[STAT-05] Binomial, Poisson and normal distributions

Miguel-Angel Canela
Associate Professor, IESE Business School

The Bernouilli distribution

The probability distribution of a 0/1 variable is called a **Bernouilli distribution**. I use here the notation

$$\pi = \mathrm{p}\big[X = 1\big], \qquad 1 - \pi = \mathrm{p}\big[X = 0\big].$$

The Bernouilli distribution is a probability model with one parameter π ($0 < \pi < 1$). I use a Greek letter for consistency, but p is more popular. Since

$$\mathrm{E}[X] = \mathrm{E}\big[X^2\big] = \pi \cdot 1 + (1 - \pi) \cdot 0 = \pi,$$

the mean and the variance are directly obtained from the parameter π, as

$$\mu = \pi, \qquad \sigma^2 = \pi(1 - \pi).$$

Statisticians call a **Bernouilli trial** one with two possible outcomes, called **success** and **failure**. Of course, these labels are arbitrary and can be interchanged. We usually code success as 1 and failure as 0, getting a Bernouili distribution.

The binomial probability formula

The probability of having exactly k successes in n (statistically) independent Bernouilli trials is given by the **binomial probability formula**,

$$\mathrm{p}\big[X = k\big] = \binom{n}{k} \pi^k \big(1 - \pi\big)^{n-k}, \qquad k = 0, 1, \dots, n.$$

The first factor on the right side of the formula is the **combinatorial number**

$$\binom{n}{k} = \frac{n!}{k!\,(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{k!}\,,$$

also called **binomial coefficient** because of its role in the Newton's binomial formula. In most cases, we are interested in **cumulative probabilities**. For instance, for the probability of getting at most 2 heads throwing 10 coins, we sum the binomial probabilities from 0 to 2. In the computer, cumulative probabilities can be obtained directly.

Example 1. Let us examine how lucky can a student be in a standardized test. Suppose that the test consists of 20 multiple choice questions, each with four possible answers. If the student guesses on each question, what is the probability of getting at least 10 questions correct?

If the student is just guessing, the probability of being right is 1/4. The probability of k successes is then

$$\binom{20}{k} \left(\frac{1}{4}\right)^k \left(\frac{3}{4}\right)^{20-k}.$$

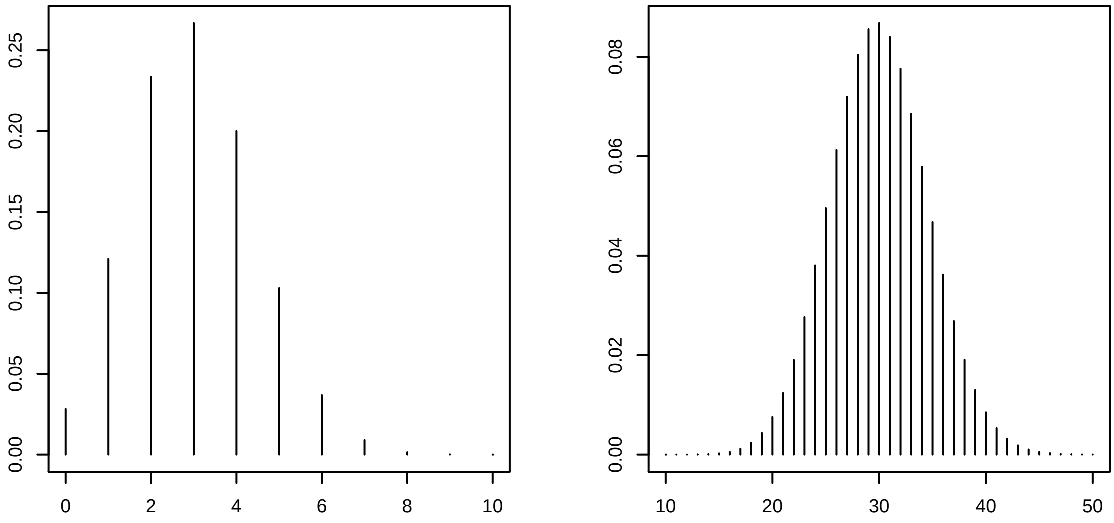

Figure 1. $\mathcal{B}(10, 0.3)$ **(left) and** $\mathcal{B}(100, 0.3)$ **(right)**

To get the probability of at least 10 successes, we sum these probabilities from $k = 10$ to $k = 20$, both included. Alternatively, we sum the probabilities from $k = 0$ to $k = 9$ and subtract the total from 1. The computer can do that for us. The result is 0.014. □

The multinomial probability formula

The extension of the binomial probability formula to trials with more than two possible outcomes is the **multinomial probability formula**. In a multinomial setting, we have a partition A_1, ..., A_r of the sample space, with probabilities π_1, ..., π_r. The probability of observing n_1 times A_1, n_2 times A_2, etc, in n independent trials, is

$$\pi\big(n_1, \dots, n_r\big) = \binom{n}{n_1 \cdots n_r} \pi_1^{n_1} \cdots \pi_r^{n_r}.$$

The first factor on the right side is the **multinomial coefficient**

$$\binom{n}{n_1 \cdots n_r} = \frac{n!}{n_1! \cdots n_r!}\,.$$

The binomial distribution

The **binomial distribution**, based on the binomial probability formula, gives, for $x = 0, 1, \dots, n$, the probability of x successes in n independent trials. It has two parameters, the number of trials n and the probability of success π. I denote the binomial distribution as $\mathcal{B}(n, \pi)$, writing $X \sim \mathcal{B}(n, \pi)$ to indicate that a variable X has this distribution. The Bernouilli distribution is then $\mathcal{B}(1, \pi)$.

The mean and the variance of the $\mathcal{B}(n, \pi)$ distribution can be calculated directly using the properties of the binomial coefficients, but it is much simpler to look at a binomial variable as the sum of n independent Bernouilli variables. The mean and the variance can then be obtained by multiplying by n those of the Bernouilli distribution. Hence, $\mu = n\pi$ and $\sigma^2 = n\pi(1 - \pi)$.

The application of the binomial distribution as a model for the probabilities related to the extraction of n units from a finite set (cards from decks, balls from urns, etc) is a classic. Since the

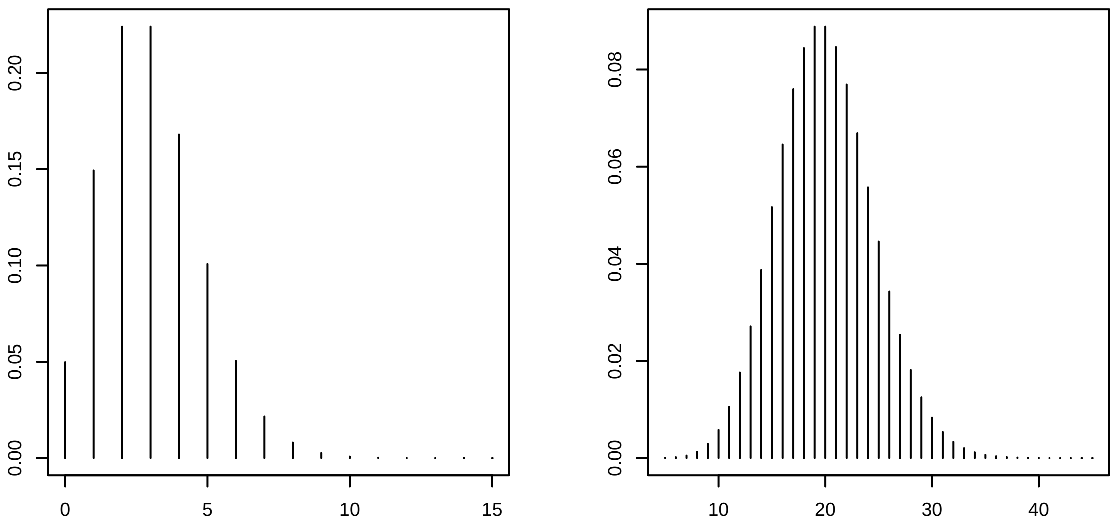

Figure 2. $\mathcal{P}(3)$ **(left) and** $\mathcal{P}(20)$ **(right)**

extractions must be statistically independent, the binomial distribution can only be used when each unit extracted is replaced before the next extraction. This is called **sampling with replacement**. The same question appears when sampling randomly from a population.

When we sample with replacement or from an infinite population, the binomial can be used as an exact model. When we sample without replacement from a big population (this means, in practice, that the population is much bigger than the sample), independence can be accepted, and the binomial is then used as an approximate model. When the population is not that big, the binomial is replaced by the **hypergeometric distribution**, not covered in this course.

The Poisson distribution

Let $\lambda > 0$. The **Poisson probability formula** is

$$\mathrm{p}\big[X = k\big] = e^{-\lambda}\,\frac{\lambda^k}{k!}\,, \qquad k = 0, 1, 2, \ldots$$

Using the Taylor expansion of the exponential function, it is easy to check that the sum of all these probabilities equals 1. So, the Poisson formula defines a probability distribution on the integers (including zero), called the **Poisson distribution**, which I denote by $\mathcal{P}(\lambda)$. It is a discrete distribution, with nonzero probability on the any integer $x = 0, 1, 2, \ldots$

The Poisson distribution is a probability model with one parameter, λ. By using again Taylor expansions, it can proved that

$$\mathrm{E}[X] = \lambda, \qquad \mathrm{E}\big[X^2\big] = \lambda + \lambda^2,$$

It follows that the mean and the variance of the $\mathcal{P}(\lambda)$ distribution are equal, $\mu = \sigma^2 = \lambda$. Although this is a very restrictive property, the Poisson distribution is very popular because of its simplicity. It is used as a model for the number of times that an event is observed in a certain context: the number of customers per hour at a service point, the number of accidents in a highway during the weekend, the number of patents per year in a company, etc. λ is then the mean number of occurrences of that event.

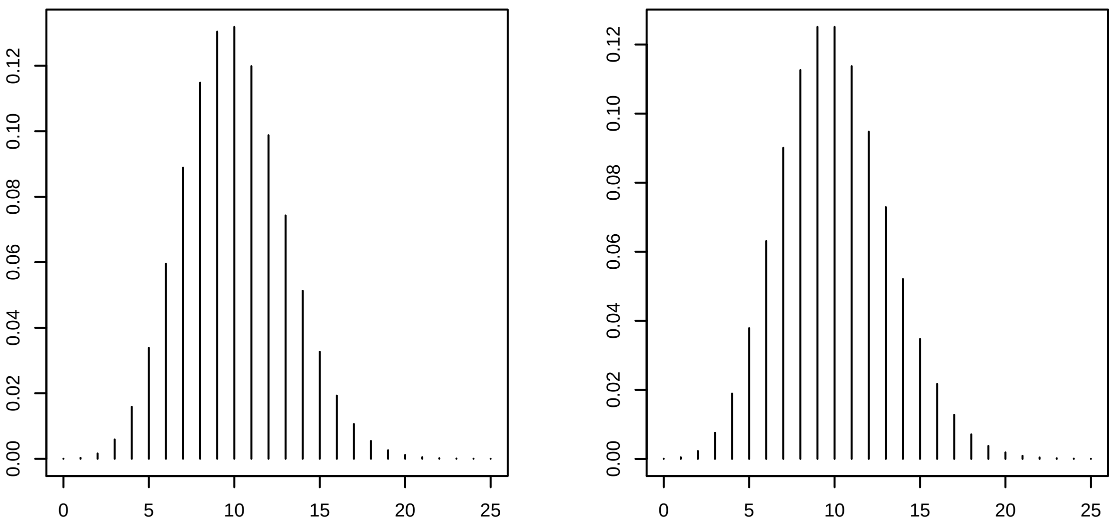

Figure 3. $\mathcal{B}(100, 0.1)$ **(left) and** $\mathcal{P}(10)$ **(right)**

Example 2. A telephone operator handles, on the average, five calls every three minutes. What is the probability that there will be no calls in the next minute? Of at least two calls?

Let X be the number of calls in a minute, and assume $X \sim \mathcal{P}(5/3)$. Then, the probability of zero calls is

$$\mathrm{p}\big[X = 0\big] = \frac{e^{-5/3}(5/3)^0}{0!} = 0.189,$$

and that of at least two calls,

$$\mathrm{p}\big[X \geq 2\big] = 1 - \mathrm{p}\big[X \leq 1\big] = 1 - \frac{e^{-5/3}(5/3)^0}{0!} - \frac{e^{-5/3}(5/3)^1}{1!} = 0.496. \ \square$$

Computation troubles with count data are, nowadays, a tale of the past, but in textbooks we may still find a second argument for the popularity of the Poisson distribution, that it can be used as an approximation for the binomial. Indeed, it can be proved that, as $n \to \infty$ and $\pi \to 0$ in such a way that $n\pi \to \lambda$, the binomial tends to the Poisson distribution. For instance, $\mathcal{B}(100, 0.03)$ can be approximated by $\mathcal{P}(3)$, or $\mathcal{B}(100, 0.1)$ by $\mathcal{P}(10)$ (Figure 3).

Finally, another nice property of the Poisson distribution is that we can add Poisson distributions under certain restrictions: the sum of two independent Poisson variables, with means λ_1 and λ_2, is a Poisson variable with mean $\lambda_1 + \lambda_2$. The proof is based on the properties of binomial coefficients.

The standard normal distribution

The **normal distribution** is statisticians' favorite distribution. I start with the **standard normal distribution**, for which a specific notation is typical in textbooks. A standard normal variables is denoted by Z, and the corresponding PDF and CDF by $\phi(z)$ and $\Phi(z)$, respectively. The formulas are

$$\phi(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}}\,, \qquad \Phi(z) = \int_{-\infty}^{z} \frac{e^{-t^2/2}}{\sqrt{2\pi}}\,dt.$$

$\sqrt{2\pi}$ is a normalization constant. The graphs of both functions are shown in Figure 4. The density curve (left) has the characteristic bell-shaped profile, with a maximum at $z = 0$ and inflection

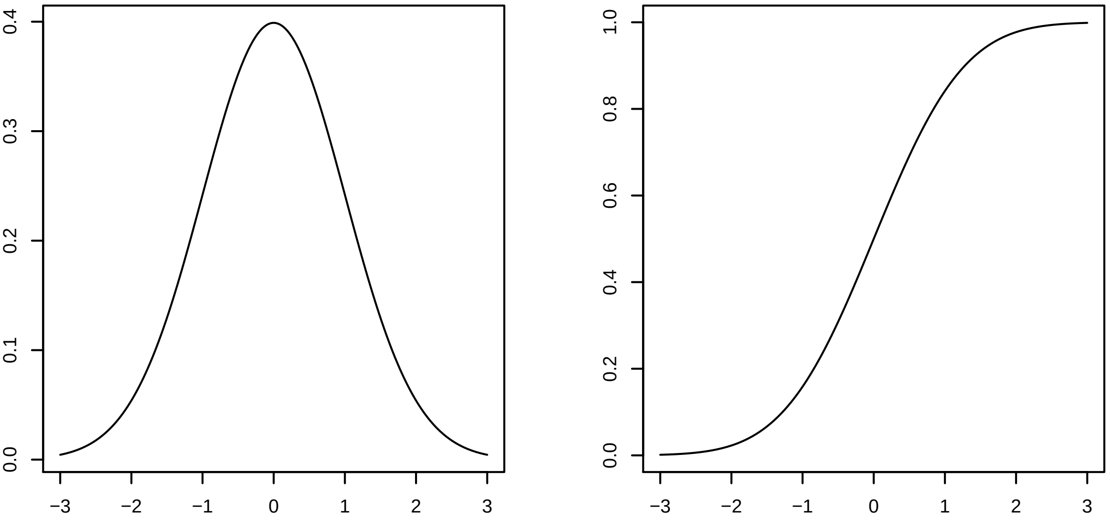

Figure 4. Standard normal PDF and CDF

points at $z = \pm 1$. Although there is no simple formula for $\Phi(z)$, it can be easily managed in the computer, by means of numerical integration.

The standard normal distribution has zero mean and unit variance (this is what standard means here). Indeed, we have

$$\mathrm{E}[Z] = \int_{-\infty}^{+\infty} \frac{z\,e^{-z^2/2}}{\sqrt{2\pi}}\,dz = \left[\frac{-e^{-z^2/2}}{\sqrt{2\pi}}\right]_{z=-\infty}^{z=+\infty} = 0.$$

Also, integrating by parts and using L'Hôpital's rule,

$$\mathrm{E}[Z^2] = \int_{-\infty}^{+\infty} \frac{z^2\,e^{-z^2/2}}{\sqrt{2\pi}}\,dz = \left[\frac{-z\,e^{-z^2/2}}{\sqrt{2\pi}}\right]_{z=-\infty}^{z=+\infty} + \int_{-\infty}^{+\infty} \frac{e^{-z^2/2}}{\sqrt{2\pi}}\,dz = 1.$$

The general normal distribution

If Z is standard normal, the linear transformation $X = \mu + \sigma Z$ defines a variable with density

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

This is the general normal distribution, denoted by $\mathcal{N}(\mu, \sigma^2)$. With this notation, the standard normal is $\mathcal{N}(0, 1)$. For the general normal distribution, the mean is μ and the standard deviation σ (this is consistent with the notation used). The density curve is still bell-shaped, with the maximum at $x = \mu$, but more or less flat, depending on σ, as shown in Figure 5. When modeling real phenomena, we search for the appropriate values of μ and σ.

The probability calculations for a normal distribution are based on the standard case, since the z-transform of a normal variable is a standard normal. More specifically, taking $z_i = (x_i - \mu)/\sigma$,

$$\mathrm{p}\big[x_1 < X < x_2\big] = \mathrm{p}\big[z_1 < Z < z_2\big] = \Phi(z_2) - \Phi(z_1).$$

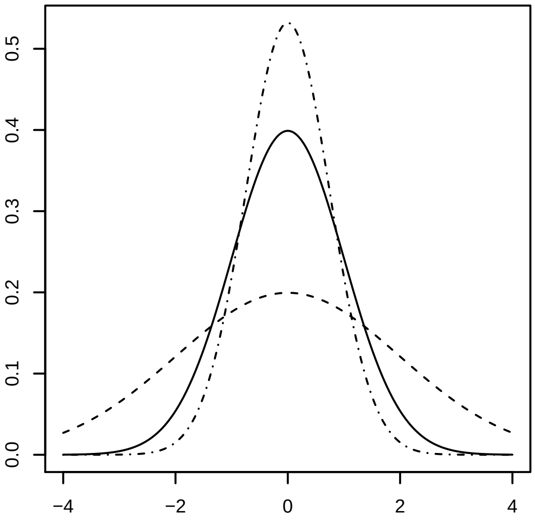

Figure 5. $\mathcal{N}(0,0.5^2)$, $\mathcal{N}(0,1)$ **and** $\mathcal{N}(0,2^2)$ **density curves**

Some useful probabilities

In spite of being based in a difficult formula, the normal distribution is very simple to manage in practice. For many applications, it suffices to know the probabilities associated to three basic intervals:

- First, $\text{p}\big[|Z| \le 1\big] = 68.27\%$. This means that, for $X \sim \mathcal{N}(\mu,\sigma^2)$, the interval defined by $\mu \pm \sigma$ contains about 2/3 of the population. An application of this formula: for the distribution of the income in a specific population, this gives an operational definition of the "middle class" in that population.
- Second, $\text{p}\big[|Z| \le 2\big] = 95.45\%$. So, the limits $\mu \pm 2\sigma$ enclose most of the population. Many applications are based on 95% limits. For instance, it is common, in the health sciences, to take the central 95% interval as "normal", the 2.5% left tail as "hypo" and the 2.5% right tail as "hyper". So, from estimates of the mean and the standard deviation of the cholesterol level in an age/gender group, we can derive an operational definition of hypercholesterolemia. The exact value for the 95% interval is $z = 1.96$.
- Third, $\text{p}\big[|Z| \le 3\big] = 99.73\%$. This means that, although a normal variable can take any value, those beyond the limits $\mu \pm 3\,\sigma$ rarely occur. This fact is used to set the limits in quality control charts.

Normal quantiles

The notation of the quantiles of the $\mathcal{N}(0,1)$ distribution, and those of the distributions derived from the normal (see later), is based on a practical convention. For $0 < \alpha < 1$, we denote by z_α the quantile $\Phi^{-1}(1-\alpha)$. Equivalently, $\text{p}\big[Z > z_\alpha\big] = \alpha$, or $\text{p}\big[-z_\alpha < Z < z_\alpha\big] = 1 - 2\alpha$. With this notation, $z_{0.025} = 1.96$.

The quantiles z_α associated to the values of α used in hypothesis testing are called **critical values**. For any probability $\alpha < 0.5$, the tails associated to z_α (the right tail, on the right of z_α, and left tail, on the left of $-z_\alpha$, have both area α).

¶ This notation is not completely universal. For some authors, z_α is the value whose one-tail area is equal to $\alpha/2$.

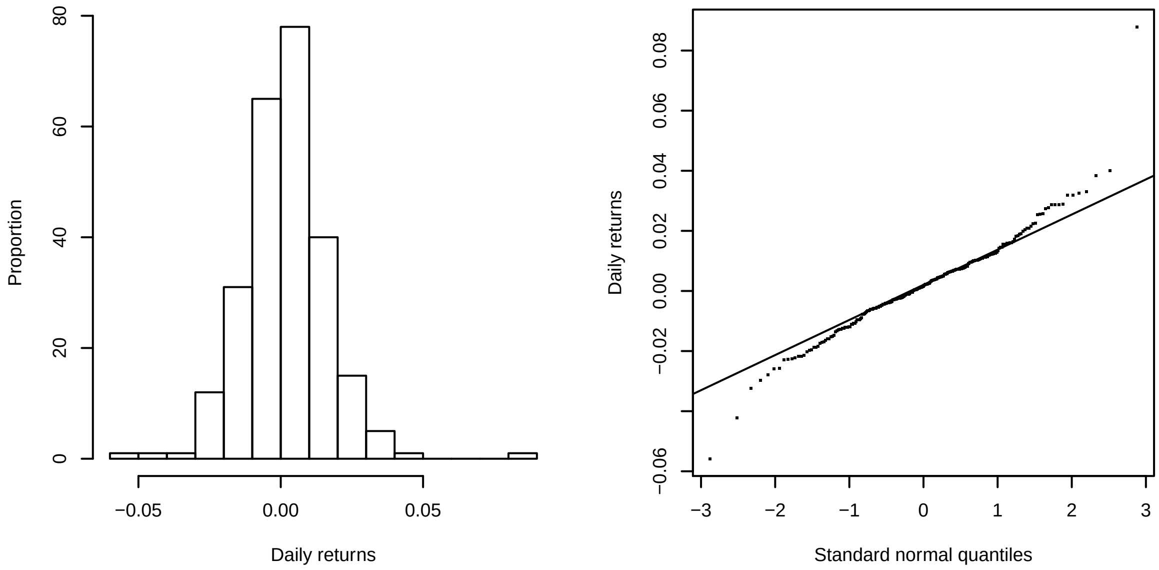

Figure 6. Histogram and normal probability plot (Example 1)

The normal probability plot

Quantile-quantile (QQ) plots are scatterplots in which the two axes correspond to quantiles of some distribution. This course only uses a special QQ plot for the normal distribution, the **normal probability plot**.

The normal probability plot is based on the fact that there is a linear relationship between a normal variable and the $\mathcal{N}(0,1)$ distribution. The idea is as follows:

(a) Take a sample of a normal distribution.

(b) Sort it, getting, say $x_1, x_2 \ldots, x_n$.

(c) Put x_i on one axis and the $\mathcal{N}(0,1)$ quantile $z_i = \Phi^{-1}(i/(n+1))$ on the other axis. This is the normal probability plot.

(d) The n points in the plot should be close to a straight line.

Example 3. Heavy (or fat) tails, are a special pattern of departure from the normal distribution, frequently found in finance. Since the normality of the returns was taken for granted in the classical portfolio theory, the persistent evidence of heavy tails found in financial returns data has been discussed many times. Nowadays, normality of returns is rarely assumed.

The `amzn` data set provides an example of this phenomenon. It contains daily OHLC (Open/High/Low/Close) data on the prices of Amazon shares for the year 2013. Figure 6 shows the histogram (left) and the normal probability plot (right) of the returns of the opening price. The distribution is reasonably symmetric, but the tails are heavier than expected in a normal distribution. I have included in the normal probability plot a straight line, chosen so that it passes through the first and third quartiles (this is the default in R, other applications fit a regression line). You may find in this graphic the traits already identified in the histogram.

The skewness and the kurtosis are

$$\mathrm{Sk} = 0.550, \qquad \mathrm{K} = 4.659,$$

in agreement with the my comments on the diagnostic plots of Figure 6.

Homework

A. A well-known gambler, the Chevalier De Méré (XVIIth century) has been associated by historians to the rise of the probability calculus, since, at his request, Pascal and Fermat developed a mathematical formulation of gambling odds. He posed to Pascal two problems connected with the games of chance. The first problem, called the **De Méré problem**, is a classic of probability textbooks, which illustrates the difficulty of managing probability calculations without a proper set of mathematical rules. Today, most students can solve this problem after a primer of probability calculus but, at De Méré's time, some famous mathematicians failed. Newton himself is said to have given a wrong solution. It also shows that common sense is not always right in probability.

The problem is: *since the probability of getting one 6 tossing a die is six times that of getting a double 6 tossing two dice, getting at least one 6 in four tosses of one die should be equally likely to getting at least one double 6 in twenty-four tosses of two dice. Is this true?*

B. This is the *birthday problem.* Which is the least number of persons required, if the probability exceeds 1/2 that two or more of them have the same birthday?