Занятие 18. 29 февраля

Иван Высоцкий

Три важных распределения

Задача 1. Игральную кость бросают 100 раз. Найдите математическое ожидание числа выпавших шестерок.
Задача 2. Игральную кость бросают до тех пор, пока не выпадет шестерка. Найдите математическое ожидание числа бросаний.
Задача 3. В отаре 240 овец. Известно, что ровно 40 из них – черные. Случайным образом отлавливаюти 100 овец. Найдите математическое ожидание числа отловленных черных овец.

Что общего в этих трех задачах? Все три эксперимента построены из отдельных испытаний с вероятностью успеха $p = 1/6$ в каждом испытании. На этом сходство более-менее заканчивается. Правда, между первой и третьей задачей можно уловить дополнительные общие черты: в обоих случаях речь идет о выборке объемом 100. Но есть и серьезное отличие. В первой задаче вероятность появления шестерки на каждой позиции одна и та же, а в третьей задаче вероятность появления черной овцы меняется по мере того, как мы узнаем цвета уже отловленных овец. Вероятность того, что первая выбранная овца будет черной, равна $40/240 = 1/6$, а затем, в зависимости от того, какая овца попалась первой, вероятность того, что вторая овца окажется черной, равна либо $39/239$, либо $40/239$.

Ясно, что мы имеем дело с тремя различными случайными величинами.

1. S «Число успехов в n одинаковых независимых испытаниях с вероятностью успеха p в каждом испытании».
2. X «Число попыток до появления первого успеха в серии одинаковых независимых испытаний с вероятностью успеха р в каждом испытании».
3. S «Число меченых элементов (успехов) в случайной выборке объема n из совокупности объема N, в которой содержится K меченых элементов».

Наша цель – найти распределение каждой из этих величин и ее математическое ожидание.

1. Биномиальное распределение. Пусть n – число испытаний в серии одинаковых независимых испытаний с вероятностью успеха p в каждом. Элементарный исход такого опыта – последовательность вида

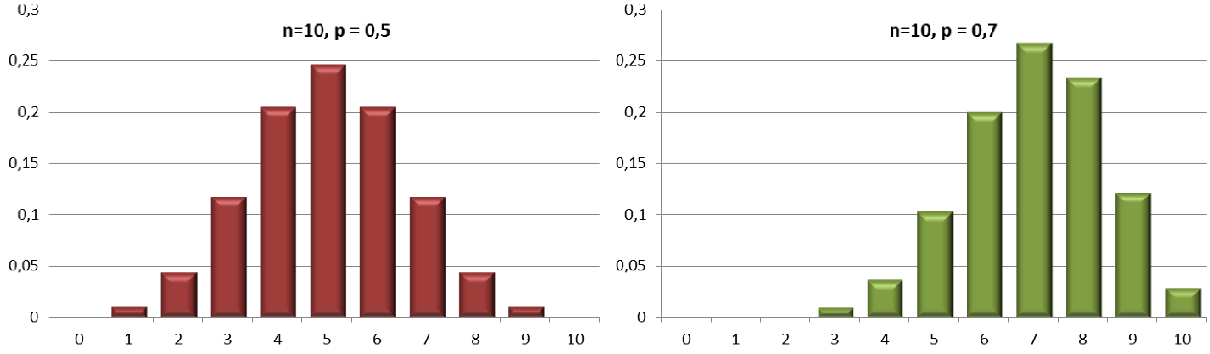
$$\underbrace{\text{УННУУУНУНУУУНУ}}_{n\ \text{букв}}$$

Буква У помечает успех, а буква Н – неудачу. Вероятность любой такой последовательности, в которой ровно k успехов и $n-k$ неудач, равна $p^k(1-p)^{n-k}$, а количество таких последовательностей равно C_n^k. Обозначая для краткости вероятность неудачи $1-p$ буквой q, получаем:

$$\mathrm{P}(S = k) = C_n^k p^k q^{n-k}.$$

Это – ***формула Бернулли***. Она дает распределение случайной величины S «Число успехов». Такое распределение называется ***биномиальным***. Числа n и p называются ***параметрами распределения***, вероятность каждого значения S зависит только от них.

Диаграмма биномиального распределения обычно имеет колоколообразную форму. На рисунке показаны диаграммы биномиальных распределений для $n=10$ и для двух различных значений p.



Перед разговором о математическом ожидании вернемся к задаче 1. Ясно, что при 100 бросках кости следует ожидать в среднем $100\cdot\frac{1}{6}=16\frac{2}{3}$ шестерок. Это интуитивное представление о доле успехов справедливо. Тогда в общем случае математическое ожидание равно np. Иными словами, должно быть верно равенство

$$\mathrm{E}S=0\cdot C_n^0p^0q^n+1\cdot C_n^1p^1q^{n-1}+2\cdot C_n^2p^2q^{n-2}+\ldots+n\cdot C_n^np^nq^0=pn.$$

Сразу скажем, что это можно вывести преобразованиями, упрощая левую часть. Но мы применим ***индикаторы***. Пусть I_k – индикатор события «В испытании номер k наступил успех». Распределение всех индикаторов одинаковое:

$$I_k\sim\begin{pmatrix}0 & 1\\ q & p\end{pmatrix}.$$

Следовательно, $\mathrm{E}I_k=0\cdot q+1\cdot p=p$. Общее число успехов это сумма числа успехов в каждом отдельном испытании: $S=I_1+I_2+I_3+\ldots+I_n$. Переходя к ожиданиям, получаем

$$\mathrm{E}S=\mathrm{E}I_1+\mathrm{E}I_2+\mathrm{E}I_3+\ldots+\mathrm{E}I_n=np.$$

2. Геометрическое распределение. Теперь число испытаний не определено. Опыт заканчивается, как только случился первый успех. Вероятность каждого из событий $X=k$ можно найти несложно:

$$\mathrm{P}(X=1)=p,\ \mathrm{P}(X=2)=qp,\ \mathrm{P}(X=3)=q^2p,\ \mathrm{P}(X=4)=q^3p,\ldots,\ \mathrm{P}(X=k)=q^{k-1}p,\ldots$$

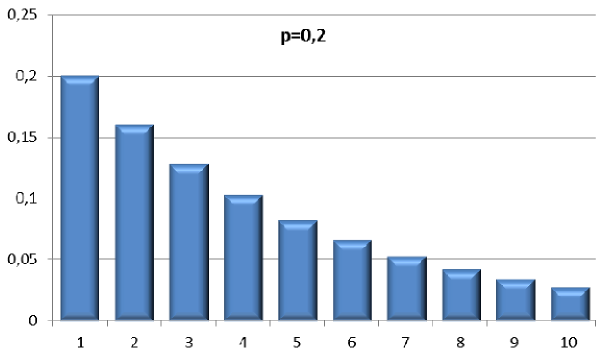
Эти формулы и дают нужное распределение. Вероятности образуют геометрическую прогрессию со знаменателем q, и сумма всех вероятностей должна равняться единице:

$$p+qp+q^2p+q^3p+\ldots=1,$$

даже несмотря на то, что слагаемых бесконечно много. Отсюда, кстати, следует важный факт алгебры – ***формула суммы геометрической прогрессии***:

$$1+q+q^2+q^3+q^4+\ldots=\frac{1}{p}=\frac{1}{1-q}.$$

Отсюда и название распределения случайной величины X – ***геометрическое распределение***. Единственный параметр распределения – вероятность успеха p. Диаграмма геометрического распределения при $p=0,2$ показана на рисунке.



А что с математическим ожиданием? Есть ли оно вообще у такого распределения? Поможет ли нам интуиция и в этом случае? Поможет. Вспомним задачу 2. Представим, что мы бросаем игральную кость много-много раз. Вероятность шестерки равна $1/6$, поэтому шестерка будет выпадать в среднем каждый шестой раз. Следовательно, среднее число бросков до выпадения шестерки равно шести.

В общем случае вероятность успеха равна p, поэтому математическое ожидание числа бросков до первого успеха равно $\mathrm{E}X=\frac{1}{p}$. Получается, что должно быть так:

$$\mathrm{E}X=p+2qp+3q^2p+4q^3p+\ldots=\frac{1}{p}.$$

Но так ли это на самом деле? Может, мы с нашей интуицией не учли какой-нибудь фактор? В левой части заменим p выражением $1-q$. Получим:

$$(1-q)+2q(1-q)+3q^2(1-q)+4q^3(1-q)+\ldots=1-q+2q-2q^2+3q^2-3q^3+4q^3-4q^4+\ldots$$

Сгруппируем слагаемые иначе:

$$1+2q-q+3q^2-2q^2+4q^3-3q^3+5q^4-4q^4+\ldots=1+q+q^2+q^3+q^4+\ldots$$

А эту сумму мы уже знаем: $\frac{1}{p}$.

В рассуждении мы допустили неаккуратность – переставляли местами слагаемые. Все знают, что от перестановки двух, трех или даже сотни слагаемых сумма не меняется. Но здесь мы переставляем бесконечное число слагаемых. Вообще говоря, от такого действия сумма может меняться и даже очень сильно (придумайте пример). Можно ли дать другое доказательство? Попробуем индикаторы. Пусть I_k – индикатор события «k испытаний окончились неудачей».

$$\mathrm{E}I_k=\mathrm{P}(I_k=1)=q^k.$$

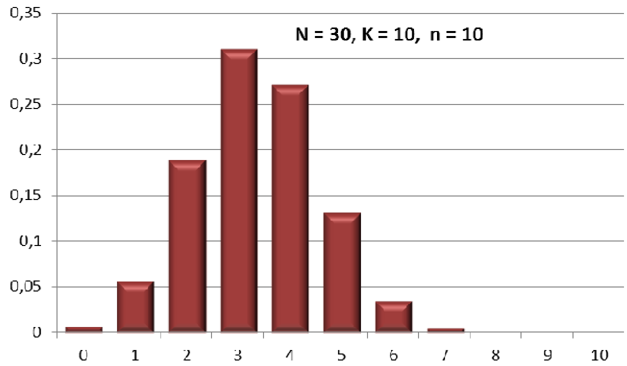
Общее число попыток на единицу больше числа неудачных попыток:

$$X=1+I_1+I_2+I_3+\ldots$$

Поэтому сразу получается

$$\mathrm{E}X=1+q+q^2+q^3+q^4+\ldots=\frac{1}{p}.$$

3. Гипергеометрическое распределение. Имеется множество из N объектов, из которых K – особые (например, черные овцы). Выбираем случайным образом n объектов. Какова вероятность того, что среди выбранных число S особых объектов (успехов) окажется равно k? Ясно, что выбрать k объектов из K и $n-k$ объектов из оставшихся $N-K$ можно $C_K^k C_{N-K}^{n-k}$ способами, тогда как общее число способов выбрать n объектов, равно C_N^n. Используя равновозможность всех комбинаций, получаем искомую вероятность:



$$\mathrm{P}(S=k)=\frac{C_K^k C_{N-K}^{n-k}}{C_N^n}.$$

Эта формула дает ***гипергеометрическое распределение*** случайной величины S.

Диаграмма гипергеометрического распределения для одного конкретного случая (параметры $N=30, K=10, n=10$) показана на рисунке. Как видим, диаграмма тоже колоколообразная, а при больших N разница между гипергеометрическим распределением и биномиальным распределением с параметром $p=\frac{K}{N}$ невелика.

Займемся математическим ожиданием. Снова попробуем применить интуицию. Кажется очевидным, что доля черных овец в выборке из 100 штук в среднем должна быть такой же, как доля черных овец во всей отаре. То есть, должно выполняться равенство $\frac{\mathrm{E}S}{100}=\frac{40}{240}$, откуда $\mathrm{E}S=100\cdot\frac{1}{6}$, а в общем случае $\mathrm{E}S=n\frac{K}{N}$, что очень напоминает формулу математического ожидания биномиального распределения. Иными словами, должно выполняться равенство

$$\mathrm{E}S=0\cdot\frac{C_K^0 C_{N-K}^{n-0}}{C_N^n}+1\cdot\frac{C_K^1 C_{N-K}^{n-1}}{C_N^n}+2\cdot\frac{C_K^2 C_{N-K}^{n-2}}{C_N^n}+\ldots+m\cdot\frac{C_K^m C_{N-K}^{n-m}}{C_N^n}=n\frac{K}{N}.$$

(число m – наибольшее возможное особых объектов в выборке, то есть минимальное из n и K, ведь число особых объектов в выборке не может превосходить ни объем выборки, ни общее число особых объектов во всей совокупности).

Пока что это равенство обосновано только нашей интуицией. Можно попытаться доказать его, преобразуя сумму в левой части, что весьма непросто. А индикаторы? Пронумеруем объекты выборки каким-либо образом. Например, в порядке отлова (если это овцы). Пусть I_k – индикатор события «k объект в выборке оказался особым». Очевидно,

$$\mathrm{E}I_k=\mathrm{P}(I_k=1)=\frac{K}{N}.$$

Дальше проделайте выкладки сами точно так же, как мы поступили в случае с биномиальным распределением. Убедитесь, что и на сей раз интуиция нас не подвела:

$$\mathrm{E}S=n\frac{K}{N}.$$

Мы познакомились с индикаторами в действии, при этом намеренно использовали индикаторы для подтверждения интуитивно ясных предположений. Однако, согласитесь, если бы вас попросили доказать любое из трех тождеств:

$$C_n^1 p^1 q^{n-1} + 2 \cdot C_n^2 p^2 q^{n-2} + \ldots + n \cdot C_n^n p^n q^0 = pn\,,$$

$$p + 2qp + 3q^2 p + 4q^3 p + \ldots = \frac{1}{p}$$

или

$$1 \cdot \frac{C_K^1 C_{N-K}^{n-1}}{C_N^n} + 2 \cdot \frac{C_K^2 C_{N-K}^{n-2}}{C_N^n} + \ldots + K \cdot \frac{C_K^K C_{N-K}^{n-K}}{C_N^n} = n\frac{K}{N}\,,$$

вы вряд ли с легкостью сумели бы справиться с этим, используя только алгебраические преобразования.

В других случаях метод индикаторов дает гораздо менее очевидные и ожидаемые результаты. На следующем занятии мы решим несколько задач, где индикаторы помогают в, казалось бы, совершенно безнадежной ситуации.