

# Week 12 notes

## Summations

### 01 Theory

In many contexts it is useful to consider random variables that are summations of a large number of variables.

#### ☞ Summation formulas: $E[X]$ and $\text{Var}[X]$

Suppose  $X$  is a large sum of random variables:

$$X = X_1 + X_2 + \cdots + X_n$$

Then:

$$E[X] = E[X_1] + E[X_2] + \cdots + E[X_n]$$

$$\text{Var}[X] = \text{Var}[X_1] + \cdots + \text{Var}[X_n] + 2 \sum_{i < j} \text{Cov}[X_i, X_j]$$

If  $X_i$  and  $X_j$  are uncorrelated (e.g. if they are independent):

$$\text{Var}[X] = \text{Var}[X_1] + \cdots + \text{Var}[X_n]$$

#### ☞ Extra - Derivation of variance of a sum

Using the definition:

$$\begin{aligned} \text{Var}[X_1 + \cdots + X_n] &= E\left[(X_1 + \cdots + X_n - (\mu_{X_1} + \cdots + \mu_{X_n}))^2\right] \\ &= E\left[\left((X_1 - \mu_{X_1}) + \cdots + (X_n - \mu_{X_n})\right)^2\right] \\ &= E\left[\sum_{i,j} (X_i - \mu_{X_i})(X_j - \mu_{X_j})\right] \\ &= \sum_{i,j} \text{Cov}(X_i, X_j) \\ &= \sum_i \text{Var}[X_i] + 2 \sum_{i < j} \text{Cov}[X_i, X_j] \end{aligned}$$

In the last line we use the fact that  $\text{Cov}[X, X] = \text{Var}[X]$  for the first term, and the symmetry property of covariance for the second term with the factor of 2.

### 02 Illustration

#### ☞ Example - Binomial expectation and variance

Suppose we have repeated Bernoulli trials  $X_1, \dots, X_n$  with  $X_i \sim \text{Ber}(p)$ .

The sum is a binomial variable:  $S_n = \sum_{i=1}^n X_i$ .

We know  $E[X_i] = p$  and  $\text{Var}[X_i] = pq$ .

The summation rule for expectation:

$$E[S_n] = \sum_{i=1}^n E[X_i] \gg \gg \sum_{i=1}^n p \gg \gg np$$

The summation rule for variance:

$$\begin{aligned} \text{Var}[S_n] &= \sum_{i=1}^n \text{Var}[X_i] + 2 \sum_{i < j} \text{Cov}[X_i, X_j] \\ &\gg \gg \sum_{i=1}^n pq + 2 \cdot 0 \gg \gg npq \end{aligned}$$

### ≡ Example - Pascal expectation and variance

Let  $X \sim \text{Pasc}(\ell, p)$ .

Let  $X_1, \dots, X_\ell$  be independent random variables, where:

- $X_1$  counts the trials until the first success
- $X_2$  counts the trials *after* the first success until the *second* success
- $X_i$  counts the trials after the  $(i-1)^{\text{th}}$  success until the  $i^{\text{th}}$  success

Observe that  $X = \sum_{i=1}^n X_i$ .

Notice that  $X_i \sim \text{Geom}(p)$  for every  $i$ . Therefore:

$$E[X_i] = \frac{1}{p} \quad \text{Var}[X_i] = \frac{1-p}{p^2}$$

Using linearity, conclude:

$$E[X] = \frac{k}{p} \quad \text{Var}[X] = \frac{kq}{p^2}$$

### ≡ Example - Multinomial covariances

Each trial of an experiment has possible outcomes labeled  $1, \dots, r$  with probabilities of occurrence  $p_1, \dots, p_r$ . The experiment is run  $n$  times.

Let  $X_i$  count the number of occurrences of outcome  $i$ . So  $X_i \sim \text{Bin}(n, p_i)$ .

Find  $\text{Cov}[X_i, X_j]$ .

#### Solution

Notice that  $X_i + X_j$  is also a binomial variable with success probability  $p_i + p_j$ . ('Success' is an outcome of either  $i$  or  $j$ .)

The variance of a binomial is known to be  $npq$  for whatever relevant  $p$  and  $q = 1 - p$ .

So we compute  $\text{Cov}[X_i, X_j]$  by solving:

$$\begin{aligned}\text{Var}[X_i + X_j] &= \text{Var}[X_i] + \text{Var}[X_j] + 2\text{Cov}[X_i, X_j] \\ n(p_i + p_j)(1 - (p_i + p_j)) &= np_i(1 - p_i) + np_j(1 - p_j) + 2\text{Cov}[X_i, X_j] \\ \gg \gg \quad \text{Cov}[X_i, X_j] &= -np_i p_j\end{aligned}$$

### ≡ Example - Hats in the air

All  $n$  sailors throw their hats in the air, and catch a random hat when they fall down.

How many sailors do you expect will catch the hat they own?

What is the variance of this number?

#### Solution

Strangely, the answers are both 1, regardless of the number of sailors. Here is the reasoning:

(1) Let  $X_i$  be an indicator of sailor  $i$  catching their own hat. So  $X_i = 1$  when sailor  $i$  catches their own hat, and  $X_i = 0$  otherwise. Thus  $X_i$  is Bernoulli with success probability  $1/n$ .

Then  $X = \sum_{i=1}^n X_i$  counts the total number of hats caught by original owners.

(2) Note that  $E[X_i] = 1/n$ .

Therefore:

$$E[X] \gg \sum_{i=1}^n E[X_i] \gg \sum_{i=1}^n \frac{1}{n} \gg 1$$

(3) Similarly:

$$\text{Var}[X] \gg \gg \sum_{i=1}^n \text{Var}[X_i] + 2 \sum_{i < j} \text{Cov}[X_i, X_j]$$

We need  $\text{Var}[X_i]$  and  $\text{Cov}[X_i, X_j]$ .

(4) Use  $\text{Var}[X_i] = E[X_i^2] - E[X_i]^2$ . Observe that  $X_i^2 = X_i$ . Therefore:

$$\text{Var}[X_i] \gg \gg \frac{1}{n} - \frac{1}{n^2} \gg \gg \frac{n-1}{n^2}$$

(5) Now for covariance:

$$\text{Cov}[X_i, X_j] = E[X_i X_j] - E[X_i]E[X_j]$$

We need to compute  $E[X_i X_j]$ .

Notice that  $X_i X_j = 1$  when  $i$  and  $j$  *both* catch their own hats, and 0 otherwise.

We have:

$$\begin{aligned} P[X_i = 1 \text{ and } X_j = 1] &= \frac{1}{n(n-1)} \\ \gg \gg E[X_i X_j] &= \frac{1}{n(n-1)} \end{aligned}$$

Therefore:

$$\text{Cov}[X_i, X_j] \gg \gg \frac{1}{n(n-1)} - \frac{1}{n} \cdot \frac{1}{n} \gg \gg \frac{1}{n^2(n-1)}$$

(6) Putting everything together back in (1):

$$\begin{aligned} \text{Var}[X] &\gg \gg \sum_{i=1}^n \text{Var}[X_i] + 2 \sum_{i < j} \text{Cov}[X_i, X_j] \\ &\gg \gg \sum_{i=1}^n \frac{n-1}{n^2} + 2 \sum_{i < j} \frac{1}{n^2(n-1)} \\ &\gg \gg \frac{n-1}{n} + n(n-1) \frac{1}{n^2(n-1)} \gg \gg 1 \end{aligned}$$

Suppose study groups of 10 are formed from a large population.

For a typical study group, how many months out of the year contain a birthday of a member of the group? (Assume the 12 months have equal duration.)

### Solution

Let  $X_i$  be 1 if month  $i$  contains a birthday, and 0 otherwise.

So we seek  $E[X_1 + \dots + X_{12}]$ . This equals  $E[X_1] + \dots + E[X_{12}]$ .

The answer will be  $12E[X_i]$  because all terms are equal.

For a given  $i$ :

$$P[\text{no birthday in month } i] = \left(\frac{11}{12}\right)^{10}$$

The complement event:

$$P[\text{at least one birthday in month } i] = 1 - \left(\frac{11}{12}\right)^{10}$$

Therefore:

$$12E[X_i] = 12 \left(1 - \left(\frac{11}{12}\right)^{10}\right) \gg \gg \text{6.97}$$

## Central Limit Theorem

### 03 Theory

#### IID variables

Random variables are called **independent, identically distributed** when they are independent and have the same distribution.

#### IID variables: Same distribution, different values

Independent variables cannot be correlated, so the values taken by IID variables will disagree on all (most) outcomes.

We do have:

$$\text{same distribution} \iff \text{same PMF or PDF}$$

### 📊 Standardization

Suppose  $X$  is any random variable.

The **standardization** of  $X$  is:

$$Z = \frac{X - \mu_X}{\sigma_X}$$

The variable  $Z$  has  $E[Z] = 0$  and  $\text{Var}[Z] = 1$ . We can reconstruct  $X$  by:

$$X = \sigma_X Z + \mu_X$$

Suppose  $X_1, X_2, \dots, X_n$  is a collection of IID random variables.

Define:

$$S_n = \sum_{i=1}^n X_i \quad Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

where:

$$\mu = E[X_i] \quad \sigma^2 = \text{Var}[X_i] \quad (\text{every } i)$$

So  $Z_n$  is the standardization of  $S_n$ .

Let  $Z$  be a standard normal random variable,  $Z \sim \mathcal{N}(0, 1)$ .

### 📋 Central Limit Theorem

Suppose  $S_n = \sum_{i=1}^n X_i$  for IID variables  $X_i$ , and  $Z_n$  are the standardizations of  $S_n$ .

Then for any interval  $[a, b] \subset \mathbb{R}$ :

$$\lim_{n \rightarrow \infty} P[Z_n \in [a, b]] = P[Z \in [a, b]] = \Phi(b) - \Phi(a)$$

We say that  $Z_n$  *converges in probability* to the standard normal  $Z$ .

Here is a good [explainer video](#) by 3blue1brown.

The distribution of *a very large sum* of IID variables is determined merely by  $\mu$  and  $\sigma^2$  from the original IID variables, while the data of higher moments fades away.

The name “**normal distribution**” is used because it arises from a large sum of repetitions of *any* other kind of distribution. It is therefore ubiquitous in applications.

### ⚠ Misuse of the CLT

It is important to learn when the CLT is applicable and when it is not. Many people (even professionals) apply it wrongly.

For example, sometimes one hears the claim that *if enough students take an exam, the distribution of scores will be approximately normal*. This is totally wrong!

### Extra - Derivation of CLT

[Derivation of Central Limit Theorem](#)

## 04 Illustration

### Exercise - Test scores distribution

Explain what is wrong with the claim that test scores should be normally distributed when a large number of students take a test.

Can you imagine a scenario with a *good* argument that test scores would be normally distributed?

(Hint: think about the composition of a single test instead of the number of students taking the test.)

### Exercise - Height follows a bell curve

The height of female American basketball players follows a bell curve. Why?

## 05 Theory

Normal approximations rely on the limit stated in the CLT to approximate probabilities for large sums of variables.

### Normal approximation

Let  $S_n = X_1 + \dots + X_n$  for IID variables  $X_i$  with  $\mu = E[X_i]$  and  $\sigma^2 = \text{Var}[X_i]$ .

The **normal approximation** of  $S_n$  is:

$$F_{S_n}(s) \approx \Phi\left(\frac{s - n\mu}{\sigma\sqrt{n}}\right)$$

For example, suppose  $X_i \sim \text{Ber}(p)$ , so  $S_n \sim \text{Bin}(n, p)$ . We know  $\mu = p$  and  $\sigma = \sqrt{pq}$ . Therefore:

$$F_{S_n}(s) \approx \Phi\left(\frac{s - np}{\sqrt{npq}}\right)$$

A rule of thumb is that the normal approximation to the binomial is effective when  $npq > 10$ .

### 🔥 Efficient computation

This CDF is *far* easier to compute for large  $n$  than the CDF of  $S_n$  itself. The factorials in  $\binom{n}{k}$  are hard even for a computer when  $n$  is large, and the summation adds another  $n$  factor to the scaling cost.

## 06 Illustration

### ≡ Example - Binomial estimation: 10,000 flips

Flip a fair coin 10,000 times. Write  $H$  for the number of heads.

Estimate the probability that  $4850 < H < 5100$ .

#### Solution

Check the rule of thumb:  $p = q = 0.5$  and  $n = 10,000$ , so  $npq = 2500 \gg 10$  and the approximation is effective.

Now, calculate needed quantities:

$$\mu = E[X_i] \gg \mu = 0.5 \gg n\mu = 5000$$

$$\sigma^2 = \text{Var}[X_i] \gg \sigma = 0.5 \gg \sigma\sqrt{n} = 50$$

Set up CDF:

$$F_H(h) = \Phi\left(\frac{h - 5000}{50}\right)$$

Compute desired probability:

$$P[4850 < H < 5100] = F_H(5100) - F_H(4850)$$

$$\gg \Phi\left(\frac{100}{50}\right) - \Phi\left(\frac{-150}{50}\right) \gg \Phi(2) - \Phi(-3)$$

$$\gg \approx 0.9772 - (1 - 0.9987) \gg \mathbf{0.9759}$$



### ≡ Example - Summing 1000 dice

About 1,000 dice are rolled.

Estimate the probability that the total sum of rolled numbers is more than 3,600.

#### Solution

Let  $X_i$  be the number rolled on the  $i^{th}$  die.

Let  $S = \sum_{i=1}^n X_i$ , so  $S$  counts the total sum of rolled numbers.

We seek  $P[S \geq 3600]$ .

Now, calculate needed quantities:

$$\mu = E[X_i] \gg \mu = 7/2 \gg n\mu = 3500$$

$$\sigma^2 = \text{Var}[X_i] \gg \sigma = \sqrt{\frac{35}{12}} \gg \sigma\sqrt{n} = \sqrt{\frac{35000}{12}}$$

Set up CDF:

$$F_S(s) = \Phi\left(\frac{s - 3500}{\sqrt{\frac{35000}{12}}}\right)$$

Compute desired probability:

$$\begin{aligned} P[S \geq 3600] &= F_S(3600) \\ &\gg \gg \Phi\left(\frac{100}{54.01}\right) \gg \gg \Phi(1.852) \approx 0.03201 \end{aligned}$$

### ≡ Exercise - Estimating $S_{1000}$

The odds of a random poker hand containing one pair is 0.42.

Estimate the probability that at least 450 out of 1000 poker hands will contain one pair.

### ≡ Exercise - Nutrition study

A nutrition review board will endorse a diet if it has any positive effect in at least 65% of those tested in a certain study with 100 participants.

Suppose the diet is bogus, but 50% of participants display some positive effect by pure chance.

What is the probability that it will be endorsed?

**Answer**

$$0.0019 = 1 - \Phi(2.9)$$

## 07 Theory

### 📖 De Moivre-Laplace Continuity Correction Formula

The normal approximation to a discrete distribution, for *integers  $a$  and  $b$  close together*, should be improved by adding 0.5 to the range on either side:

$$\begin{aligned} P[a \leq S_n \leq b] &\approx P[a - 0.5 \leq \sigma\sqrt{n}Z + n\mu \leq b + 0.5] \\ &\approx \Phi\left(\frac{b + 0.5 - n\mu}{\sigma\sqrt{n}}\right) - \Phi\left(\frac{a - 0.5 - n\mu}{\sigma\sqrt{n}}\right) \end{aligned}$$

## 08 Illustration

### ≡ Example - Continuity correction of absurd normal approximation

Let  $S_n$  denote the number of sixes rolled after  $n$  rolls of a fair die. Estimate  $P[S_{720} = 113]$ .

**Solution**

We have  $S_n \sim \text{Bin}(720, 1/6)$ , and  $np = 120$  and  $\sqrt{npq} = 10$ .

The usual approximation, since  $Z$  is continuous, gives an estimate of 0, which is useless.

Now using the continuity correction:

$$\begin{aligned} &P[113 \leq S_{720} \leq 113] \\ &\approx \Phi\left(\frac{113 + 0.5 - 120}{10}\right) - \Phi\left(\frac{113 - 0.5 - 120}{10}\right) \\ &\approx \Phi(-0.65) - \Phi(-0.75) \approx 0.0312 \end{aligned}$$

The exact solution is 0.0318, so this estimate is quite good: the error is 1.9%.