W12 - Examples

Summations

Binomial expectation and variance

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

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

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

The summation rule for expectation:

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

The summation rule for variance:

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

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 $X_i \sim \mathrm{Bin}(n, p_i)$.

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

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 $\mathrm{Cov}[X_i, X_j]$ by solving:

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

Hats in the air

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 $E[X_i] = 1/n$.

Therefore:

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

---

(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:

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

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

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):

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

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

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

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:

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

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

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

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

Using linearity, conclude:

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

Central Limit Theorem

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.)

Height follows a bell curve

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

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:

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

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

Set up CDF:

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

Compute desired probability:

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

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

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

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:

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

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

Set up CDF:

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

Compute desired probability:

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

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

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

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

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 $S_n \sim \mathrm{Bin}(720, 1/6)$, 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%.