Topic 3: Random Variables and Important Probability Distributions

Optional Reading: Chapter 4

Xiner Zhou

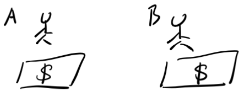
Department of Statistics

University of California, Davis

- What is Random Variable and its Distribution?
- Discrete random variable
- Important discrete distribution
- Continuous random variable
- Important continuous distribution

We have been using “verbal description” about experiment & event, when calculating probabilities.

But such tedious verbal description soon becomes convoluted and obscure how the quantities of interest are related.



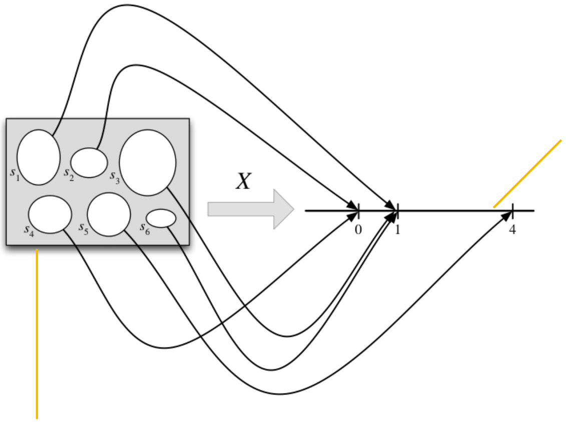
e.g. gambler A’s from Europe, gambler B is from US, what’s the probability that gambler A’s wealth is more than the gambler B’s wealth?

=> P(gambler A’s wealth * exchange rate < gambler B’s wealth)

But, wouldn’t be nice if we could just say something like this? $P(X_A < X_B)$

The notion of random variable will allow us to do exactly this!

Given an experiment with sample space S, a random variable (r.v.) is a function from the sample space S to the real numbers.



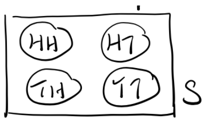
r.v. X records the assigned value

Sample space has 6 elements/simple events, each has an assigned value from {0,1,4} coming from some characteristics that we are interested in

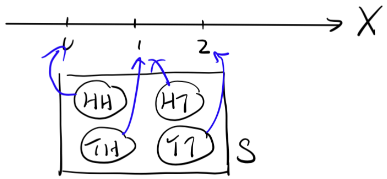
- The randomness in random variable come from the fact that we have a random experiment
- This definition is abstract but fundamental

Each r.v. is a numerical summary of some aspect of experiment that we are interested to study

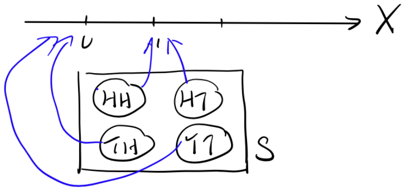
- (Two coins) Consider an experiment where we toss a fair coin twice.



· X = number of tails



· $X = \begin{cases} 1, & \text{if first toss is head} \\ 0, & \text{otherwise} \end{cases}$



- Examples:

  - Number of defects on a randomly selected piece of furniture

  - SAT score for a randomly selected college applicant

  - Number of telephone calls received by a crisis intervention hotline during a randomly selected time period

We assign each possible outcome of an experiment a number, this quantification process results in random variables.

Associated with each value of a random variable is the "probability" or "likelihood" of the random variable taking that value.

This information is contained in the probability distribution of the random variable.

For a random variable, X, its probability distribution lists:

- What values x of X can occur
- The chance (probability or likelihood) of each value x that can occur.

Once we know the probability distribution of the random variable, probability of any events can be calculated quite straightforward!

The possible set of values that a random variable could take could be either discrete or continuous.

- A r.v. X is said to be discrete if:

  X can take finite number, or infinite by countably many (1,2,3…) of values

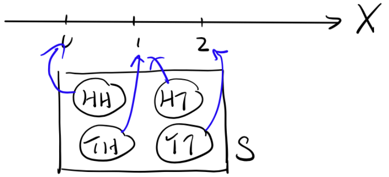
- Probability Mass Function (PMF) is the probability distribution for

  discrete random variables.

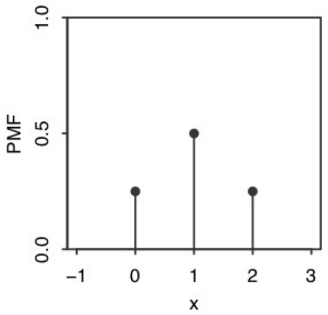
  $P_X(x) = P(X = x)$ for each possible value of r.v. X

(Two coins) Consider an experiment where we toss a fair coin twice.

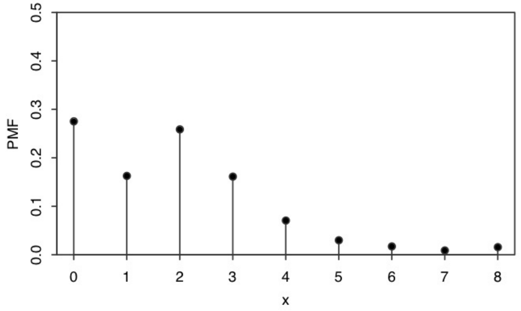
- X = number of tails



$$P_X(0) = P(X = 0) = 1/4$$
$$P_X(1) = P(X = 1) = 1/2$$
$$P_X(2) = P(X = 2) = 1/4$$



# of children in US households



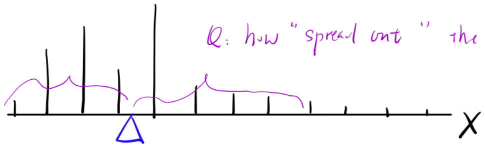
X= # children in a household in US

The distribution of a r.v. gives us full information about probabilities of any events associated with this r.v.

But often, to manage/look at the whole distribution is too much and unnecessary.

We want a few number that summarize key aspects of the distribution.

Q: how "spread out" the distribution is? ⟹ variation, standard deviation



Q: where is the "center" / "average" ? ⟹ expected value or mean

Expectation / Expected Value/ Mean of a discrete r.v. which takes values in $\{x_1, x_2, \dots\}$ is:

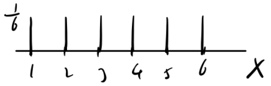
$$\mu = E(X) = \sum_x xP(X = x)$$

Values x

PMF at x

Expectation of X is a weighted sum of possible values that X can take, weight being the probability that r.v. X =x

X: result of rolling a fair 6-sided die



$$E(x) = \frac{1}{6} \times 1 + \frac{1}{6} \times 2 + \cdots + \frac{1}{6} \times 6 = 3.5$$

- Variance of a discrete r.v. X is:

$$\sigma^2 = Var(X) = E(X - EX)^2 = \sum_x (x - EX)^2 P(X = x)$$

Variability/dispersion: How far away X is from its center, on average

Weighted sum of squared distance between each possible value x to the center mean E(X),
Weight being the probability of X=x

- The square root of variance is called standard deviation (SD),

  which measures variability in the original unit.

$$SD(X) = \sqrt{\mathrm{Var}(X)}$$

Why study famous distributions that have designate names?

The most famous distributions in Statistics all have stories,

Which explain why they are so often used as models for data.

They serve as models for a large number of applications.

Thinking about these named distributions in terms of their stories, not by their distribution formula!

What’s important is understanding the stories, it helps pattern recognition.

See two problems are essentially identical in structure, just in difference cloths!

The Binomial r.v. and its distribution

Story: A coin-tossing experiment is a simple example of binomial random variable.

- Toss a coin for n times
- The tosses are independent
- Each toss, I either get a Head or a Tail
- The probability of getting a Head is p, which is the same for all tosses; the probability of getting a Tail is $q=1-p$
- We are interested in the total number of Heads I get in the fixed n tosses

Denote $X=$ total number of heads

=> X is a binomial random variable whose distribution is called "Binomial Distribution"

We usually denote $X \sim \text{Bin}(n,p)$, n is fixed quantity, p is the only parameter of the distribution.

Many practical problems are essentially the same structure as coin-toss.

- Political polls used to predict voter preferences in elections
    - Each sampled voter = a coin
    - Each voter either votes for Rep or Dem = Head or Tail
        - The proportion of voters who favor one party = probability of getting a head

- A sociology is interested in the proportion of elementary school teachers who are men
- A soft drink marketer is interested in the proportion of cola drinkers who prefer her brand
- A geneticist is interested in the proportion of the population who possess a gene linked to Alzheimer's disease

A Binomial experiment has 5 characteristics:

1. The experiment consists of n identical trials

2. The trails are independent

3. Each trial results in one of two outcomes.

We usually call the one outcome that we care about "a success", and the other "a failure"

4. The probability of success on a single trail is equal to p, which is the same for all trials.

The probability of failure is then q=1-p

5. We are interested in the number of successes observed during a fixed n trials,

denote the number of successes as X, which is the binomial random variable.

We denote $X \sim \mathrm{Bin}(n, p), X = 0,1,2,\ldots n$

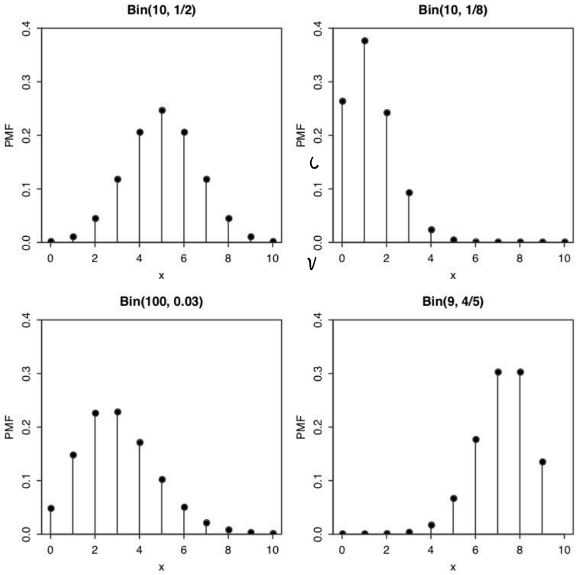
What is good about using binomial as the model to describe the problems?
- Only one parameter describe all possible question of interested

Probability Mass Function to describe its Probability Distribution:

$$p(x) = C_k^n p^k q^{n-k} = \frac{n!}{k!(n-k)!} p^k q^{n-k}$$

Where:

- p= Probability of a success on a single trial
- q=1-p
- n= Number of trails
- x= number of successes in n trials



Mean: $\mu = np$
Variance: $\sigma^2 = npq$
Standard deviation: $\sigma = \sqrt{npq}$

Find $\mathrm{P}(\mathrm{X} = 2)$ for a binomial random variable X with $\mathrm{n} = 10$ and $\mathrm{p} = 0.1$.

P(X=2) is the probability of observing 2 successes and 8 failures in a sequence of 10 trials.
You might observe 2 successes first, followed by 8 failures:
S,S,FFFFFFFF

This particular sequence has probability $= ppqqqqqqqq = p^2q^8$

But, is this the only sequence that results in X=2 successes? NO.
How many such sequences are there? This is a combination problem where we try to find which 2 positions to put the 3 successes, while the rest are failures.

The "10 choose 2" gives the total possible configurations of 2 successes+8 failures.
That is the idea of the binomial distribution formula!

$$P(x = 2) = C_2^{10}(.1)^2(.9)^{10-2}$$

$$= \frac{10!}{2!(10-2)!}(.1)^2(.9)^8 = \frac{10(9)}{2(1)}(.01)(.430467) = .1937$$

A market research firm hires operators to conduct telephone surveys. The computer randomly dials a telephone number, and the operator asks the respondent whether or not he has time to answer some questions. Let X be the number of telephone calls made until the first respondent is willing to answer the operators questions. Is this a binomial experiment? Explain.

No. X should be the total number of success events out of n trials, in this case, total number of respondents willing to answer out of n phone calls.

A home security system is designed to have a 99% reliability rate.

Suppose that nine homes equipped with this system experience an attempted burglary.

Find the probabilities of these events:

a. At least one of the alarms is triggered.

b. More than seven of the alarms are triggered.

c. Eight or fewer alarms are triggered.

Story: A coin-tossing experiment is a simple example of binomial random variable.

- Toss a coin for n times: homes equipped with this system with an burglary; n=9
- The tosses are independent: the happenings of burglary in different homes are independent
- Each toss, I either get a Head or a Tail: alarm or no alarm
- The probability of getting a Head is p, which is the same for all tosses; the probability of getting a Tail is q=1-p: p=99%
- We are interested in the total number of Heads I get in the fixed n tosses: X=number of alarms triggered out of 9

X ~ Bin(9,0.99)

$$P(X \geq 1) = 1 - P(X = 1) = 1 - C_1^9 0.99 * 0.01^8$$

$$P(X > 7) = C_7^9 0.99^7 * 0.01^2 + C_8^9 0.99^8 * 0.01^1 + C_9^9 0.99^9 * 0.01^0$$

$$P(X \leq 8) = 1 - P(X > 8) = 1 - C_9^9 0.99^9 * 0.01^0$$

Over a long period of time, it has been observed that a professional basketball player can make a free throw on a given trial with probability equal to .8. Suppose he shoots four free throws.

1. What is the probability that he will make exactly two free throws?

2. What is the probability that he will make at least one free throw?

Solution:
“trial”: a single free throw
“Success”: hit the basket
If assume: player’s chance of making free throw does not change from shot to shot,
Then, the number of times that he makes the free throw is a binomial random variable

$$
\begin{aligned}
P(x=2) &= C_2^4(.8)^2(.2)^2 \\
&= \frac{4!}{2!2!}(.64)(.04) = \frac{4(3)(2)(1)}{2(1)(2)(1)}(.64)(.04) = .1536
\end{aligned}
$$

$$
\begin{aligned}
P(\text{ at least one }) &= P(x \geq 1) = p(1) + p(2) + p(3) + p(4) \\
&= 1 - p(0) \\
&= 1 - C_0^4(.8)^0(.2)^4 \\
&= 1 - .0016 = .9984.
\end{aligned}
$$

Would you rather take a multiple-choice or a full recall test? If you have absolutely no knowledge of the material, you will score zero on a full recall test. However, if you are given five choices for each question, you have at least one chance in five of guessing correctly! If a multiple-choice exam contains 100 questions, each with five possible answers, what is the expected score for a student who is guessing on each question? Within what limits will the no-knowledge scores fall?

The expected score is E(X)=np=100(.2)=20

According to empirical rule:

- about 95% of score will lie within 2 standard deviation of the mean
- about 99% of score will lie within 3 standard deviation of the mean

The standard deviation for binomial r.v. is $\sigma = \sqrt{npq} = \sqrt{100(.2)(.8)} = 4$

So, the limits of no-knowledge scores fall between

- about 95% of score will lie within [12, 28]
- about 99% of score will lie within [8,32]

The “guessing” option gives better score than 0 score in a full recall test, but still, will not pass the exam.