Statistical Method
Probability
Discrete Probability Distributions
Continuous Probability Distributions

I-Chen Lee, STAT, NCKU

September 19, 2023

Textbooks

- Sections 3 and 4 of
  Maurits; Van Den Heuvel Kaptein (Edwin). (2022). *Statistics for Data Scientist: An Introduction to Probability, Statistics, and Data Analysis.*
- Sections 4, 5, 6, 7 of
  Akinkunmi, M. (2019). *Introduction to statistics using R. Synthesis Lectures on Mathematics and Statistics*, 11(4), 1-235.

Overview

Univariate Data

- Univariate data is used to describe one characteristic or variation from observation to observation.
- To describe patterns, some ways including graphical methods, measures of central tendency, and measures of variability could be conducts.
    - Graphical methods: bar chart, histogram, pie chart,...
    - Location: mean/median/mode
    - Dispersion: variance/standard deviation/range,...
- The whole picture of univariate dataset could be addressed by a distribution.

Textbook: Akinkunmi, M. (2019). *Introduction to statistics using R. Synthesis Lectures on Mathematics and Statistics*, 11(4), 1-235.

Random variables

Types of a random variable are:

- Discrete.
  - X denotes the point of a die, and $x = 1, 2, 3, 4, 5, 6$.
- Continuous.
  - X denotes the body weight, and $0 < x < \infty$.
  - X denotes the temperature, and $-\infty < x < \infty$.

What is the relationship with nominal, ordinal, and cardinal variables?

Sample space & event

Sample space, S: a set including all possible events.

- Discrete random variables
  - Roll a die: $S = \{1, 2, 3, 4, 5, 6\}$.
  - Flip two coins: $S = \{(\text{head, head}), (\text{head, tail}), (\text{tail, head}), (\text{tail, tail})\}$.
- Continuous random variables (We also call domain of random variable)
  - $S = \{X : -\infty < x < \infty\}$
  - $S = \{X : x > 0\}$
  - $S = \{X : a \leq x \leq b\}$

Event, A: any subset of the sample space.

- The point is even when rolling a die: $A = \{2, 4, 6\}$.
- At least one head when flipping two coins: $A = \{(\text{head, head}), (\text{head, tail}), (\text{tail, head})\}$.
- The age is located between [30, 40]: $A = \{x : 30 \leq x \leq 40\}$.

Axioms of probability

$P(A)$ denotes the probability of the event A.

$$P(A) = \frac{\text{number of outcome with A}}{\text{number of possible outcome}}.$$

- $0 \leq P(A) \leq 1$.
- $P(S) = 1$.
- For any sequence of mutually exclusive events $A_1, A_2, \ldots$

$$P\left(\bigcup_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i), \; n = 1, 2, ..., \infty.$$

Example: A dice

A fair die is rolled once, what is the probability of: (i) Rolling an even number? (ii) Rolling an odd number?

Solution:

$$\text{Probability (even number)} = \frac{\textit{number of even number}}{\textit{number of possible outcomes}}.$$

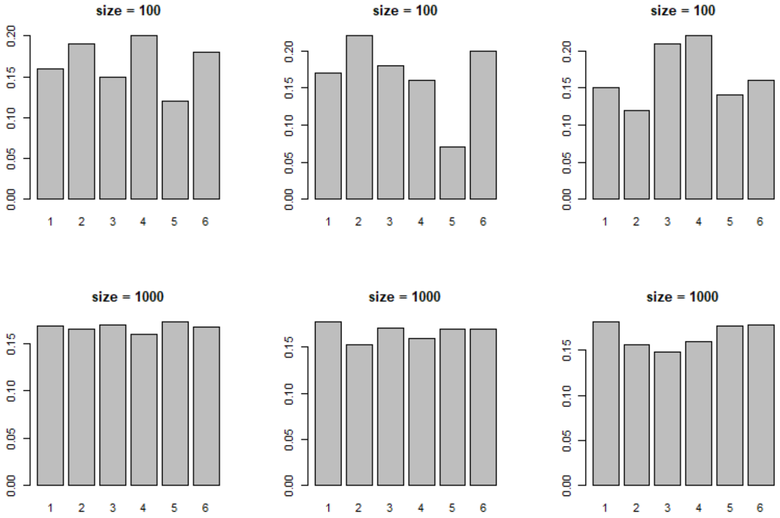
Let A be set of even number, then $A = \{2, 4, 6\}$ and B be a set of odd numbers in a tossing of a die, the $B = \{1, 3, 5\}$ and the sample space, $\mathbb{S} = \{1, 2, 3, 4, 5, 6\}$.

(i) $P(A) = \frac{3}{6}$.

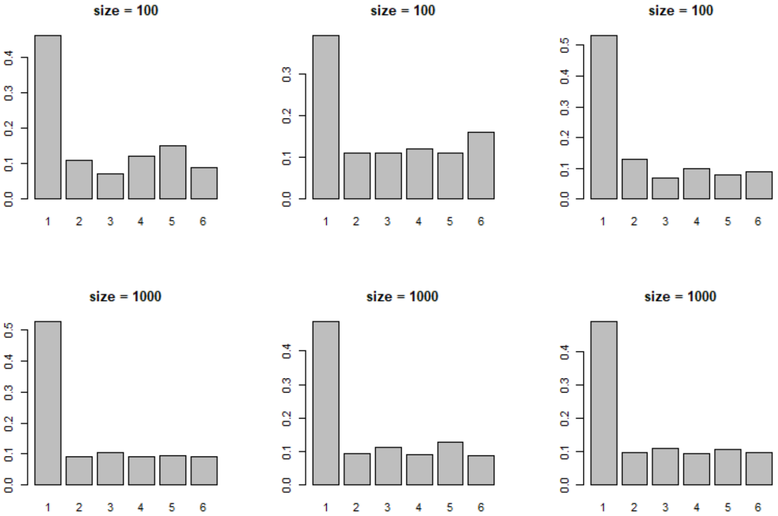
(ii) $P(B) = \frac{3}{6}$.

Do the exact experiment to show if the probability is right!
Do the simulation in R to show if the probability is right!
Question: What if it is not a fair dice? probability is 0.5, 0.1, 0.1, 0.1, 0.1, 0.1?

Example: A fair dice



Example: An unfair dice



Exercise in class

Consider rolling two fair dice; calculate the probability of getting a sum of their outcomes when the experiment is performed in 100 times.

- Sample space?
- Give the probability of each event.
- What is the distribution?
- With 10000 times, are the probabilities similar to the theoretical ones?

Discrete probability distributions

A discrete probability distribution satisfies two conditions:

$$0 \leq P(X = x) \leq 1, \; x \in S.$$

and

$$\sum_{x \in S} P(X = x) = 1.$$

Probability mass function

- Discrete random variable X:

$$p(x) = P\{X = x\}.$$

- $p(x)$ is called the probability mass function (pmf). If the random variable takes on one of the possible values $x_1, x_2, \ldots$, then we have

$$\sum_{i=1}^{\infty} p(x_i) = 1.$$

- Example: X denotes the point of a fair die, and

$$p(x) = P\{X = x\} = \frac{1}{6}, \ x = 1, 2, 3, 4, 5, 6.$$

$$\sum_{x=1}^{6} p(x) = 1.$$

Example

Experiment: toss a fair coin two times.
Sample space: $S = \{HH, HT, TH, TT\}$.
Random variable X is the number of tosses showing heads.
Thus, $X : S \to \mathbb{R}$

$$
\begin{aligned}
X &= (HH) = 2 \\
X &= (HT) = (TH) = 1 \\
X &= (TT) = 0 \\
X &= \{0, 1, 2\}.
\end{aligned}
$$

That is, random variable X takes a range of values 0, 1, and 2. Hence, the pmf is given by:

$$P(X = 0) = \frac{1}{4}, \ P(X = 1) = \frac{1}{2}, \quad \text{and} \quad P(X = 2) = \frac{1}{4}.$$

Expectation and Variance

The expectation or expected value fo X is also called the mean of X and denoted by $E(X)$.

- Discrete random variable X:

$$E(X) = \sum_{i=1}^{\infty} x_i P\{X = x_i\}.$$

If X is a random variable with mean μ, then the variance of X, denoted by

$$Var(X) = E\left[(X-\mu)^2\right] = \sum_{i=1}^{\infty}(x_i - \mu)^2 P\{X = x_i\}.$$

- Alternative formula: $Var(X) = E(X^2) - [E(X)]^2$.
- If a and b are constants, then $Var(aX + b) = a^2 Var(X)$.

Examples of expectation

- X denotes the point of a fair die, and the mean of X is

$$E(X) = \sum_{x=1}^{6} xP\{X = x\} = \frac{1}{6}\sum_{x=1}^{6} x = 3.5.$$

- Let I be an indicator random variable for the event A, expressed by

$$I = \begin{cases} 1 & \text{if } A \text{ occurs} \\ 0 & \text{if } A \text{ does not occur.} \end{cases}$$

Then, $E(X) = 1P(A) + 0P(A^c) = P(A)$, where A^c is the complement of event A.

Exercise in R

Defective(X)	0	1	2	3	4	5
$P(X)$	1/15	1/6	3/10	1/5	2/15	2/15

In R, calculate the following items:

- expected value of the distribution
- variance of the distribution

Property of expectation

- A function g of a random variable X is also a random variable.
- The expectation of the function $g(X)$ is

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

- If a and b are constants, then $E(aX + b) = aE(X) + b$.
- If there are n random variables, X_1, X_2, …, then

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} E(X_i).$$

Well known distributions

- Discrete uniform distribution
  - Sample space: $S = \{1, 2, ..., m\}$.
  - $P(X = i) = \dfrac{1}{m}$, where $i = 1, ..., m$.
  - $E(X) = \dfrac{1}{m}\dfrac{m(m+1)}{2} = \dfrac{(m+1)}{2}$.
- Bernoulli distribution ($X \sim Ber(p)$)

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1 - p. \end{cases}$$

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

  - Sample space: $S = \{0, 1\}$.
  - $E(X) = 1 \times p + 0 \times (1 - p)$.

Binomial distribution

If $X \sim Bin(n, p)$, then

- X means that the total number of n Bernoulli trials having 1.
- Sample space: $S = \{0, 2, ..., n\}$.
- The probability mass function is

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k},$$

where $j = 0, 1, \ldots, n$.

- The expectation is

$$\sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} = \cdots = np.$$

- The variance is

$$\sum_{k=0}^{n} (k - np)^2 \binom{n}{k} p^k (1-p)^{n-k} = \cdots = np(1-p).$$

Binomial distribution in R

Generate a random sample from a binomial distribution:

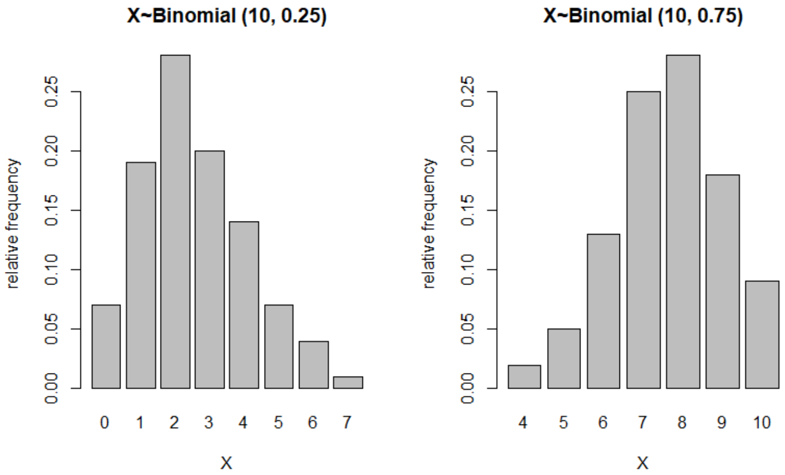
We shall use the R function to simulate a binomial random variables with given parameters.

The function `rbinom()` is used to generate n independent binomial random variables. The general form is `rbinom (n, size, prob)`.

`n`	number of random sample
`size`	number of trials
`prob`	probability of success.

Binomial distribution in R

Generate a random sample from two binomial distributions:



Assignment

The probability of having typhoid fever after drinking from a well water in a villager is 0.45. If ten villagers drank out of the well, what is the probability that:

(a) one villager will have typhoid fever;

(b) two villagers will have typhoid fever;

(c) at most two villagers will have typhoid fever;

(d) at least two villagers will have typhoid fever; and

(e) hence, calculate the number of villagers that expected to have typhoid fever.

Continuous distributions

The random variable X takes any real value within a specified range. Typical examples of continuous random variables are weight, temperature, height, lifetime, and price.

- What is the probability of an event having an exact value? For example, what is the probability that the height is 160cm in a class?
- A continuous random variable can take with their associated probabilities along "the range of values".
- Similar idea from the bar chart (for discrete variable) to the histogram (for continuous variable).
- The bar chart demonstrates the probability mass function $P(X = x)$.
- The histogram demonstrates the probability density function $f(x)$.

Recall: Bar chart and histogram

Univariate Analysis
Histogram

Cardinal variables

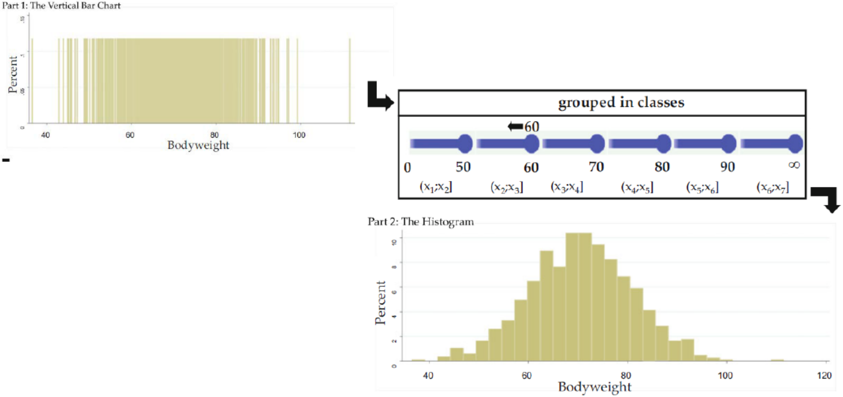


Fig. 3.7 Using a histogram to classify data

Continuous distributions

A discrete distribution X has:

$$P(a \leq X \leq b) = \sum_{x=a}^{b} P(X = x).$$

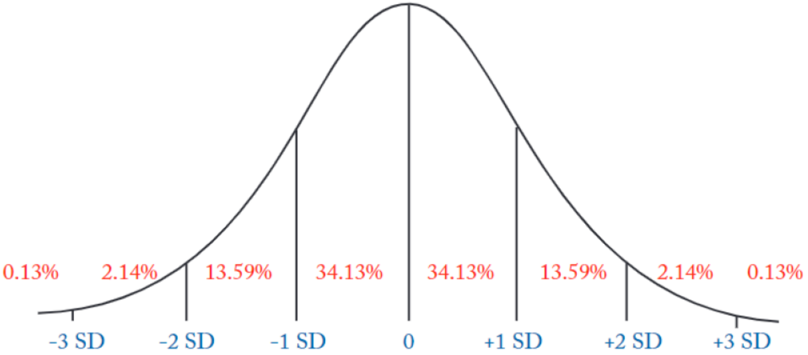
A continuous distribution X has:

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx.$$

Note that the probability is zero if the continuous random variable X is not bounded an interval:

$$P(X = a) = \int_{a}^{a} f(x)dx = 0.$$

The normal distribution



Density estimation (empirical)

The kernel density estimation is a nonparametric way to estimate the probability density function.

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right),$$

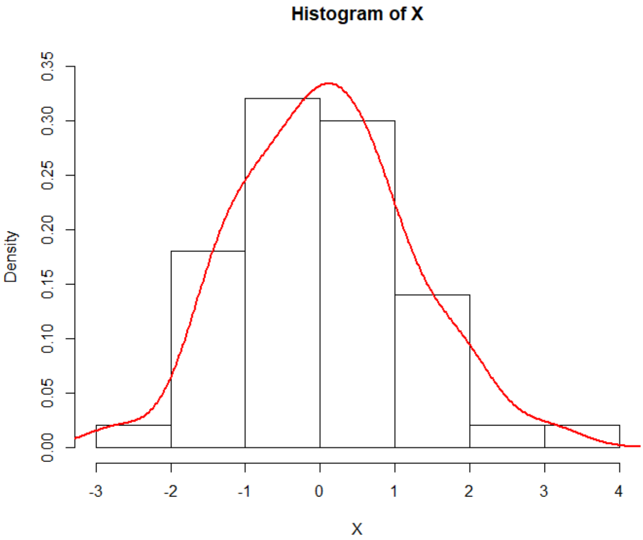
where h is the bandwidth and $K(\cdot)$ is the kernel function.

- Gaussian kernel:

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{u^2}{2}\right\}.$$

$$h = 1.06\hat{\sigma} n^{-1/5} \text{ or } h = 0.9 \min\left\{\hat{\sigma}, \frac{\mathrm{IQR}}{1.34}\right\} n^{-1/5}.$$

Illustration of density estimation



Cumulative distribution function

The cumulative distribution function (CDF), F, of the random variable X is defined for any real number x by

$$F(x) = P\{X \leq x\}.$$

- $0 \leq F(x) \leq 1$, where $x \in S$ and S is the domain of the random variable X.
- Continuous random variables X:
  There is a non-negative function $f(x)$ defined for all real numbers x and having the property that for any set C of real numbers

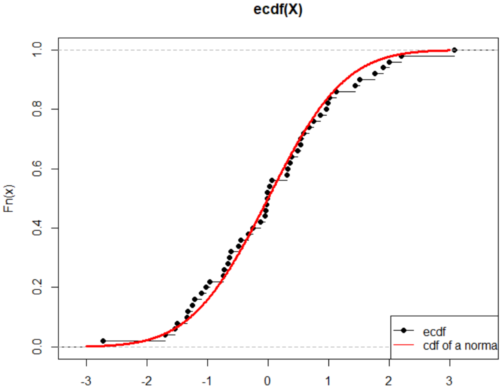
$$P\{X \in C\} = \int_C f(x)dx.$$

- $f(x)$ is called the probability density function (pdf).

Empirical cumulative distribution function (ecdf)

$$\hat{F}(x) = \frac{\text{number of samples } \leq x}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{x_i \leq x},$$

where $\mathbf{1}$ is an indicator function.



Probability density function

- The relationship between the cumulative distribution function $F(\cdot)$ a dn the probability density $f(\cdot)$ is

$$F(x) = P\{X \in (-\infty, x)\} = P\{X \leq x\} = \int_{-\infty}^{x} f(s)ds.$$

-

$$\frac{d}{dx}F(x) = f(x).$$

  or

$$f(x) = F'(x).$$

- A continuous distribution X has:

$$P(a \leq X \leq b) = \int_{a}^{b} f(x)dx = F(b) - F(a) = P(X \leq b) - P(X \leq a).$$

Expectation, $E(X)$

The expectation or expected value fo X is also called the mean of X and denoted by $E(X)$.

- Discrete random variable X:

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

- Continuous random variables X:

$$E(X) = \int_{-\infty}^{\infty} xf(x)dx.$$

Examples of expectation (conti.)

- If the probability density function of X is given by

$$f(x) = \begin{cases} 3x^2 & \text{if } 0 < x < 1 \\ 0 & otherwise, \end{cases}$$

then

$$E(X) = \int_0^1 x \cdot 3x^2 dx = \int_0^1 3x^3 dx = \frac{3}{4}.$$

Uniform distribution

If a random variable X follows the Uniform distribution (a, b), then

- The parameters are a and b.
- The parameters a and b are for the minimum and maximum.
- The pdf is

$$f(x) = \frac{1}{b-a}, \quad a \leq x \leq b.$$

- The cdf is

$$F(x) = \int_a^x \frac{1}{b-a} ds = \frac{x-a}{b-a}, \quad a \leq x \leq b.$$

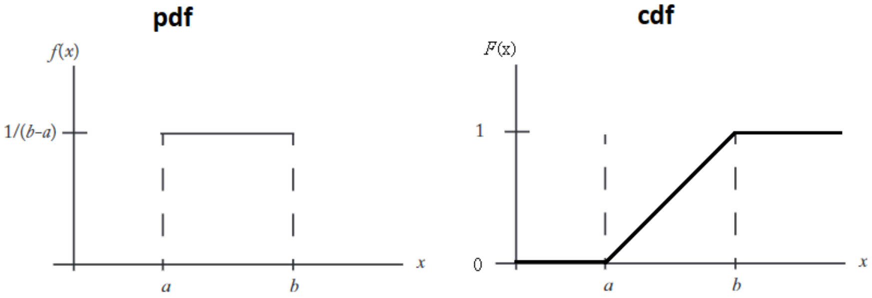
- The expectation is

$$E(X) = \int_a^b x \frac{1}{b-a} dx = \frac{a+b}{2}.$$

- The variance is

$$Var(X) = \int_a^b \left(x - \frac{a+b}{2}\right)^2 \frac{1}{b-a} dx = 12(b-a)^2.$$

Uniform distribution: $U(a, b)$



Normal distribution

If a random variable X follows the normal distribution (μ, σ^2), then

- The parameters are μ and σ^2.
- The pdf is

$$f(x)\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/(2\sigma^2)}, \quad -\infty < x < \infty.$$

- The cdf is

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}}e^{-(s-\mu)^2/(2\sigma^2)}ds, \quad -\infty < x < \infty$$

- The expectation is

$$E(X) = \int_{-\infty}^{\infty} x\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/(2\sigma^2)}dx = \mu.$$

- The variance is

$$Var(X) = \int_{-\infty}^{\infty} (x-\mu)^2\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/(2\sigma^2)}dx = \sigma^2.$$

Normal distribution

Properties:

- The parameter μ is for the mean (or location) of the distribution, and the parameter σ^2 is for the variance (or dispersion) of distribution.
- The mean, median, and mode have the same value.
- The curve of the pdf is symmetric.
- The total area under the curve is 1. (why?)
- About 68% (95% and 99%) of the area of a normal distribution is within one (two and three) standard deviation σ of the mean μ.

Standard normal score (z-score)

The standard normal score is the standardized value of a normally distributed random variable and it is usually referred as z-score. Let X be from a $N(\mu, \sigma^2)$, and the z-score is

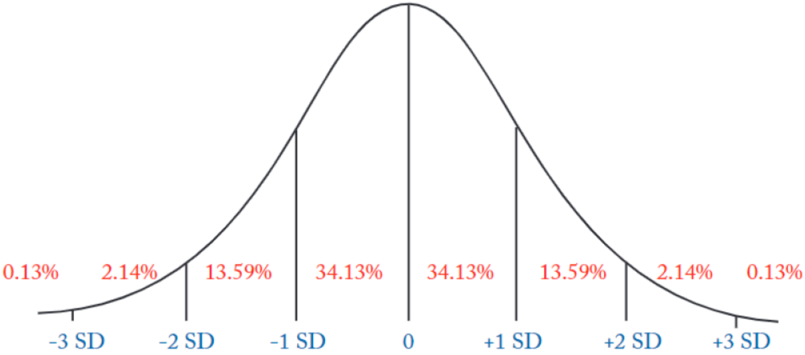
$$Z = \frac{X - \mu}{\sigma}.$$

- Is Z is a random variable? What is its distribution?
- It is a version of **standardization**.
- It is approximately normal with mean of 0 and standard deviation of 1.
- The pdf of Z is

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty,$$

  which includes no parameter, and it is called a standard normal distribution.

The normal distribution



Practice in R

Let the scores be normally distributed, and the mean is 65 and the standard deviation is 12.5. What is the percentage of scores above 70? How to do it in R?

.

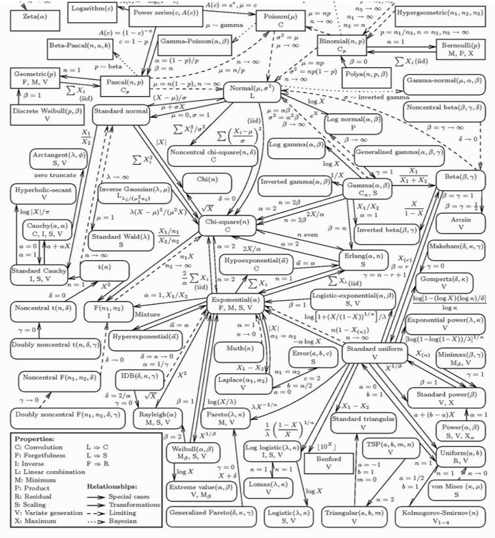
Some random variables

Binomial, Poisson, Geometric, Negative Binomial, Hypergeometric, Uniform, Discrete Uniform, Normal, Cauchy, t, F, χ^2, Exponential, Gamma, Beta, Dirichlet, Bernoulli, Bivariate Normal, ...

(Do you remember the cdf, pmf (pdf), expectation, and variance of each distribution?)

- Discrete:
- Continuous:
  - $x > 0$:
  - $-\infty < x < \infty$:
  - $0 < x < 1$:
- Multivariate:

Relationship between distributions



Relationship between distributions (I)

A random variable X:

- Bernoulli: $P(X = 1) = p$ and $P(X = 0) = 1 - p$.
- Binomial: If X_1, X_2, ..., X_n are from the Bernoulli distribution, then $S = \sum_{i=1}^{n} X_i$ is the Binomial distribution, denoted by $Bin(n, p)$.
- Geometric: Let X denote the total number of the first success. (How to draw from the Bernoulli?)
- Negative Binomial: Let X denote the total number of the first r successes. (How to draw from the Bernoulli?)

Relationship between distributions (II)

A random variable X:

- If X_1, X_2, ..., X_n are from the exponential distribution with rate parameter λ, $Exp(\lambda)$, then $S = \sum_{i=1}^{n} X_i$ is the gamma distribution with shape n and rate λ.
- What is the relationship between the gamma distribution and the χ^2 distribution?
- What is the rate of an exponential distribution equivalent to the $\chi^2(2)$ distribution?

Relationship between distributions (III)

A random variable X:

- If X_1, X_2, ..., X_n are from the normal distribution, $N(\mu, \sigma^2)$, then

$$Z_i = \frac{X_i - \mu}{\sigma} \sim N(0, 1).$$

- Z_1^2, Z_2^2, ..., Z_n^2 are from the χ^2 distribution with degree of freedom 1.
- $S = \sum_{i=1}^{n} Z_i^2$ is the χ^2 distribution with degree of freedom n.
- If Y_1, ..., Y_n are from the t-distribution with the degree of freedom ν, what is the value of ν such that the t-distribution can be approximated to a standard normal distribution?

Thinking univariate data together with probability distributions helps good guess of your data

	Univariate data	Probability
Distribution	Bar chart/histogram	pmf/pdf
Distribution	empirical cdf	cdf
Measure of location	mean/median	expectation/median
Measure of dispersion	variance	variance

Homework 1

- Give the expressions of given distributions.
- Match data to the given distributions (Multiple selection).

Examples are as the following pages:

Question 1: Normal distribution

If a random variable X follows the normal distribution (μ, σ^2), then

- The parameters are μ and σ^2.
- The pdf is

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

- The cdf is

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(s-\mu)^2/(2\sigma^2)} ds, \quad -\infty < x < \infty.$$

- The expectation is $E(X) = \mu$.
- The variance is $Var(X) = \sigma^2$.

Question 2: Given a dataset and distributions

- Empirical Distributions with given distributions

1 Normal($\mu = 2$, $\sigma = 0.5$)

2 Gamma(shape = 8, rate = 4)