Lab 4. Normal Random Variables

Objectives

- Normal distribution in R
- Related statistics, properties, and simulation

The normal, a continuous distribution, is the most important of all the distributions. It is widely used and even more widely abused. Its graph is bell-shaped. You see the bell curve in almost all disciplines. Some of these include psychology, business, economics, nursing, and, of course, statistics. Most IQ scores are normally distributed. Many bodily measurements fit a normal distribution. The normal distribution is extremely important, but it cannot be applied to everything in the real world. In this lab, you will study the normal distribution, the standard normal distribution, and applications associated with them.

Normal random variable – Continuous Case

$$X \sim \text{Normal}\left(\mu, \sigma^2\right)$$

Then,

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

When $\mu = 0$ and $\sigma = 1$, we get the *standard* normal distribution with the pdf

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

Here is R code to draw the pdf and cdf plot of the standard normal distribution.

```
x <- seq(-3,3,0.02)
pdf_x <- dnorm(x,0,1)
plot(x,pdf_x,type="l",col=4,ylim=c(0,1))
lines(x,cdf_x,type="l",lty=2,col=2)
abline(v=0,lty=2,col=3)
abline(h=0.5,lty=2,col=3)
legend(locator(1),c("pdf","cdf"),lty=1:2,col=c(4,2))
```

Z-score, standard score, standard unit

If X is a normally distributed random variable and $X \sim N\left(\mu, \sigma^2\right)$, then $z = \dfrac{x-\mu}{\sigma}$ is $N(0,1)$.

Proof. The cdf of Z is

$$\begin{aligned} P(Z \leq z) &= P\left(\frac{X-\mu}{\sigma} \leq z\right) = P\left(X \leq z\sigma + \mu\right) \\ &= \int_{-\infty}^{z\sigma+\mu} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \mathrm{d}x \end{aligned}$$

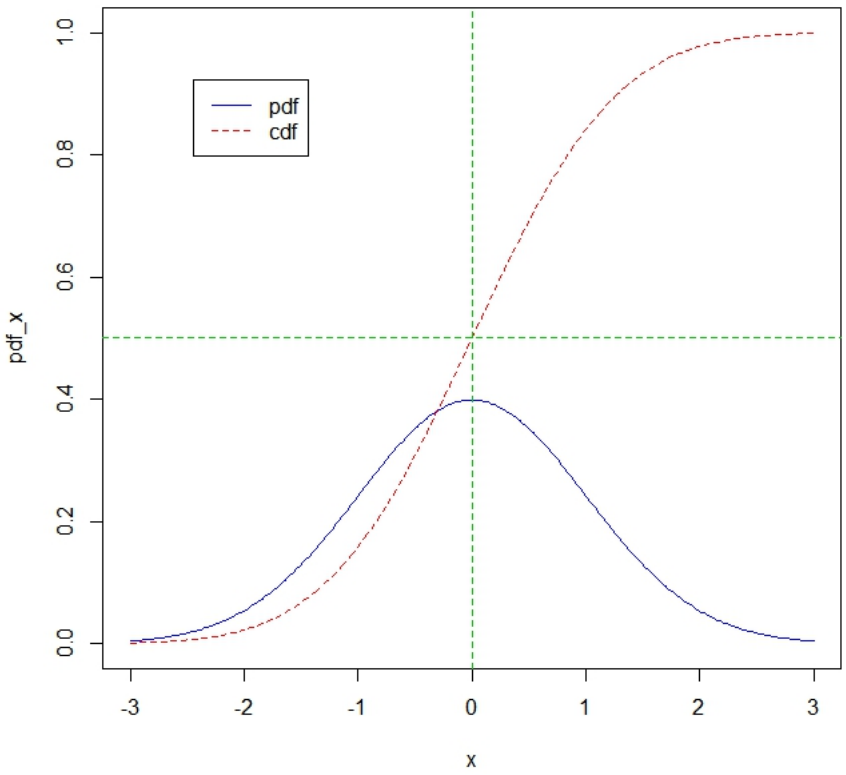

Figure 1: pdf and cdf of the standard normal random variable ($\mu = 0, \sigma = 1$)

Let $w = \dfrac{x - \mu}{\sigma}$

$$P(Z \leq z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-w^2/2} \, \mathrm{d}z \quad \sim \quad \text{cdf of } N(0,1)$$

□

In 2012, 1,664,479 students took the SAT exam. The distribution of scores in the verbal section of the SAT had a mean $\mu = 496$ and a standard deviation $\sigma = 114$. Let X be an SAT exam verbal section score in 2012. Then $X \sim N(496, 114^2)$.

1. Find the Z score of someone who scored 325. What's the student's percentile?
2. Find $P(X \geq 500)$.
3. About how many students have scored 500 or above?

```
(z1 <- (325-496)/114)
pnorm(z1)
z2 <- (500-496)/114
pnorm(z2)
1-pnorm(z2)
1664479*(1-pnorm(z2))
```

The Empirical Rule

If X is a random variable and has a normal distribution with mean μ and standard deviation σ, then the Empirical Rule says:

- About 68% of the x values lie between -1σ and $+1\sigma$ of the mean μ.
- About 95% of the x values lie between -2σ and $+2\sigma$ of the mean μ.
- About 99.7% of the x values lie between -3σ and $+3\sigma$ of the mean μ.

Notice that almost all the x values lie within three standard deviations of the mean.

- The z-scores that capture middle 68% are -1 and $+1$.
- The z-scores that capture middle 95% are -2 and $+2$.
- The z-scores that capture middle 99.7% are -3 and $+3$.

The empirical rule is also known as the "68-95-99.7 rule."

```
pnorm(1)-pnorm(-1)
pnorm(2)-pnorm(-2)
pnorm(3)-pnorm(-3)
```

Calculating probabilities

Suppose that the average number of hours a household personal computer is used for entertainment is two hours per day. Assume the times for entertainment are normally distributed and the standard deviation for the times is half an hour.

1. Find the probability that a household personal computer is used for entertainment between 1.8 and 2.75 hours per day.
2. Find the maximum number of hours per day that the bottom quartile of households uses a personal computer for entertainment.
3. Find the 80th percentile of this distribution, and interpret it in a complete sentence.

```
pnorm(2.75,2,0.5)-pnorm(1.8,2,0.5)
qnorm(0.25,2,0.5)
qnorm(0.8,2,0.5)
```

If you square z numbers, you get χ^2 with df=1, i.e., $Z^2 \sim \chi^2(1)$

If X is a normally distributed random variable and $X \sim N\left(\mu, \sigma^2\right)$, then $Z^2 = \left(\frac{X-\mu}{\sigma}\right)^2$ is $\chi^2(1)$.

Proof. Let $V = Z^2$, then the cdf of V is

$$P(V \le v) = P\left(Z^2 \le v\right) = P\left(-\sqrt{v} \le Z \le \sqrt{v}\right)$$
$$= 2\int_0^{\sqrt{v}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2}\,\mathrm{d}z$$

Let $z = \sqrt{y}$, then $\mathrm{d}z = \frac{1}{2\sqrt{y}}\,\mathrm{d}y$, and

$$G(v) = P(V \le v) = 2 \cdot \frac{1}{2}\int_0^y \frac{1}{\sqrt{2\pi y}} e^{-y/2}\,\mathrm{d}y = \int_0^y \frac{1}{\sqrt{2\pi y}} e^{-y/2}\,\mathrm{d}y$$

$$g(v) = G'(v) = \frac{1}{\sqrt{2\pi v}} e^{-v/2} \quad \sim \quad \Gamma\left(\frac{1}{2}, 2\right) = \chi^2(1) \text{ since } \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$$

□

```
z <- rnorm(1000,0,1)
mean(z^2); var(z^2); sd(z^2)  #Check if these stats are the same as from chi^2 (1)
x <- seq(0,12,0.02)
pdf_x <- dchisq(x,1)
plot(density(z^2),ylim=c(0,1),xlab=expression(z^2),main="")
lines(x,pdf_x,lty=2,col=2)
legend(locator(1),c(expression(z^2),expression(paste(chi^2,"(1)"))),lty=1:2,col=1:2)
```

The moment-generating-function $M(t) = exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$

Let's show the graphs of the three different normal distributions: $N(0,1), N(-1,1)$ and $N(2,1)$.

```
t <- seq(-2,2,0.01)
M01 <- exp(0*t + (1*t^2/2))
M_11 <- exp((-1*t) + (1*t^2/2))
M21 <- exp((2*t) + (1*t^2/2))
plot(M01~t,lty=1,col=1,type="l",ylab="",ylim=c(0,5))
lines(M_11~t,lty=2,col=2,type="l")
lines(M21~t,lty=3,col=4,type="l")
text(locator(1),c("N(0,1)"),cex=0.8)
text(locator(1),c("N(-1,1)"),cex=0.8,col=2)
text(locator(1),c("N(2,1)"),cex=0.8,col=4)
```

Selected Problems

1. At a police station in a large city, calls come in at an average rate of four calls per minute. Assume that the time that elapses from one call to the next has the exponential distribution. Take note that we are concerned only with the rate at which calls come in, and we are ignoring the time spent on the phone. We must also assume that the times spent between calls are

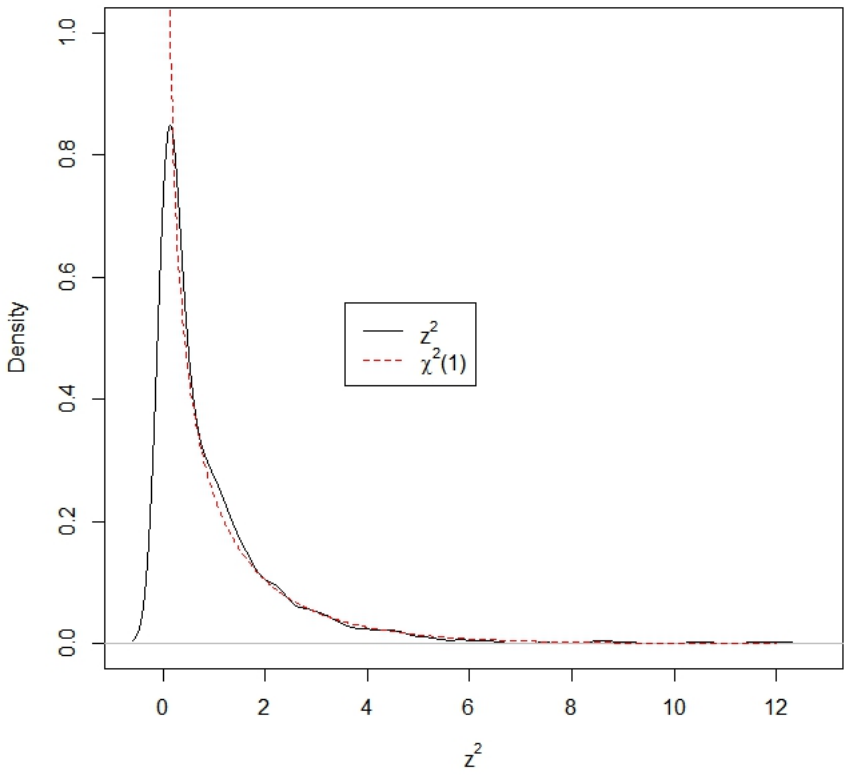

Figure 2: $Z^2 \sim \chi^2(1)$

independent. This means that a particularly long delay between two calls does not mean that there will be a shorter waiting period for the next call. We may then deduce that the total number of calls received during a time period has the Poisson distribution.

(a) Find the average time between two successive calls.

(b) Find the probability that after a call is received, the next call occurs in less than ten seconds.

(c) Find the probability that exactly five calls occur within a minute.

(d) Find the probability that less than five calls occur within a minute.

(e) Find the probability that more than 40 calls occur in an eight-minute period.

2. Cars arrive at a toolbooth at a mean rate of 5 cars every 10 minutes according to a Poisson process. Find the probability that the toll collector will have to wait longer than 26.30 minutes before collecting the eighth toll.

3. An aluminum screen 2 feet in width has, on average, three flaws in a 100-foot roll.

   (a) Define an appropriate random variable X and state the distribution and its relevant

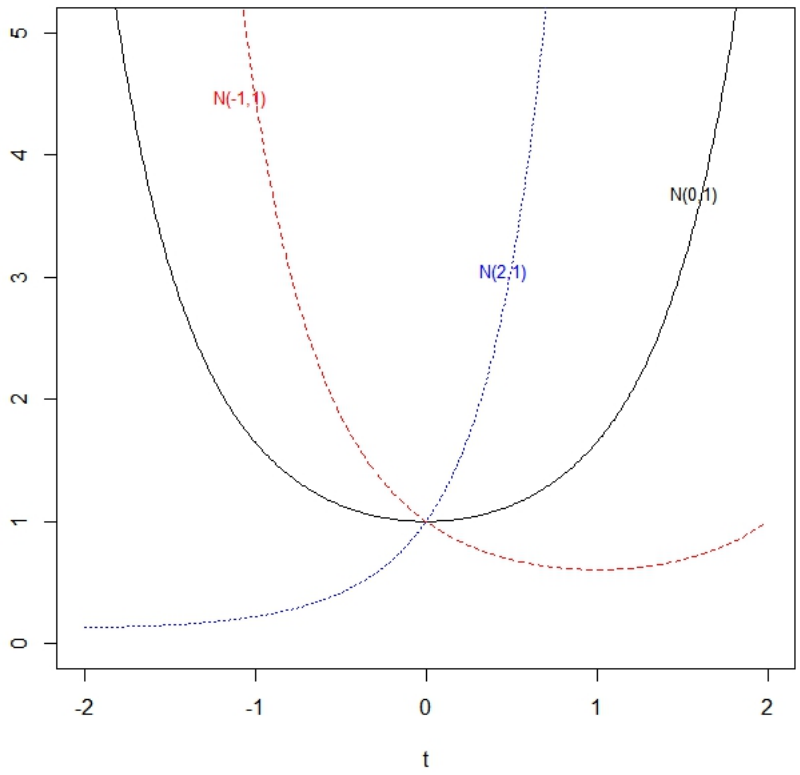

Figure 3: mgf of three normal distributions

parameters.

(b) Find the probability that the first 40 feet in a roll contains no flaws.

(c) What assumptions did you make to solve the part (b)?

(d) Plot a graph of the pdf of X.

4. A bakery sells rolls in units of a dozen. The demand X (in 1,000 units) for rolls has a gamma distribution with parameters $\alpha = 3$, $\theta = 0.5$, where θ is in units of days per 1,000 units of rolls. It costs \$2 to make a unit that sells for \$5 on the first day when the rolls are fresh. Any leftover units are sold on the second day for \$1. How many units should be made to maximize the expected value of the profit?