STAT 390 A
Statistical Methods in Engineering and Science

Week 3 Lectures – Part 1 – Spring 2023

Discrete Random Variables and Probability Distributions

Alexander Giessing

Department of Statistics

University of Washington

April 14, 2023

Outline

Bernoulli Distribution

- The Bernoulli distribution is used to model a (single) Bernoulli trial.
- "Success" and "Failure" are encoded as 1 and 0, respectively.

Bernoulli distribution

A discrete random variable X has a Bernoulli distribution with parameter p, where $0 \leq p \leq 1$, if its probability mass function is given by

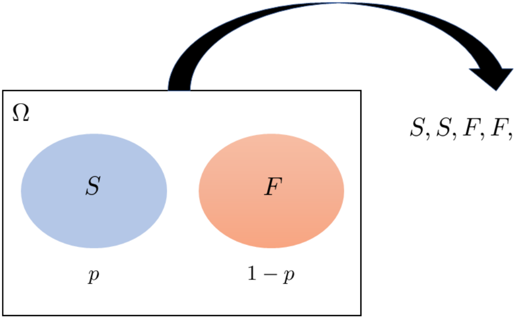
$$p_X(1) = P(X = 1) = p \quad \text{and} \quad p_X(0) = P(X = 0) = 1 - p.$$

We denote this distribution by $Ber(p)$.

- **Fact:** $\mathrm{E}[X] = p$ and $\mathrm{Var}(X) = p(1 - p)$.

Bernoulli Trials

- Bernoulli trials are (probably) the simplest experiments, i.e. "abstract" version of a coin toss.
- Each trial has two outcomes: "Success" or "Failure".
- For each trial $P(\text{"Success"}) = p$.



$S, S, F, F, F, F, S, \ldots$

Examples for Bernoulli Trials

Which of the following are Bernoulli trials?

- Roll a die 100 times and observe whether getting even or odd spots.
- Observe daily temperature and see whether it is above or below frezzing temperature.
- Sample 1000 people and observe whether they belong to upper, middle or low socio economic status (SES) class.

Outline

Binomial Distribution

- Suppose we want to model the number of successes in a sequence of n independent and identical Bernoulli trials.
- The random variable in question is $X = Y_1 + Y_2 + \ldots Y_n$, where $Y_i \sim_{iid} Ber(p)$, $i = 1, \ldots, n$ and iid = "indendent and identically disributed"

Binomial distribution

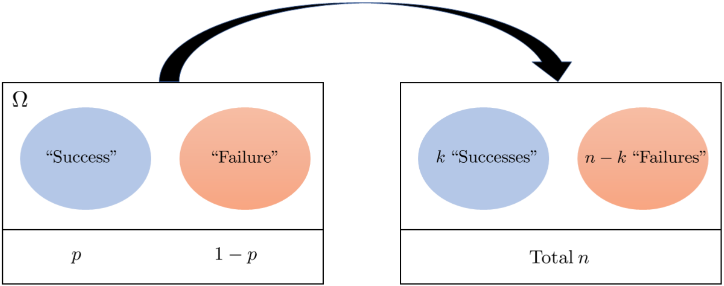
A discrete random variable X has a Binomial distribution with parameters n and p, where $n = 1, 2, \ldots,$ and $0 \leq p \leq 1$, if its probability mass function is given by

$$p_X(k) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{for } k = 0, 1, \ldots, n.$$

We denote this distribution by $Bin(n, p)$.

- **Fact:** $\mathrm{E}[X] = np$ and $\mathrm{Var}(X) = np(1-p)$.

Binomial Distribution (Cont.)



- The probability of "Success" or "Failure" stays the same as we sample n times, since we return/ replace each draw.

Examples for Binomial Random Variables

According to Mendelian's theory, a cross fertilization of related species of red and white flowered plants produces offsprings of which 25% are red flowered plants. Suppose we cross five pairs of flowers.

- Let $X = \#$ red flowered plants. Then, $X \sim Bin(5, 0.25)$.
- What is the probability of no red flower?

$$P(X = 0) = (1 - p)^5 = 0.75^5 = 0.237.$$

- What is the probability of no more than 3 red flowered plants?

$$P(X \leq 3) = P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) = ...$$

Stop! There's a shorter way.

$$P(X \leq 3) = 1 - P(X \geq 4) = 1 - P(X = 4) - P(X = 5) = 0.984.$$

Outline

Hypergeometric Distribution

- In survey sampling, one usually draws without replacement.

HYPERGEOMETRIC DISTRIBUTION

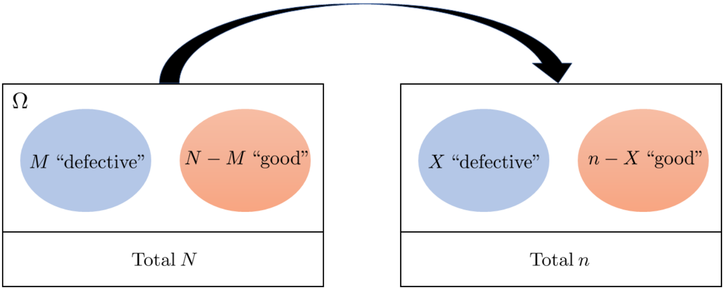
A discrete random variable X has a Hypergeometric distribution with parameters N, M,and n, if its probability mass function is given by

$$p_X(x) = P(X = x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}}, \text{ for } \max\{0, n-N+M\} \leq x \leq \min\{n, M\}.$$

We denote this distribution by $HyperGeo(N, M, n)$.

- **Fact:** $\mathrm{E}[X] = np$ and $\mathrm{Var}(X) = \frac{N-n}{N-1}np(1-p)$, where $p := \frac{M}{N}$.

Hypergeometric Distribution (Cont.)



- When sampling with replacement, the distribution is binomial.
- When sampling a small fraction from a large population, like a poll, the trials are nearly Bernoulli.

Examples for Hypergeometric Random Variables

Suppose that there are N deer in a region. Capture 5 of them, tag them and then release them back. Now, among 10 newley aptured deer, X of them are tagged.

- **Probabilistic question:** If $N = 25$, how likely is it that $X = 2$?

  The distribution of X is hypergeo. with $N = 25$, $M = 5$, $n = 10$, i.e. $X \sim HyperGeo(25, 5, 10)$. Hence,

  $$P(X = 2) = \frac{\binom{5}{2}\binom{25-5}{10-2}}{\binom{25}{10}} = 0.385.$$

- **Statistical question:** What is the population size if we observe 3 tagged animal?
  - Note that $\mathrm{E}[X] = n\frac{5}{N}$.
  - In the absence of more information, an observed value of X is a reasonable estimate of $\mathrm{E}[X]$. Since we observe $X = 3$, we compute

    $$3 = n\frac{5}{N} \Leftrightarrow \widehat{N} = \frac{5n}{3} = 16.66 \text{ or } 17.$$

  (Nota bene: We will refine and further justify this estimation procedure in the second half of the course!)

Outline

Geometric Distribution

- The distribution of waiting times until a first "Success" occurs, where $P(\text{"Success"}) = p$ and $P(\text{"Failure"}) = 1 - p$.

Geometric distribution

A discrete random variable X has a Geometric distribution with parameter p, where $0 \leq p \leq 1$, if its probability mass function is given by

$$p_X(k) = P(X = k) = (1-p)^{k-1}p \quad \text{for } k = 1, 2, \ldots.$$

We denote this distribution by $Geo(p)$.

- **Fact:** $\mathrm{E}[X] = \frac{1}{p}$ and $\mathrm{Var}(X) = \frac{1-p}{p^2}$.

Examples for Geometric Random Variables

Suppose that we want to investigate the number of menstrual cylces it takes women to become pregnant, measured from the moment they have decided to become pregnant. Let X be the number of cyles up to pregnancy.

- **Probabilistic question:** If the probability of becoming pregnant in any cycle is 1/7 independent of the previous cycles, what is the probability that a woman becomes pregnant in the third cycle?

  Since $X \sim Geo(1/7)$ it follows that

$$P(X = 3) = \left(\frac{6}{7}\right)^2 \left(\frac{1}{7}\right) = 0.105.$$

- **Statistical question:** If we observed the majority of women getting pregnant in their fourth menstrual cycle, what would be a reasonable estimate of the probability p of becoming pregnant in any cycle?
  - Note that $\mathrm{E}[X] = \frac{1}{p}$.
  - Again, any observed value of X is a reasonable estimate of $\mathrm{E}[X]$. Since we observe $X = 4$, we conclude

$$4 = \frac{1}{p} \Leftrightarrow \widehat{p} = 0.25.$$

Outline

Poisson Distribution

- Useful for modeling rare events, e.g. # of traffic accidents, # of houses damaged by fire; # of customer arrivals to a service center.
- A Poisson distribution is a reasonable model if all of the following holds:
  - independence: # of events occurring in any time interval is independent of # of events in any other non-overlaping interval.
  - rare: It is almost impossible for two or more events simultaneously.
  - constant rate: the average # of occurrences per time unit is constant, denoted by λ.

POISSON DISTRIBUTION

A discrete random variable X has a Poisson distribution with parameter $\lambda > 0$ if its probability mass function is given by

$$p_X(k) = P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \quad \text{for } k = 0, 1, 2, \ldots.$$

We denote this distribution by $Pois(\lambda)$.

- **Fact:** $\mathrm{E}[X] = \lambda$ and $\mathrm{Var}(X) = \lambda$.