

Chapter 7: Theoretical Probability Distributions

- Random Variables
- Probability Distributions
- Binomial Distribution
- Poisson Distribution
- Normal Distribution (Bell-Shaped Curve)
- Calculation with Normal Distribution

Random Variable: a random variable is a variable that takes values with certain probability.

- Discrete random variable: only takes finite or countable many number of values.
- Continuous random variable: can take any value within a specified interval or continuum.

Probability Distribution

Example Toss a coin once. There are 2 possible outcomes: head and tail. Let

$$X = \begin{cases} 1 & \text{if it is a Head} \\ 0 & \text{if it is a Tail} \end{cases}$$

Suppose the probability of a head is $p = 1/2$. Then

$$P(X = 1) = p$$

$$P(X = 0) = 1 - p.$$

Example Toss a coin twice. The set of all possible outcomes is

$$S = \{HH, HT, TH, TT\}.$$

Let X be the number of heads. Then X can take values 0, 1 and 2. The probabilities are

$$P(X = 0) = (1 - p)^2$$

$$P(X = 1) = 2p(1 - p)$$

$$P(X = 2) = p^2.$$

Example Let X be a discrete random variable that represents the live birth order of each child born to a woman residing in the US in 1986 [*Vital and Health Statistics, 1986*].

Probab. dist. of X		Cumulative probab. dist.
x	$P(X=x)$	$P(X \leq x)$
1	0.416	0.416
2	0.330	0.746
3	0.158	0.904
4	0.058	0.962
5	0.021	0.983
6	0.009	0.992
7	0.004	0.996
8+	0.004	1.000
Total	1.000	

Binomial Distribution

Example: Flip a coin 3 times. Suppose that the probability of head is p . Let X be the number of heads out of 3 flips.

Outcome	Probability	No. of Heads
HHH	ppp	3
HHT	$pp(1-p)$	2
HTH	$p(1-p)p$	2
HTT	$p(1-p)(1-p)$	1
THH	$(1-p)pp$	2
THT	$(1-p)p(1-p)$	1
TTH	$(1-p)(1-p)p$	1
TTT	$(1-p)(1-p)(1-p)$	0

From this table, we see that

$$P(X = 0) = (1 - p)^3$$

$$P(X = 1) = 3p(1 - p)^2$$

$$P(X = 2) = 3p^2(1 - p)$$

$$P(X = 3) = p^3$$

Example: Suppose that in a certain population 52% of all recorded births are males. We interpret this to mean that the probability of a recorded male is 0.52. If we randomly select 5 birth records from this population, what is the probability that exactly 3 of the records will be male birth?

Assumptions for Binomial distribution

1. There are a fixed number of trials n , each of which results in one of mutually exclusive outcomes.
2. The outcomes of the trials are independent.
3. The probability of success p is constant for each trial.

General formula for Binomial distribution

For a binomial random variable X resulted from n trials with the probability of success p , the probability that there are exactly x successes of n outcomes is

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

The mean and variance of the Binomial distribution

For a Binomial random variable $X \sim \text{Binomial}(n, p)$,

$$E(X) = np,$$

$$\text{Var}(X) = np(1 - p).$$

Poisson Distribution: distribution of rare events

Example 0.1 In a study of suicides, Gibbons et al. (1990, Amer. J. Epidemiology, 132, S183-191) found that the monthly distribution of adolescent suicides in Cook county, Illinois, between 1977 and 1987 closely follow a Poisson distribution with parameter $\lambda = 2.75$. That is, for any randomly selected month during this decade, the probability of x suicides is

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

For example, the probability of three adolescent suicides in a month is

$$P(X = 3) = e^{-2.75} \frac{2.75^3}{3!} = 0.2216.$$

The probability of either 3 or 4 suicides is

Poisson distribution: approximation to Binomial(n, p)

When n is large and p small, then the Binomial distribution can be approximated by a Poisson distribution. Specifically, let $\lambda = np$, then

$$P(X = x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \approx e^{-\lambda} \frac{\lambda^x}{x!}.$$

This approximation is satisfactory is $n \geq 20$ and $p \leq 0.05$, and is quite accurate if $n \geq 100$ and $np \leq 10$.

Example: Let X be a random variable representing the number of motor vehicle accidents per 10,000 people each year in the US, where the probability that an individual is involved in an accident is 0.00024. Strictly speaking, X is a binomial random variable with parameters $n = 10,000$ and $p = 0.00024$. But here we approximate the distribution of X by a Poisson distribution with parameter $\lambda = np = 10000 \times 0.00024 = 2.4$.

X	Binomial	Poisson
0	0.09069	0.0907
1	0.21771	0.2177
2	0.26129	0.2613
3	0.20904	0.2090
4	0.12541	0.1254
5	0.06019	0.0602
6	0.02407	0.0241

Assumptions underlying the Poisson distribution

1. The probability that a single event occurs in an interval is proportional to the length of the interval.
2. Theoretically, within a finite interval an infinite number of occurrences of the event are possible.
3. The events occur independently both within the same interval and between consecutive intervals.

Normal distribution

The normal distribution is perhaps the most important distribution in statistics. This is because many measurements have normal distributions, at least approximately. This is due to the central limit theorem (CLT) effect. Roughly, the CLT says that, if many factors contribute to the value of a measurement, and none of the factors dominant the others, then this measurement has an approximate normal distribution.

For example, in quantitative genetics, many continuous traits (height, weight, etc.) tend to have normal distributions. One model to explain this phenomenon is that, these traits are influenced by multiple genes, each of these genes have a small effect on the traits (polygenic effect). This model was put forward by R.A. Fisher in his 1918 paper “The correlation between relatives on the supposition of Mendelian inheritance.” Fisher examined a mathematical model of multifactorial inheritance. This model shows that the variation in continuous traits can be maintained by the independent segregation of multiple genes.

Galton was immensely impressed with the observation that many natural phenomenon follows the normal distribution (“law of frequency of error”). He called it “the supreme law of unreason.”

The form of the normal density function:

A random variable X has a normal distribution if its probability density function is

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

Notationally, this is usually written as

$$X \sim N(\mu, \sigma^2).$$

A useful property: if $X \sim N(\mu, \sigma^2)$, then

$$Z \equiv \frac{X - \mu}{\sigma} \sim N(0, 1).$$

Some useful numbers associated with the standard normal distribution

z	Area in the right tail
0.00	0.500
1.65	0.049
1.96	0.025
2.58	0.005
3.00	0.001

Example (probability calculation using standard normal table): Let X be a random variable that represents the systolic blood pressure of the population of 18- to 74-year-old males in the US. The distribution of X is approximately normal with mean 129 millimeters of mercury and standard deviation 19.8 millimeters of mercury. We can use the fact that

$$Z = \frac{X - 129}{19.8} \sim N(0, 1)$$

to do some calculations of interest.

(i) Find the value that cuts off the upper 2.5% of systolic blood pressures.

(ii) Find the value that cuts off the lower 2.5% of systolic blood pressures.

(iii) Find the proportion of the men who have systolic blood pressures greater than 150 mm Hg.

(iv) Find the proportion of the men who have systolic blood pressures greater than 100 mm Hg.

Example: For the population of men who are not taking corrective medication, diastolic blood pressure (DBP) is distributed as $N(80.7 \text{ mm Hg}, 9.2^2)$. For the men who are taking antihypertensive drugs, DBP is distributed as $N(94.9, 11.5^2)$. Suppose we are interested in determining whether a man is taking the antihypertensive medication solely on the basis of his DBP reading, and our goal is to identify 90% of the individuals who are taking medication.

(i) What value of DBP should be used as the lower cutoff value?

(ii) What is the proportion of the men with normal DBP will be incorrectly identified as the medication users?

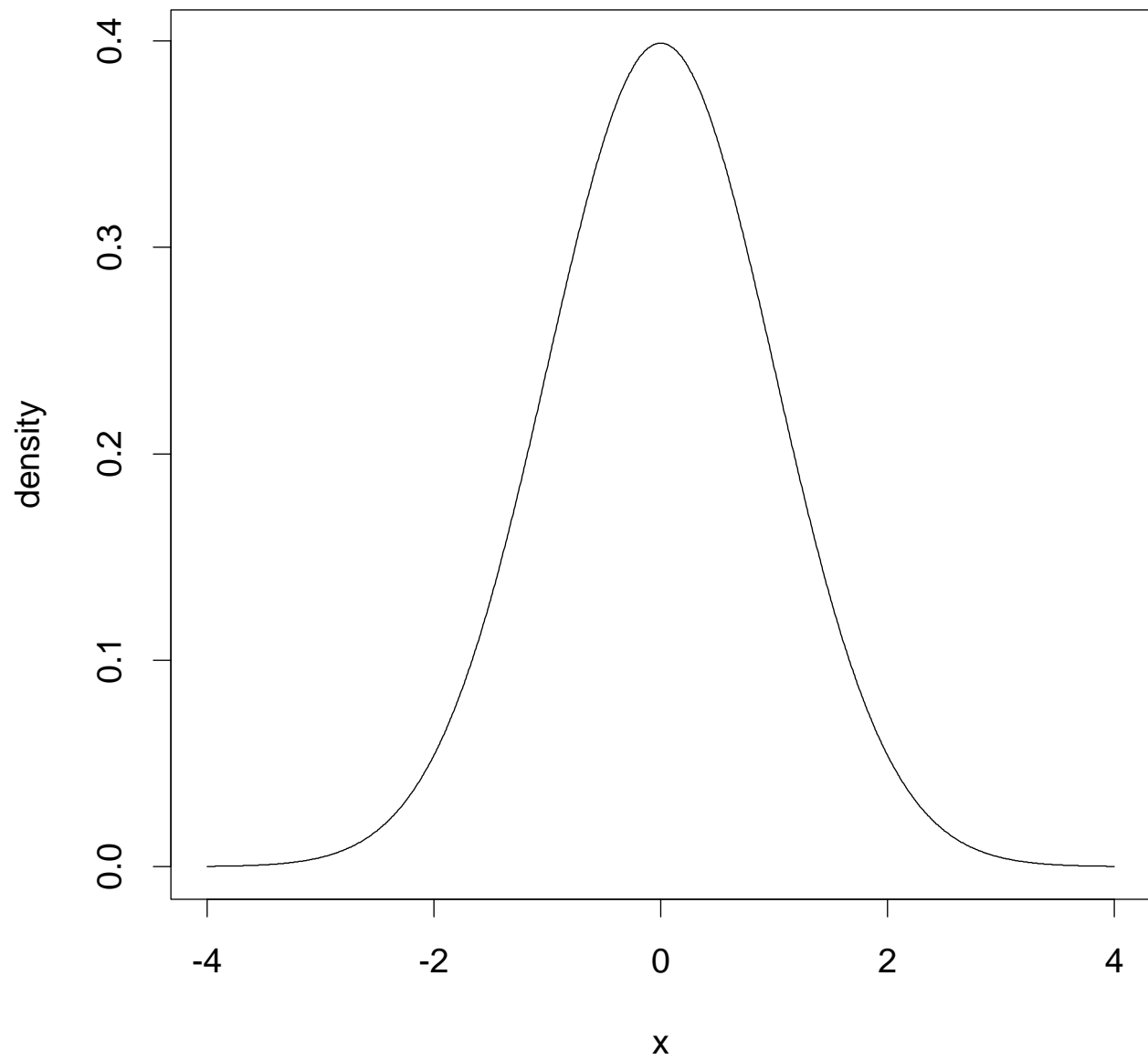
Comments on the normal distribution and the Central Limit Theorem by Galton.

I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the “law of frequency of error” [the normal distribution]. Whenever a large sample of chaotic elements is taken in hand and marshaled in the order of their magnitude, this unexpected and most beautiful form of regularity proves to have been latent all along.

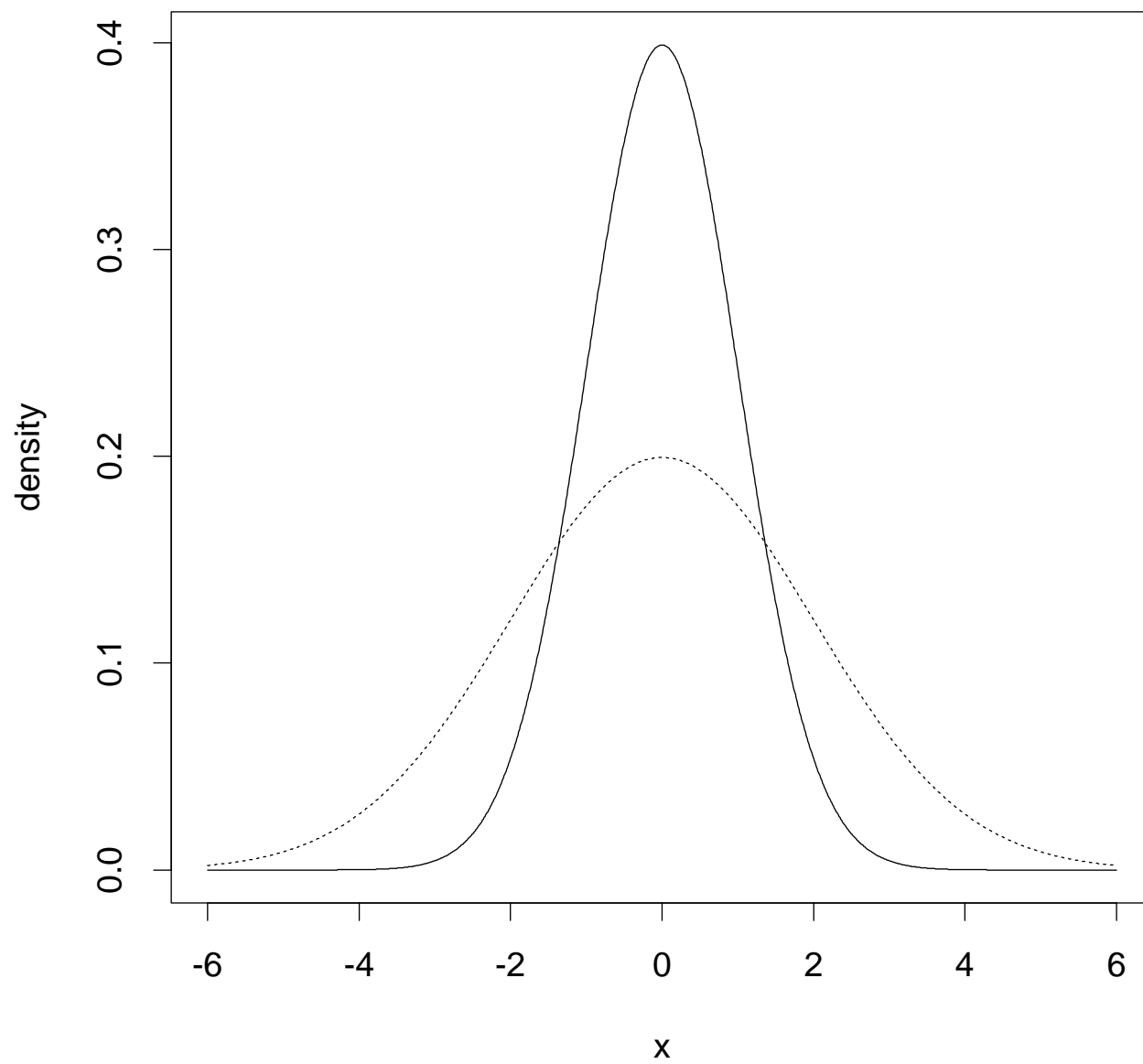
The law would have been personified by the Greeks if they had known of it. It reigns with serenity and complete self-effacement amidst the wildest confusion. The larger the mob and the greater the apparent anarchy, the more perfect is its sway. It is the supreme law of unreason.

—by Francis Galton [1822-1911], the inventor of regression and a pioneer in the application of statistics to biology.

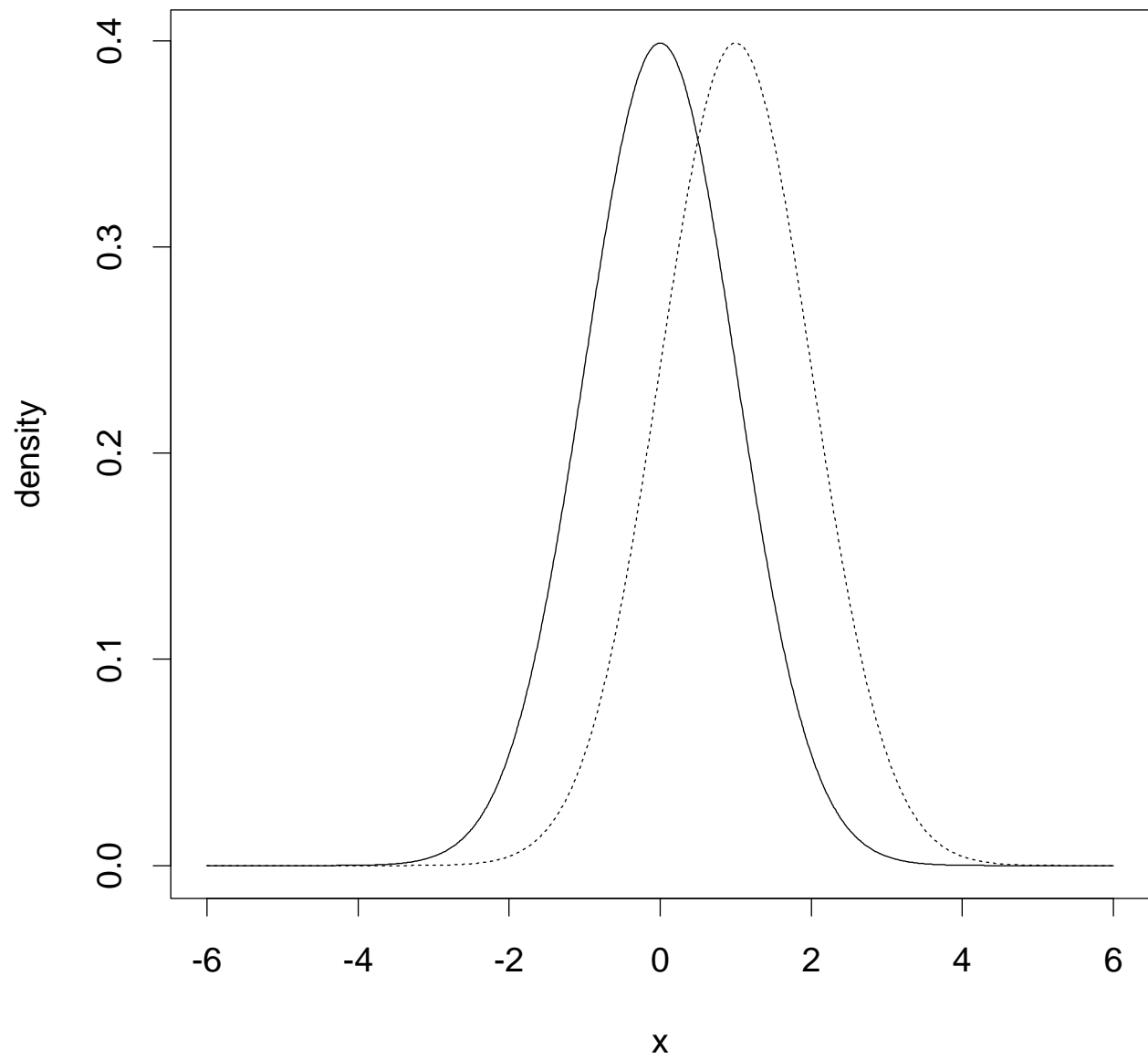
Standard Normal Density: $N(0,1)$



Normal Densities: $N(0,1)$ and $N(0, 4)$



Normal Densities: $N(0,1)$ and $N(1,1)$



Normal Densities: $N(0,1)$ and $N(1, 4)$ 