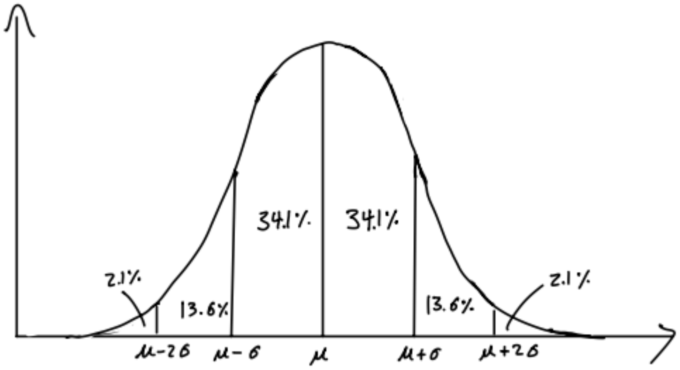

Figure 2: Probabilities in the Normal distribution

If X has the probability density $g(x)$ given in Eq. 19, then we write

$$X \sim \text{Normal}(\mu, \sigma^2) \tag{20}$$

It can be shown that $E[X] = \mu$ and $\text{var}(X) = \sigma^2$, hence the Normal distribution has the mean and variance. But BE CAREFUL: Sometimes (in code or in math) normal random variables are parameterized by the mean and variance, so one would instead write Normal(μ, σ). Both conventions are used, so always check.

Example 10 (Calculating probabilities for Normal distribution)**.** We use the curve above to calculate probabilities of events in the Normal distribution. Suppose

$$Y \sim \text{Normal}(5, 4)$$

Question: What is (approximately) $P(Y > 7)$?

Solution: Note that $5 + 2 = 7$, so this is asking how likely it is that a Normal variable is greater than 1 standard deviation above the mean. This about $13.5 + 2 = 15.5\%$. We can always easily compute these probabilities in python as well.

Question: What is

$$P(Y > 3|Y < 7)?$$

Solution: In this case we would use

$$P(Y > 3|Y < 7) = \frac{P(Y > 3, Y < 7)}{P(Y < 7)}$$

Notice that $3 = 5 - 2 = \mu - \sigma$ and we already saw $7 = 5 + 2 = \mu + \sigma$, so $P(Y < 7) \approx 0.839$ and $P(Y > 3, Y < 7) \approx 0.682$, thus the result is about 0.81.

2.2.4 The central limit theorem and sample distribution

We now have the formalism in place to state the Central Limit Theorem (CLT) in more precise terms.

Theorem 2. *Let X_i be a sequence of iid random variables and let*

$$E[X_i] = \mu, \quad \text{var}(X_i) = \sigma^2$$

and set

$$S_N = \sum_{i=1}^{N} X_i.$$

Then

$$P\left(\frac{S_N - N\mu}{\sqrt{N\sigma^2}} < z\right) \to P(Z < z) \tag{21}$$

Where

$$Z \sim \text{Normal}(0, 1)$$

The normal variable Z with zero mean and variance one is called a standard normal random variable. Since evaluations of the CDF of a standard normal random variable appear so often, we use the shorthand

$$\Phi(z) = P(Z < z)$$

and write $\phi(z)$ for the pdf.

Example 11 (Binomial)**.** Let

$$Y \sim \text{Binomial}(N, q)$$

Question: Assume N is even and use the central limit theorem to approximate $P(Y < N/2)$ with a Normal distribution. How does the accuracy depend on N and q?

Solution: Using that $\mu = E[X_i] = q$ and $\sigma^2 = \text{var}(X_i) = q(1-q)$, we find that the normal approximation to Y is

$$P\left(\frac{Y - Nq}{\sqrt{N\sigma^2}} < z\right) \to P(Z < z)$$

for

$$Z \sim \text{Normal}(0, 1).$$

Now we write

$$\begin{aligned}
P\left(Y < N/2\right) &= P\left(Y - Nq < N/2 - Nq\right) \\
&= P\left(\frac{Y - Nq}{\sqrt{Nq(1-q)}} < \frac{N/2 - Nq}{\sqrt{Nq(1-q)}}\right) \\
&= P\left(\frac{Y - Nq}{\sqrt{Nq(1-q)}} < \sqrt{N}\frac{1-2q}{2\sqrt{q(1-q)}}\right) \\
&\to P\left(Z < \sqrt{N}\frac{1-2q}{2\sqrt{q(1-q)}}\right) = \Phi\left(\sqrt{N}\frac{1-2q}{2\sqrt{q(1-q)}}\right)
\end{aligned}$$

We can compute this in Python, both by generating samples and using the CDF function.

```python
import numpy as np
from scipy.stats import norm, binom

# parameters
N, q = 200, 0.3
trials = 100000

# exact probability via binomial CDF
exact = binom.cdf(N//2 - 1, N, q)

# CLT approximation using Normal CDF (no continuity correction)
z = np.sqrt(N) * (1 - 2*q) / (2 * np.sqrt(q * (1 - q)))
clt_approx = norm.cdf(z)

# Monte Carlo estimate
samples = np.random.binomial(N, q, size=trials)
mc_estimate = np.mean(samples < N/2)

print(f"Exact:       {exact:.4f}")
print(f"CLT approx:  {clt_approx:.4f}")
print(f"Monte Carlo: {mc_estimate:.4f}")
```

Note on iid assumption: One of the most important things to recognize about the CLT when it comes to application is that the assumptions that the X_i are independent are not that important, so long as they are not too correlated. Even though the precise quantitive statement of the CLT won't when there are correlations, the sum will still be well approximated by a Normal distribution.

2.2.5 Properties of Normal random variables

Linear transformations of Normal random variables: Suppose

$$Z \sim \text{Normal}(0,1)$$

and define

$$X = \sigma Z + \mu \tag{22}$$

Then

$$\begin{aligned} P(X < x) &= P(\mu + \sigma Z < x) = P\left(Z < \frac{x-\mu}{\sigma}\right) \\ &= \int_{-\infty}^{\frac{x-\mu}{\sigma}} \phi(z) dz \end{aligned}$$

Now set

$$u = \mu + \sigma z \implies dz = \frac{du}{\sigma}.$$

When $z \to -\infty$, we have $u \to -\infty$, and when $z = \frac{x-\mu}{\sigma}$, we have $u = x$. Therefore,

$$\begin{aligned} P(X < x) &= \int_{-\infty}^{x} \phi\left(\frac{u-\mu}{\sigma}\right) \frac{1}{\sigma} du \\ &= \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(u-\mu)^2}{2\sigma^2}\right) du. \end{aligned}$$

We have shown that

$$X \sim \text{Normal}(\mu, \sigma^2).$$

when X is given by Eq. 22. In particular, any Normal random variable is obtained via a linear transformation of a standard normal.

With this understanding of how to linearly transform a Normal random variable, we can see that the CLT can be informally stated as

$$S_N \approx S_{\text{CLT}} \sim \text{Normal}(N\mu, N\sigma^2)$$

More generally,

$$X \sim \text{Normal}(\mu_x, \sigma_x)$$

Now consider

$$Y = aX + b$$

At this point it should make sense that Y is also normal. Taking the average of both sides,

$$E[Y] = a\mu + b$$

and

$$\text{var}(Y) = \text{var}(aX) + \text{var}(b)$$

Form the formula for variance, we know $\text{var}(aX) = a^2\text{var}(X)$. Also, $\text{var}(b) = 0$ So

$$Y \sim \text{Normal}(a\mu_x + b, a^2\sigma_x^2).$$

Note that in going from Z to X and X to Y, we are just multiplying and shifting everything. Think about what this does to the histogram. The process of going from X to Z is called standardizing. For any variable X the standardized variable is defined as

$$Z = \frac{X - \mu_x}{\sigma_x}$$

Transforming X to a standard Normal is equivalent to measuring X in units of standard deviations. For example, if we make a histogram of X, all this transformation does is change the X axis to units of standard deviations from the mean.

Theorem 3 (Special case of Theorem 4.6.1 in [Evans and Rosenthal, 2004]). *Let*

$$X_1 \sim \text{Normal}(\mu_1, \sigma_1^2)$$
$$X_2 \sim \text{Normal}(\mu_2, \sigma_2^2)$$

be independent, then

$$aX_1 + bX_2 + d \sim \text{Normal}\left(a\mu_1 + b\mu_2 + d, a^2\sigma_1^2 + b^2\sigma_2^2\right)$$

2.3 Examples of linear regression models

Equipped with the normal distribution and it properties, we can begin to explore linear regression models, but we focus on properties of the model at the moment and not statistical inference aspect until the next unit. A (single-predictor) linear regression model is a model of the form

$$Y|X \sim \text{Normal}(\beta_0 + \beta_1 X, \sigma^2). \tag{23}$$

here X is a called a predictor and Y is the response variable. We begin with some examples.

Example 12 (Linear regression with a binary predictor: the difference of means). Let

$$X \sim \text{Bernoulli}(q)$$
$$Y|X \sim \text{Normal}(\beta_1 X + \beta_0, \sigma^2)$$

Question: What is the marginal mean of Y? How would you estimate β_1? (not using least squares if you know it)

Solution: Another way to write the model is

$$Y = \beta_1 X + \beta_0 + \sigma Z \tag{24}$$

and taking the expectation gives

$$E[Y] = \beta_1 q + \beta_0 \tag{25}$$

You might notice that

$$\beta_1 = E[Y|X = 1] - E[Y|X = 0] \tag{26}$$

hence, we could estimate β_1 by computing the conditional mean within each group.

Example 13 (Linear regression model with Normal predictor). Let

$$X \sim \text{Normal}(\mu_x, \sigma_x^2)$$
$$Y|X \sim \text{Normal}(\beta_1 X + \beta_0, \sigma^2)$$

Question: What is the marginal distribution of Y? What is $E[XY]$? How does this compare to $E[X]E[Y]$?

Solution: We know that

$$Y|X = \beta_1 X + \beta_0 + Z, \quad Z \sim \text{Normal}(0, \sigma^2)$$

Thus, the marginal distribution of Y is the sum of two Normal random variables with mean and variance $(\beta_1\mu_x + \beta_0, a\sigma_x^2)$ and $(0, \sigma^2)$ respectively. As we saw in the previous unit

$$Y \sim \text{Normal}(\beta_1\mu_x + \beta_0, \beta_1^2\sigma_x^2 + \sigma^2)$$

To compute $E[XY]$, we note that

$$E[XY|X=x] = E[xY|X=x] = xE[Y|X=x]$$

therefore

$$E[XY] = E[XE[Y|X]] = E[X(\beta_1 X + \beta_0)] = \beta_1 E[X^2] + \beta_0 E[X]$$

Using

$$E[X^2] = \text{var}(X) + E[X]^2 = \sigma_x^2 + \mu_x^2$$

Therefore

$$E[XY] = \beta_1\sigma_x^2 + \beta_1\mu_x^2 + \beta_0\mu_x$$

On the other hand,

$$E[X]E[Y] = \mu_x(\beta_1\mu_x + \beta_0) = \beta_1\mu_x^2 + \beta_0\mu_x$$

The difference between the two is the additional term $\beta_1\sigma_x^2$, which we picked up from the variance of x.

A Additional discussion of continuous random variables (optional)

A.1 Exponential distribution

Suppose we want to model that time before a component of a machine fails. We will assume that the rate of failure – that is, the chance that it fails per unit time – is a constant λ. In other words, for a small time interval dt, the probability for the component to fail in a small time interval $[t, t+dt)$ given that it has not yet failed is λdt. Or, in mathematical notation If T is the time of failure, then the density of $f_T(t)$ is

$$f_T(t) = \lambda e^{-\lambda t}.$$

T is an exponentially distributed random variable, and we write

$$T \sim \text{Exponential}(\lambda).$$

An exponential variable has mean $E[T] = 1/\lambda$ and variance $\text{var}(T) = 1/\lambda^2$.

Example 14 (Heterogeneous failure rate)**.** Suppose that the machine is defective with probability 0.1. We can introduce a variable X which indicates whether the machine is defective and will fail with a rate 10. In other words, our model is

$$X \sim \text{Bernoulli}(0.1)$$
$$T|(X=x) \sim \text{Exponential}(x10 + (1-x))$$

Question: What is $E[T]$? What about $\text{var}(T)$? Does T follow an exponential distribution?

Solution: Using the tower property of expectation

$$\begin{aligned} E[T] &= E[E[T|X]] \\ &= E[T|X=0]P(X=0) + E[T|X=1]P(X=1) \\ &= 1\cdot(1-0.1) + \frac{0.1}{10} = 0.9 + 0.01 = 0.91 \end{aligned}$$

The variance

$$\text{var}(T) = E[T^2] - E[T]^2$$

and

$$E[T^2] = E[E[T^2|X]] = (1-0.1) \times E[T^2|X=0] + 0.1 \times E[T^2|X=1]$$

Note that, from the variance formula and the fact that T is exponential,

$$E[T^2|X=1] = \mathrm{var}(T|X=1) + E[T|X=1]^2 = \frac{1}{10^2} + \frac{1}{10^2} = \frac{2}{10^2}$$

Hence

$$E[T^2] = E[E[T^2|X]] = 0.9 \cdot 2 + 0.1 \cdot \frac{2}{10^2} = 1.802$$

$$\mathrm{var}(T) = 1.802 - 0.91^2 = 0.9739$$

If T is exponential, then

$$\lambda = \frac{1}{E[T]} = \frac{1}{0.91}.$$

We know that $\mathrm{var}(T) = E[T]^2$ for an exponential distribution, but

$$\frac{1}{\lambda^2} = 0.91^2 = 0.8281 \neq 0.9739 = \mathrm{var}(T).$$

A.2 Conditional probability and expectation with Continuous variables

The definition of expected value can be generalized to continuous variables by replacing the sums with integrals. That is, for a variable X with density f_X, we have

$$E[X] = \int x f_X(x) dx$$

Suppose X and Y are two variables on the sample spaces $S_X = S_Y = \mathbb{R}$. Then we can define a joint density $f_{X,Y}(x,y)$. From this, we can compute things like

$$P(X > x, Y > y) = \int_x^\infty \int_y^\infty f_{X,Y}(x,y) dx dy$$

If X and Y are independent, then $f_{X,Y}(x,y) = f_X(x) f_Y(y)$ and

$$\begin{aligned} P(X > x, Y > y) &= \int_x^\infty \int_y^\infty f_{X,Y}(x,y) dx dy \\ &= \int_x^\infty f_X(x) dx \int_y^\infty f_Y(y) dy = P(X > x) P(Y > y) \end{aligned}$$

Exercises

Exercise 1 (Computing conditional averages ❑)**:** Suppose we have some data representing samples of a pair of random variables (Y_1, Y_2):

$$\{(1,2),(1,2),(3,1),(1,4),(3,3),(2,2),(1,5)\}$$

Compute the following both by hand and with Python.

(a) $E[Y_1]$

(b) $E[Y_1|Y_2 = 2]$

(c) $E[Y_2|Y_1 = 1]$

(d) $E[Y_2|Y_1 > 1]$

Exercise 2 (❑)**:** Do Exercises 3.1.3, 3.1.4, 3.1.10, 3.1.14 in [Evans and Rosenthal, 2004] and for each one check your answer using simulations.

Exercise 3 (Independence and conditional expectation ❑)**:** Let X and Y be two random variables with (discrete) sample spaces S_X and S_Y. (you can find these in the textbook, but give them a try yourself first).

(a) Show that if X and Y are independent $E[X|Y = y] = E[X]$ and $E[Y|X = x] = E[Y]$ for all $x \in S_X$ and $y \in S_Y$. You may assume S_X and S_Y have a finite number of elements, e.g. $S_X = \{1, 2, 3, 4\}$.

(b) Prove the tower property of expectation, which says that

$$E[X] = \sum_{y \in S_Y} E[X|Y = y]P(Y = y)$$

This is sometimes stated as $E[X] = E[E[X|Y]]$ where the inner expectation is interpreted as a random variable depending on the value of Y.

(c) Show that if X and Y are independent, then

$$\mathrm{var}(X + Y) = \mathrm{var}(X) + \mathrm{var}(Y)$$

Exercise 4 (Verifying variance formula for Bernoulli variable ❑)**:** Using Python, verify the formula for the variance

$$\mathrm{Var}(Y) = q(1 - q)$$

This is a bit vague, but part of the exercise is to think about how you might justify this formula. In particular, what type of plot should you make? Use AI to help with coding if needed, but think about what you are trying to achieve first.

Exercise 5 (Conditioning with continuous variables ❑)**:** Let

$$Z_1 \sim \mathrm{Normal}(0, 1)$$
$$Z_2 \sim \mathrm{Normal}(1, 2)$$

Compute each of the following using Python

(a) $P(Z_1 + Z_2 > 3)$

(b) $P(Z_1 + Z_2 > 3|Z_1 < -1)$

(c) $P(Z_2 Z_1 > 0|Z_1 + Z_2 < 4)$

Exercise 6 (General linear transformation)**:** Suppose $X \sim \text{Normal}(5, 9)$. Define Y = -2X + 7.

(a) Find the distribution of Y (mean and variance).

(b) Estimate $P(Y > 0)$ by hand and check with simulations.

Exercise 7: Do Exercise 2.4.2 in [Evans and Rosenthal, 2004] using simulations. You can also check your answer using calculus if you wish.

Exercise 8: The **random walk** is a foundational model in nearly every area of science. It describes the "motion" of a variable which moves randomly over time without any memory of its past. Einstein developed a theory of the motion of microscopic particles based on random walks and they have been used as rudimentary models of stock prices.

We can define a random walk as follows. Let $X_0 = 0$ and define X_k for $k = 1, 2, 3, \dots$ by the recursive formula

$$X_{k+1} = X_k + \Delta(2U_k - 1) \tag{27}$$

where Δ is a constant and

$$U_k \sim \text{Bernoulli}(1/2)$$

are iid random variables.

We can think of X_k as the position of a person who is randomly walking with 50-50 chance of the moving to the left or right by Δ at each time-step. The entire sequence $X_0, X_1, X_2, \dots$ is referred to as the path of the random walker.

(a) Write a python function simulaterw(Delta,K) which simulates a random walk for N steps. Yours code should return the entire path in a numpy array. Make some plots of X_k vs. k.

(b) What are $E[X_k|X_{k-1} = 2]$ and $E[X_k]$?

(c) Using the central limit theorem, derive an approximation of the **mean squared displacement**

$$\text{MSD}(X_k) = E[X_k^2]$$

(you might notice this is just another name for the variance that is used in the context of random walks) Verify your approximation by plotting $\text{MSD}(X_k)$ as a function of N.

References

[Evans and Rosenthal, 2004] Evans, M. J. and Rosenthal, J. S. (2004). Probability and statistics: The science of uncertainty. Macmillan.