This lectures of chapter 4 are based on the textbook for

STAT3025Q (Statistical Methods)

Probability and Statistics for Engineering and Sciences

University of Connecticut

9th Edition

Jay L. Devore

CENGAGE Learning

Chapter 4

Continuous Random Variables and Probability Distributions

Section 1: Probability Density Functions

A random variable X is continuous if:

(1) Possible values comprise either a single interval on the number line (for some $A < B$, any number x between A and B is a possible value) or a union of disjoint intervals.

(2) $P(X = c) = 0$ for any number c that is a possible value of X.

Example

If a chemical compound is randomly selected and its pH "X" is determined, then X is a continuous random variable because any pH value between 0 and 14 is possible.

If more is known about the compound selected for analysis, then the set of possible values might be a subinterval of $[0, 14]$, such as $5.5 \leq x \leq 6.5$, but X would still be a continuous random variable.

Probability Distributions for Continuous Variables

Let X be a continuous random variable. Then a probability distribution or probability density function (pdf) of X is a function $f(x)$ such that for any two numbers a and b with $a \leq b$, the probability that X takes on a value in the interval $[a, b]$ is:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

Graphically, it can be represented as Figure 4.2 (page 143).

The probability density function must satisfy the following two conditions:

(1) $f(x) \geq 0$ for all x

(2) $\int_{-\infty}^{\infty} f(x)dx = area\ under\ the\ entire\ graph\ of\ f(x) = 1$

Example

One possible pdf for X is $f(x;A,B) = \begin{cases} \frac{1}{360} & 0 \le x \le 360 \\ 0 & otherwise \end{cases}$

- $f(x;A,B)$ is a probability density function since:
  (1) $f(x) \ge 0$ for all x
  (2) The area under the density curve is just the area of a rectangle$=$ $height.base =$ $(1/360)(360) = 1.$
- $P(90 \le X \le 180) = \int_{90}^{180} \frac{1}{360} dx = \frac{x}{360} |_{x=90}^{x=180} = \frac{180}{360} - \frac{90}{360} = 0.25.$
- $P(0 \le X \le 90 \text{ } or \text{ } 90 \le X \le 180) = P(0 \le X \le 90) + P(90 \le X \le 180) =$

$\int_{0}^{90} \frac{1}{360} dx + \int_{90}^{180} \frac{1}{360} dx = \frac{x}{360} |_{x=0}^{x=90} + \frac{x}{360} |_{x=90}^{x=180} =$

$\left(\frac{90}{360} - \frac{0}{360}\right) + \left(\frac{180}{360} - \frac{90}{360}\right) = 0.25 + 0.25 = 0.5.$

Note that:

The probability that X lies in some interval between a and b does not depend on whether the lower limit a or the upper limit b is included in the probability calculation:

$P(a \le X \le b) = P(a < X < b) = P(a < X \le b) = P(a \le X < b).$

Exercise 4 (page 146)

Let X denote the vibratory stress (psi) on a wind turbine blade at a particular wind speed in a wind tunnel. The article "Blade Fatigue Life Assessment with Application to VAWTS" (J. of Solar Energy Engr., 1982: 107–111) proposes the Rayleigh distribution, with pdf

$$f(x;\theta) = \begin{cases} \frac{x}{\theta^2}.\, e^{-x^2/(2\theta^2)} & x > 0 \\ 0 & otherwise \end{cases}$$

(a) Verify that $f(x;\theta)$ is a legitimate pdf.

Answer

1. It is non-negative for $x > 0$.

2. $\int_0^\infty \frac{x}{\theta^2} e^{-x^2/(2\theta^2)} dx = -e^{-x^2/(2\theta^2)}|_0^\infty = -e^{-\infty^2/(2\theta^2)} - \left(-e^{-0^2/(2\theta^2)}\right) = -0 - (-1) = 1.$

(b) Suppose $\theta = 100$. What is the probability that X is at most 200? Less than 200? At least 200?

Answer

$P(X \le 200) = \int_0^{200} \frac{x}{\theta^2} e^{-x^2/(2\theta^2)} dx = -e^{-x^2/(2\theta^2)}|_0^{200} = -e^{-200^2/\left(2(100^2)\right)} - \left(-e^{-0^2/\left(2(100^2)\right)}\right) = -e^{-2} - (-e^{-0}) = -0.1353 + 1 = 0.8647.$

$P(X < 200) = \int_0^{200} \frac{x}{\theta^2} e^{-x^2/(2\theta^2)}\, dx = -e^{-x^2/(2\theta^2)}|_0^{200} = -e^{-200^2/\left(2(100^2)\right)} - \left(-e^{-0^2/\left(2(100^2)\right)}\right) = -e^{-2} - (-e^{-0}) = -0.1353 + 1 = 0.8647$ [same answer as above].

$P(X \ge 200) = 1 - P(X < 200) = 1 - \int_0^{200} \frac{x}{\theta^2} e^{-x^2/(2\theta^2)}\, dx = 1 - 0.8647 = 0.1353.$

(c) What is the probability that X is between 100 and 200 (assuming $\theta = 100$)?

Answer

$P(100 \le X \le 200) = \int_{100}^{200} \frac{x}{\theta^2} e^{-x^2/(2\theta^2)}\, dx = -e^{-x^2/(2\theta^2)}|_{100}^{200} = \left[-e^{-200^2/\left(2(100^2)\right)}\right] - \left[-e^{-100^2/\left(2(100^2)\right)}\right] = -e^{-2} - (-e^{-0.5}) = -0.1353 + 0.6065 = 0.4712.$

(d) Give an expression for $F(x) = P(X \le x)$.

Answer

$F(x) = P(X \le x) = \int_0^x \frac{y}{\theta^2} e^{-y^2/(2\theta^2)}\, dy = -e^{-y^2/(2\theta^2)}|_0^x = -e^{-x^2/\left(2\theta^2\right)} - \left(-e^{-0^2/\left(2\theta^2\right)}\right) = 1 - e^{-x^2/\left(2\theta^2\right)}.$

Section 2: Cumulative Distribution Functions and Expected Values

The Cumulative Distribution Function $F(x)$ for a Continuous random Variable X

$$F(x) = P(X \leq x) = \int_{y=-\infty}^{x} f(x)dy$$

Graphically, for each x, $F(x)$ is the area under the density curve to the left of x. Figure 4.5 (page 148).

Using $F(x)$ to Compute Probabilities

Let X be a continuous random variable with pdf $f(x)$ and cdf $F(x)$. Then for any number a,

- $P(X > a) = 1 - F(a)$.
- If a and b are two numbers such that $a < b$, then $P(a \leq X \leq b) = F(b) - F(a)$.

Obtaining $f(x)$ from $F(x)$

Proposition

If X is a continuous random variable with pdf $f(x)$ and cdf $F(x)$, then at every x, $f(x) = F'(x)$.

Percentiles of a Continuous Distribution

Let p be a number between 0 and 1. The $(100p)th$ percentile of the distribution of a continuous random variable X, denoted by $\eta(p)$, is defined as:

$$p = F[\eta(p)] = \int_{y=-\infty}^{\eta(p)} f(y)dy$$

Graphically Figure 4.10 (page 151), $\eta(p)$ is that value on the measurement axis such that $100p\%$ of the area under the graph of $f(x)$ lies to the left of $\eta(p)$ and $100(1-p)\%$ lies to the right.

For example: $\eta(0.75)$ "the $75th$ percentile", is such that the area under the graph of $f(x)$ to the left of $\eta(0.75)$ is 0.75.

Note that:

The median of a continuous distribution, denoted by $\tilde{\mu}$, is the $50th$ percentile, so $0.5 = F(\tilde{\mu})$. That is, half the area under the density curve is to the left of $\tilde{\mu}$, and half is to the right of it.

Exercise 11 (page 154)

Let X denote the amount of time a book on two-hour reserve is actually checked out, and suppose the cdf is

$$F(x) = \begin{cases} 0 & x < 0 \\ \frac{x^2}{4} & 0 \leq x < 2 \\ 1 & 2 \leq x \end{cases}$$

(a) Calculate $P(X \leq 1)$.

Answer

$P(X \leq 1) = F(1) = \frac{1^2}{4} = \frac{1}{4}$.

(b) Calculate $P(0.5 \leq X \leq 1)$.

Answer

$P(0.5 \leq X \leq 1) = F(1) - F(0.5) = \frac{1^2}{4} - \frac{0.5^2}{4} = 0.1875$.

(c) Calculate $P(X > 1.5)$.

Answer

$P(X > 1.5) = 1 - P(X \leq 1.5) = 1 - F(1.5) = 1 - \frac{1.5^2}{4} = 0.4375$.

(d) What is the median checkout duration?

Answer

The median checkout duration $\tilde{\mu}$, solve $0.5 = F(\tilde{\mu}) = \frac{\tilde{\mu}^2}{4}$. Thus, $0.5 = \frac{\tilde{\mu}^2}{4} \rightarrow \tilde{\mu}^2 = 2 \rightarrow \tilde{\mu} = \sqrt{2} = 1.4142$.

(e) Obtain the probability density function $f(x)$.

Answer

$f(x) = F'(x) = \frac{2x}{4} = \frac{x}{2}$. Thus,

$$f(x) = \begin{cases} \frac{x}{2} & 0 \leq x < 2 \\ 0 & otherwise \end{cases}$$

Expected Values

- $E(X) = \mu_X = \int_{-\infty}^{\infty} x.f(x)dx$
- $E[h(X)] = \mu_{h(X)} = \int_{-\infty}^{\infty} h(x).f(x)dx$. Thus, if $h(X) = aX + b$, then $E[h(X)] = E[aX + b] = E(aX) + E(b) = aE(X) + b = a\mu_X + b$.

Example

Two species are competing in a region for control of a limited amount of a certain resource. Let $X =$ the proportion of the resource controlled by species 1 and suppose X has pdf

$$f(x) = \begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & otherwise \end{cases}$$

$$h(x) = \begin{cases} 1 - X & if\ 0 \leq X < 1/2 \\ X & if\ 1/2 \leq X \leq 1 \end{cases}$$

Calculate $E[h(X)]$.

Answer

$E[h(X)] = \mu_{h(X)} = \int_{-\infty}^{\infty} h(x).f(x)dx = \int_0^{1/2}(1 - X).1dx + \int_{1/2}^{1}(X).1dx = X|_{x=0}^{x=1/2} - \frac{X^2}{2}|_{x=0}^{x=1/2} + \frac{X^2}{2}|_{x=1/2}^{x=1} = \frac{3}{4}$.

Variance and Standard Deviation

$$V(X) = \sigma_X^2 = E[(X - \mu)^2] = \int_{-\infty}^{\infty}(x - \mu)^2.f(x)dx$$

Also, $V(X) = \sigma_X^2 = E(X^2) - [E(X)]^2$

$$SD(X) = \sigma_X = \sqrt{V(X)} = \sqrt{\sigma_X^2}$$

Note that:

$E(aX + b) = aE(X) + E(b) = aE(X) + b.$

$V(aX + b) = a^2V(X) = a^2\sigma_X^2.$

Exercise 11 continued (page 154)

(f) Calculate $E(X)$.

Answer

$E(X) = \int_0^2 x\, f(x) dx = \int_0^2 x \left(\frac{x}{2}\right) dx = \frac{1}{2}\frac{x^3}{3}|_0^2 = \frac{1}{6}(2^3 - 0^3) = \frac{8}{6} = \frac{4}{3}.$

(g) Calculate σ_X^2 and σ_X.

Answer

$\sigma_X^2 = E(X^2) - [E(X)]^2.$

$E(X^2) = \int_0^2 x^2\, f(x) dx = \int_0^2 x^2 \left(\frac{x}{2}\right) dx = \frac{x^4}{2\times 4}|_0^2 = \frac{1}{8}(2^4 - 0^4) = 2.$

$\sigma_X^2 = 2 - \left(\frac{4}{3}\right)^2 = \frac{2}{9} = 0.2222.$

$\sigma_X = \sqrt{\sigma_X^2} = \sqrt{0.2222} = 0.4714.$

(h) Define $h(X) = X^2$. Calculate $E[h(X)]$.

Answer

$E[h(X)] = E(X^2) = 2.$

Exercise 11 continued

Calculate $E(2X + 3)$ and $V(2X + 3)$.

Answer

$E(2X + 3) = 2E(X) + E(3) = 2 \times \left(\frac{4}{3}\right) + 3 = 5.6667.$

$V(2X + 3) = a^2 V(X) = a^2 \sigma_X^2 = 2^2(0.2222) = 0.8888.$

Section 3: The Normal Distribution

The normal distribution is the most important one in all of probability and statistics. The density curve is bell-shaped and therefore symmetric.

A continuous random variable X follows a normal distribution with parameters μ and σ (or σ^2) "X~N(μ,σ) or X~N(μ, σ^2)" where $-\infty < \mu < \infty$ and $\sigma > 0$, if the pdf of X has the following form: $f(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}$ where $-\infty < x < \infty$, e denotes the base of the natural logarithm system ≈ 2.71828,and $\pi \approx 3.14159$.

Note that:

- $f(x;\ \mu, \sigma)\ \geq 0$, and $\int_{-\infty}^{\infty} f(x;\ \mu, \sigma) = 1$.
- $E(X) = \mu$
- $V(X) = \sigma^2$

Figure 4.13 (page 157) presents graphs of $f(x;\ \mu, \sigma)$ for several different (μ, σ) pairs.

The Standard Normal Distribution

The computation of $P(a \leq X \leq b)$ when X is a normal random with parameters μ, and σ can be calculated as $\int_{x=a}^{b} f(x; \mu, \sigma)dx = \int_{x=a}^{b} \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}dx$.

$\mu = 0$ and $\sigma = 1$ result in standard normal distribution that can be used to calculate probabilities based on standard normal distribution table.

A random variable Z follows standard normal distribution if its pdf has the following form:

$f(z; 0{,}1) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad -\infty < z < \infty$

The cdf $\Phi(z) = P(Z \leq z)$ is the area under the standard normal density curve to the left of z, for $z = -3.49, -3.48, \ldots, 3.48, 3.49$. Figure 4.14 (page 158) illustrates the type of cumulative area (probability) tabulated in Table A.3.

In general, the $(100p)th$ percentile is identified by the row and column of Appendix Table A.3 in which the entry p is found.

For example:

- The 99th percentile of the standard normal distribution is that value on the horizontal axis such that the area under the z curve to the left of the value is 0.9900. Here 0.9901 lies at the intersection of the row marked 2.3 and column marked 0.03, so the $99th$ percentile is (approximately) $z = 2.33$ as Figure 4.17 (page 160).
- By symmetry, the first percentile $=\ -2.33$ (1% lies below the first and also above the $99th$) as Figure 4.18 (page 160).
- The $95th$ percentile, we look for 0.9500 inside the table. Although 0.9500 does not appear, both 0.9495 and 0.9505 do, corresponding to $z\ = 1.64$ and 1.65, respectively. Since 0.9500 is halfway between the two probabilities that do appear, we will use 1.645 as the $95th$ percentile and -1.645 as the $5th$ percentile.

Note that:

Z_α: Denotes the value on the z axis for which α of the area under the z curve lies to right of Z_α. Graphically, as Figure 4.19 (page 160).

- $P(Z \leq 1.25) = \Phi(1.25) = 0.8944$.
- $P(Z \leq -1.25) = \Phi(-1.25) = 0.1056$.
- $P(Z > 1.25) = 1 - P(Z \leq 1.25) = 1 - \Phi(1.25) = 1 - 0.8944 = 0.1056$.
- $P(-0.38 \leq Z \leq 1.25) = P(Z \leq 1.25) - P(Z \leq -0.38) = \Phi(1.25) - \Phi(-0.38) = 0.8944 - 0.3520 = 0.5424$.
- $z_{0.05}$ *w*here 5% of the area under the curve to the right of it $(the\ 95th\ percentile) = 1.645$.

Nonstandard Normal Distributions

When $X \sim N(\mu, \sigma^2)$, probabilities involving X are computed by "standardizing" as $Z = \frac{(X-\mu)}{\sigma}$.

If $X \sim N(\mu, \sigma)$, then $Z = \frac{(X-\mu)}{\sigma} \sim$ standard normal distribution "$N(0,1)$" and thus:

$P(a \leq X \leq b) = P\left(\frac{a-\mu}{\sigma} \leq \frac{X-\mu}{\sigma} \leq \frac{b-\mu}{\sigma}\right) = P\left(\frac{a-\mu}{\sigma} \leq Z \leq \frac{b-\mu}{\sigma}\right) = \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$.

$P(X \leq a) = P\left(\frac{X-\mu}{\sigma} \leq \frac{a-\mu}{\sigma}\right) = P\left(Z \leq \frac{a-\mu}{\sigma}\right) = \Phi\left(\frac{a-\mu}{\sigma}\right)$.

$P(X \geq b) = P\left(\frac{X-\mu}{\sigma} \geq \frac{b-\mu}{\sigma}\right) = P\left(Z \geq \frac{b-\mu}{\sigma}\right) = 1 - P\left(Z < \frac{b-\mu}{\sigma}\right) = 1 - \Phi\left(\frac{b-\mu}{\sigma}\right)$.

Note that:

• If $\mu = 100$ and $\sigma = 15$, then $x = 130$ corresponds to $z = \frac{130-100}{15} = 2$. That is, 130 is 2 standard deviations above (to the right of) the mean value.

• If the population distribution of a variable is (approximately) normal, then

1. Roughly 68% of the values are within one standard deviation of the mean.
2. Roughly 95% of the values are within two standard deviations of the mean.
3. Roughly 99.7% of the values are within three standard deviations of the mean.

• $(100p)th$ percentile for $N(\mu, \sigma) = \mu + [(100p)th$ percentile for standard normal$](\sigma)$.

Example

The amount of distilled water dispensed by a certain machine is normally distributed with mean value 64 oz and standard deviation 0.78 oz.

What container size c will ensure that overflow occurs only 0.5% of the time?

If X is a continuous random variable denotes the amount dispensed, the desired condition is that $P(X > c) = 0.005$.

$P(X > c) = 0.005 \Rightarrow P(X \leq c) = 1 - 0.005 = 0.995$. Thus, c is the 99.5^{th} percentile of the $N(64, 0.78) \Rightarrow c = \mu + [99.5th$ percentile for standard normal$](\sigma) = 64 + (2.58)(0.78) = 66\ oz$.

The Normal Distribution and Discrete Populations

Approximating the Binomial Distribution

Let X be a binomial random variable based on n trials with success probability p. Then if the binomial probability histogram is not too skewed, X has approximately $N(\mu = np, \sigma = \sqrt{npq})$.

$P(X \leq x) \approx \Phi\left(\frac{x+0.5-np}{\sqrt{npq}}\right)$ provided that $np \geq 10$ and $nq \geq 10$.

Exercise 34 (page 167)

The article "Reliability of Domestic-Waste Biofilm Reactors" (J. of Envir. Engr., 1995: 785–790) suggests that substrate concentration (mg/cm^3) of influent to a reactor is normally distributed with $\mu = 0.30$ and $\sigma = 0.06$.

(a) What is the probability that the concentration exceeds 0.50?

Answer

$$P(X > 0.50) = 1 - P(X \leq 0.50) = 1 - P\left(\frac{X-\mu}{\sigma} \leq \frac{0.50-\mu}{\sigma}\right)$$
$$= 1 - P\left(Z \leq \frac{0.50-0.30}{0.06}\right) = 1 - P(Z \leq 3.33) = 1 - 0.9996 = 0.0004.$$

(b) What is the probability that the concentration is at most 0.20?
Answer
$$P(X \leq 0.20) = P\left(\frac{X-\mu}{\sigma} \leq \frac{0.20-\mu}{\sigma}\right) = P\left(Z \leq \frac{0.20-0.30}{0.06}\right)$$
$$= P\left(Z \leq \frac{0.20-0.30}{0.06}\right) = P(Z \leq -1.67) = 0.0475.$$

(c) How would you characterize the largest 5% of all concentration values?

Answer

We want the $95th$ percentile, c, of this normal distribution, so that 5% of the values are higher. The $95th$ percentile of the standard normal distribution table satisfies $\Phi(z) = 0.95$, which is corresponding to the value 1.645 (using standard normal table).

$$c = \mu + \sigma\,(z_{0.05}) = 0.30 + 0.06(1.645) = 0.3987.$$

The largest 5% of all concentration values $=$ 0.3987 mg/cm^3.

Exercise 45 (page 168)

A machine that produces ball bearings has initially been set so that the true average diameter of the bearings it produces is 0.500 in. A bearing is acceptable if its diameter is within 0.004 in. of this target value. Suppose, however, that the setting has changed during the course of production, so that the bearings have normally distributed diameters with mean value 0.499 in. and standard deviation 0.002 in.

Answer

What percentage of the bearings produced will not be acceptable?

The percentage of the bearings produced will not be acceptable$= P(X < 0.496 \; or \; X > 0.504)$

$$= P(X < 0.496) + P(X > 0.504) = P\left(\frac{X-\mu}{\sigma} < \frac{0.496-\mu}{\sigma}\right) + P\left(\frac{X-\mu}{\sigma} > \frac{0.504-\mu}{\sigma}\right)$$

$$P\left(Z < \frac{0.496-0.499}{0.002}\right) + P\left(Z > \frac{0.504-0.499}{0.002}\right) = P(Z < -1.5) + P(Z > 2.5)$$

$P(Z < -1.5) + [1 - P(Z \leq 2.5)] = 0.0668 + [1 - 0.9938] = 0.073$. Thus, 7.3% of the bearings will be unacceptable.

Exercise 50 (page 169)

In response to concerns about nutritional contents of fast foods, McDonald's has announced that it will use a new cooking oil for its french fries that will decrease substantially trans fatty acid levels and increase the amount of more beneficial polyunsaturated fat. The company claims that 97 out of 100 people cannot detect a difference in taste between the new and old oils. Assuming that this figure is correct (as a long-run proportion), what is the approximate probability that in a random sample of 1000 individuals who have purchased fries at McDonald's, for the following:

(a) At least 40 can taste the difference between the two oils? Interpret it.

Answer

Let X denote the number of people in the sample of 1000 who can taste the difference, so $X \sim Binomial(1000, 0.03)$.

Because $\mu = np = 1000(0.03) = 30 > 10$ and $nq = 1000(0.97) = 970 > 10$.

Thus, X is approximately $N(\mu, \sigma)$ where

$\mu = np = 1000(0.03) = 30$ and $\sigma = \sqrt{np(1-p)} = \sqrt{1000 \times 0.03 \times 0.97} = 5.3944$.

We can calculate approximate probabilities using a continuity correction.

$P(X \geq 40) = 1 - P(X < 40) = 1 - P(X \leq 39) = 1 - P\left(\frac{X-\mu}{\sigma} \leq \frac{39+0.5-30}{5.3944}\right) = 1 -$

$P(Z \leq 1.76) = 1 - 0.9608 = 0.0392$.

Thus, the approximate probability that in a random sample of 1000 individuals who have purchased fries at McDonald's at least 40 can taste the difference between the two oils = 0.0392.

(b) At most 5% can taste the difference between the two oils? Interpret it.

Answer:

$5\%\ of\ 1000 = 50$. Thus,

$P(X \leq 50) = P\left(\frac{X-\mu}{\sigma} \leq \frac{50+0.5-30}{5.3944}\right) = P(Z \leq 3.8) \approx 1$.

Thus, the approximate probability that in a random sample of 1000 individuals who have purchased fries at McDonald's at most 5% can taste the difference between the two oils ≈ 1.

Section 4: The Exponential and Gamma Distributions

There are many practical situations in which the variable of interest to an investigator might have a skewed distribution.

The Exponential Distribution

The family of exponential distributions provides probability models that are very widely used in engineering and science disciplines. **For example:** Model the distribution of component lifetime. If X is a continuous random variable follows exponential distribution, then the pdf has the following form:

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & otherwise \end{cases}$$

Graphically, Figure 4.26 (page 171).

$$E(X) = \mu = \frac{1}{\lambda}$$

$$Var(X) = \sigma_X^2 = \frac{1}{\lambda^2}$$

The cdf has the following form:

$$F(x;\alpha) = P(X \leq x) = \begin{cases} \int_{y=0}^{x} \lambda e^{-\lambda y} \ \ dy \\ 0 \quad otherwise \end{cases} = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & otherwise \end{cases}$$

Note that:

- $P(X > x) = 1 - P(X \leq x) = 1 - \left[1 - e^{-\lambda x}\right] = e^{-\lambda x}$
- $P(a \leq X \leq b) = F(b;\lambda) - F(a;\lambda) = e^{-(0.1667)(5)} - e^{-(0.1667)(15)} = e^{-\lambda a} - e^{-\lambda b}$

The Gamma Function

It is defined as $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x}\, dx$ where $\alpha > 0$.

Note that:

- For any $\alpha > 1$, $\Gamma(\alpha) = (\alpha - 1). \Gamma(\alpha - 1)$.
- For any positive integer, n, $\Gamma(n) = (n - 1)!$
- $\Gamma(1/2) = \sqrt{\pi}$

The standard gamma distribution ($\boldsymbol{\beta = 1}$)

If X is a continuous random variable follows standard gamma distribution, then the pdf has the following form:

$$f(x;\alpha) = \begin{cases} \dfrac{x^{\alpha-1} e^{-x}}{\Gamma(\alpha)} & x \geq 0 \\ 0 & otherwise \end{cases}$$

It satisfies the two requirements to be a pdf since:

1. $f(x;\alpha) \geq o$
2. $\int_0^\infty f(x;\alpha) dx = \int_0^\infty \frac{x^{\alpha-1} e^{-x}}{\Gamma(\alpha)} dx = 1$

Figure 4.27(b) (page 173) presents graphs of the standard gamma pdf.

The cdf is $F(x;\alpha) = P(X \leq x) = \begin{cases} \int_{y=0}^{x} \frac{y^{\alpha-1} e^{-y}}{\Gamma(\alpha)} dy & x > 0 \\ 0 & otherwise \end{cases}$

There is a table of $F(x;\ \alpha)$ available; in Appendix Table A.4. that can be used to calculate probabilities.

Example

Suppose the reaction time X of a randomly selected individual to a certain stimulus has a standard gamma distribution with $\alpha = 2$. Calculate the following probabilities:

$P(3 \leq X \leq 5)$.

$P(X > 4)$.

Answer

Using the table (A.4) of standard gamma distribution, then:

$P(3 \leq X \leq 5) = F(5; 2) - F(3; 2) = 0.960 - 0.801 = 0.159$.

$P(X > 4) = 1 - P(X \leq 4) = 1 - F(4; 2) = 1 - 0.908 = 0.092$.

The Gamma Distribution

If X is a continuous random variable follows gamma distribution, then the pdf has the following form:

$$f(x;\alpha,\beta) = \begin{cases} \frac{x^{\alpha-1}e^{-x/\beta}}{\beta^{\alpha}\Gamma(\alpha)} & x \geq 0 \\ 0 & otherwise \end{cases} \quad \alpha > 0 \text{ and } \beta > 0$$

$E(X) = \mu = \alpha\beta$

$Var(X) = \sigma_X^2 = \alpha\beta^2$

Note that:

- The parameter β is called the scale parameter because values other than 1 either stretch or compress the pdf in the x direction. Graphically, Figure 4.27(a) (page 173) illustrates the graphs of the gamma pdf $f(x;\ \alpha,\beta)$ for several (α,β) pairs.
- The exponential distribution results from taking $\alpha = 1$ and $\beta = \frac{1}{\lambda}$.
- The incomplete gamma function can also be used to compute probabilities involving nonstandard gamma distributions as: $P(X \leq x) = F(x;\alpha,\beta) = F\left(\frac{x}{\beta};\alpha\right)$ where $F(.;\alpha)$ is the incomplete gamma function.

Example

Assume that X follows gamma distribution with $\alpha = 8$ and $\beta = 15$.

- Find the value of μ and σ.
- Calculate $P(X \geq 30)$.

Answer

- $E(X) = \mu = \alpha\beta = (8)(15) = 120$

$Var(X) = \sigma_X^2 = \alpha\beta^2$=$(8)(15)^2 = 1800 \Rightarrow \sigma_X = \sqrt{1800} = 42.4264$.

- $P(X \geq 30) = 1 - P(X < 30) = 1 - P(X \leq 30) = 1 - F\left(\frac{30}{15};8\right) = 1 - 0.001 = 0.999$.

Exercise

Evaluate $\Gamma(6)$.

Answer

$\Gamma(6) = 5! = 5 \times 4 \times 3 \times 2 \times 1 = 120.$

Exercise 61 (page 175)

Data collected at Toronto Pearson International Airport suggests that an exponential distribution with mean value 2.725 hours is a good model for rainfall duration (Urban Stormwater Management Planning with Analytical Probabilistic Models, 2000, p. 69).

(a) What is the probability that the duration of a particular rainfall event at this location is at least 2 hours?

Answer

Mean$= \mu = \frac{1}{\lambda} = 2.725 \Rightarrow \lambda = \frac{1}{\mu} = \frac{1}{2.725} = 0.3670.$

$P(X \geq 2) = 1 - P(X < 2) = 1 - P(X \leq 2) = 1 - F(2;\lambda) = 1 - \left[1 - e^{-\lambda x}\right] = e^{-\lambda x} =$
$e^{-(0.3670 \times 2)} = 0.48.$

(b) What is the probability that the duration of a particular rainfall event at this location is at most 3 hours?

Answer

$P(X \leq 3) = F(3;\lambda) = 1 - e^{-\lambda x} = 1 - e^{-(0.3670 \times 3)} = 0.6675.$

(c) What is the probability that the duration of a particular rainfall event at this location is between 2 and 3 hours?

Answer

$P(2 \leq X \leq 3) = P(X \leq 3) - P(X \leq 2) = F(3;\lambda) - F(2;\lambda)$

$= e^{-(0.3670 \times 2)} - e^{-(0.3670 \times 3)} = 0.1475.$

Section 5: Other Continuous Distributions

The Weibull Distribution

If X is a continuous random variable follows lognormal distribution if the continuous random variable $Y = \ln(X)$ has a normal distribution. The pdf of lognormal random variable when $ln(X)$ is normally distributed with parameters μ and σ is:

$$f(x;\alpha,\beta) = \begin{cases} \frac{\alpha x^{\alpha-1} e^{-(x/\beta)^{\alpha}}}{\beta^{\alpha}} & x \geq 0 \\ 0 & otherwise \end{cases} \qquad \alpha > 0 \text{ and } \beta > 0$$

$$E(X) = \mu = \beta\Gamma\left[1 + \left(\frac{1}{\alpha}\right)\right]$$

$$Var(X) = \sigma_X^2 = \beta^2\left\{\Gamma\left[1 + \left(\frac{2}{\alpha}\right)\right] - \left(\Gamma\left[1 + \left(\frac{1}{\alpha}\right)\right]\right)^2\right\}$$

Note that:

- Graphically, both α and β can be varied to obtain a number of different density curves [Figure 4.28 (page 177)].
- When $\alpha = 1$, the pdf of Weibull reduces to the exponential distribution with $\lambda = \frac{1}{\beta}$.
- When $\alpha = 1$ and $\beta = \frac{1}{\lambda}$, the pdf of gamma reduces to exponential distribution.

The cdf of a Weibull random variable (with parameters α and β) $= P(X \leq x) = F(x;\alpha,\beta)$ has the form:

$$F(x;\alpha,\beta) = P(X \leq x) = \begin{cases} 1 - e^{-(x/\beta)^{\alpha}} & x \geq 0 \\ 0 & otherwise \end{cases}$$

Note that:

- $P(X > x) = 1 - P(X \leq x) = 1 - \left[1 - e^{-(x/\beta)^{\alpha}}\right] = e^{-(x/\beta)^{\alpha}}$.
- $P(a \leq X \leq b) = P(X \leq b) - P(X \leq a) = e^{-(a/\beta)^{\alpha}} - e^{-(b/\beta)^{\alpha}}$.

Example:

Suppose that X has a Weibull distribution with $\alpha = 2$ and $\beta = 10$, then:

$P(X \leq 5) = 1 - e^{-(5/10)^2} = 1 - 0.7788 = 0.2212$.

$P(X > 5) = 1 - P(X \leq 5) = e^{-(5/10)^2} = 0.7788$.

$P(5 \leq X \leq 10) = P(X \leq 10) - P(X \leq 5) = e^{-(5/10)^2} - e^{-(10/10)^2} = 0.4109$.

The Lognormal Distribution

If X is a continuous random variable follows lognormal distribution if the continuous random variable $Y = \ln(X)$ has a normal distribution. The pdf of lognormal random variable when $ln(X)$ is normally distributed with parameters μ and σ is:

$$f(x;\mu,\sigma) = \begin{cases} \frac{e^{-[\ln(X)-\mu]^2/(2\sigma^2)}}{\sqrt{2\pi}\sigma x} & x \geq 0 \\ 0 \quad otherwise & \end{cases} \quad -\infty < \mu < \infty \text{ and } \sigma > 0$$

Be careful; the parameters μ and σ are not the mean and standard deviation of X but of $ln(X)$.

Figure 4.30 (page 180) illustrates graphs of the lognormal pdf which is positive skew.

$$E(X) = \mu_X = e^{\mu+(\sigma^2/2)}$$

$$Var(X) = \sigma_X^2 = \left(e^{2\mu+\sigma^2}\right).\left(e^{\sigma^2} - 1\right)$$

Because $ln(X)$ has a normal distribution, the cdf of X can be expressed in terms of the cdf $\Phi(z)$ of a standard normal random variable Z.

$$\begin{cases} F(x;\mu,\sigma) = P(X \leq x) = P[\ln(X) \leq \ln(x)] = P\left(Z \leq \frac{\ln(x)-\mu}{\sigma}\right) = \Phi\left(\frac{\ln(x)-\mu}{\sigma}\right) & x \geq 0 \\ 0 & otherwise \end{cases}$$

Note that:

$\mu_{\ln(X)} = E[\ln(X)] = \mu$ and $\sigma_{\ln(X)} = \sigma$.

Exercise 79 (page 183)
Assuming that X = nonpoint source load of total dissolved solids could be modeled with a lognormal distribution having mean value 10,281 kg/day/km and a coefficient of variation $CV = 0.40$ ($CV = \sigma_X/\mu_X$).
Note that:
$\mu_X = 10{,}281$ and since $CV = \sigma_X/\mu_X$, $0.40 = \sigma_X/\mu_X \Rightarrow \sigma_X = 0.40\mu_X = 0.40(10{,}281) = 4{,}112.4$.

(a) What are the mean and standard deviation values of $ln(X)$?
Answer
X follows lognormal distribution with mean= $(\mu_X) = 10{,}281$ and standard deviation $(\sigma_X) = 4{,}112.4 \Rightarrow$ this information will be used to fin the values of mean of $ln(X) = \mu$ and standard deviation of $ln(X) = \sigma$ as follow:

$$\mu_X = E(X) = e^{\mu+\sigma^2/2} = 10{,}281 \Rightarrow \mu_X^2 = \left(e^{\mu+\sigma^2/2}\right)^2 = e^{2\mu+\sigma^2}$$

$$Var(X) = (\sigma_X)^2 = (4{,}112.4)^2 = 16{,}911{,}833.76 = \left(e^{2\mu+\sigma^2}\right)\left(e^{\sigma^2}-1\right)$$

$$\frac{\left(e^{2\mu+\sigma^2}\right)\left(e^{\sigma^2}-1\right)}{e^{2\mu+\sigma^2}} = e^{\sigma^2}-1 = \frac{(4{,}112.4)^2}{(10{,}281)^2} \Rightarrow e^{\sigma^2} = \frac{(4{,}112.4)^2}{(10{,}281)^2}+1 = 1.16$$

$$\Rightarrow \sigma = \sigma_{\ln(X)} = \sqrt{\ln(1.16)} = 0.3852.$$

$$\mu_X = e^{\mu+\sigma^2/2} = 10{,}281 \Rightarrow e^{\mu+0.3852^2/2} = 10{,}281 \Rightarrow$$

$$\ln\left(e^{\mu+0.3852^2/2}\right) = \ln(10{,}281) \Rightarrow \mu + 0.3852^2/2 = \ln(10{,}281) \Rightarrow$$

$$\mu = \ln(10{,}281) - 0.3852^2/2 = 9.1639.$$

(b) What is the probability that X is at most $15{,}000\ kg/day/km$?
Answer

$$P(X \le 15{,}000) = P[\ln(X) \le \ln(15{,}000)] = P\left[\frac{\ln(X)-\mu}{\sigma} \le \frac{\ln(15{,}000)-\mu}{\sigma}\right]$$

$$P\left(Z \le \frac{9.6158-9.1639}{0.3852}\right) = P\left(Z \le \frac{9.6158-9.1639}{0.3852}\right) = P(Z \le 1.17) = 0.8790.$$

(c) What is the probability that X is at least $15{,}000\ kg/day/km$?

Answer

$$P(X \geq 15{,}000) = P[\ln(X) \geq \ln(15{,}000)] = 1 - P[\ln(X) \leq \ln(15{,}000)]$$

$$= 1 - P\left[\frac{\ln(X) - \mu}{\sigma} \leq \frac{\ln(15{,}000) - \mu}{\sigma}\right]$$

$$= 1 - P\left(Z \leq \frac{9.6158 - 9.1639}{0.3852}\right) = 1 - P\left(Z \leq \frac{9.6158 - 9.1639}{0.3852}\right) = 1 - P(Z \leq 1.17) = 1 - 0.8790 = 0.121.$$

The Beta Distribution

The beta distribution provides positive density only for X in an interval of finite length.

If X is a continuous random variable follows beta distribution with parameters α and β (both are positive), then the pdf of beta has the following form:

$$f(x;\alpha,\beta,A,B) = \begin{cases} \left(\frac{1}{B-A}\right).\left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha).\Gamma(\beta)}\right).\left(\frac{x-A}{B-A}\right)^{\alpha-1}.\left(\frac{B-x}{B-A}\right)^{\beta-1} & A \leq x \leq B \\ 0 & otherwise \end{cases}$$

Figure 4.32 (page 181) illustrates graphs of the beta pdf.

$$E(X) = \mu = A + (B-A).\left(\frac{\alpha}{\alpha+\beta}\right)$$

$$Var(X) = \sigma_X^2 = \frac{(B-A)^2(\alpha\beta)}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Note that:

If $A = 0$ and $B = 1$, then standard beta distribution results. The standard beta distribution has the following form:

$$f(x;\alpha,\beta) = \begin{cases} \left(\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha).\Gamma(\beta)}\right).(x)^{\alpha-1}.(1-x)^{\beta-1} & 0 \leq x \leq 1 \\ 0 & otherwise \end{cases}$$

$$E(X) = \mu = \left(\frac{\alpha}{\alpha+\beta}\right)$$

$$Var(X) = \sigma_X^2 = \frac{(\alpha\beta)}{(\alpha+\beta)^2(\alpha+\beta+1)}$$

Exercise 84 (page 183)

Suppose the proportion X of surface area in a randomly selected quadrat that is covered by a certain plant has a standard beta distribution with $\alpha = 5$ and $\beta = 2$.

(a) Compute $E(X)$.
Answer

$E(X) = \frac{\alpha}{\alpha+\beta} = \frac{5}{5+2} = 0.7143.$

(b) Compute $P(X \leq\ 0.2)$.

Answer

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\times\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1}$$

$P(X \leq\ 0.2) = \int_0^{0.2}\frac{\Gamma(5+2)}{\Gamma(5)\times\Gamma(2)}x^{5-1}(1-x)^{2-1}dx$ (the minimum value of X is 0 since X follows standard beta distribution).

$= \frac{\Gamma(5+2)}{\Gamma(5)\times\Gamma(2)} \times \left(\int_0^{0.2} x^4\,(1-x)dx\right) = \frac{\Gamma(5+2)}{\Gamma(5)\times\Gamma(2)} \times \left(\int_0^{0.2} x^4\,dx - \int_0^{0.2} x^5\,dx\right) = 0.0015.$

(c) What is the expected proportion of the sampling region not covered by the plan?

Answer

Since X is the proportion covered by the plant, then is the proportion not covered by the plant = $1 - X$.

$E(1 - X) = E(1) - E(X) = 1 - 0.7143$ [$uning\ part\ (a)$] = 0.2857.

Section 6: Probability Plots

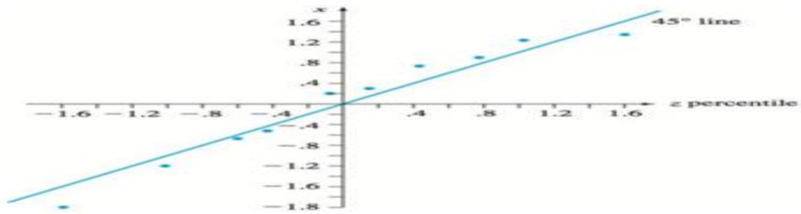
An effective way to check a distributional assumption is to construct what is called a probability plot. The essence of such a plot is that:

1. If the distribution on which the plot is based on is correct, the points in the plot should fall close to a straight line.

2. If the actual distribution is quite different from the one used to construct the plot, the points will likely depart substantially from a linear pattern.

Example

Based on the following plot, is it plausible that the random variable measurement error has a standard normal distribution?

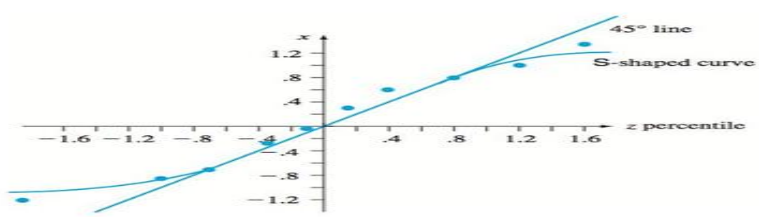


Answer

The line fits the points very well. The plot suggests that the standard normal distribution is a reasonable probability model for measurement errors.

Example

Based on the following plot, is it plausible that the random variable measurement error has a standard normal distribution?



The plot has a well-defined S-shaped appearance. Thus, this plot indicates that the standard normal distribution would not be a plausible choice for the probability model that gave rise to these observed measurement errors.