Transformation of Random Variables in Probability

Probability Theory is the branch of science which deals with random features of our world.

Probability Theory is instrumental in physics, biology, economics,
finance, etc.

Here we study how the definite integral and the fundamental theorem of Calculus is used in Probability Theory.

PROBLEM: Find the derivative of the function

$$\int_{2x}^{3x} \frac{u^2 - 1}{u^2 + 1} du.$$

SOLUTION: Look at a more general situation. Let

$$Q(x) = \int_a^{g(x)} f(u) du.$$

Find the derivative $Q'(x)$. Denote

$$F(z) = \int_a^z f(u) du$$

Note that $Q(x) = F(g(x))$ and also $F'(z) = f(z)$.
Using the chain rule, we obtain

$$Q'(x) = F'(g(x))g'(x) = f(g(x))g'(x).$$

To finish the problem, note that

$$\int_{2x}^{3x} \frac{u^2-1}{u^2+1}du = \int_0^{3x} \frac{u^2-1}{u^2+1}du - \int_0^{2x} \frac{u^2-1}{u^2+1}du.$$

Then, applying the formula

$$\frac{d}{dx}\int_a^{g(x)} f(u)du = f(g(x))g'(x)$$

obtain

$$\begin{aligned}
\frac{d}{dx}\int_{2x}^{3x} \frac{u^2-1}{u^2+1}du &= \frac{d}{dx}\int_0^{3x} \frac{u^2-1}{u^2+1}du - \frac{d}{dx}\int_0^{2x} \frac{u^2-1}{u^2+1}du \\
&= \frac{3[(3x)^2-1]}{(3x)^2+1} - \frac{2[(2x)^2-1]}{(2x)^2+1} \\
&= \frac{27x^2-3}{9x^2+1} - \frac{8x^2-2}{4x^2+1}.
\end{aligned}$$

How does this relate to probability theory?

Let us start from something familiar

A random variable is called **discrete** if it takes finite number of values, say n. Let $A < X < B$.
For the discrete random variable X, one can define its distribution:

Distribution of a discrete random variable

Value of X	x_1	x_2	x_3	$\ldots$	x_{n-1}	x_n
Probability	p_1	p_2	p_3	$\ldots$	p_{n-1}	p_n

CUMULATIVE DISTRIBUTION FUNCTION (CDF)

One can calculate any probability associated with X, $A < X < B$. For example,

$$P(a < X \leq b) = p_k + p_{k+1} + \cdots + p_{k+s}$$

where $a < x_k$ and $x_{k+s} \leq b$.
The cumulative distribution function (cdf) $F_X(x)$ of X is the probability that $X \leq x$:

$$F_X(x) = P(X \leq x) = p_1 + p_2 + \cdots + p_l, \quad x_l \leq x$$

Observe that $F_X(A) = 0$, $F_X(B) = 1$ and $0 \leq F_X(x) \leq 1$.
The cdf of X can be used for evaluating probabilities:

$$P(a < X \leq b) = F(b) - F(a).$$

TRANSFORMATIONS

Let $Y = g(X)$ and $y_i = g(x_i)$.

Distribution of Y

Value of X	y_1	y_2	y_3	$\cdots$	y_{n-1}	y_n
Probability	p_1	p_2	p_3	$\cdots$	p_{n-1}	p_n

Probabilities: $\quad P(c < Y \leq d) = p_l + p_{l+1} + \cdots + p_{l+r}.$

Here the sum is taken over i's such that $c < y_i = g(x_i) < d$.

The cdf of Y is

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = p_l + p_{l+1} + \cdots + p_{l+r}$$

The sum is taken over i such that $y_i = g(x_i) \leq y$.

THE MEAN

The mean of the random variable X is

$$EX = x_1 p_1 + x_2 p_2 + \cdots + x_n p_n.$$

The mean of the random variable Y can be found as

$$EY = g(x_1) p_1 + g(x_2) p_2 + \cdots + g(x_n) p_n.$$

The mean of the random variable indicates the average value the random variable takes.

EXAMPLE

Consider a discrete random variable X

Distribution of X

Value of X	1	2	4	6	8
Probability	0.1	0.3	0.05	0.25	0.3

Probability $P(2 < X \leq 6) = 0.05 + 0.25 = 0.3$.

Find $P(2 < X \leq 6)$ using the cdf $F_X(x)$

$F_X(2) = 0.1 + 0.3 = 0.4$

$F_X(6) = 0.1 + 0.3 + 0.05 + 0.25 = 0.7$

. $\Longrightarrow P(2 < X \leq 6) = F_X(6) - F_X(2)$

Also, $EX = 1*0.1+2*0.3+4*0.05+6*0.25+8*0.3 = 4.8$

EXAMPLE: CONTINUATION

Let $Y = X^2$. The distribution of Y has the form:

Distribution of Y

Value of Y	1	4	16	36	64
Probability	0.1	0.3	0.05	0.25	0.3

Note that $F_Y(4) = F_X(2) = 0.4$, $F_Y(36) = F_X(6) = 0.7$
$P(4 < Y \leq 36) = P(2 < X \leq 6) = 0.3$.

Also, $EY = 1*0.1+4*0.3+16*0.05+36*0.25+64*0.3 = 1^2*0.1+2^2*0.3+4^2*0.05+6^2*0.25+8^2*0.3 = 30.3$.

AIRLINE EXAMPLE

Airlines frequently overbook the flights.

Suppose, for a plane with 100 seats, an airline takes 110 reservations.

X is the number of people with reservations who show up

From past experiences, we know the distribution of X

Distribution of X

i	1	2	3	4	5	6	7	8
x_i	95	96	97	98	99	100	101	102
$p(x_i)$	0.05	0.10	0.12	0.14	0.24	0.17	0.06	0.04

i	9	10	11	12	13	14	15	16
x_i	103	104	105	106	107	108	109	110
$p(x_i)$	0.03	0.02	0.01	0.005	0.005	0.005	0.0037	0.0013

AIRLINE PROFIT

Suppose that the airline charges \$200 per ticket
So, it is profitable to sell more tickets
Every extra passenger who shows up for the flight but does not have a seat cost airline \$250
X is the number of people with reservations who show up
Y is the airline profit from overbooking

$$Y = g(X) = 10 * 200 - (X - 100) * 250 * I(X > 100)$$

Here $I(X > 100)$ is the indicator function
$I(X > 100) = 1$ if $X > 100$; $\quad I(X > 100) = 0$ if $X \leq 100$

GROUP PROBLEM SOLVING (10 MINUTES)

Construct the probability distribution of Y
(use Table 5 on page 7–5).

Calculate the average profit.

Distribution of the amount of the profit

i	1	2	3	4	5	6	7	8
y_i	2000	2000	2000	2000	2000	2000	1750	1500
$p(y_i)$	0.05	0.10	0.12	0.14	0.24	0.17	0.06	0.04

i	9	10	11	12	13	14	15	16
y_i	1250	1000	750	500	250	0	-250	-500
$p(x_i)$	0.03	0.02	0.01	0.005	0.005	0.005	0.0037	0.0013

Here $0.05 + 0.10 + 0.12 + 0.14 + 0.24 + 0.17 = 0.82$

Hence, the table can be rewritten like this:

Distribution of the amount of the profit

i	1	2	3	4	5	6
y_i	2000	1750	1500	1250	1000	750
$p(y_i)$	0.82	0.06	0.04	0.03	0.02	0.01

i	7	8	9	10	11
y_i	500	250	0	-250	-500
$p(x_i)$	0.005	0.005	0.005	0.0037	0.0013

The cdf of Y is

$$F_Y(y) = P(Y \leq y) = P(g(X) \leq y) = \sum_{i:g(x_i)\leq y} p_i.$$

The average profit is

$$\begin{aligned} EY &= \sum_i g(x_i)p_i = \sum_i y_i p_i \\ &= 2000 * 0.82 + 1750 * 0.06 + 1500 * 0.04 + 1250 * 0.03 \\ &+ 1000 * 0.02 + 750 * 0.01 + 500 * 0.005 + 250 * 0.005 \\ &+ 0 * 0.005 - 250 * 0.0037 - 500 * 0.0013 \end{aligned}$$

QUESTION

What is the probability that the airline does not have any extra expenses due to overbooking?

How can this be expressed via the value of the random variables X?

What is the probability that overbooking is profitable? How can this be expressed via the value of the random variable Y?

FIVE MINUTE PAPER

Would overbooking be still profitable if the ticket cost $150? Explain.

SOLUTION TO FIVE MINUTE PAPER

Distribution of the amount of the profit

i	1	2	3	4	5	6	7	8
y_i	1500	1500	1500	1500	1500	1500	1250	1000
$p(y_i)$	0.05	0.10	0.12	0.14	0.24	0.17	0.06	0.04

i	9	10	11	12	13	14	15	16
y_i	750	500	250	0	-250	-500	-750	-1000
$p(x_i)$	0.03	0.02	0.01	0.005	0.005	0.005	0.0037	0.0013

Again $0.05 + 0.10 + 0.12 + 0.14 + 0.24 + 0.17 = 0.82$.

The average profit is

$$\begin{aligned} EY &= 1500 * 0.82 + 1250 * 0.06 + 1000 * 0.04 + 750 * 0.03 \\ &+ 500 * 0.02 + 250 * 0.01 + 0 * 0.005 - 250 * 0.005 \\ &- 500 * 0.005 - 750 * 0.0037 - 1000 * 0.0013 = 1374.675 \end{aligned}$$

CONTINUOUS RANDOM VARIABLES

A random variable X which may take any value in the finite interval (A, B) is called **continuous**.

Then, the probability of any particular value x is zero:
$P(X = x) = 0$ for any x

How can one calculate the probabilities?
Calculate the cumulative distribution function (cdf) of X.

THE CUMULATIVE DISTRIBUTION FUNCTION (CDF)

The cdf of X at x is $F_X(x) = P(X \leq x)$.
Use cdf to calculate $P(a < X \leq b) = F_X(b) - F_X(a)$.
One can also evaluate the "instantaneous" probability that $X = x$ as

$$\begin{aligned} f_X(x) &= \lim_{\Delta x \to 0} \frac{P(x < X \leq x + \Delta x)}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{F_X(x + \Delta x) - F_X(x)}{\Delta x} = \frac{dF_X(x)}{dx}. \end{aligned}$$

Function $f_X(x)$ is called the probability density function (pdf) of X.

Analogy: the cdf $F_X(x)$ is the distance
. the pdf $f_X(x)$ is the velocity

THE PDF AND THE MEAN

The properties of pdf $f_X(x) = F'_X(x)$ for $A < X < B$

1. $f_X(x) \geq 0$ since $F_X(x)$ is increasing
2. $\int_A^x f_X(z)dz = F_X(x)$ since $F_X(x) = f_X(x) = 0$ for $x \leq A$
3. $\int_a^b f_X(z)dz = P(a < X < b)$ since

   $$P(a < X < b) = F_X(b) - F_X(a) = \int_A^a f_X(z)dz - \int_A^b f_X(z)dz$$

The mean of X can be evaluated as

$$EX = \int_A^B x f_X(x)dx.$$

EXAMPLE

Let the pdf of X be $f_X(x) = 2(x+1)/3, \quad 0 < x < 1.$
Then the cdf $F_X(x)$ can be calculated as

$$F_X(x) = \int_0^x \frac{2(z+1)}{3}\, dz = \frac{x^2+2x}{3}, \quad 0 < x < 1,$$

$F_X(x) = 0$ for $x \le 0$ and $F_X(x) = 1$ for $x \ge 1$.
Using the expression for $F_X(x)$, one can find

$$\begin{aligned} P(X \le 1/2) &= F(1/2) = 5/12; \\ P(1/4 < X \le 1/2) &= F(1/2) - F(1/4) = 5/12 - 3/16. \end{aligned}$$

The expectation of X can be obtained as

$$EX = \int_0^1 \frac{2z(z+1)}{3} dz = \int_0^1 \frac{2z^2+2z}{3} dz = 5/9.$$

GROUP PROBLEM SOLVING (10 MINUTES)

The amount of time T a student at UCF spends working on Calculus at home every week has the pdf

$$f_T(t) = \frac{12t - t^2}{288}, \quad 0 \le t \le 12, \quad f_T(t) = 0, \quad \text{otherwise.}$$

Find the cdf $F_T(t)$ of T using formula

$$F_T(t) = \int_0^t f_T(z)dz$$

Using the expression for $F_T(t)$, find the probability that a student spends less than 4 hours studying Calculus.

ANSWERS

$$\begin{aligned} F_T(t) &= \int_0^t f_T(z)dz = \int_0^t \frac{12z - z^2}{288} dz \\ &= \left[\frac{6z^2}{288} - \frac{z^3}{3 * 288}\right]_0^t = \frac{18t^2 - t^3}{864} \\ P(T \leq 4) &= F(4) = 0.2593 \end{aligned}$$

FIVE MINUTE PAPER

Find the expectation of T using formula

$$ET = \int_0^{12} t\, f_T(t) dt.$$

What does this expectation show?

SOLUTION TO FIVE MINUTE PAPER

$$\begin{aligned} ET &= \int_0^{12} t\, f_T(t)dt = \frac{1}{288}\int_0^{12}(12t^2 - t^3)dt \\ &= \frac{1}{288}\left[4t^3 - 0.25t^4\right]_0^{12} = 6. \end{aligned}$$

The fact that $ET = 6$ means that on the average students spend 6 hours a week on Calculus

TRANSFORMATION OF VARIABLES

Let X have the pdf $f_X(x)$ and the cdf $F_x(x)$.
The variable of interest is $Y = g(X)$.

Let $y = g(x)$ be a monotone (increasing or decreasing) function of x.
In both cases, $g(x)$ has an inverse function $x = h(y)$.

Objective: evaluate the cdf $F_y(y)$ and the pdf $f_Y(y)$.

EVALUATION OF CDF AND PDF

Let $g(x)$ be an increasing function of x.
Then $h(y)$ is also an increasing function and

$$\begin{aligned} F_Y(y) &= P(Y \le y) = P(g(X) \le y) \\ &= P(h(g(X)) \le h(y)) = P(X \le h(y)) = F_X(h(y)), \end{aligned}$$

To find the pdf $f_Y(y)$ of Y, take the derivative of both sides

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) = \frac{d}{dy} F_X(h(y)) \\ &= \frac{d}{dy} \int_A^{h(y)} f_X(x)dx \quad \text{(since } F_X(z) = \int_A^z f_X(x)dx\text{)} \\ &= f_X(h(y))h'(y) \quad \text{(by fundamental theorem of Calculus).} \end{aligned}$$

Hence,

$$f_Y(y) = f_X(h(y))h'(y).$$

EVALUATION OF CDF AND PDF (CONTINUATION)

Now, let $g(x)$ be a decreasing function.
Then $h(y)$ is also decreasing and

$$\begin{aligned} F_Y(y) &= P(Y \le y) = P(g(X) \le y) \\ &= P(h(g(X)) \ge h(y)) = P(X \ge h(y)) \\ &= 1 - P(X < h(y)) = 1 - F_X(h(y)). \end{aligned}$$

Taking the derivative of both sides, obtain

$$f_Y(y) = -f_X(h(y))h'(y).$$

THE PDF AND THE MEAN

We have

$$\begin{aligned} f_Y(y) &= f_X(h(y))h'(y) \quad \text{if } h(y) \text{ is increasing} \\ f_Y(y) &= -f_X(h(y))h'(y) \quad \text{if } h(y) \text{ is decreasing} \end{aligned}$$

$$f_Y(y) = f_X(h(y))|h'(y)|.$$

The expectation of Y can be evaluated as

$$EY = \int_C^D y f_Y(y) dy$$

where $C < Y < D$.

THE MEAN (CONTINUATION)

The expectation of Y can also be found as

$$EY = \int_A^B g(x) f_X(x) dx.$$

Show that the expressions are equal

Consider the case when $g(x)$ is increasing.

Then $C = g(A)$, $D = g(B)$.

Use substitution: $h(y) = x$, $h'(y)dy = dx$, $y = g(x)$, $h(g(A)) = A$, $h(g(B)) = B$.

$$\begin{aligned} EY &= \int_C^D y f_Y(y) dy = \int_{g(A)}^{g(B)} y f_X(h(y)) h'(y) dy \\ &= \int_A^B g(x) f_X(x) dx. \end{aligned}$$

EXAMPLE

Consider variable X with the pdf

$$f_X(x) \;=\; 2(x+1)/3, 0 < x < 1.$$

Let $Y = X^2$, i.e. $g(x) = x^2$ and $h(y) = \sqrt{y}$, $0 < y < 1$.
Recall that $F_X(x)$

$$F_X(x) = \int_0^x \frac{2(z+1)}{3}\, dz = \frac{x^2+2x}{3}, \quad 0 < x < 1,$$

The cdf: $$\begin{aligned} F_Y(y) &= P(Y \le y) = P(X^2 \le y) = P(X \le \sqrt{y}) \\ &= F_X(\sqrt{y}) = (y + 2\sqrt{y})/3. \end{aligned}$$

The pdf of Y calculated by taking the derivative of $F_Y(y)$:

$$f_Y(y) = F_Y'(y) = (1 + y^{-1/2})/3.$$

EXAMPLE

The pdf of Y calculated by using formula $f_Y(y) = f_X(h(y))h'(y)$:

$$f_Y(y) = f_X(\sqrt{y}) * (\sqrt{y})' = \frac{2(\sqrt{y}+1)}{3}\ \frac{y^{-1/2}}{2} = \frac{1+y^{-1/2}}{3}.$$

The expectation of Y can be found as

$$EY = \frac{1}{3}\int_0^1 y(1+y^{-1/2})dy = \frac{1}{3}\int_0^1 (y+y^{1/2})dy = \frac{7}{18},$$

or

$$EY = EX^2 = \int_0^1 x^2\ \frac{2(x+1)}{3}dx = \frac{2}{3}\int_0^1 (x^3+x^2)dx = \frac{7}{18}.$$

HOMEWORK PROJECT

Recall that the amount of time T a student at UCF spends working on Calculus at home every week has the pdf

$$f_T(t) = \frac{12t - t^2}{288}, \quad 0 \le t \le 12, \quad f_T(t) = 0, \quad \text{otherwise}$$

The percentage of the final grade Y, $0 \le Y \le 100$, is related to the time a student spends working on Calculus at home as

$$Y = 100\sqrt{T/12}, \quad \text{i.e.} \quad g(t) = 100\sqrt{t/12}$$

HOMEWORK PROJECT

1. Using the expression for $F_T(t)$, find the probability that a student works less than 6 hours a week, works less than three hours a week, works more than 9 hours a week.

2. Find the inverse $t = h(y)$ of $y = g(t)$.

3. Find the cdf $F_Y(y)$ of the final percentage Y using the reasoning we employed in the general case. Write the explicit espression for $F_Y(y)$ as a final answer.

4. Using $F_Y(y)$, find the probability that a student has less than 70% (passes the course), more than 90% (gets an A).

5. Find the pdf $f_Y(y)$ of the final percentage Y by taking the derivative of $F_Y(y)$. Compare the result of differentiation with $f_Y(y)$ found using formula $f_Y(y) = -f_X(h(y))h'(y)$.

6. Graph the pdfs $f_T(t)$ and $f_Y(y)$ (on separate graphs). Is $f_T(t)$ symmetric? Is $f_Y(y)$ symmetric? What does this mean?

7. Find the expectation of Y using formulae

$$EY = \int_C^D y f_Y(y)dy \quad \text{and} \quad EY = \int_A^B g(x) f_X(x)dx.$$

Check that both formulae give the same result.

8. How much work does the student need to put in order to get at least a "B" in the course?