Math 3501 - Probability and Statistics I

2.4 - The binomial distribution

Binomial distribution

Let X equal the number of observed successes in n Bernoulli trials.

The pmf of X is

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, \ldots, n$$

These probabilities are called *binomial probabilities*, and the random variable X is said to have a *binomial distribution*.

Notation and Terminology:

- A binomial distribution is denoted by $b(n, p)$ and we sometimes write

$$X \sim b(n, p)$$

to denote that the distribution of X is $b(n, p)$

- The constants n and p are called the parameters of the binomial distribution

Summary

A binomial experiment satisfies the following properties:

1. A Bernoulli (success-failure) experiment is performed n times, where n is a (non-random) constant.

2. The Bernoulli trials are independent.

3. The probability of success on each trial is a constant p; the probability of failure is $q = 1 - p$.

4. The random variable X equals the number of successes in the n trials.

$X \sim bi(n, p)$ means $X =$ # successes in n independent Bernoulli trials, each with prob. of success p

Remark

Let $X \sim b(n, p)$. *Then*

$$Y = n - X \sim b(n, 1-p) \ .$$

Interpretation: If $X \sim b(n, p)$ represents the number of observed successes in n independent Bernoulli trials, then $Y = n - X$ is the corresponding number of observed failures.

Example

Suppose that the probability of germination of a beet seed is 0.8 and the germination of a seed is called a success.

Suppose that the germination of one seed is independent of the germination of another seed.

If we plant 10 seeds, find the probability that:

a) • exactly 4 germinate;

b) • at most 8 germinate;

c) • at least 3 germinate.

Define the r.v. X = # seeds (out of 10) that germinate

$\begin{cases} n = 10 \text{ (\# of repetitions)} \\ p = 0.8 \text{ (prob of success on each repetition)} \end{cases}$

We know that $X \sim b(10, 0.8)$ (n, p)

a) $P(X = 4) = \binom{10}{4} (0.8)^4 \cdot (0.2)^6 \approx \ldots$ (calculator)

b) $P(X \leq 8) = ?$

We can find this number in multiple ways:

$$P(X \leq 8) = \sum_{x=0}^{8} \binom{10}{x} \cdot (0.8)^x \cdot (0.2)^{10-x} =$$

looks similar to a sum we evaluated in our last class using a table

but table only lists values of p up to 0.5

$X \leq 8$ means $X = 0, 1, 2, 3, \ldots, 8$ (at most 8 successes)

is equivalent to

$Y = 10 - X \geq 2$, i.e., we observe at least two failures

number of failures

since $Y = 10 - X$ then $X \leq 8 \iff \underbrace{10 - X}_{Y} \geq 2$

But we know that $Y = 10 - X \sim bi(10, 0.2)$

prob. of failures

$$P(X \leq 8) = P(Y \geq 2) = 1 - P(Y < 2) = 1 - \underbrace{P(Y \leq 1)}_{\text{read from table}} = 1 - 0.3758 = \ldots$$

$n = 10$, $p = 0.2$, $x = 1$

e) $P(X \geq 3) = P(Y \leq 7) = \sum_{y=0}^{7} \binom{10}{y} \cdot (0.2)^y \cdot (0.8)^{10-y} = \rightarrow 0.9999$

table

$n = 10$
$x = 7$
$p = 0.2$

$X \sim b(10, 0.8)$ ← there is no table

$Y = 10 - X \sim b(10, 0.2)$ ← we can use table

failures

$X = 3, 4, 5, \ldots, 10 \Leftrightarrow Y = 0, 1, 2, \ldots, 7$

$X \geq 3 \Leftrightarrow -X \leq -3 \Leftrightarrow 10 - X \leq 10 - 3 \Leftrightarrow Y \leq 7$

Moment generating function of binomial distribution

Let $X \sim b(n, p)$. The mgf of X is given by

$$M(t) = E\left(e^{tX}\right) = \left[(1-p) + pe^{t}\right]^{n}, \quad -\infty < t < \infty .$$

Proof:

$$M(t) = E\left[e^{tX}\right] = \sum_{x=0}^{n} e^{tx} \cdot f(x) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^{x} (1-p)^{n-x} =$$

X takes values on $\{0, 1, 2, \ldots, n\}$

$$= \sum_{x=0}^{n} \binom{n}{x} (pe^{t})^{x} \cdot (1-p)^{n-x} = (pe^{t} + 1 - p)^{n} = (1 - p + pe^{t})^{n}$$

$a = pe^{t}$, $b = 1 - p$

Recall: $(a+b)^{n} = \sum_{x=0}^{n} \binom{n}{x} a^{x} \cdot b^{n-x}$

converges for all $t \in \mathbb{R}$ (finite # of terms)

we can use this to find $E[X] = M'(0)$, $E[X^2] = M''(0)$

Binomial distribution mean and variance

Differentiating

$$M(t) = \left[(1-p) + pe^t\right]^n$$ ← from previous slide

we obtain (chain rule)

$$M'(t) = n\left[(1-p) + pe^t\right]^{n-1}\left(pe^t\right)$$ (product rule + chain rule)

and

$$M''(t) = n(n-1)\left[(1-p) + pe^t\right]^{n-2}\left(pe^t\right)^2 + n\left[(1-p) + pe^t\right]^{n-1}\left(pe^t\right)$$ ←

$M''(0) = n(n-1)\left[1-p+p\right]^{n-2} \cdot p^2 + n\left[1-p+p\right]^{n-1} \cdot p = n(n-1)p^2 + np$

Evaluating at $t = 0$, we find that

$$\mu = E(X) = M'(0) = np$$

$M'(0) = n\left[1-p+p\right]^{n-1} \cdot p$

and

$$\sigma^2 = E\left(X^2\right) - [E(X)]^2 = M''(0) - [M'(0)]^2$$
$$= n(n-1)p^2 + np - (np)^2 = np(1-p)$$

$E[X^2] = M''(0)$

$E[X^2]$ $\quad (E[X])^2$

$n^2p^2 - np^2 + np - n^2p^2 = np(1-p)$

Conclusion: If $X \sim b(n, p)$:

- $\mu = E[X] = np$
- $\sigma^2 = \text{Var}(X) = npq$

Intuition: when p is the probability of success on each trial, the expected number of successes in n trials is np.

Special case: when $n = 1$, X has a Bernoulli distribution and we have

- $M(t) = (1 - p) + pe^t$
- $\mu = E[X] = p$ ✓
- $\sigma^2 = \text{Var}(X) = pq$ ✓

set $n=1$

↳ Bernoulli distribution is the same as $b(1, p)$ ($n=1$)

Example

Suppose that observation over a long period of time has disclosed that, on the average, 1 out of 10 items produced by a certain process is defective.

Select five items independently from the production line and test them. Let X denote the number of defective items among the $n = 5$ items.

Determine the mean and variance of X.

5 repetitions of independent Bernoulli trials.

We're counting the # of defective item.

We're told that probability of having 1 defective item is $p = 0.1$

We're told 1 out of 10 items, on average, are defective

X = # defective items (out of 5 inspected) $\quad X \sim b(5, 0.1)$ (n, p)

mean $\mu = E[X] = n \cdot p = 5 \cdot (0.1) = 0.5$

variance $\sigma^2 = Var(X) = n \cdot p \cdot q = 5 \, (0.1) \cdot (0.9) = 0.45$

Math 3501 - Probability and Statistics I

2.5 - The hypergeometric distribution

Random experiment

Consider a collection of $N = N_1 + N_2$ similar objects split into two dichotomous classes (e.g. red balls vs blue balls):

- $N_1 > 0$ belong to one class (red balls)
- $N_2 > 0$ belong to the other class (blue balls)

A collection of n objects, where $1 \leq n \leq N_1 + N_2$, is selected from these N objects at random and without replacement.

Define the random variable:

X = number of objects selected that belong to the first class .

Question: Determine $P(X = x)$, where the nonnegative integer x satisfies

$$x \leq n\,, \quad x \leq N_1\,, \quad \text{and} \quad n - x \leq N_2\,,$$

that is, the probability that exactly x of these n objects belong to the first class and $n - x$ belong to the second.

Strategy: We can select:

- x objects from the first class in $\binom{N_1}{x}$ ways;
- $n - x$ objects from the second class in $\binom{N_2}{n-x}$ ways;
- By the multiplication principle, the product

$$\binom{N_1}{x}\binom{N_2}{n-x}$$

  equals the number of ways the joint selection operation can be performed;
- there are $\binom{N}{n}$ ways of selecting n objects from $N = N_1 + N_2$ objects.

Conclusion:

$$P(X = x) = \frac{\binom{N_1}{x}\binom{N_2}{n-x}}{\binom{N}{n}},$$

where the space of X is the set S of nonnegative integers x such that

$$x \leq n\,, \quad x \leq N_1\,, \quad \text{and} \quad n - x \leq N_2\,.$$

Hypergeometric distribution

We say that a random variable X has a *hypergeometric distribution* with parameters N_1, N_2, and n, if its pmf is of the form

# of objects of type 1

# of objects of type 2

$N = N_1 + N_2$

sample size

$$f(x) = \frac{\binom{N_1}{x}\binom{N_2}{n-x}}{\binom{N}{n}},$$

where x is an integer such that

$$\max\{n - N_2, 0\} \leq x \leq \min\{n, N_1\}.$$

$x \leq n$

$x \leq N_1$

$n - x \leq N_2 \iff x \geq n - N_2$

$x \geq 0$

Notation: $X \sim HG(N_1, N_2, n)$.

We do not need to specify N since we know that $N = N_1 + N_2$

Example

A small pond has 50 fish, ten of which have been tagged. If a fisherman's catch consists of seven fish selected at random and without replacement, determine the probability that exactly two tagged fish are caught.

50 fish → 10 are tagged
N → 40 are not tagged

→ Two types of objects:
$N_1 = 10$ tagged fish
$N_2 = 40$ non tagged fish
} 50 fish in total

sample size $n = 7$

Define the r.v.: X = # of tagged fish caught (out of 7)

The pmf of X is $f(x) = P(X = x) = \dfrac{\binom{10}{x}\cdot\binom{40}{7-x}}{\binom{50}{7}}$, $\max\{-33, 0\} \le x \le \min\{7, 10\}$

$\max\{n - N_2, 0\} \le x \le \min\{n, N_1\}$

$0 \le x \le 7$

$$f(2) = P(X = 2) = \frac{\binom{10}{2}\cdot\binom{40}{5}}{\binom{50}{7}}$$

Example

A lot consisting of 100 fuses is inspected by the following procedure:

Five fuses are chosen at random and tested; if all five blow at the correct amperage, the lot is accepted.

Suppose that the lot contains 20 defective fuses. Determine the probability of accepting the lot.

Lot has 100 fuses → 20 defective $N_1 = 20$

↑ $N = 100$

→ 80 are not defective $N_2 = 80$

Pick a sample of $n = 5$ fuses (sample size is $n = 5$)

Define the r.v X = # of defective fuses out of the 5 selected (without replacement)

Lot is accepted if $X = 0$. We want to find $P(X=0)$ knowing that $X \sim HG(20, 80, 5)$ (N_1, N_2, n)

and so $P(X=0) = \dfrac{\binom{20}{0}\cdot\binom{80}{5}}{\binom{100}{5}}$

Hypergeometric distribution mean and variance

Let $X \sim HG(N_1, N_2, n)$. Then:

- $\mu = E[X] = n \cdot \frac{N_1}{N}$
- $\sigma^2 = \mathrm{Var}(X) = n \cdot \frac{N_1}{N} \cdot \frac{N_2}{N} \cdot \frac{N-n}{N-1}$

Before drawing

N_1 objects of type 1
N_2 objects of type 2
} out of $N = N_1 + N_2$

← identical to the mean of $b(n,p)$ with p replaced by $\frac{N_1}{N}$

# of objects selected

like p

like q

last factor adjusts the variance to sampling without replacement

with $\frac{N-n}{N-1} \leq 1$

⇒ "no replacement" reduces variance

Recall variance of $b(n,p)$ is $\sigma^2 = n \cdot p \cdot q$

Example (previous example continued)

$N = 100$ fuses

$N_1 = 20$ defective

$N_2 = 80$ non defective

$n = 5$ objects selected without replacement

$X \sim HG(20, 80, 5)$

↕ # of defective items found in a sample of size 5 obtained without replacement

Mean of X is

$$\mu = E[X] = n \cdot \frac{N_1}{N} = 5 \cdot \frac{20}{100} = 1$$

Variance of X is

$$\sigma^2 = Var(X) = n \cdot \frac{N_1}{N} \cdot \frac{N_2}{N} \cdot \frac{N-n}{N-1} = 5 \cdot (0.2)(0.8) \frac{95}{99} \approx \ldots$$