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 $\mathrm{Var}[X]$

Suppose X is a large sum of random variables:

$$X \;=\; X_1 + X_2 + \cdots + X_n \quad = \sum_{i=1}^{n} X_i$$

Then:

$$E[X] \;=\; E[X_1] + E[X_2] + \cdots + E[X_n] \quad = \sum_{i=1}^{n} E[X_i]$$

$$\mathrm{Cov}[X_1+\cdots+X_n,\, X_1+\cdots+X_n]$$

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

$$\xrightarrow{n=2} \mathrm{Var}[X_1] + \mathrm{Var}[X_2] + 2\,\mathrm{Cov}[X_1, X_2]$$

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

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

$$(x+y+z)^2 = x^2+y^2+z^2 + xy + xz + yx + yz + zx + zy$$

$$= x^2+y^2+z^2 + 2(xy + xz + yz)$$

$$(x_1+\dots x_n)^2 = x_1^2+\dots+x_n^2 \; 2\begin{pmatrix} x_1x_2 + x_1x_3 + \dots x_1x_n \\ + \; x_2x_3 + \dots x_2x_n \\ \vdots \\ x_{n-1}x_n \end{pmatrix}$$

$$= 2\sum_{i<j} x_i x_j$$

Extra - Derivation of variance of a sum

Using the definition:

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

In the last line we use the fact that $\mathrm{Cov}[X, X] = \mathrm{Var}[X]$ 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 $X_1, \ldots, X_n$ with $X_i \sim \text{Ber}(p)$.

The sum is a binomial variable: $S_n = \sum_{i=1}^{n} X_i$. $S_n \sim Bin(n, p)$

$$E[S_n^2] = \sum_{k=0}^{n} k^2 \binom{n}{k} p^k q^{n-k}$$

$$= npq + (np)^2$$

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

The summation rule for expectation:

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

The summation rule for variance:

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

$$\gg\gg \quad \sum_{i=1}^{n} pq + 2 \cdot 0 \quad \gg\gg \quad npq$$

Example - Pascal expectation and variance

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

Let $X_1, \ldots, X_\ell$ 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 $(i-1)^{th}$ success until the i^{th} success

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

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

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

Using linearity, conclude:

$$E[X] = \frac{\overset{\ell}{\cancel{k}}}{p} \qquad \text{Var}[X] = \frac{\overset{\ell}{\cancel{k}}q}{p^2}$$

Example - Multinomial covariances

$$\binom{n}{r_1\ r_2 \dots r_k} = \frac{n!}{r_1!\, r_2! \dots r_k!}$$

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.) $X_i + X_j \sim Bin(n, p_i + p_j)$

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:

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

Example - Hats in the air

$X = X_1 + \dots + X_n$

$X_i = \begin{cases} 1 & \text{"i" catches own hat} \\ 0 & \text{"i" catches other hat} \end{cases}$

All n sailors throws 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 $E[X_i] = 1/n$. $E[X_i] = 1 \cdot p + 0 \cdot q = 1 \cdot \frac{1}{n}$

Therefore:

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

---

(3) Similarly:

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

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

---

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

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

---

(5) Now for covariance:

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

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:

$\frac{1}{n-1} = P[X_j = 1 \mid X_i = 1]$

$$P[X_i = 1 \text{ and } X_j = 1] \quad = \quad \frac{1}{n(n-1)}$$

$= P[X_i = 1] \cdot P[X_j = 1 \mid X_i = 1]$

$P[A] \cdot P[B \mid A] = P[AB]$

$$\gg\gg \quad E[X_i X_j] \quad = \quad \frac{1}{n(n-1)}$$

Therefore:

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

---

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

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

$2\sum_{i<j} 1 = \sum_{i \neq j} 1 = \#\{(i,j) \mid i \neq j,\ i,j = 1, \dots, n\}$

$= n \cdot (n-1)$

$$\gg\gg \quad \sum_{i=1}^{n} \frac{n-1}{n^2} + 2\sum_{i<j} \frac{1}{n^2(n-1)}$$

$$\gg\gg \quad \frac{n-1}{n} + n(n-1)\frac{1}{n^2(n-1)} \quad \gg\gg \quad 1$$

Months with a birthday

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 $E[X_1 + \cdots + X_{12}]$. This equals $E[X_1] + \cdots + E[X_{12}]$.

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

For a given i:

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

The complement event:

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

Therefore:

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

Central Limit Theorem

03 Theory

IID variables

(IID)

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:

$$\text{same distribution} \quad \Longleftrightarrow \quad \text{same PMF or PDF}$$

Standardization

Suppose X is any random variable.

The **standardization** of X is:

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

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

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

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

Define:

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

where:

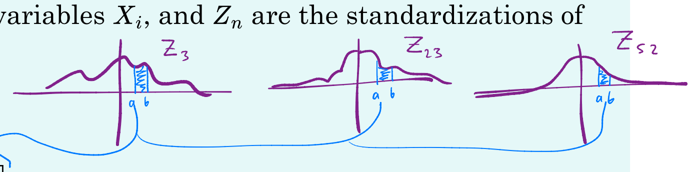
$$\mu \;=\; E[X_i] \qquad \sigma^2 \;=\; \mathrm{Var}[X_i] \qquad \text{(every } i\text{)}$$

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.



w/

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

$$\lim_{n\to\infty} P\Big[\, Z_n \in [a,b] \,\Big] \quad = \quad P\big[\, Z \in [a,b] \,\big] \;=\; \Phi(b) - \Phi(a)$$

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 μ and σ^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 $S_n = X_1 + \cdots + X_n$ for IID variables X_i with $\mu = E[X_i]$ and $\sigma^2 = \mathrm{Var}[X_i]$.

The **normal approximation** of S_n is:

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

For example, suppose $X_i \sim \mathrm{Ber}(p)$, so $S_n \sim \mathrm{Bin}(n, p)$. We know $\mu = p$ and $\sigma^2 = pq$. Therefore:

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

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.

$\sigma^2 = pq$

Now, calculate needed quantities:

$E[X_i] = \mu$
$E[S = \Sigma X_i] = n\mu$

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

$Var[X_i] = \sigma^2$
$Var[\Sigma X_i] = n\sigma^2$

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

$\sqrt{n\sigma} = \sigma\sqrt{n}$
$=$ std dev of ΣX_i

Set up CDF:

$$F_H(h) \quad \overset{\approx}{\neq} \quad \Phi\left(\frac{h-5000}{50}\right)$$

Compute desired probability:

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

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

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

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:

$E[X_i^2] - E[X_i]^2$

$\hookrightarrow 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + 3^2 \cdot \frac{1}{6} + \dots + 6^2 \cdot \frac{1}{6}$

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

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

Set up CDF:

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

Compute desired probability:

$$P[\, S \geq 3600 \,] \quad = 1 - F_S(3600)$$

$$\gg\gg \quad \Phi\left(\frac{100}{54.01}\right) \quad \gg\gg \quad \Phi(1.852) \approx \cancel{0.03201}$$

$1 - 3\% = 97\%$

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:

$$P[\,a \leq S_n \leq b\,] \quad \approx \quad P\left[\,a - 0.5 \leq \sigma\sqrt{n}\,Z + n\mu \leq b + 0.5\right]$$

I.e. use
[a-0.5, b+0.5]

$$\approx \Phi\left(\frac{b + 0.5 - n\mu}{\sigma\sqrt{n}}\right) - \Phi\left(\frac{a - 0.5 - n\mu}{\sigma\sqrt{n}}\right)$$

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:

$$P[\,113 \leq S_{720} \leq 113\,]$$

$$\approx \quad \Phi\left(\frac{113 + 0.5 - 120}{10}\right) - \Phi\left(\frac{113 - 0.5 - 120}{10}\right)$$

$$\approx \Phi(-0.65) - \Phi(-0.75) \approx 0.0312$$

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