Distributions, Moments, and VaR

January 25, 2013

Table of contents

Basics

Statistics allows us to model things in real life which occur with some uncertainty.

Definition 1.1

An **event** is a set of outcomes that occur with some probability.

Definition 1.2

A **random variable** is a variable which can take on multiple values. The set of values that it can take is called its **sample space** (often denoted S), and the probabilities corresponding to each value is defined by its distribution.

An Example

Example 1.3

We toss a coin which lands 'heads' with probability p, and 'tails' with probability $1 - p$. Let X be a random variable which equals 1 if the coin toss resulted in 'heads' and 0 otherwise. What is the sample space of X?

An Example

An Example

Answer

One coin toss results in either $X = 1$ (if heads) or $X = 0$ (if tails). Thus, the sample space of X is $\{0, 1\}$.

Distributions

Definition 1.4

A **distribution** describes the relationship between a set of values and their associated probabilities. For a **discrete distribution**, the sample space S is the set of discrete numbers, whereas for a **continuous distribution**, S is comprised of continuous interval(s).

Distributions

Definition 1.5

A **probability mass function** (PMF) $f(x)$ for a discrete random variable X is defined as:

$$f(x) := Pr(X = x)$$

A **probability density function** (PDF) $f(x)$ for a continuous random variable X expresses a relative probability of x.

A **cumulative distribution function** (CDF) $F(x)$ applies to both continuous and discrete random variables:

$$F(x) := Pr(X \leq x)$$

Relationship between CDF and PDF

Formulas

$$f(x) = \frac{d}{dx}F(x) \quad (1.1)$$

$$F(x) = \int_{-\infty}^{x} f(t)dt \quad (1.2)$$

Rules

1. If X is a continuous random variable, $Pr(X = x) = 0$ for any x.
2. If X is a discrete random variable with sample space $S = \{x_1, \ldots, x_k\}$, then $\sum_{x \in S} f(x) = \sum_{i=1}^{k} f(x_i) = 1$.
3. If X is a continuous random variable with sample space S, where S is possibly the real line, or any subset of it, then $\int_S f(x)dx = 1$.
4. The CDF is a function that always increases monotonically from 0 to 1.

Example

Example 1.6

Let X be a discrete random variable with a sample space of $\{1, 2, 3, 4\}$. Compute the probability that X is between 2 and 4, inclusive. Compute the same probability assuming X is a continuous random variable which takes on values in the interval $[1, 4]$.

Example

Example

Answer

We can use the PDF to compute the the probability that X is between 2 and 4, inclusive, by summing up $f(x)$ for values in this given interval.

$$Pr(2 \leq X \leq 4) = \sum_{x=2}^{4} f(x)$$

Note that if X was, instead, a continuous random variable which takes on values in the interval $[1, 4]$, then we would have computed the same probability as

$$Pr(2 \leq X \leq 4) = \int_{2}^{4} f(x)dx$$

Uniform Distribution

Definition 1.7

Let X be a **discrete uniform** random variable over integers in the interval $\{a, \dots, b\}$. Then,

$$f_X(x) = \frac{1}{b - a + 1} \qquad x \in \{a, \dots, b\}$$

Let Y be a **continuous uniform** random variable over the interval (a, b). Then,

$$f_Y(y) = \frac{1}{b - a} \quad y \in (a, b)$$

Example

Example 1.8

Let $X \sim$Unif$(\{1, 2, 3, 4\})$. Let $Y \sim$Unif$(1, 4)$. Compute $Pr(2 \leq X \leq 4)$ and $Pr(2 \leq Y \leq 4)$.

Example

Example

Answer

The formulas we use will come straight out of Example 1.6.

$$F(4) - F(1) = Pr(2 \leq X \leq 4) = \sum_{x=2}^{4} f_X(x) = \sum_{x=2}^{4} \frac{1}{4} = 0.75$$

$$F(4) - F(2) = Pr(2 \leq Y \leq 4) = \int_2^4 f_Y(y)dy = \int_2^4 \frac{1}{4}dy = 0.5$$

Empirical Models

Definition 1.9

An **empirical model** or **empirical distribution** is a discrete distribution derived from an observed dataset of n data points $\{x_1, \ldots, x_n\}$, where each observed value is assigned a probability of $1/n$. Formally, the empirical PDF and CDF are given by,

$$f_n(x) = \frac{\#\ \{\text{data points equal to } x\}}{n}$$

$$F_n(x) = \frac{\#\ \{\text{data points} \leq x\}}{n}$$

Mean

Definition 1.10

The **mean** or **expectation** of a distribution (or of a random variable X following the distribution) is given by

$$E(X) = \sum_{x \in S} x f(x) \qquad \text{(for a discrete distribution)}$$

$$E(X) = \int_S x f(x) dx \qquad \text{(for a continuous distribution)}$$

Example

Example

Let $Y \sim$Unif$(1,4)$. Compute $E(Y)$.

Example

Example

Answer

Recall that

$$f_Y(y) = \frac{1}{b-a} = \frac{1}{4-1} = \frac{1}{3} \quad y \in (1,4)$$

Using Definition 1.10, we have

$$E(Y) = \int_1^4 y \cdot \frac{1}{3} dy = \frac{1}{3} \left. \frac{y^2}{2} \right|_1^4 = 2.5$$

Expectation of a Function

Definition 1.11

The **expectation of function of a random variable** $g(X)$ is given by

$$E(g(X)) = \sum_{x \in S} g(x)f(x) \qquad \text{(for a discrete distribution)}$$

$$E(g(X)) = \int_S g(x)f(x)dx \qquad \text{(for a continuous distribution)}$$

Percentiles

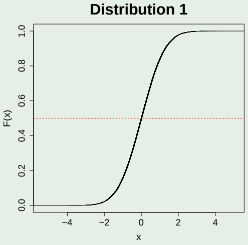
Definition 1.12

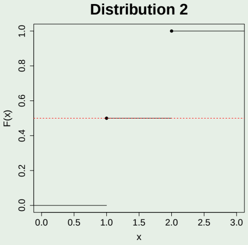
The $100p$-**th percentile** of a distribution is any value of π_p such that the CDF $F(\pi_p^-) \leq p \leq F(\pi_p)$.

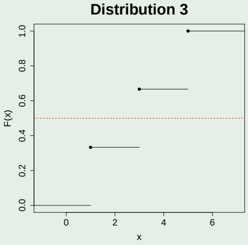
The **median** is defined to be the 50th percentile, $\pi_{0.5}$.

Example

Median Example







(a) Continuous Normal(0,1) distribution. Median $= 0$.

(b) Discrete Unif({1,2}) distribution. The set of medians is an interval $[1, 2)$.

(c) Discrete Unif({1,3,5}) distribution. Median $= 3$.

Moments

Definition 1.14 and Definition 1.16

The k-**th raw moment** of a random variable X is defined to be $E[X^k]$, and is sometimes denoted μ_k'.

The k-**th central moment** of a random variable X is $\mu_k = E[(X - \mu)^k]$.

Variance

Variance

One example of a central moment is **variance** σ^2 or $Var(X)$ (and rarely denoted μ_2), is given by

$$Var(X) = \sigma^2 = E[(X - \mu)^2)]$$

Standard deviation is defined by $\sigma = \sqrt{Var(X)}$.

A convenient formula

$$Var(X) = E[(X - \mu)^2)] = \mu_2 = E[X^2] - E[X]^2 \quad (1.8)$$

Example

Example

Suppose X is uniformly distributed on (a, b). Calculate the variance of X.

Example

Example

Answer

Using the same approach as in the earlier example, $E(X) = (b+a)/2$.

$$E(X^2) = \int_a^b x^2 f(x)dx = \frac{1}{b-a}\int_a^b x^2 dx = \frac{a^2+ab+b^2}{3}$$

$$\begin{aligned} Var(X) &= E(X^2) - E(X)^2 \\ &= \frac{a^2+ab+b^2}{3} - \frac{a^2+2ab+b^2}{4} \\ &= \frac{(b-a)^2}{12} \end{aligned}$$

Normal Distribution

Definition 1.20

Let $X \sim \text{Normal}(\mu, \sigma^2)$. Then,

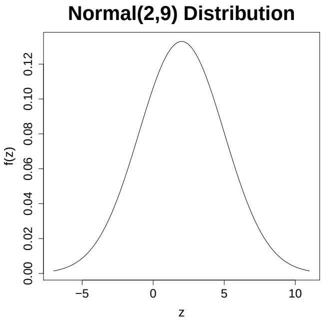
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

$$E(X) = \mu$$

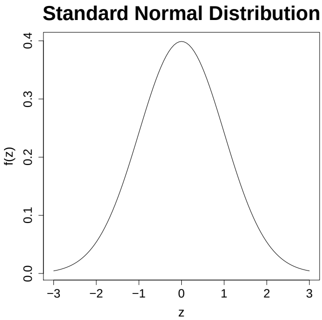
$$Var(X) = \sigma^2$$

A **standard normal distribution** is a $\text{Normal}(0, 1)$ distribution. A standard normal random variable is often denoted as Z.

Normal Distribution



(a) PDF of $X \sim \text{Normal}(2, 9)$



(b) PDF of Z

Remarks

Useful Fact

Let $X \sim \text{Normal}(\mu, \sigma^2)$. Then, we can *standardize* X in such a way that we get a standard normal distribution. To do so, we define Z as follows:

$$Z = \frac{X - \mu}{\sigma} \qquad (1.9)$$

Then, $Z \sim \text{Normal}(0, 1)$.

$\Phi(z)$ and $\phi(z)$

The CDF of a normal distribution has no closed form. Therefore, we use tables for the standard normal CDF $\Phi(z)$ (included on the Exam 4/C Tables) to compute normal probabilities. $\phi(z)$ denotes the PDF of a standard normal distribution.

Generating Functions

Definition 1.21

For a random variable X, the **moment generating function** (MGF) of X (denoted $M_X(t)$) is equal to $E(e^{tx})$.

Furthermore, for discrete variables, the **probability generating function** (PGF) (denoted by $P_X(z)$) is equal to $E(z^X)$.

Useful Formulas

Formula (1.10)

$$M_X(t) = E(e^{tX}) = E\left((e^t)^X\right) = P_X(e^t)$$

Useful Formulas

Formula (1.11)

$$P_X(z) = E(z^X) = E\left((e^{\ln z})^X\right) = E(e^{(\ln z)X}) = M_X(\ln z)$$

Useful Formulas

Formula

$$Pr(S = k) = \frac{d^k}{dz^k} \frac{P_S(z)}{k!} \bigg|_{z=0}$$

Useful Formulas

Formula

$$E(X^k) = \frac{d^k}{dt^k} M_X(t)\bigg|_{t=0}$$

Example

Example

Compute the MGF for the Poisson distribution, using the PGF given in the Exam 4/C Tables.

Example

Example

Answer

The Exam 4/C Tables give the PGF to be $P_X(z) = e^{\lambda(z-1)}$. Thus,

$$M_X(t) = P_X(e^t) = e^{\lambda(e^t - 1)}$$

Formulas

Let $\{X_1, \ldots X_k\}$ be a sequence of random variables, and let S_k denote their sum.

Results

$$E(S_k) = \sum_{i=1}^{k} E(X_i)$$

If all X_i are independent, then

$$Var(S_k) = \sum_{i=1}^{k} Var(X_i)$$

Formulas

Let $\{X_1, \ldots X_k\}$ be a sequence of random variables, and let S_k denote their sum.

Theorem 1.23

If all X_i are independent, then

$$M_{S_k}(t) = \prod_{i=1}^{k} M_{X_i}(t)$$

$$P_{S_k}(z) = \prod_{i=1}^{k} P_{X_i}(z)$$

Example

MGF Example

Find the distribution of S_k where each $X_i \sim \text{gamma}(\alpha_i, \theta)$ and all the X_i's are independent.

MGF Example

MGF Example

Answer

We apply Theorem 1.23.

$$M_{S_k}(t) = \prod_{i=1}^{k} M_{X_i}(t) = \prod_{i=1}^{k} \frac{1}{(1-\theta t)^{\alpha_i}} = \frac{1}{(1-\theta t)^{\sum \alpha_i}}$$

Central Limit Theorem

Theorem 1.25

Let $\{X_1, \ldots X_k\}$ be a sequence of random variables, and let S_k denote their sum, i.e. $S_k = X_1 + X_2 + ... + X_k$. Under certain nice conditions (which usually can be assumed for the actuarial exam!),

$$\lim_{k \to \infty} \frac{S_k - E(S_k)}{\sqrt{Var(S_k)}} \rightsquigarrow N(0, 1)$$

where $\rightsquigarrow$ means convergence in distribution.

Central Limit Theorem

The corollary below is extremely important and serves as the basis for confidence intervals and hypothesis testing (which we cover later on).

Corollary 1.26

Assume that the X_i are independent and identically distributed (iid), such that $E(X_i) = \mu$ and $Var(X_i) = \sigma^2$ for all i. Then the Central Limit Theorem implies that:

1. S_k approximately follows a Normal$(k\mu, k\sigma^2)$ distribution.
2. $\bar{X} = S_k/k$ approximately follows a Normal$(\mu, \sigma^2/k)$ distribution.

Example

Example 1.27

Suppose we ask 100 actuarial students about whether they believe that they will pass Exam 4 on their first sitting. Let S_{100} denote the number of people who say yes. In fact, we could write out $S_{100} = \sum_{i=1}^{100} X_i$, where each $X_i \sim \text{Binomial}(m = 1, q)$ independently. If the true q is 0.4 (40% of all actuarial students sitting for Exam 4 for the first time believe they will pass it the first time), then what is the probability that our poll resulted in at least 50 people believing they will pass?

Ignore the continuity correction for normal approximation.

Example

Example

Answer

$\bar{X}$ approximately follows a Normal$(\mu,\sigma^2/k)$ = Normal$(0.4, 0.24/100)$ distribution.

$$
\begin{aligned}
P(\bar{X} > 0.5) &= P\left(\frac{\bar{X} - 0.4}{\sqrt{0.24/100}} \geq \frac{0.5 - 0.4}{\sqrt{0.24/100}}\right) \\
&= P\left(Z \geq 2.04\right) \\
&= 1 - P(Z < 2.04) \\
&= 1 - \Phi(2.04)
\end{aligned}
$$

From the standard normal distribution table, we find that $\Phi(2.04) = 0.9793$, so $P(\bar{X} > 0.5) = 1 - 0.9793 = 0.0207$.