

### 2.2.3 Normal distributions

The *normal*, or *Gaussian* distribution is the most important and widely-used random variable in statistics.

**def'n:**  $X \sim N(\mu, \sigma^2)$  if  $X$  has pdf

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp \left\{ -\frac{1}{2\sigma^2} (x - \mu)^2 \right\} = (2\pi)^{-0.5} \sigma^{-1} e^{-0.5(x-\mu)^2/\sigma^2},$$

on  $R = (-\infty, \infty) = \mathbb{R}$ .

We require  $\sigma > 0$  for this to be a valid probability model.  $\mu$  is called the *mean* and  $\sigma^2$  is the variance of  $X$ ; more on this later.

**Why is the normal distribution so important?**

- The **Central Limit Theorem** tells us that measurements which are the sum of lots of independent, smaller measurements are normally distributed, e.g. atmospheric or underwater “white noise.” See Example A & Figures 2.14 and 2.15 on pp. 55-56.
- Measurement error is often approximately normal. So is IQ, height, blood pressure, etc. Often measurements that are not approximately normal, such as salary, can be transformed to be approximately normal, using  $\log(\cdot)$ ,  $\sqrt{\cdot}$ , etc.
- Normality is the fundamental assumption in many statistical techniques and models, including one and two-sample problems and  $t$ -tests, ANOVA, regression, ANCOVA, mixed models, etc.
- Often estimators for unknown population parameters (e.g.  $\theta$  in exponential data,  $(\mu, \sigma^2)$  in normal data), are approximately normal in large samples.

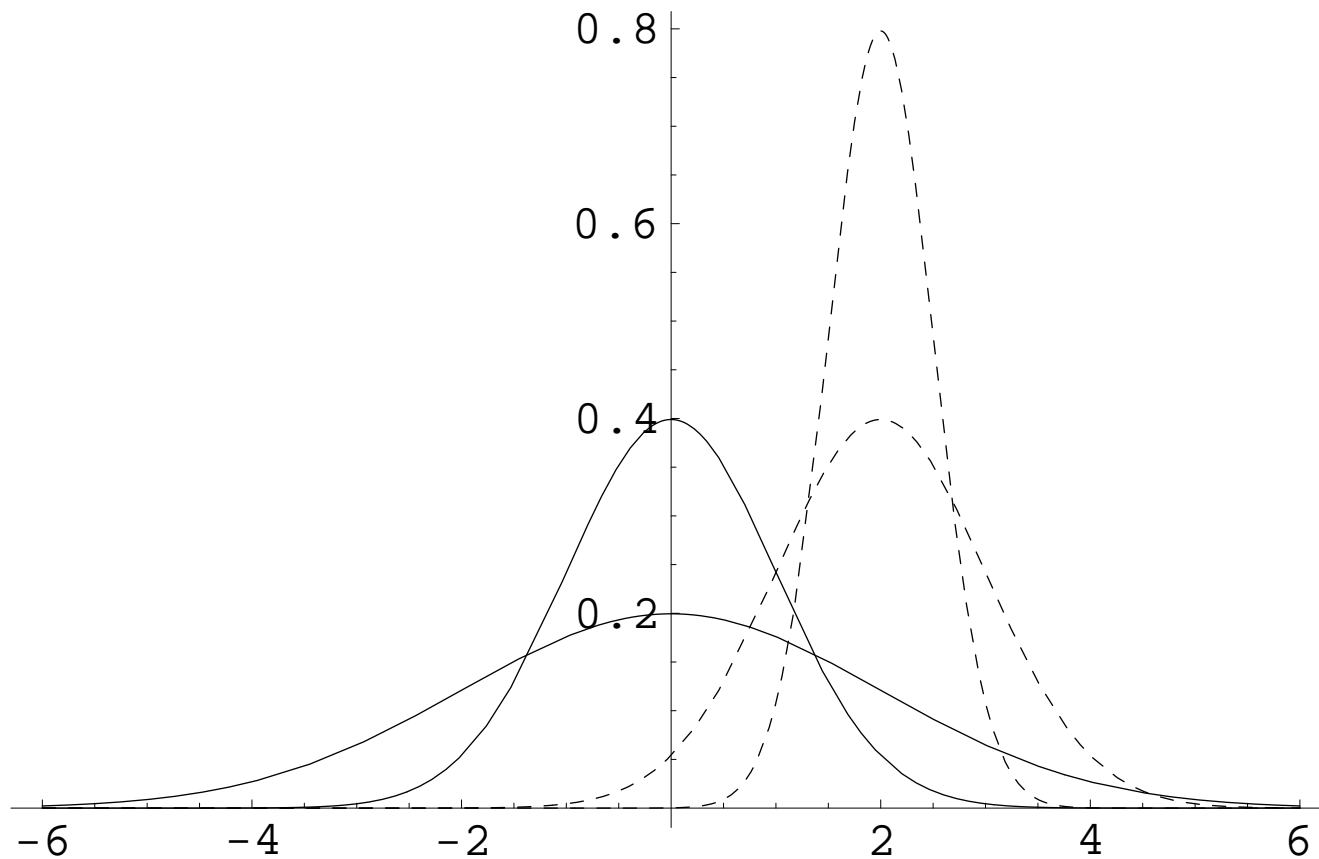


Figure 1:  $N(0, 1^2)$ ,  $N(0, 2^2)$ ,  $N(2, 1^2)$  and  $N(2, (\frac{1}{2})^2)$  densities.

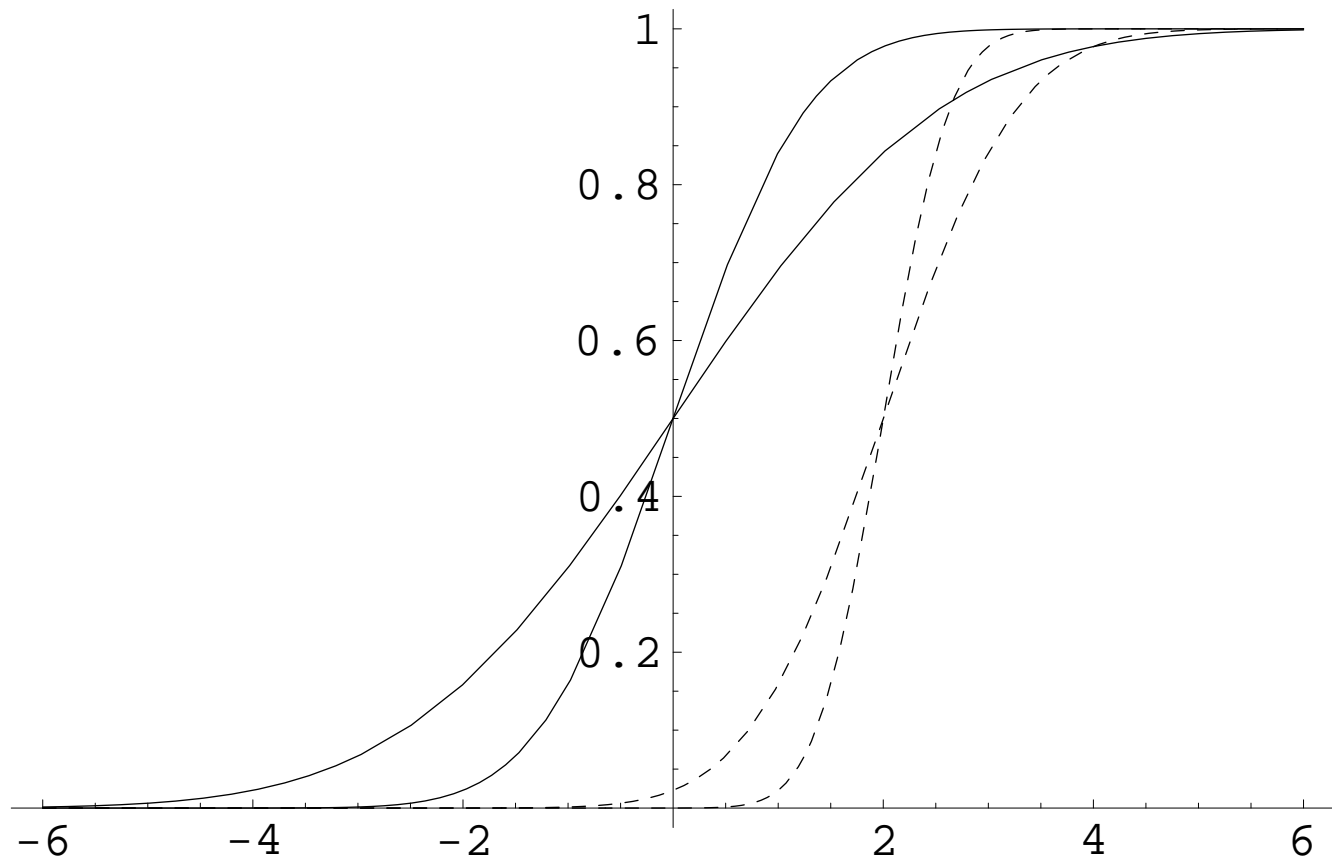


Figure 2:  $N(0, 1^2)$ ,  $N(0, 2^2)$ ,  $N(2, 1^2)$  and  $N(2, (\frac{1}{2})^2)$  cdf's.

## Comments:

- For  $X \sim N(\mu, \sigma^2)$ , the parameter  $\mu$  is the center of the density, i.e. the highest point, and the density is symmetric about  $\mu$ . The parameter  $\sigma$  controls how “spread out” the density is, with larger values of  $\sigma$  making  $f(x)$  more spread out.
- One can show that for  $X \sim N(\mu, \sigma^2)$ ,

$$P(\mu - \sigma \leq X \leq \mu + \sigma) \approx 0.683.$$

$$P(\mu - 2\sigma \leq X \leq \mu + 2\sigma) \approx 0.955.$$

$$P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) \approx 0.997.$$

- There is no closed form for the cdf  $F(x)$ . We’ll talk about how to compute probabilities from a normal random variable in the next section.

## 2.3 Functions of a random variable

If  $X$  is a random variable, we can think of taking some function of  $X$ , say  $Y = g(X)$ . Because  $X$  is random and has a distribution,  $Y$  is also random and has a distribution.

Example: Let  $X$  be the temperature in degrees Celsius on a random December day in St. Paul.  $Y = \frac{9}{5}X + 32$  is the temperature in degrees Fahrenheit. If  $X$  has a pdf  $f_X(x)$ , then  $Y$  will also have a pdf  $f_Y(y)$  that is *induced*, or obtained from  $f_X(x)$ . It is often easier to initially work with cdf's.

**Useful result:** If  $R_X$  is the range of  $X$ , then the range of  $Y = g(X)$  is  $R_Y = \{g(x) : x \in R_X\}$ .

For discrete random variables, one needs to simply compute the range of  $Y = g(X)$ , denoted  $R_Y = \{g(x) : x \in R_X\}$  and associated probabilities  $p_Y(y) = P(Y = y) = P(g(X) = y)$ .

Example: Let  $X$  be the number of children out of  $n = 4$  with Tay-Sachs from a previous example; recall  $X \sim \text{bin}(4, 0.25)$ . Let  $Y = 1$  if one or more children have Tay-Sachs and  $Y = 0$  if no children have Tay-Sachs.

$$p_Y(0) = P(X = 0) = 0.75^4 \approx 0.316.$$

$$p_Y(1) = P(X \in \{1, 2, 3, 4\}) = 1 - P(X = 0) \approx 0.685.$$

So  $Y \sim \text{Bern}(0.685)$ .

Example. Let  $X$  be discrete with  $P(X = j) = \frac{|j|+1}{11}$  for outcomes  $R_X = \{-2, -1, 0, 1, 2\}$ . Let  $Y = X^2$ . The range of  $Y$  is  $R_Y = \{g(x) : x \in R_X\} = \{0, 1, 4\}$ . The pmf of  $Y$  over these three values is calculated

$$p_Y(0) = P(Y = 0) = P(X^2 = 0) = P(X = 0) = \frac{1}{11}.$$

$$\begin{aligned} p_Y(1) &= P(Y = 1) = P(X^2 = 1) \\ &= P(X = -1 \text{ or } X = 1) = P(X = -1) + P(X = 1) = \frac{4}{11}. \end{aligned}$$

$$\begin{aligned} p_Y(4) &= P(Y = 4) = P(X^2 = 4) \\ &= P(X = -2 \text{ or } X = 2) = P(X = -2) + P(X = 2) = \frac{6}{11}. \end{aligned}$$

For continuous  $X$  and  $Y = g(X)$  the trick is to write the unknown cdf of  $Y$  in terms of the known cdf of  $X$  then compute  $F_Y(y)$ . One can differentiate to get the density if needed,  $f_Y(y) = F'_Y(y)$ .

Example: Let  $X \sim U(0, 1)$  &  $Y = \sqrt{X}$ . What's the distribution of  $Y$ ?

The range of possible outcomes for  $X$  is  $R_X = (0, 1)$ . Think of taking the square root of each number in  $R_X$ ; the range of  $Y$  is

$R_Y = \{\sqrt{x} : x \in R_X\} = (0, 1)$  too. A plot of  $Y$  versus  $X$  helps.

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(\sqrt{X} \leq y) \\ &= P(X \leq y^2) \\ &= F_X(y^2) = y^2 \end{aligned}$$

on  $R_Y = (0, 1)$  because  $F_X(x) = x$ . Note then  $f_Y(y) = F'_Y(y) = 2y$  on  $R_Y$ . What kind of random variable is  $Y$ ?

Example: Say  $X \sim U(0, 1)$ . What is the distribution of  $Y = -\log(X)$ ?

$$\begin{aligned}F_Y(y) &= P(Y \leq y) \\&= P(-\log(X) \leq y) \\&= P(X \geq e^{-y}) \\&= 1 - F_X(e^{-y}) \\&= 1 - e^{-y}.\end{aligned}$$

The derivative gives the pdf of  $Y$ ,

$$f_Y(y) = F'_Y(y) = \frac{d}{dy}[1 - e^{-y}] = e^{-y},$$

on  $R_Y = (0, \infty)$ . So  $Y \sim \exp(1)$ .

**Proposition:** If  $X$  has density  $f_X(x)$ , then  $Y = a + bX$  where  $a$  and  $b > 0$  are known constants has density  $f_Y(y) = \frac{1}{b} f_X\left(\frac{y-a}{b}\right)$ .

Let's prove this one:

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(a + bX \leq y) \\ &= P(bX \leq y - a) \\ &= P\left(X \leq \frac{y - a}{b}\right) \\ &= F_X\left(\frac{y - a}{b}\right). \end{aligned}$$

Now use the chain rule to get

$$f_Y(y) = F'_Y(y) = \frac{1}{b} f_X\left(\frac{y - a}{b}\right).$$

We have two immediate results for normal distributions. The first one comes in handy when we consider the *delta method* later on.

**Corollary:** Let  $X \sim N(\mu, \sigma^2)$ ,  $a$  be any constant, and  $b > 0$ . If  $Y = a + bX$  then  $Y \sim N(a + \mu, b^2\sigma^2)$ .

This one relates *any*  $X \sim N(\mu, \sigma^2)$  to a standard normal  $Z \sim N(0, 1)$ :

**Corollary:** Let  $X \sim N(\mu, \sigma^2)$ . If  $Z = \frac{X - \mu}{\sigma}$  then  $Z \sim N(0, 1)$ .

The standard normal  $Z \sim N(0, 1)$  is important enough to have its own symbol for its cdf, the capital Greek letter “phi,”  $\Phi(z) = P(Z \leq z)$  where  $Z \sim N(0, 1)$ . Every statistics textbook written has a table of  $\Phi(z)$  probabilities in it (see p. A7).

Example: Let  $S$  be the systolic blood pressure of a randomly selected American adult. Assume that approximately  $S \sim N(127, 15^2)$  (Wolf et al., 1997 *Journal of Human Hypertension*; Wolf-Maier et al., 2003 *Journal of the American Medical Association*).

Hypertension is defined to be prolonged resting blood pressure over 140/90. What is the probability of randomly selecting an adult with with systolic blood pressure over 140?

$$\begin{aligned} P(S > 140) &= P(S - 127 > 140 - 127) \\ &= P\left(\frac{S - 127}{15} > \frac{140 - 127}{15}\right) \\ &= P(Z > 0.87) = 1 - P(Z \leq 0.87) \\ &= 1 - \Phi(0.87) \approx 1 - 0.81 = \mathbf{0.19}. \text{ (p. A7)} \end{aligned}$$

Some estimates put hypertension prevalence at about 1 in 5; does this agree with what we computed above?

Let  $Z \sim N(0, 1)$ . What is the distribution of  $Y = Z^2$ ? We know  $R_Y = \{z^2 : -\infty < z < \infty\} = (0, \infty)$ . So where  $y > 0$ ,

$$\begin{aligned}F_Y(y) &= P(Y \leq y) \\&= P(Z^2 \leq y) \\&= P(-\sqrt{y} \leq Z \leq \sqrt{y}) \\&= \Phi(\sqrt{y}) - \Phi(-\sqrt{y}).\end{aligned}$$

To get the pdf we differentiate

$$\begin{aligned}f_Y(y) &= F'_Y(y) = \Phi'(\sqrt{y}) - \Phi'(-\sqrt{y}) \\&= (2\pi)^{-0.5} 0.5y^{-0.5} \exp(-0.5y) + (2\pi)^{-0.5} 0.5y^{-0.5} \exp(-0.5y) \\&= \frac{1}{\sqrt{2\pi}} y^{-0.5} \exp(-0.5y)\end{aligned}$$

So  $Y \sim \text{gamma}(0.5, 0.5)$ .

The general approach of finding the cdf  $F_Y(y)$  of  $Y = g(X)$  and differentiating can be formalized into a proposition.

$$\begin{aligned} F_Y(y) &= P(Y \leq y) \\ &= P(g(X) \leq y) \\ &= P(X \leq g^{-1}(y)) \\ &= F_X(g^{-1}(y)). \end{aligned}$$

The third equality holds if  $g(x)$  is strictly increasing or decreasing over  $R_X$ . Now use the chain rule to differentiate and get the following proposition.

**Proposition:** Let  $X$  be a continuous random variable with range that's an interval  $R_X = (a, b)$  and let  $g(x)$  be a strictly increasing or strictly decreasing function defined on  $R_X$ . Then the density of  $Y = g(X)$  is given by

$$f_Y(y) = \left| \frac{dg^{-1}(y)}{dy} \right| f_X(g^{-1}(y)),$$

on  $R_Y = \{g(x) : x \in R_X\}$ .

Examples of strictly increasing or decreasing functions on  $(a, b)$  are

- $\log(x)$  on  $(0, \infty)$
- $\sqrt{x}$  on  $(0, \infty)$ ;  $x^a$  on  $(0, \infty)$  for any  $a \neq 0$ .
- $e^x$  on  $(-\infty, \infty)$

A final result that is extraordinarily useful for generating random numbers is

**Proposition:** Let  $F(x)$  be any cdf. Let  $U \sim U(0, 1)$  and  $X = F^{-1}(U)$ . Then  $X$  is distributed according to  $F(x)$ .

Most programming languages have automated routines to generate  $U(0, 1)$  random variables.

For example, if you want to simulate a random draw from an  $\exp(1)$  distribution, compute  $F^{-1}(x) = -\log(x)$ , then simulate  $U \sim U(0, 1)$  and set  $X = -\log(U)$ . Look at the slide 10.

The next slide is a histogram of  $n = 10,000$  independent draws from a  $U(0, 1)$  distribution. The following slide is a histogram where each  $U(0, 1)$  draw is transformed to  $-\log(U)$ .

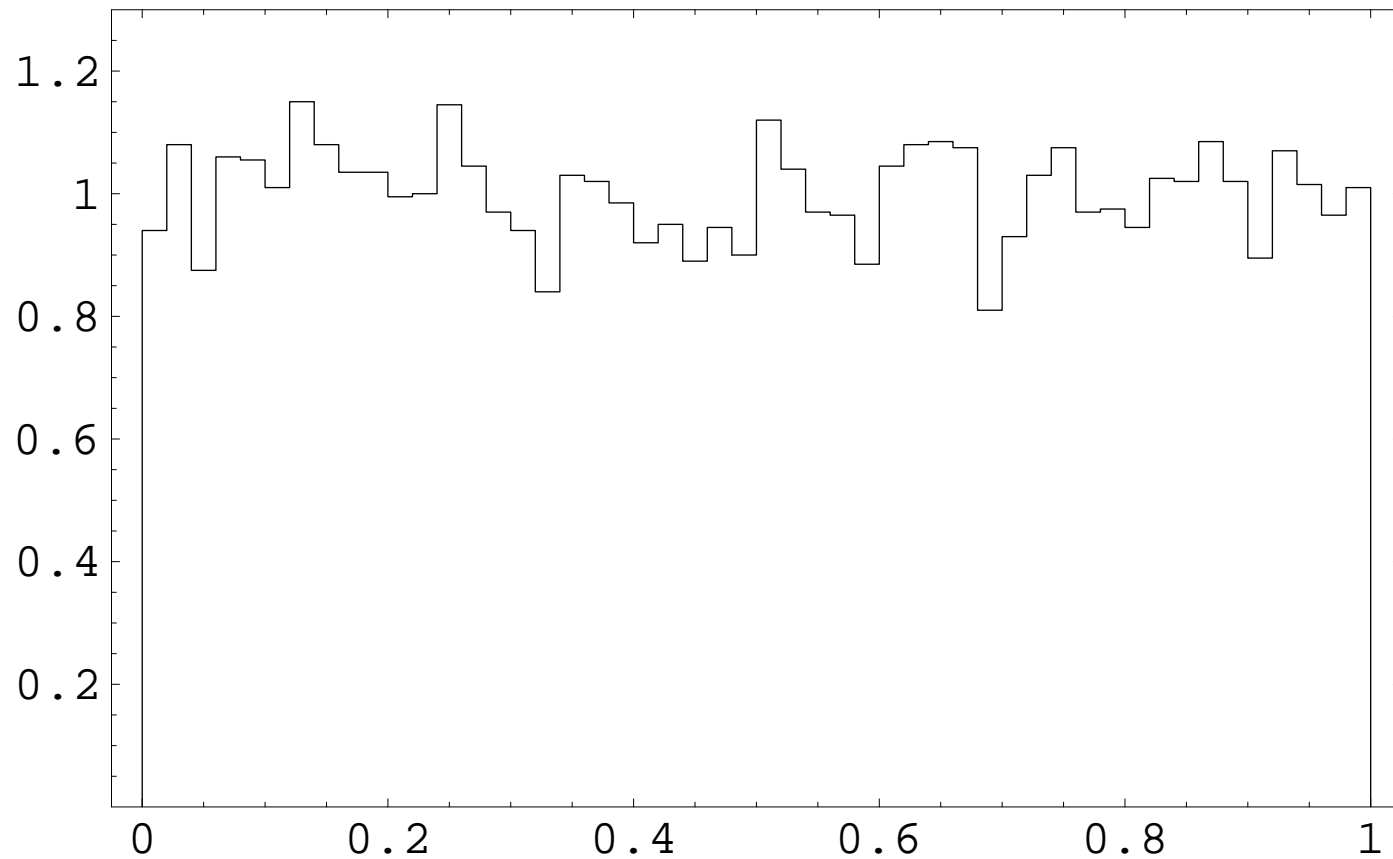


Figure 3: Histogram of  $U_1, U_2, \dots, U_{10000} \stackrel{iid}{\sim} U(0, 1)$ .

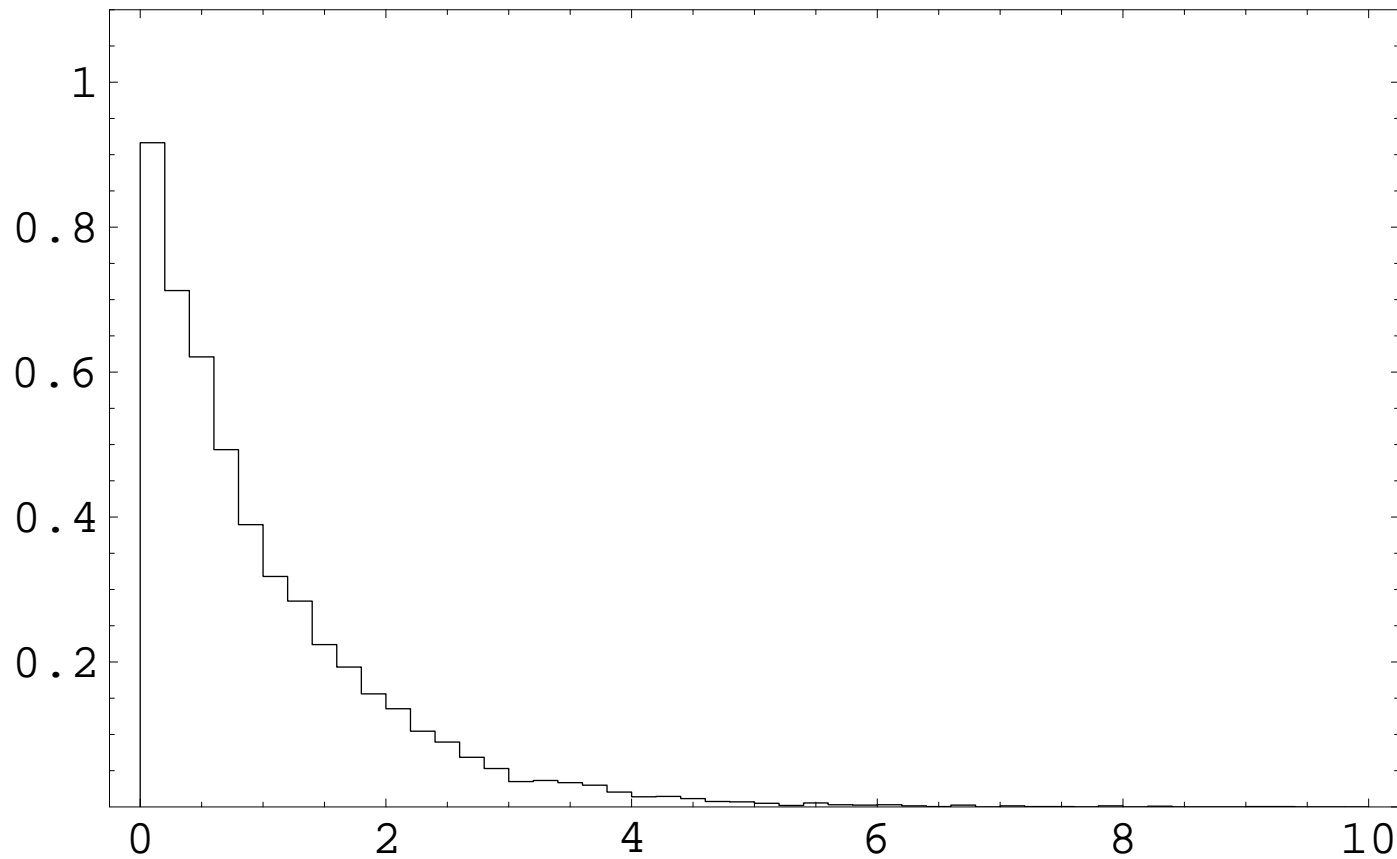


Figure 4: Histogram of  $X_1, X_2, \dots, X_{10000} \stackrel{iid}{\sim} \exp(1)$ .

Homework 4, Chapter 2: 52\*, 53, 54\*, 55, 56\* (find the cdf, not pdf), 57, 58\*, 59, and additional exercises posted on the course webpage.

\*hand in Thursday Oct. 2.