PS 2010: 7. Random Variable 2

Qing Chang

Department of Political Science
University of Pittsburgh

Fall 2023

- Midterm Exam
- Homework 4 due Oct 17th

Agenda

- Discrete and continuous distribution
- Joint probability
- Covariance and Correlation

Distribution

- Distribution for discrete random variable
  - Bernoulli and Binomial
  - Negative Binomial
  - Poisson
- Distribution for continuous random variable
  - Uniform
  - Normal
  - Chi-square
  - Student t

Bernoulli and Binomial

- A random variable that takes two distinct values is called a **Bernoulli Random Variable**
- Success equals to 1 with probability p, and fail 0 with 1-p
- Example: flip a coin, election, classification
- **Binomial Random Variable** is the sum of n independently and identically distributed Bernoulli random variables.
- The probability distribution of this random variable is called binomial distribution.
- Example: conflict early warning system for the next ten times

Binomial

The probability mass function (PMF) of a binomial random variable with success probability p and n trials is given by

$$f(x) = P(X = x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

The cumulative distribution function (CDF) can be written as

$$F(x) = P(X \leq x) = \sum_{k=0}^{x} \binom{n}{k} p^k (1-p)^{n-k},$$

for $x = 0, 1, \ldots, n$.

Example

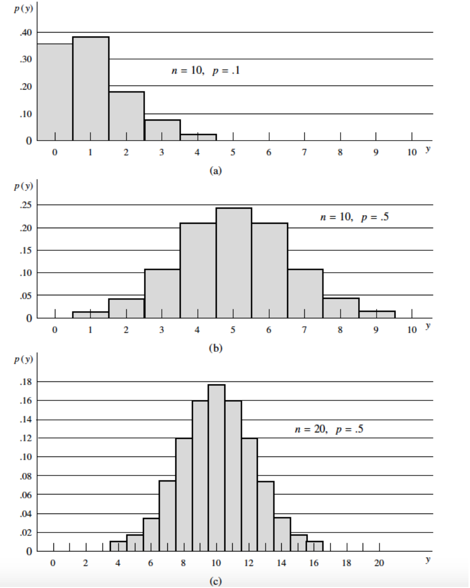
- The probability that a certain kind of component will survive a shock test is $3/4$. Find the probability that exactly 2 of the next 4 components tested survive.
- 49% of Americans likes Trump, if we randomly select 10 people from public, what is the probability of at least 9 people like Trump
- In order to verify the accuracy of their financial accounts, companies use auditors on a regular basis to verify accounting entries. The company's employees make erroneous entries 5% of the time. Suppose that an auditor randomly checks three entries. Find the probability distribution for Y , the number of errors detected by the auditor.

Mean and Variance of Binomial

Let X be a binomial random variable based on n trials and success probability p, then

$$\mu = E(X) = np \text{ and } \sigma^2 = V(X) = np(1-p)$$

Binomial Distribution



Negative Binomial

- Negative binomial also count events, but this time we care about the fixed number of success occur
- The number of X trials required to produce k successes is a negative binomial random variable
- A binomial random variable X has the following pdf:

$$f(x) = P(X = x) = \left(\begin{array}{c} x-1 \\ k-1 \end{array} \right) p^k (1-p)^{x-k}.$$

- And the mean $\mu = E(X) = \frac{k}{p}$, and $\sigma^2 = V(X) = \frac{k(1-p)}{p^2}$

Example

- A conflict early warning system can predict successfully 55% percent of potential conflicts in the experimental stage. Now it handed over to another team for testing, and it can pass the testing if the system can make 4 success predictions in 7 testings.
- What is the probability that system pass the test at the 6th test
- What is the probability that the system pass the test

Example

When Qing plays chess against his favorite computer program, he wins with probability 0.60, loses with probability 0.10, and 30% of the games result is a draw.

- Find the probability that Qing's fifth win happens when he plays his eighth game.
- Find the probability that Qing wins 7 games, if he plays 10 games.

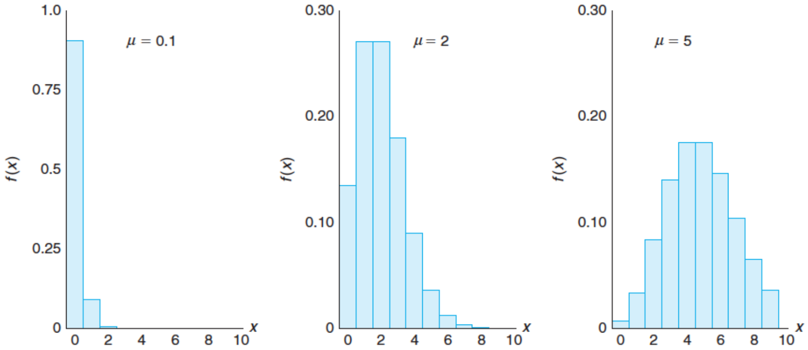
Poisson Distribution

- The number X of events occurring in one time interval is called a Poisson random variable
- Example: conflicts happening in one year, # of people dead due to Covid in a day
- The Poisson distribution for random variable X with mean λ, has the following pdf:

$$f(x) = p(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

- The means of Poisson distribution is $\mu = \lambda$ and $V(x) = \lambda$

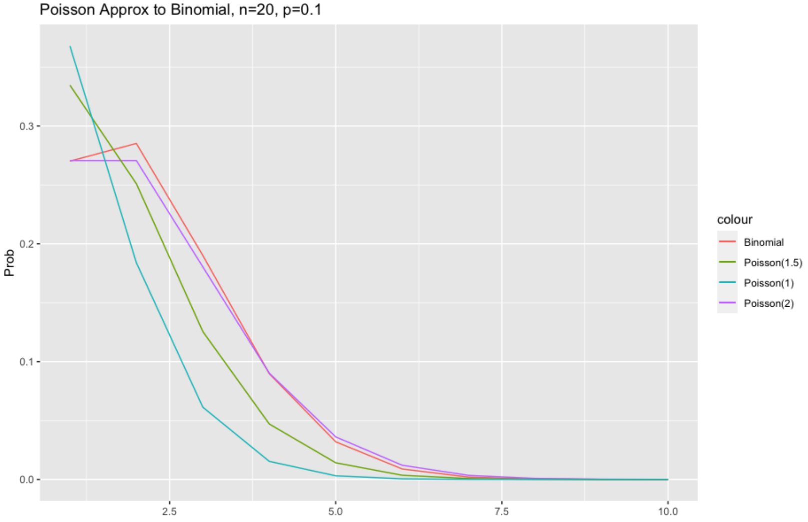
Poisson Example



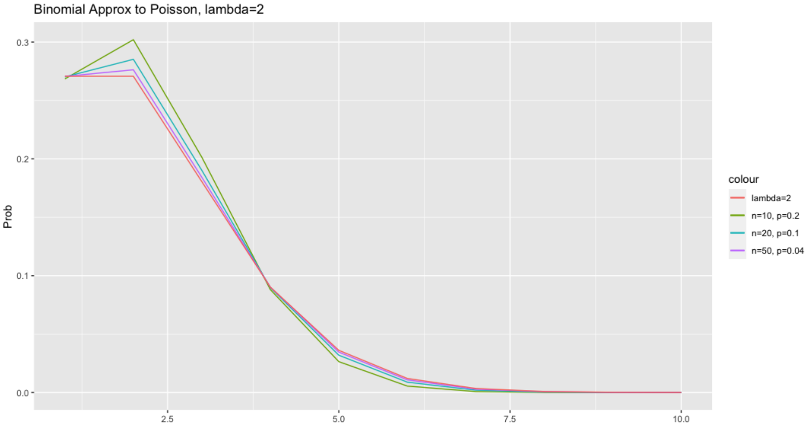
Example

- The average conflict of a year is 5, what is the probability that 6 conflicts happened in a year?
- The average number of bills passed by congress is 36 a year, what is the probability that 45 bills passed in a year?

Approximation of Binomial Distribution by a Poisson Distribution



Approximation of Poisson Distribution by a Binomial Distribution



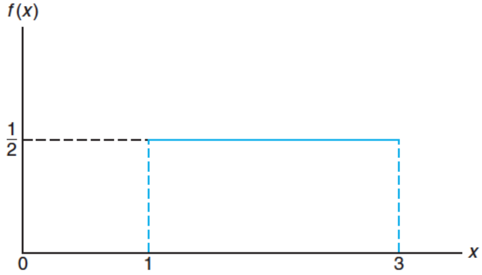
Uniform Distribution

- A very simple continuous probability distribution
- Pdf for a uniform random variable X on the interval [a,b]:

$$f(x) = \begin{cases} \frac{1}{B-A}, & A \leq x \leq B \\ 0, & \text{elsewhere} \end{cases}$$

- The density function forms a rectangle with base B-A and constant height $\frac{1}{B-A}$
- The mean of uniform random variable is $\mu = \frac{A+B}{2}$ and $\sigma^2 = \frac{(B-A)^2}{12}$

Uniform Distribution



Example

Hillman library require students to reserve study room for no more than 4 hours per person. In fact, it can be assumed that the length X of a reservation has a uniform distribution on the interval [0, 4].

- What is the probability density function?
- What is the probability that any given conference lasts at least 3 hours?

Normal Distribution

- Normal or Gaussian distribution takes the value from $-\infty$ to ∞
- The density function of a normal random variable X, with mean μ and variance σ^2 is

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, -\infty \leq x \leq \infty$$

- There is no analytical tractable form of cdf for a normal distribution.
- However, you can always calculate the probability between interval [a,b] as the area under the normal distribution curve

Standard Normal

- X is normal random variable with mean μ and standard deviation σ, denoted as $N(\mu, \sigma^2)$
- If a normal distribution has mean 0 and variance 1, then it is called **standard normal distribution**
- If c is a constant number, then Z $=$ X $+$ c is also a normal distribution with Z $\sim N(\mu + c, \sigma^2)$
- If Z $=$ cX, then Z $\sim N(c\mu, (c\sigma)^2)$
- $z - score = \frac{X-\mu}{\sigma} \sim N(0, 1)$

Standard Normal

- For a standard normal, 1 standard deviation from the mean covers 68% of the observations

$$P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.68$$

- 2 standard deviations from the mean covers 95%

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.95$$

- z-score locates a point by how many in units of the standard deviation from the mean.
  Very useful when construct confidence interval and hypothesis testing
- A simple way to do standardization for variables

Example

- Given that X has a normal distribution with $\mu = 300$ and $\sigma = 50$, find the probability that X assumes a value greater than 362.
- Suppose the final grade for this class is normally distributed with $\mu = 75$ and $\sigma = 10$, what fraction of the scores lies between 80 and 90?
- If your instructor want 90% of students get an A and the rest of them are Bs, how should the instructor set the cutoff point?

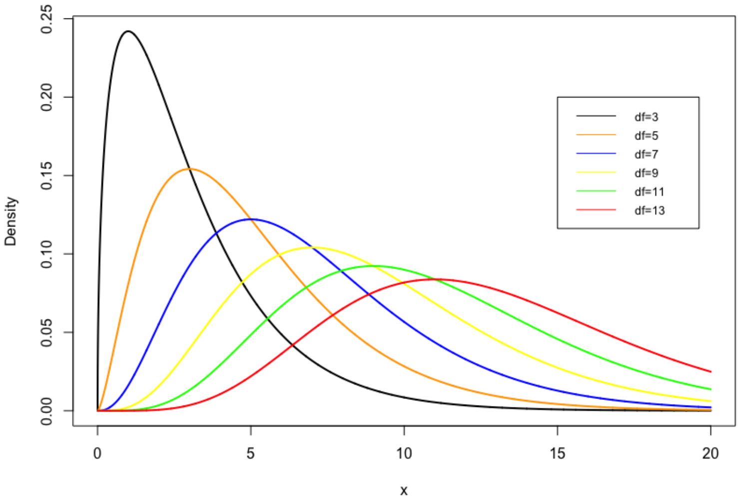
Chi-square

Several other commonly seen distributions:

- Chi-square, χ^2, distribution: If $Z_1, \ldots, Z_k$ are independent standard normal random variables, then

$$X^2 = Z_1^2 + \cdots + Z_k^2$$

  has a χ^2 distribution with k degrees of freedom.
- χ^2 distribution converges to normal distribution
- Useful for hypothesis testing for multivariate



0.25
0.20
0.15
0.10
0.05
0.00
Density
0
5
10
15
20
x
df=3
df=5
df=7
df=9
df=11
df=13

Student t

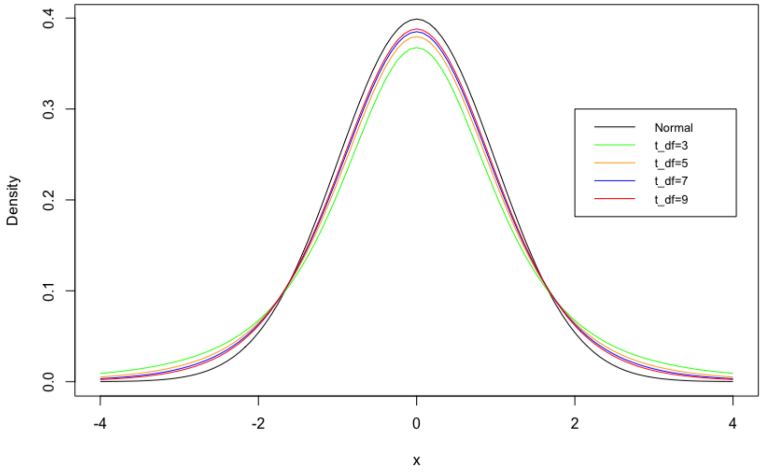
If Z is a standard normal random variable and if X^2 is a χ^2 random variable with k degrees of freedom, then

$$T = \frac{Z}{\sqrt{X^2/k}}$$

has a t distribution with k degrees of freedom.

- t densities are symmetric, bell-shaped, and centered at 0 just like the standard normal density, but are more spread out (higher variance).
- As the degrees of freedom increases, the t distributions converge to the standard normal.
- t distributions will be useful for statistical inference for one or more populations of quantitative variables.

t distribution



Joint Probability: discrete random variable

- Extend from one dimension random variable to two dimension.
- For example: toss two dices, p(x1 = head, x2 = head)
- **Joint probability distribution** for discrete random variable: $f(x, y) = P(X = x, Y = y)$
  - $f(x, y) \geq 0$ for all (x, y),
  - $\sum_x \sum_y f(x, y) = 1$
  - $P(X = x, Y = y) = f(x, y)$

Example

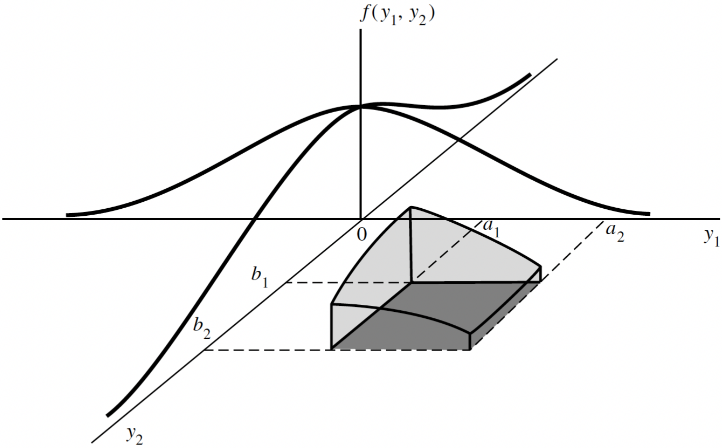
- Toss two fair dices, what is $P(2 \leq Y_1 \leq 3, 1 \leq Y_2 \leq 2)$
- Two ballpoint pens are selected at random from a box that contains 3 blue pens, 2 red pens, and 3 green pens. If X is the number of blue pens selected and Y is the number of red pens selected, find the joint probability function f(x, y), and the table for this joint distribution

Joint Probability: continuous random variable

The function $f(x, y)$ is a joint density function of the continuous random variables X and Y if

- $f(x, y) \geq 0$, for all (x, y),
- $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dxdy = 1$,
- $P[(X, Y) \in A] = \int\int_A f(x, y) dxdy$, for any region A in the xy plane.

Figure



Example

X and Y are two continuous random variable that have joint pdf:

$$f(x,y) = \begin{cases} \frac{2}{5}(2x+3y), & 0 \le x \le 1, 0 \le y \le 1 \\ 0, & \text{elsewhere} \end{cases}$$

- Verify that $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)dxdy = 1$,
- what is $P(0 < x < \frac{1}{2}, \frac{1}{4} < y < \frac{1}{2})$

Marginal Distribution

The marginal distributions of X alone and of Y alone are

$$g(x) = \sum_y f(x, y) \quad \text{and} \quad h(y) = \sum_x f(x, y)$$

for the discrete case, and

$$g(x) = \int_{-\infty}^{\infty} f(x, y)dy \text{ and } h(y) = \int_{-\infty}^{\infty} f(x, y)dx$$

for the continuous case.

- Example: Calculate the marginal distribution for the last two examples.

Conditional Probability Distribution

Let X and Y be two random variables, discrete or continuous. The conditional distribution of the random variable Y given that $X = x$ is

$$f(y \mid x) = \frac{f(x,y)}{g(x)}, \text{ provided } g(x) > 0.$$

Similarly, the conditional distribution of X given that $Y = y$ is

$$f(x \mid y) = \frac{f(x,y)}{h(y)}, \text{ provided } h(y) > 0.$$

- Generally speaking, conditional probability $= \frac{Joint}{Marginal}$

If we wish to find the probability that the discrete random variable X falls between a and b when it is known that the discrete variable $Y = y$, we evaluate

$$P(a < X < b \mid Y = y) = \sum_{a<x<b} f(x \mid y)$$

where the summation extends over all values of X between a and b. When X and Y are continuous, we evaluate

$$P(a < X < b \mid Y = y) = \int_a^b f(x \mid y)dx$$

Example

- For the ballpoint pens example, find the conditional distribution of X, given that Y = 1, and use it to determine $P(X = 0 \mid Y = 1)$.

- Two random variables X and y have the following joint distribution:

$$f(x, y) = \begin{cases} 10xy^2, & 0 < x < y < 1 \\ 0, & \text{elsewhere} \end{cases}$$

- Find the marginal distribution for x and y and the conditional density $f(y \mid x)$

- What is $P\left(Y > \frac{1}{2} \mid X = 0.25\right)$?

Covariance

- **Definition** Covariance of two random variable:
  $Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$

  $$Cov(X, Y) = E(XY - XE(Y) - YE(X) + E(X)E(Y))$$
  $$Cov(X, Y) = E(XY) - E(X)E(Y) - E(Y)E(X) + E(X)E(Y)$$
  $$Cov(X, Y) = E(XY) - E(X)E(Y)$$

- **Rules**
  - $Cov(X, Y) = 0$ if X and Y are independent
  - $Cov(X, X) = V(X)$
  - $Cov(aX, Y) = aCov(X, Y)$
  - $Cov(X_1 + X_2, Y) = Cov(X_1, Y) + Cov(X_2, Y)$
  - $V(X + Y) = V(X) + V(Y) + 2Cov(X, Y)$

Correlation

- Covariance gives you relationship between two variables, but you don't know the strength of the relationship.
- **Definition:** Correlation of X and Y is
$$Corr(X, Y) = \rho(X, Y) = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}$$
- $-1 \leq \rho(X, Y) \geq 1$
- Correlation does not care what units you use for X and Y, so if $a > 0$ and $c > 0$, then $\rho(aX + b, cY + d) = \rho(X, Y)$
- Two event independent implies $\rho(X, Y) = 0$
- However, if $\rho(X, Y) = 0$, is this means two event independent?