Statistics for the Sciences

Probability Distributions

Xuemao Zhang
East Stroudsburg University

January 18, 2025

Outline

- Probability
- Discrete Probability Distributions
  - Mean and Variance
  - Binomial Distributions
  - Poisson Distributions
  - Hypergeometric Distribution
- Continuous Probability Distributions
  - Mean and Variance
  - Uniform Distribution
  - Exponential Distribution
  - Normal Distribution
  - Chi-Square Distribution
  - t-Distribution
  - F-Distribution
- Lab: Calculating Probabilities

Probability

- Probability can be considered as relative frequency for population data (numerical), discrete or continuous.
- A random variable is a variable whose possible values are numerical outcomes of a random phenomenon.
  - A random phenomenon is generally random selection from a population
  - A random variable can take on different values, each associated with a certain probability
- Example: Define a random variable X as the outcome of rolling a balanced die. X can take on any of the values $\{1, 2, 3, 4, 5, 6\}$.

Probability

- The probability of **independent** events A and B happening is found by multiplying their probabilities.
- Example: You roll a pair of dice; one red and the other green. What is the probability of rolling a five on the red die and an even number on the green die?

Probability

- A compound event means that multiple `single` events are happening at the same time or one after another: like rolling two dice
- Sometimes, probability problems involving compound events will need counting.
  - We usually want to focus on the probability of a single case, then add the probabilities of all different cases.
- Example: Flip a balance coin 4 times, what is the probabillity of obtaining exactly two heads.

- Binomial Coefficent: The number of distinct combinations of n distinct objects that can be formed, taking them r at a time, denoted by C_r^n or $\binom{n}{r}$ is

$$\binom{n}{r} = \frac{n!}{r!(n-r)!}$$

Discrete Probability Distributions

- Discrete Random Variables: These take on a finite or countably infinite number of values.
- Discrete probability distribution lists all possible values of a discrete random variable X and the probability associated with each value x.
- The probability mass function (PMF)

$$P(x) = P(X = x)$$

  must satisfy two conditions:
  - $0 \leq P(x) \leq 1$ for any x
  - $\sum_x P(x) = 1$
    - $\sum_x$ means sum over all x values

Mean and Variance

- Let X be a discrete random variable with pmf $P(x)$. Then the expected value of X , denoted by $E(X)$ or μ, is defined to be

$$\mu = E(X) = \sum_{x} xP(x).$$

- Let X be a discrete random variable with pmf $p(x)$. If g is a function, then

$$E[g(X)] = \sum_{x} g(x)p(x)$$

Mean and Variance

- Especially, if $g(x) = (x - \mu)^2$, it defines the **variance** of X,

$$\sigma^2 = Var(X) = E[(X - EX)^2] = \sum_x (x - \mu)^2 p(x)$$

  - Short cut formula $\sigma^2 = E(X^2) - (EX)^2 = \sum_x x^2 p(x) - \mu^2$.
  -
  - $\sigma = \sqrt{\sigma^2}$ is called the standard deviation of X.

Binomial Distribution

- Bernoulli Trials
  - Each trial results in one of two outcomes: success, S, or failure, F.
  - The trials are independent. (The outcome of any individual trial does not affect the probabilities in the other trials.)
  - The probability of a success p remains the same in all trials.
- The **binomial random variable** is defined as the number of successes out of n independent Bernoulli trials.

Binomial distribution

The probability mass function of the binomial random variable Y is given by

$$p(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \qquad x = 0, 1, 2, \ldots, n, 0 \le p \le 1.$$

Note. $\binom{n}{x}$ is the number of outcomes with exactly x successes among n trials.

Binomial Distribution

Binomial Mean and variance

Let Y be a binomial random variable based on n trials and success probability p. Then

$$\mu = E(Y) = np \text{ and } \sigma^2 = Var(Y) = np(1-p).$$

Poisson Distribution

Poisson experiment

(1) Consists of an **infinite** number of identical trials.

(2) Each trial results in one of two outcomes: success, S, or failure, F.

(3) The trials are independent.

(4) The probability of a success (p) is the same for all trials.

- **Note** Conditions 2 to 4 are Bernoulli trials.
- A Poisson random variabl is the number of successes (X) observed in a Poisson experiment.

Poisson Distribution

- Poisson random variables describe the number of events, that occur over a specified interval (a period of time or, space, distance, area, volume or some similar unit) during which an average of λ such events can be expected to occur.

pmf

A random variable X is said to have a Poisson probability distribution if and only if its probability mass function is given by

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda}, x = 0, 1, 2, \dots, \lambda > 0,$$

where $e \approx 2.71828$ is a constant.

Poisson Mean and variance

If Y is a Poisson random variable with parameter λ, then

$$\boxed{\mu = E(Y) = \lambda \text{ and } \sigma^2 = Var(Y) = \lambda.}$$

Hypergeometric Distribution

- Suppose that a population contains a finite number of elements N that posses one of two characteristics, say red and white. Thus r of the elements might be red and $N - r$ is white.
  - A sample of n elements is randomly selected from the population and define Y to be the number of red elements in the sample.
  - This random variables Y is said to have a $\{$**hypergeometric distribution**$\}$**.**

Hypergeometric experiment

(1) Sample Space (polulation) is finite.

(2) Each trial results in one of two outcomes: success, S, or failure, F.

(3) The trials are dependent.

(4) The probability of a success for each trial is different.

(5) We are interested in the number of successes in sample size n.

Hypergeometric Distribution

Hypergeometric pmf

A random variable Y is said to have a hypergeometric probability distribution if and only if its probability mass function is given by

$$p(y) = \frac{\binom{r}{y}\binom{N-r}{n-y}}{\binom{N}{n}},$$

where y is an integer $0, 1, 2, \ldots, n$, subject to the restrictions $y \leq r$ and $n - y \leq N - r$.

Hypergeometric Mean and variance

If Y is a hypergeometric random variable, then

$$\boxed{\mu = E(Y) = \frac{nr}{N} \text{ and } \sigma^2 = Var(Y) = n\left(\frac{r}{N}\right)\left(\frac{N-r}{N}\right)\left(\frac{N-n}{N-1}\right).}$$

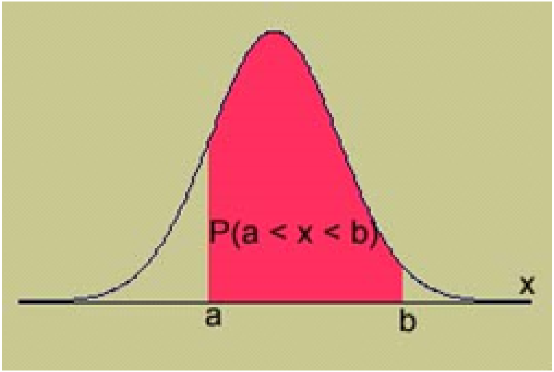
Hypergeometric Distribution

Example The size of animal populations are often estimated by using a capture-recapture method. In this method, k animals are captured, tagged, and then released into the population. Some time later n animals are captured, and Y, the number of tagged animals among the n, is noted. The probabilities associated with Y are a function of N, the number of animals in the population, so the observed value Y contains information on this unknown N. Suppose that $k = 10$ animals are tagged and then released. A sample of $n = 6$ animals is then selected at random from the same population. Find $P(Y = 1)$ as function of N. What value of N will maximize $P(Y = 1)$.

Continuous Probability Distributions

- Continuous Random Variables: These take on an all possible values on a real line interval.
- Probability density function (PDF) $f(x) \geq 0$ describes the probability distribution of a continuous random variable.
- The PDF $f(x)$ must satisfy the following properties
  - $f(x) \geq 0$ for any $x \in R$
  - $\int\limits_{-\infty}^{\infty} f(x)dx = 1$
    - Total area under the density curve is 1.

Continuous Probability Distributions



- $P(a \leq x \leq b) =$ area under the curve between a and b.
- There is no probability attached to any single value of x. That is, $P(x = a) = 0$.

Mean and Variance

- Let X be a continuous random variable with pdf $f(x)$. Then the expected value of X , denoted by $E(X)$ or μ, is defined to be

$$\mu = E(X) = \int_{-\infty}^{\infty} tf(t)dt.$$

- Let X be a continuous random variable with pdf $f(x)$. If g is a function, then

$$E[g(X)] = \int_{-\infty}^{\infty} g(t)f(t)dt.$$

- Let X be a continuous random variable. The variability is characterized by its variance.

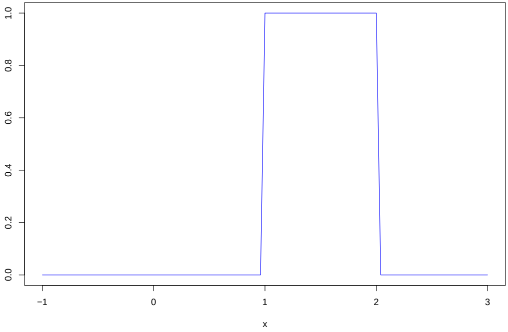
$$\sigma^2 = Var(X) = E[(X-\mu)^2] = \int_{-\infty}^{\infty} (t-\mu)^2 f(t)dt = E(X^2) - \mu^2.$$

  - Again, σ is called the standard deviation of X.

Uniform Distribution

- Uniform distribution: an even probability for all data values. It is not common for real data.

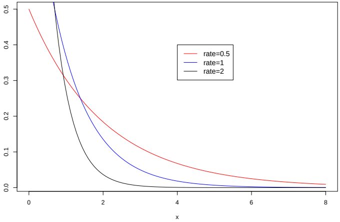
$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & otherwise \end{cases}$$



Exponential Distribution

- The pdf of an exponential distribution is given by

$$f(x) = \begin{cases} \beta e^{-\beta x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$



- If $Y \sim exponential(\beta)$, then

$$E(Y) = \frac{1}{\beta} \text{ and } Var(Y) = \frac{1}{\beta^2}.$$

Exponential Distribution

- The exponential distribution is often used to model the time between events in a Poisson process.
- Let T be the waiting time until the next event of a Poisson process. The waiting time T has an exponential distribution. That is, if

$$X \sim Poisson(\lambda)$$

  and let T be the time to the first occurrence (waiting time), then

$$T \sim Exponential(\lambda)$$

Normal Distribution

A random variable Y is said to have a normal probability distribution if and only if, for $\sigma > 0$ and $-\infty < \mu < \infty$, the pdf of Y is

$$\boxed{f(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}, -\infty < y < \infty.}$$

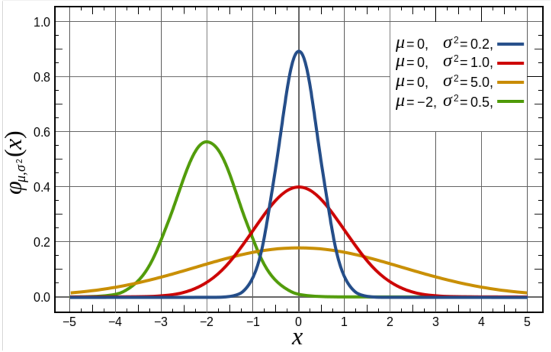
- If Y is a normally distributed random variable with parameters μ and σ, then

$$\boxed{E(Y) = \mu \text{ and } Var(Y) = \sigma^2.}$$

- Let $Y \sim N(\mu, \sigma^2)$. Then

$$\boxed{Z = \frac{Y-\mu}{\sigma} \sim N(0,1).}$$

Normal Distribution



1. Mean $= \mu$; Standard deviation $= \sigma$.
2. Symmetric about $x = \mu$.
3. Total area under the curve is 1.

Chi-square Distribution

- Let $Y_1, \ldots, Y_n$ be a random sample of size n from a normal distribution with mean μ and variance σ^2. Then $Z_i = (Y_i - \mu)/\sigma$ are independent, standard normal random variables, $i = 1, 2, \ldots, n$, and

$$\sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{Y_i - \mu}{\sigma} \right)^2$$

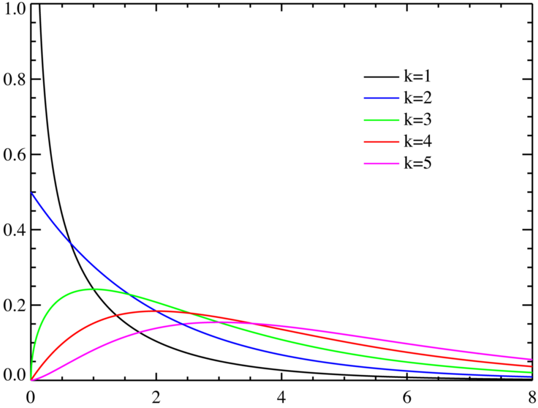
  has a χ^2 distribution with n degrees of freedom (df).

- Let $Y_1, \ldots, Y_n$ be a random sample of size n from a normal distribution with mean μ and variance σ^2. Let $S^2 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1}$ be the sample variance. Then

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{\sigma^2}$$

  has a χ^2 distribution with $n-1$ degrees of freedom (df). Also, $\overline{Y}$ and S^2 are independent random variables.

Chi-square Distribution



1. The values of chi-square can be zero or positive, but it cannot be negative.
2. The chi-square distribution is not symmetric, unlike the Normal distributions. As the number of degrees of freedom increases, the distribution approaches a Normal distribution and thus becomes more symmetric.

Student's t-distribution

- t-distribution is proposed by W.S. Gosset in 1908. Due to Gosset's pseudonym "Student", it is known as "Student's t-distribution".
- Let Z be a standard normal random variable and let W be a χ^2-distributed variable with v df. If Z and W are independent, then

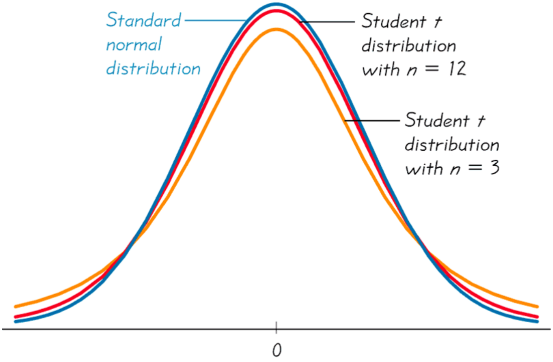
$$T = \frac{Z}{\sqrt{W/v}}$$

  is said to have a t-distribution with v df.
- Let $Y_1, \ldots, Y_n$ be a random sample of size n from a normal distribution with mean μ and variance σ^2. Then

$$\frac{\overline{Y} - \mu}{S/\sqrt{n}}$$

  has Student's t-distribution with $n - 1$ degrees of freedom.

Student's t-distribution



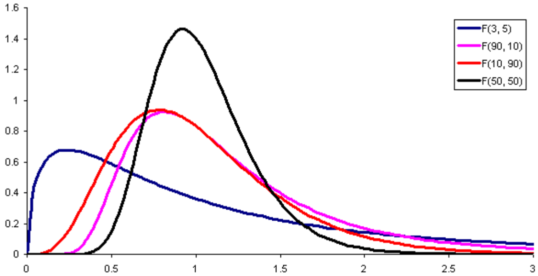
1. The density curves of the t-distribution look quite similar to the standard normal curve.
2. The spread of the t-distributions is a bit bigger than that of the standard normal curve.
3. As df gets bigger, the $t(df)$ density curve gets closer to the standard normal density curve.

F-Distribution

- Let W_1 and W_2 be independent χ^2-distributed random variables with v_1 and v_2 df, respectively. Then,
$$F = \frac{W_1/v_1}{W_2/v_2}$$
is said to have an F distribution with v_1 numerator degrees of freedom and v_2 denominator degrees of freedom.

- Let $X_1, \ldots, X_n$ be a random sample from a $N(\mu_X, \sigma_X^2)$ population, and let $Y_1, \ldots, Y_m$ be a random sample from an independent $N(\mu_Y, \sigma_Y^2)$ population. Then
$$F = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$$
has an F-distribution with $n-1$ numerator degrees of freedom and $m-1$ denominator degrees of freedom.

F-Distribution



1. The F distribution is not symmetric.
2. Values of the F distribution cannot be negative.
3. The exact shape of the F distribution depends on the two different dfs: Numerator df and Denominator df.

Lab

- We click `Transform` $\rightarrow$ `Compute Variable...` to calculate probabilities of various distributions. We may need to understand the concept of `CDF`.

cumulative distribution function

The cumulative distribution function or CDF of a random variable X, denoted by $F_X(x)$, is defined by

$$F_X(x) = P(X \leq x) \text{ for all } x.$$

Lab

- Example 1. Suppose the height of this plant species is normally distributed with a mean (μ) of 150 cm and a standard deviation (σ) of 20 cm. We want to find the probability that a randomly selected plant has a height between 140 cm and 160 cm.

Lab

- Example 2. Assume that the average number of mutations in the gene of interest is 3 mutations per 1000 bacteria. You want to find the probability of observing exactly 5 mutations in a sample of 1000 bacteria.
  - Hint: use Poisson distribution

License

This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.