The Calculus of Statistics

Blaise Whitesell Jeremiah Gelb

May 26, 2015

1 Probability Distribution Functions

1.1 Probability Density Functions

A probability density function (**PDF**) describes the relative likelihood $f(x)$ of possible outcomes for a continuous random variable.

- The probability of any x occurring is either positive or zero, so a PDF can never be negative.
$$f(x) \geq 0 \quad \text{for all } x.$$
- The area under the curve must equal 1
$$\int_a^b f(x)\,\mathrm{d}x = 1$$
where a and b are the lower and upper bounds, often $-\infty$ and ∞.

1.2 Cumulative Distribution Functions

A cumulative distribution function (**CDF**) gives the probability $F(x)$ that the outcome of a continuous random variable will be less than or equal to x.

- The CDF is related to the PDF by this integral:
$$F(x) = \int_a^x f(t)\,\mathrm{d}t \quad \text{for } a \leq x \leq b.$$
- A CDF is monotonically increasing on the interval (a, b).
$$F(a) = 0 \qquad F(b) = 1$$

1.3 Using Calculus

To find the probability of an outcome in a certain range, one can integrate the PDF over that interval (c, d) contained in (a, b).

$$\int_c^d f(x)\,\mathrm{d}x \quad = F(d) - F(c)$$

This gives the area under the curve, corresponding to the probability.

2 The Uniform Distribution

2.1 Definition

A uniform distribution describes a continuous random variable where every outcome on an interval (a, b) is equally likely.

$$f(x) = \kappa$$

2.2 Implications

- The distribution looks like a rectangle with length $b - a$ and height κ
- Area $= 1 = (b - a) \cdot \kappa$

$$\int_a^b \frac{1}{b-a}\,\mathrm{d}x = \frac{x}{b-a}\bigg|_a^b = \frac{b}{b-a} - \frac{a}{b-a} = \frac{b-a}{b-a} = 1$$

- Therefore, we have an explicit definition for κ:

$$\kappa = \frac{1}{b-a}$$

2.3 General form

The uniform distribution has two parameters: the bounds a and b.

(PDF)
$$f(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \le x \le b, \\ 0 & \text{otherwise.} \end{cases}$$

(CDF)
$$F(x) = \begin{cases} 0 & \text{for } x < a, \\ \frac{x-a}{b-a} & \text{for } a \le x < b, \\ 1 & \text{for } x \ge b. \end{cases}$$

The CDF increases linearly on the interval (a, b).

Problem 1. In the game "Spin to Win", the player spins a giant roulette wheel with 200 spaces. One is marked "Win", one is marked "Lose", and the remaining 198 are marked "Spin Again". What is the probability of losing on the first spin?

Solution. There are 200 spaces, exactly one of which corresponds to losing immediately. The continuous range of values can be modeled as a uniform distribution from 0° to 360°.

$$1/200 \times 360^\circ = 1.8^\circ$$

$$\int_0^{1.8} \frac{1}{360}\,\mathrm{d}\theta = \frac{\theta}{360}\bigg|_0^{1.8} = \frac{1.8}{360} - \frac{0}{360} = 0.005$$

Interestingly enough, this is equivalent to 1/200. □

3 The Exponential Distribution

3.1 Introduction

An exponential distribution can be used to model events that occur independently at a constant average rate.

Problem 2. 250 kids at a frat party are randomly getting sick at a continuous rate of 20% per hour, beginning at 12:00AM. What is the probability that Chad gets sick between 2:00AM and 3:00AM?

Solution. Let $t = 0$ at 12:00AM. The number of kids who are not sick is given by $250e^{-.2t}$, and the number of kids who are sick is given by $250 - 250e^{-.2t}$. Therefore, the proportion of kids who are sick after t hours is

$$\frac{250 - 250e^{-.2t}}{250} \quad = 1 - e^{-.2t}$$

This is equivalent to the probability of being sick after t hours, which is the

CDF. To find the PDF, simply take a derivative.

$$\int_0^t f(x)\,\mathrm{d}x = 1 - e^{-.2t}$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_0^t f(x)\,\mathrm{d}x = \frac{\mathrm{d}}{\mathrm{d}t}(1 - e^{-.2t})$$

$$f(x) = .2e^{-.2t}$$

Now that we know the PDF, we integrate it between $t = 2$ and $t = 3$.

$$\int_2^3 .2e^{-.2t} = -e^{-.2t}\Big|_2^3 = -e^{-.6} - (-e^{-.4}) = .1215$$

Note that we could also have found this answer by finding the difference between $F(3)$ and $F(2)$, without using the PDF at all.

$$\left(1 - e^{-.2t}\right)\Big|_{t=3} - \left(1 - e^{-.2t}\right)\Big|_{t=2} = .1215$$

Using either of these methods, the probability that Chad will get sick between 2:00AM and 3:00AM is .1215. □

3.2 General Form

The exponential distribution has one parameter k, where $0 < k < 1$. It is known as the rate parameter.

(PDF) $$f(t) = ke^{-kt}$$

(CDF) $$F(t) = 1 - e^{-kt}$$

It can be shown that the area under the PDF is 1.

$$\lim_{b\to\infty}\int_0^b ke^{-kt} = \lim_{b\to\infty} -e^{-kt}\Big|_0^b = \lim_{b\to\infty} -e^{-k(b)} - \left(-e^{-k(0)}\right) = 0 - (-1) = 1$$

4 Properties of Random Variables

4.1 Expected Value

The expected value of a random variable (also known as the mean) can be thought of intuitively as the long-run average of its outcomes. For a discrete random

variable, it is the probability-weighted average of all possible outcomes, which may be calculated directly.

Problem 3. What is the expected value for the roll of a six-sided die?

Solution. The possible outcomes are 1, 2, 3, 4, 5, and 6, each occurring with probability $\frac{1}{6}$.

$$\frac{1}{6}(1) + \frac{1}{6}(2) + \frac{1}{6}(3) + \frac{1}{6}(4) + \frac{1}{6}(5) + \frac{1}{6}(6) = 3.5$$

□

The general formula for the expected value of a discrete random variable is

$$\sum_{i=1}^{k} x_i p_i$$

where p_i is the probability of outcome x_i of k possible outcomes.

The formula for expected value can be extended to continuous random variables by replacing the sum with an integral.

$$\mu = \int_a^b x \cdot f(x)\,\mathrm{d}x$$

It follows that this is the mean of the PDF $f(x)$.

Problem 4. Returning to this frat party, at what time can Chad expect to get sick?

Solution. Use the formula for expected value of a continuous random variable.

(Setup) $$\int_0^\infty x \cdot \left(.2e^{-.2x}\right)\,\mathrm{d}x$$

(Integrate by parts) $u = x \quad \mathrm{d}u = 1\,\mathrm{d}x \quad \mathrm{d}v = .2e^{-.2x}\,\mathrm{d}x \quad v = -e^{-.2x}$

(Antiderivative) $$-xe^{-.2x} - \int -e^{-.2x}\,\mathrm{d}x \quad = -xe^{-.2x} - 5e^{-.2x}$$

(Improper integral) $$\lim_{b\to\infty} (-5 - x)\left(e^{-.2x}\right)\Big|_0^b \quad = 5$$

Chad can expect to get sick around 5:00AM. □

4.2 Median

The median M is the value of x for which

$$P(x \leq M) = .5 \quad \text{and} \quad P(x \geq M) = .5$$

It follows that the median is an M such that

$$\int_a^M f(x)\,\mathrm{d}x = \frac{1}{2}$$

Problem 5. At what time are half the kids at the frat party sick?

Solution. Solve for the upper bound for the integral using the Fundamental Theorem of Calculus.

$$\begin{aligned}
\int_0^M .2e^{-.2t}\,\mathrm{d}t &= 0.5 \\
-e^{-.2t}\Big|_0^M &= 0.5 \\
-e^{-.2M} - \left(-e^{-2(0)}\right) &= 0.5 \\
-e^{-.2M} + 1 &= 0.5 \\
e^{-.2M} &= 0.5 \\
-.2M &= \ln 0.5 \\
M &= 3.4657
\end{aligned}$$

Half the kids at the party will be sick by 3:28AM. □

4.3 Variance

Variance is a measure of the spread of a distribution. It is used to describe the average distance of an outcome from the expected value.

A naive method of calculating variance would be

(Spurious) $$\int_a^b (x-\mu)f(x)\,\mathrm{d}x$$

The problem with this method is that negative and positive values of $(x - \mu)$ cancel each other out. Instead we can compute

(Absolute deviation) $$\int_a^b |x - \mu| \, f(x) \, \mathrm{d}x$$

(Variance) $$\mathrm{Var}(x) = \int_a^b (x - \mu)^2 f(x) \, \mathrm{d}x$$

The second method, in which the term $(x - \mu)$ is squared, is preferred for reasons which are beyond the scope of this text. Note that squaring the values gives a greater weight to outcomes that are further from the expected value.

The square root of the variance is known as the **standard deviation**.

(Wow!) $$\sqrt{\mathrm{Var}(x)} = \sigma$$

Problem 6. Find the standard deviation of the time someone at this frat party gets sick.

Solution. Solve for the variance using the formula.

$$\begin{aligned}
\int_0^M .2e^{-.2t} \, \mathrm{d}t &= 0.5 \\
-e^{-.2t}\Big|_0^M &= 0.5 \\
-e^{-.2M} - \left(-e^{-2(0)}\right) &= 0.5 \\
-e^{-.2M} + 1 &= 0.5 \\
e^{-.2M} &= 0.5 \\
-.2M &= \ln 0.5 \\
M &= 3.4657
\end{aligned}$$

Half the kids at the party will be sick by 3:28AM. □

5 The Normal Distribution

5.1 Description

A Normal distribution is a bell-shaped curve with special properties that are useful in statistics. It is a continuous distribution over the set of real numbers described by two parameters μ and σ, which correspond to its mean and standard deviation.

(PDF) $$f(x) = N(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Proof. The area Q under a Normal PDF is unity.

Show that $$Q = 1 = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \, \mathrm{d}x$$

Let $$u = \frac{x-\mu}{\sigma} \qquad \mathrm{d}u = \frac{\mathrm{d}x}{\sigma} \qquad \mathrm{d}x = \sigma \, \mathrm{d}u$$

$$Q = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}u^2} \, \mathrm{d}u$$

$$Q^2 = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}u^2} \, \mathrm{d}u \cdot \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{1}{2}v^2} \, \mathrm{d}v$$

$$Q^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(u^2+v^2)} \, \mathrm{d}u \, \mathrm{d}v$$

Convert to polar form $$u = r\cos\theta \qquad v = r\sin\theta \qquad u^2 + v^2 = r^2$$

Jacobian $$\begin{vmatrix} \frac{\mathrm{d}u}{\mathrm{d}r} & \frac{\mathrm{d}u}{\mathrm{d}\theta} \\ \frac{\mathrm{d}v}{\mathrm{d}r} & \frac{\mathrm{d}v}{\mathrm{d}\theta} \end{vmatrix} = \begin{vmatrix} \cos\theta & r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = \left| -r\sin^2\theta - r\cos^2\theta \right| = r$$

$$Q^2 = \frac{1}{2\pi} \int_0^{2\pi} \int_0^{\infty} e^{-\frac{1}{2}r^2} r \, \mathrm{d}r \, \mathrm{d}\theta$$

Separate and cancel

$$Q^2 = \frac{1}{2\pi} \int_0^{2\pi} \mathrm{d}\theta \cdot \int_0^{\infty} e^{-\frac{1}{2}r^2} r \, \mathrm{d}r$$

$t = r^2 \qquad \mathrm{d}t = 2r \, \mathrm{d}r$

$$Q^2 = \int_0^{\infty} \frac{1}{2} e^{-\frac{1}{2}t} \, \mathrm{d}t$$

$$Q^2 = \lim_{b \to \infty} \left(-e^{-\frac{1}{2}t} \right) \Big|_0^b$$

$$Q^2 = 1$$

$$Q = 1$$

□