statistics

FOR

DATA SCIENCE

UNIT-2

feedback/corrections: vibha@pesu.pes.edu

Vibha Masti

Probability

Outcome – result of a single trial in an experiment

Event – Collection of outcomes of an experiment; subset of a sample space

Sample space – set of all possible outcomes of an experiment

CONDITIONAL PROBABILITY

Probability that A occurs given that B has occurred is called the conditional probability of A given B

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad P(B) \neq 0$$

Q1. **Example :** In a group of 100 sports car buyers, 40 bought alarm systems, 30 purchased bucket seats, and 20 purchased an alarm system and bucket seats. If a car buyer chosen at random bought an alarm system, what is the probability they also bought bucket seats?

$$P(\text{bucket} \mid \text{alarm}) = \frac{P(\text{bucket} \cap \text{alarm})}{P(\text{alarm})} = \frac{20}{40} = 0.5$$

$$= 50\%$$

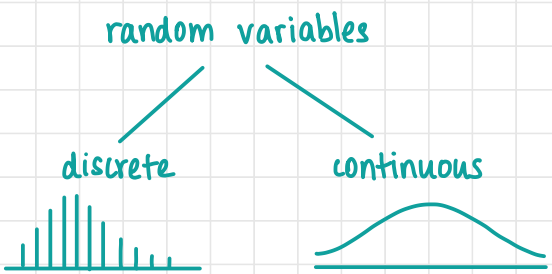
Mutually exclusive events

- no common outcomes
- $P(A \cup B) = P(A) + P(B)$ if A & B are mutually exclusive events

Random Variables

Random variable

- outcome of an experiment expressed as a number
- example: no. of heads in a series of 3 coin tosses, number of girls born in 4 deliveries etc.



- SS is finite or countable
- $P(a < X < b) = \sum_{a<t<b} f(t)$

- SS is uncountably infinite
- $P(a < X < b) = \int_a^b f(t)\,dt$

PROBABILITY MASS FUNCTION / PROBABILITY DISTRIBUTION

$$p(x) = P(X = x)$$

X = random variable
x = value of random variable

eg: $p(2) = P(X = 2)$

1) $0 < p(x) < 1$

2) $\sum p(x) = 1$

Q2. The number of flaws in a 1-inch length of copper wire varies from wire to wire. Overall, 48% of the wires produced have no flaws.39% have one flaw. 12%have two flaws and 1% have three flaws.

Write the Probability distribution of X, where X represents the no. of flaws in thewire.

Let X = no. of flaws and $P(X)$ be the probability of no. of flaws

X	P(X)
0	0.48
1	0.39
2	0.12
3	0.01
	1.00

CUMULATIVE DISTRIBUTION FUNCTION

- probability that random variable is less than or equal to a given value
- $F(x) = P(X \leq x) = \sum_{t \leq x} p(t)$

Q3. For Q2, write cdf

X	$P(X \leq x)$
0	0.48
1	0.87
2	0.99
3	1.00

MEAN OF DISCRETE RANDOM VARIABLE

- Let X be a DRV with Probability Mass Function $p(x) = P(X = x)$
- The mean of X is given by

$$\mu_X = \sum_x x\, P(X = x)$$

over all possible x's

VARIANCE OF DISCRETE RANDOM VARIABLE

$$\sigma_x^2 = \sum_x (x - \mu_X)^2 P(X = x)$$

$$\sigma_x^2 = \sum_x x^2 P(X = x) - \mu_X^2$$

Q4. A resistor in a certain circuit is specified to have a resistance in the range 99 Ω–101 Ω. An engineer obtains two resistors. The probability that both of them meet the specification is 0.36, the probability that exactly one of them meets the specification is 0.48, and the probability that neither of them meets the specification is 0.16. Let X represent the number of resistors that meet the specification. Find the probability mass function, and the mean, variance, and standard deviation of X.

X	P(X)
0	0.16
1	0.48
2	0.36

$\mu_X = 0.48 + 0.72 = 1.20$

$\sigma_X^2 = \sum x^2 P(X=x) - \mu_X^2$

$= (0.48 + 4 \times 0.36) - 1.20^2$

$= 1.92 - 1.44 = 0.48$

$\sigma_X^2 = 0.48$

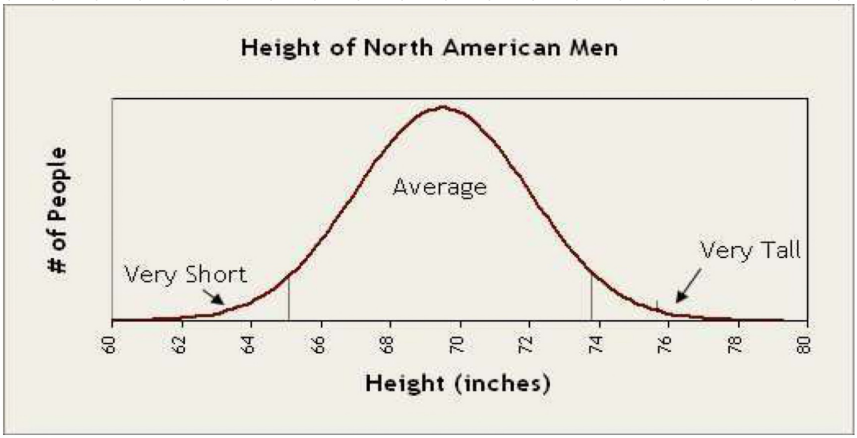
$\sigma_X = 0.693$

Probability Histogram

- PMF can be represented as a histogram for discrete values
- Area of a rectangle centered at x is $P(X=x)$

Continuous Random Variables

- described by a density curve (probability density function)



Probability Density Function

$f(x) = pdf$

$$\int_{-\infty}^{\infty} f(x)\,dx = 1$$

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

$$= \int_{a}^{b} f(x)\,dx$$

Qs. A hole is drilled in a sheet-metal component, and then a shaft is inserted through the hole. The shaft clearance is equal to the difference between the radius of the hole and the radius of the shaft. Let the random variable X denote the clearance, in millimeters. The probability density function of X is

$$f(x) = \begin{cases} 1.25(1 - x^4) & 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

Components with clearances larger than 0.8 mm must be scrapped. What proportion of components are scrapped?

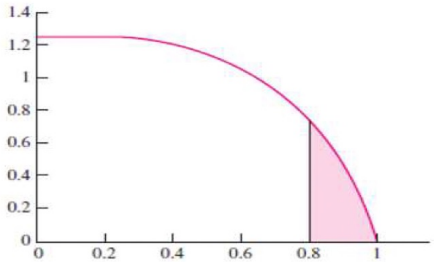


FIGURE 2.13 Graph of the probability density function of X, the clearance of a shaft. The area shaded is equal to $P(X > 0.8)$.

$$P(X>0.8) = \int_{0.8}^{\infty} f(x)\,dx = \int_{0.8}^{1} 1.25(1-x^4)\,dx = \left[1.25x - \frac{1.25x^5}{5}\right]_{0.8}^{1}$$

$$= 0.08192$$

Q c. Find CDF for above example

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

if $x \le 0$,

$$F(x) = \int_{-\infty}^{x} 0 = 0$$

if $0 < x < 1$

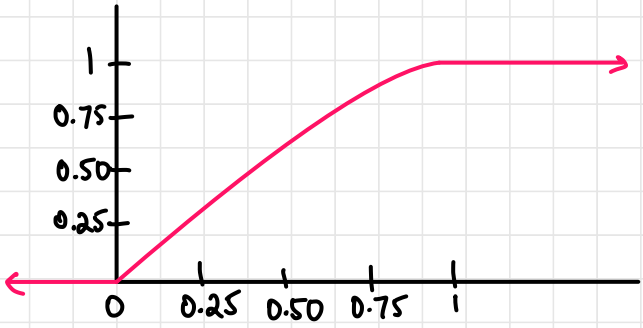
$$F(x) = \int_0^x 1.25(1-t^4)\,dt = 1.25\left[t - \frac{t^5}{5}\right]_0^x = 1.25x - 0.25x^5$$

if $x \geq 1$

$$F(x) = 0 + \int_0^1 f(t)\,dt + \int_1^x 0\,dt = 1$$

$$F(x) = \begin{cases} 0 & , \quad x \leq 0 \\ 1.25x - 0.25x^5 & , \quad 0 < x < 1 \\ 1 & , \quad x \geq 1 \end{cases}$$

Plot of $F(x)$



MEAN OF CONTINUOUS RANDOM VARIABLE

also called expected value

$$E(X) = \mu_x = \mu_x = \int_{-\infty}^{\infty} x\, f(x)\, dx$$

VARIANCE OF CONTINUOUS RANDOM VARIABLE

$$\sigma_x^2 = \int_{-\infty}^{\infty} (x - \mu_x)^2 f(x)\, dx$$

$$\sigma_x^2 = \int_{-\infty}^{\infty} x^2 f(x)\, dx - \mu_x^2$$

Q7. For Q5, find μ_x, σ_x^2, σ_x

$$\mu_x = \int_0^1 1.25(1-x^4)\, x\, dx = \frac{5}{12} = 0.417$$

$$\sigma_x^2 = \int_0^1 1.25(x^2 - x^6)\, dx - \mu_x^2 = \frac{5}{21} - \frac{25}{144} = 0.0645$$

$$\sigma_x = 0.254$$

PERENTILE & MEDIAN OF CRV

Percentile

$$F(x_p) = P(X \leq x_p) = \int_{-\infty}^{x_p} f(x)\,dx = \frac{p}{100}$$

p^{th} percentile

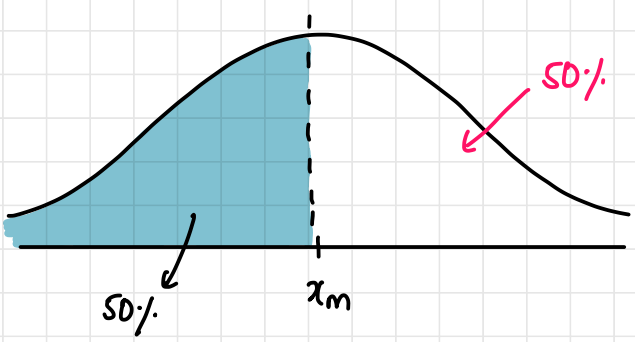
p % of the distribution lies below x_p

Median

Let the median be x_m

$$F(x_m) = P(X \leq x_m) = \int_{-\infty}^{x_m} f(x)\,dx = 0.5$$

50% of the distribution lies below x_m



Q8. For a random variable X,

$$f(x) = \begin{cases} cx^3 & 2 \le x \le 4 \\ 0 & \text{otherwise} \end{cases}$$

(a) What value of c makes it a legitimate probability dist?
(b) What is $P(X>3)$
(c) What is $P(X \le 2.7)$
(d) Median
(e) Mean, variance
(f) cdf

(a) $$\int_{-\infty}^{\infty} f(x)\,dx = 1 \Rightarrow \int_{2}^{4} cx^3\,dx = 1$$

$$c\left[\frac{x^4}{4}\right]_2^4 = 1$$

$$c\left(4^4 - 2^4\right) = 4 \Rightarrow c(256-16) = 4$$

$$c = \frac{4}{240} = \frac{1}{60} \Rightarrow f(x) = \begin{cases} \frac{x^3}{60}, & 2 \le x \le 4 \\ 0, & \text{else} \end{cases}$$

(b) $$P(X>3) = \int_{3}^{\infty} f(x)\,dx = \int_{3}^{4} \frac{x^3}{60}\,dx = 0.729$$

(c) $$P(X \le 2.7) = \int_{2}^{2.7} \frac{x^3}{60}\,dx = 0.155$$

(d) Median = x_m

we know x_m must lie b/w 2 & 4 as 100% of the data lies in that range

$$0.5 = \int_2^{x_m} \frac{x^3}{60}\,dx = \left[\frac{x^4}{60\times 4}\right]_2^{x_m}$$

$$120 = [x_m^4 - 16]$$
$$136 = x_m^4$$

$$x_m = 3.415$$

(e) mean, variance

$$\mu_x = \int_2^4 \frac{x^4}{60}\,dx = \frac{248}{75} = 3.31$$

$$\sigma_x^2 = \int_2^4 \frac{x^5}{60}\,dx - \left(\frac{248}{75}\right)^2 = \frac{56}{5} - \left(\frac{248}{75}\right)^2 = 0.266$$

(f) cdf $= \int_2^x \frac{t^3}{60}\,dt = \frac{1}{60}\left[\frac{t^4}{4}\right]_2^x = \frac{1}{60}\left(\frac{x^4}{4} - 4\right)$ $\quad [2 \le x \le 4]$

$x < 2 : 0$
$x > 4 : 1$

LINEAR FUNCTIONS OF RANDOM VARIABLES

1) Addition / Subtraction

$$X \longrightarrow X + c$$

- $\mu_X \longrightarrow \mu_X + c$
- $\sigma_X^2 \longrightarrow \sigma_X^2$

2) Multiplication / Division

$$X \longrightarrow cX$$

- $\mu_X \longrightarrow c\mu_X$
- $\sigma_X^2 \longrightarrow c^2 \sigma_X^2$

3) Linear Combination of Random Variables

$$Y = a_1 X_1 + a_2 X_2 + \ldots + a_n X_n$$

$$\mu_Y = a_1 \mu_{X_1} + a_2 \mu_{X_2} + \ldots + a_n \mu_{X_n}$$

$$\sigma_Y^2 = a_1^2 \sigma_{X_1}^2 + a_2^2 \sigma_{X_2}^2 + \ldots + a_n^2 \sigma_{X_n}^2$$

eg: $Y = X_1 - X_2$

$\sigma_Y^2 = \sigma_{X_1}^2 \oplus \sigma_{X_2}^2$

Q9. The molarity of a solute in solution is defined to be the number of moles of solute per liter of solution (1 mole $= 6.02 \times 10^{23}$ molecules). If the molarity of a stock solution of concentrated sulfuric acid (H_2SO_4) is X, and if one part of the solution is mixed with N parts water, the molarity Y of the dilute solution is given by $Y = X/(N+1)$. Assume that the stock solution is manufactured by a process that produces a molarity with mean 18 and standard deviation 0.1. If 100 mL of stock solution is added to 300 mL of water, find the mean and standard deviation of the molarity of the dilute solution.

$\mu_X = 18 \quad \sigma_X = 0.1 \quad \sigma_X^2 = 0.01$

$Y = \frac{X}{4} \Rightarrow \mu_Y = \frac{18}{4} = 4.5 \; ; \; \sigma_Y^2 = \frac{0.01}{16} \Rightarrow \sigma_Y = 0.025$

Independent & Identically Distributed Variables

- If $X_1, X_2, \ldots, X_n$ are independent variables all with identical distributions, they are IID variables
- Same mean & variance but outcome of one observation does not affect outcome of others
- eg: dice rolling, casino games, coin tossing etc

Q10. If X & Y are independent variables and E(X) = 9.5, E(Y) = 6.8, SD(X) = 0.4 and SD(Y) = 0.1, find mean & SD of the following

(a) $3X = Z \Rightarrow \mu_Z = 3 \times 9.5 = 28.5$, $\sigma_Z = 3 \times 0.4 = 1.2$

(b) $Y - X = Z \Rightarrow \mu_Z = -2.7$, $\sigma_Z = \sqrt{0.4^2 + 0.1^2} = 0.412$

(c) $X + 4Y = Z \Rightarrow \mu_Z = 9.5 + 4 \times 6.8 = 36.7$, $\sigma_Z = \sqrt{0.4^2 + 4^2 \times 0.1^2} = 0.566$

Chebyshev's Inequality

- The probability that a random variable differs from its mean by k standard deviations or more is never greater than $1/k^2$

$$P(|X-\mu_X| \geq k\sigma_X) \leq \frac{1}{k^2} \quad (\text{only } k>1)$$

- At least $1-\frac{1}{k^2}$ of data must fall with k standard devs.
- For $k=2$, $1-1/k^2 = 3/4 = 75\%$ or at least 75% of data within 2 SDs

Discrete
PROBABILITY DISTRIBUTIONS

Bernoulli distribution

Bernoulli Trial

- single trial
- 2 outcomes: success or failure
- P(success) = p and P(failure) = 1-p = q
- eg flipping a coin

For any random variable X (probability mass function p(x))

X = 1 if success occurs, probability = p
X = 0 if failure occurs, probability = 1-p

then $X \sim \text{Bernoulli}(p)$

$p(x)$ for any $x \notin \{0,1\} = 0$

Mean & Variance

If $X \sim \text{Bernoulli}(p)$

Mean

$$\mu_X = \sum_{x=0}^{1} x\, p(x) = 0(1-p) + 1(p)$$

$$\boxed{\mu_X = p}$$

Variance

$$\sigma_X^2 = \sum_{x=0}^{1} (x-\mu)^2 p(x)$$

$$= \sum_{x=0}^{1} x^2 p(x) + \sum_{x=0}^{1} (\mu^2 - 2\mu x)\, p(x)$$

$$= \sum_{x=0}^{1} x^2 p(x) + \mu^2 \sum_{x=0}^{1} p(x) - 2\mu \sum_{x=0}^{1} x\, p(x)$$

$$= \sum_{x=0}^{1} x^2 p(x) - \mu^2$$

$$= 0^2(1-p) + 1^2(p) - p^2$$

$$= p - p^2 = p(1-p)$$

$$\boxed{\sigma_X^2 = p(1-p)}$$

binomial distribution

- n Bernoulli trials
- Probability of success remains the same for each trial

$$X \sim Bin(n,p)$$

Probability Mass Function

$P(X=x) =$ no. of arrangements of x successes in n trials $\cdot p^x (1-p)^{n-x}$

$$p(x) = P(X=x) = \begin{cases} \frac{n!}{x!\,(n-x)!} p^x (1-p)^{n-x} & x = 0, 1, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

Q11. Find pmf of $X \sim Bin(10, 0.4)$. Find $P(X=5)$.

$n = 10 \quad p = 0.4$

$$p(x) = \begin{cases} \frac{10!}{(10-x)!\,x!} 0.4^x\, 0.6^{10-x} \end{cases}$$

$$P(X=5) = p(5) = \frac{10!}{5!\,5!} 0.4^5\, 0.6^5 = 0.2007$$

Mean & Variance

$\mu = np \qquad \sigma^2 = np(1-p)$

Q12. Ten percent of the components manufactured by a certain process are defective. A component is chosen at random. Let X = 1 if the component is defective, and X = 0 otherwise. What is the distribution of X?

$p = 0.1$ (probability of success)

$X \sim \text{Bernoulli}(0.1)$

Q13. A coin is flipped 10 times. Let X be the no.of.heads that appear. What is the probability distribution of X?

$n = 10 \quad p = 0.5$

$X \sim \text{Binomial}(10, 0.5)$

Sample Proportion

- Estimated value of p

$$\hat{p} = \frac{X}{n}$$

X: no. of successes
n: no. of trials

- Not equal to p; just an estimate

Q14. A quality engineer is testing the calibration of a machine that packs ice cream into containers. In a sample of 20 containers, 3 are underfilled. Estimate the probability p that the machine underfills a container.

$n = 20$
$x = 3$

$$\hat{p} = \frac{x}{n} = \frac{3}{20} = 0.15 = 15\%$$

UNCERTAINTY IN SAMPLE PROPORTION

- Bias and uncertainty

Uncertainty

$$\text{uncertainty} = \sigma_{\hat{p}}$$

$$\sigma_x = \sqrt{np(1-p)}$$

$$\hat{p} = \frac{x}{n}$$

$$\sigma_{\hat{p}} = \sigma_{\frac{x}{n}} = \frac{\sigma_x}{n}$$

$$\sigma_{\hat{p}} = \frac{\sqrt{np(1-p)}}{n}$$

$$\boxed{\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}}$$

Bias

unintentional favouring of one outcome over the other in the population

expected value - true value

$$E(\hat{p}) - p$$
$$\mu_{\hat{p}} - p$$

$$\mu_{\hat{p}} = \mu_{\frac{x}{n}} = \frac{\mu_x}{n} = \frac{np}{n} = p$$

$$\text{bias} = p - p = 0$$

$$\text{bias} = 0$$

Q15. A quality engineer takes a random sample of 100 steel rods from a days production, and finds that 92 of them meet specifications.

1. Estimate the proportion of the days production that meets specifications.
2. Find the uncertainty in the estimate.
3. Estimate the no.of.rods that must be sampled to reduce the uncertainty to 1%?

1. $\hat{p} = \frac{92}{100} = 0.92$

2. $\sigma_{\hat{p}} = \sqrt{\frac{0.92 \times 0.08}{100}} = 0.027$

3. $0.01 = \sqrt{\frac{0.92 \times 0.08}{n}}$

$n = 736$

Q16. The safety commissioner in a large city wants to estimate the proportion of buildings in the city that are in violation of fire codes. A random sample of 40 buildings is chosen for inspection, and 4 of them are found to have fire code violations. Estimate the proportion of buildings in the city that have fire code violations, and find the uncertainty in the estimate

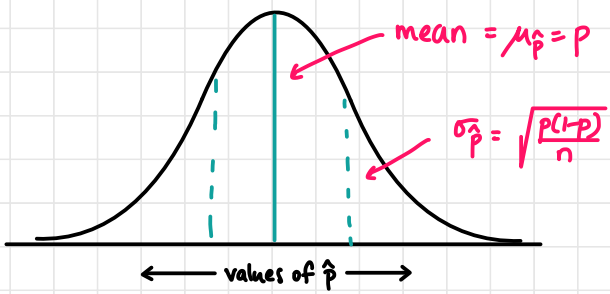
$n = 40 \quad x = 4$

$$\hat{p} = \frac{x}{n} = \frac{4}{40} = 0.1$$

$$\sigma_{\hat{p}} = \sqrt{\frac{0.1 \times 0.9}{40}} = 0.047$$

SAMPLING DISTRIBUTION OF THE sample proportion

- As n increases, sampling distribution of $\hat{p}$ becomes more normal
- Most accurate when p is close to 0.5



Poisson Distribution

- Occurences of a rare event during a specific interval (time, distance, area, volume
- Frequency of successes (low frequency, large population)
- Approximation of Binomial distribution with large n and small p

Q17. A mass contains 10000 atoms of a radioactive substance. The probability that a given atom will decay in a one minute period is 0.0002. Let X represent the no. of atoms that decay in 1 min. Write as Binomial distribution and find mean.

$$X \sim (10000, 0.0002)$$

mean is $\mu_X = (10000)(0.0002) = 2$

If n changes to 5000 and p changes to 0.0004. Let Y = random variable.

$$Y \sim (5000, 0.0004)$$

$$\mu_Y = 5000 \times 0.0004 = 2$$

The product np for X & Y are the same.

$$P(X=3) = \frac{10000!}{3!\,9997!}(0.0002)^3(0.9998)^{9997} = 0.180465091$$

$$P(Y=3) = \frac{5000!}{3!\,4997!}(0.0004)^3(0.9996)^{4997} = 0.180483143$$

- For large n and small p, we let $\lambda = np$

$$\frac{n!}{x!\,(n-x)!}\,p^x(1-p)^{n-x} \approx e^{-\lambda}\frac{\lambda^x}{x!}$$

Poisson distribution

$$p(x) = P(X=x) = \begin{cases} e^{-\lambda}\dfrac{\lambda^x}{x!}, & x \text{ is non-negative int} \\ 0, & \text{otherwise} \end{cases}$$

$X \sim \text{Poisson}(\lambda)$

mean $= \mu_X = \lambda$ variance $= \sigma_X^2 = \lambda$

per unit (freq)

Q18. If $X \sim \text{Poisson}(3)$, compute $P(X=2)$, $P(X=10)$, $P(X=0)$, $P(X=-1)$ and $P(X=0.5)$

$\lambda = 3$

$$P(X=2) = e^{-3}\frac{3^2}{2!} = 0.224$$

$$P(X=10) = e^{-3}\frac{3^{10}}{10!} = 0.00081$$

$$P(X=0) = e^{-3}\frac{3^0}{0!} = 0.0498$$

$P(X=0.5) = P(X=-1) = 0$

Q19. If electricity power failures occur according to a Poisson distribution with an average of 3 failures every twenty weeks, calculate the probability that there will not be more than one failure during a particular week.

$\lambda = \frac{3}{20}$ failures a week

$P(\text{not more than 1}) = P(X=0) + P(X=1)$

$$P(X=x) = e^{-\lambda} \frac{\lambda^x}{x!}$$

$$P(X=0) = e^{-3/20} \frac{(3/20)^0}{0!} + e^{-3/20} \frac{(3/20)^1}{1!}$$

$$= e^{-3/20}\left(1 + \frac{3}{20}\right) = 0.9898$$

Estimate a Rate

$$\hat{\lambda} = \frac{x}{t}$$ → events / → time

uncertainty & bias

bias

$$\mu_{\hat{\lambda}} - \lambda = \mu_{\frac{x}{t}} - \lambda = \frac{\mu_x}{t} - \lambda = \frac{\lambda t}{t} - \lambda$$

bias = 0

uncertainty

$$\sigma_{\hat{\lambda}} = \sigma_{\frac{x}{t}} = \frac{\sqrt{\lambda t}}{t} = \sqrt{\frac{\lambda}{t}}$$

Normal DISTRIBUTION

- Gaussian distribution / Bell curve
- Mean μ and standard deviation σ

normal distribution

pdf

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

for $-\infty < x < \infty$

- Any value is ~68% likely to be within 1 standard deviation of the mean $(\mu-\sigma, \mu+\sigma)$
- ~95% likely to be within 2 standard deviations of the mean $(\mu-2\sigma, \mu+2\sigma)$
- ~99.7% likely to be within 3 standard deviations of the mean $(\mu-3\sigma, \mu+3\sigma)$

Standard NORMAL DISTRIBUTION

- Normal distribution with $\mu=0$ and $\sigma=1$
- Random variable: $Z \sim N(0,1)$
- Probabilities calculated by using transformations to the standard normal variate z using z-table

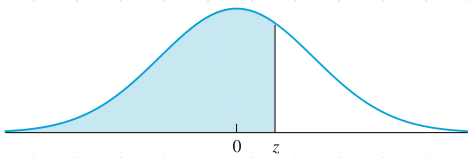
$$X \sim N(\mu, \sigma^2)$$

$$z = \frac{x-\mu}{\sigma}$$

standardising / z-scores

$$Z \sim N(0,1)$$

Z-table



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
−3.6	.0002	.0002	.0001	.0001	.0001	.0001	.0001	.0001	.0001	.0001
−3.5	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002	.0002
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998
3.5	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998	.9998
3.6	.9998	.9998	.9999	.9999	.9999	.9999	.9999	.9999	.9999	.9999

Q20. Let Z ~ N(0, 1) . Find a constant c for which

a) $P(Z \geq c) = 0.1587$

b) $P(c \leq Z \leq 0) = 0.4772$

c) $P(-c \leq Z \leq c) = 0.8664$

a) $P(Z < c) = 1 - 0.1587 = 0.8413 \Rightarrow c = 1.00$

b) $P(Z < c) = 0.5 - 0.4772 = 0.0228 \Rightarrow c = -2.00$

c) $P(Z \leq c) = \frac{0.8664}{2} + 0.5 = 0.9332 \Rightarrow c = 1.50$

Q21. The lifetime of a battery is in a certain application is normally distributed with mean 16 hours, standard deviation 2 hours.
a) What is the probability that a battery will last more than 19 hours?
b) Find the 10th percentile of the lifetimes.
c) A particular battery lasts 14.5 hours. What percentile is its lifetime on?

$$X \sim (16, 2^2)$$

a) $P(X > 19)$

$$z = \frac{19-16}{2} = 1.5$$

$$P(z > 1.5) = 1 - P(z < 1.5) = 1 - 0.9332$$
$$= 0.0668$$

b) area $= 0.1$

$$z = -1.28$$
$$x = -1.28 \times 2 + 16 = -2.56 + 16$$
$$= 13.44$$

c) $x = 14.5$

$P(X < 14.5) = ?$

$$z = \frac{14.5-16}{2} = -0.75$$

$P(z < -0.75) = 0.2266 = 22.66^{th}$ percentile

Linear functions of normal random variables

If $X \sim N(\mu_x, \sigma_x^2)$ and $Y = aX + b$

$$Y \sim N(a\mu_x + b,\ a^2\sigma_x^2)$$

Q22. A light fixture holds two lightbulbs. Bulb A is a type whose lifetime is normally distibuted with mean 800 hours and standard deviation 100 hours. Bulb B has a lifetime that is normally distibuted with mean 900 hours and standard deviation 150 hours. Assume the lifetimes of the bulbs are independent.

1) What is the probability Bulb B lasts longer than bulb A?
2) What is the probability Bulb B lasts 200 hours more than bulb A?
3) Another light fixture holds only one bulb. A bulb of type A is installed, and when it burns out, a bulb of type B is installed. What is the probability that the total lifetime of the two bulbs is more than 2000 hours?

$A \sim N(800, 100^2)$
$B \sim N(900, 150^2)$

1) $P(B-A) > 0$

$Y = B - A$
$Y \sim N(100, 150^2 + 100^2)$
$Y \sim N(100, 180.277^2)$

$$z = \frac{0 - 100}{180.277} = \frac{-2}{\sqrt{13}} = -0.5547$$

$$P(Z > -0.5547) = 1 - P(Z < -0.5547)$$
$$= 1 - 0.2912$$
$$= 0.7088$$

2) $P(B-A) > 200$

$Y = B - A$

$Y \sim N(100, 150^2 + 100^2)$

$Y \sim N(100, 180.277^2)$

$$Z = \frac{200-100}{180.277} = \frac{100}{180.277} = +0.5547$$

$P = 0.2912$

3) $P(A+B) > 2000$

$Y = A + B$

$Y \sim N(1700, 180.277^2)$

$$Z = \frac{2000-1700}{180.277} = \frac{300}{180.277} = 1.6641$$

$$P(Z > 1.6641) = 1 - P(Z < 1.6641)$$
$$= 1 - 0.9515$$
$$= 0.0485$$

Student's t- Distribution

- Samples of a full population
- Larger sample size ⟶ normal distribution
- Theoretical probability distribution — symmetrical, bell-shaped, similar to standard normal curve
- Degrees of freedom - another parameter

$$df = \text{sample size} - 1$$

- As df increases, approaches standard normal distribution (after df = 30, almost identical)
- t-score calculated like z-score

Q23. A random sample of size 10 is drawn from a normal distribution.
a) Find $P(t > 1.833)$
b) Find $P(t > 1.5)$

$df = 9$

(a) $P(t > 1.833) = 0.05$

(b) $P(t > 1.5)$ = b/w 0.05 and 0.10

Q24. Computers from a particular company are found to last on average for three years without any hardware malfunction, with standard deviation of two months. At least what percent of the computers last between 31 months and 41 months?

Chebyshev's inequality

$\mu = 36$ months
$\sigma = 2$ months

within $k = 2.5$ SDs

$1 - \frac{1}{2.5^2} = 84\%$ at least

random NUMBER GENERATION

Inverse transform technique

Direct transformation for the Normal Distribution

Convolution Method

Acceptance / Rejection Method

- All assume that uniformly distributed numbers in $[0,1]$ exists

Inverse transform technique

- cdf is $F(x)$ where x is the random variable
- Set $F(x) = R$ where R is a uniformly distributed random variable in $[0,1]$
- Solve $F(x) = R$ for X in terms of R (in range of X)
- Generate uniform random numbers $R_1, R_2, R_3 \ldots$ and compute X_i by

$$X_i = F^{-1}(R_i)$$

Q25. Generate RVs for given cdf

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

- Range of X : $[a, b]$

$$R = \frac{X-a}{b-a}$$

$$R(b-a) + a = X \qquad \text{where} \quad R \in [0,1]$$

$$\boxed{X_i = a + R_i(b-a)}$$

Generation of Bernoulli RVs

$$P(X=1) = p$$
$$P(X=0) = 1-p$$

- A Bernoulli RV can only take up 2 values (0 and 1) and the probability of getting $X=1$ is p
- Generate U from $U(0,1)$
- If $U \le p$, $X=1$; else $X=0$

Generation of Binomial RVs

$$P(X=i) = \binom{n}{i} p^i (1-p)^{n-i}$$

$$P(X=i) = {}^nC_i \, p^i (1-p)^{n-i}$$

for i successes in n independent Bernoulli trials

- Generate n Bernoulli(p) RVs (using the technique shown above) $Y_1, Y_2, Y_3, \ldots, Y_n$
- Set $X = Y_1 + Y_2 + \ldots + Y_n$ (sum of n Bernoulli RVs)
- If $Y_i = 1$, number of successes out of n increases
- Total no. of successes $= \sum_{i=1}^{n} Y_i$

Generation of Poisson RVs

$X \sim \text{Poisson}(\lambda)$

$$P(X=i) = \frac{\lambda^i e^{-\lambda}}{i!}$$

i number of events in a unit time interval

Method 1

- Generate exponential inter-event times $Y_1, Y_2 \ldots$ with mean 1
- Let I be the smallest index such that

$$\sum_{i=1}^{I+1} Y_i > \lambda$$

- Set $X = I$

Method 2

- Generate $U(0,1)$ RVs $U_1, U_2 \ldots$
- Let N be the smallest index such that

$$\prod_{i=1}^{N+1} U_i < e^{-\lambda}$$

Steps

1. Set $i=0$, $P=1$
2. Generate U_{i+1} from $U(0,1)$ and replace P with $P \cdot U_{i+1}$
3. If $P < e^{-\lambda}$, accept $N=i$ and go to step 1.

   Else, reject $N=i$ and increment i and return to step 2.

   Upon completion of step 3, $P = \prod_{i=1}^{N+1} U_i$

- If $N=n$, then $n+1$ RVs are requested. So, the average number is given by $E(N+1) = \lambda+1$

Q26. Generate 3 Poisson variates with mean $\lambda = 0.2$ for the random numbers $R = 0.4357, 0.4146, 0.8353, 0.9952, 0.8004$

$$e^{-\lambda} = e^{-0.2} = 0.8187$$

1. For X_1

$i = 0$, $P = 1$, $R = 0.4357$

(a) $P = 0.4357 \times 1 = 0.4357 < 0.8187$ ✓
(b) Accept $N = 0$
(c) $X_1 = 0$

2. For X_2

$i = 0$, $P = 1$, $R = 0.4146$

(a) $P = 1 \times 0.4146 = 0.4146 < 0.8187$ ✓
(b) Accept $N = 0$
(c) $X_2 = 0$

3. For X_3

$i = 0$, $P = 1$, $R_1 = 0.8353$, $R_2 = 0.9952$, $R_3 = 0.8004$

(a) $P = 1 \times 0.8353 = 0.8353 > 0.8187$ ✗ $i = 0$
(b) $P = 0.8353 \times 0.9952 = 0.8313 > 0.8187$ ✗ $i = 1$
(c) $P = 0.8313 \times 0.8004 = 0.6654 < 0.8187$ ✓ $i = 2$
(d) Accept $N = 2$
(e) $X_3 = 2$

Poisson numbers : 0, 0, 2

Generation of Normal RVs

Acceptance / Rejection Technique

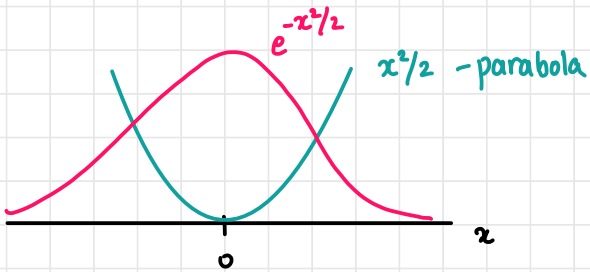
- If $X \sim N(0,1^2)$, the pdf of X is

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

reason for $1/\sqrt{2\pi}$:

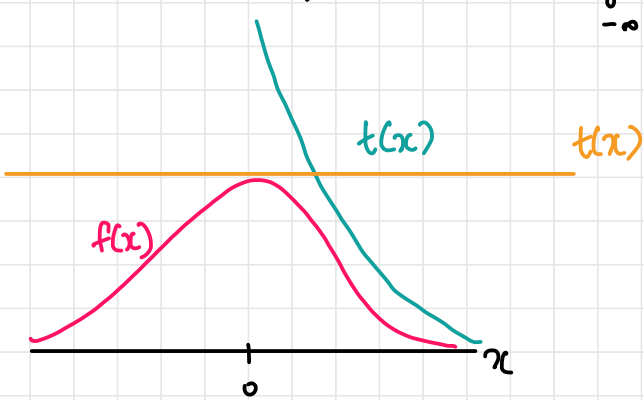
$$\int_{-\infty}^{\infty} e^{-x^2/2}\,dx = \sqrt{2\pi}$$

and $\int_{-\infty}^{\infty} f(x)\,dx = 1$



- Let a function $t(x)$ majorise pdf $f(x) \Rightarrow t$ is NOT a density

$t(x) \geq f(x) \quad \forall x$ $\qquad$ $\left(\int_{-\infty}^{\infty} t(x)\,dx \geq 1\right)$



- total area

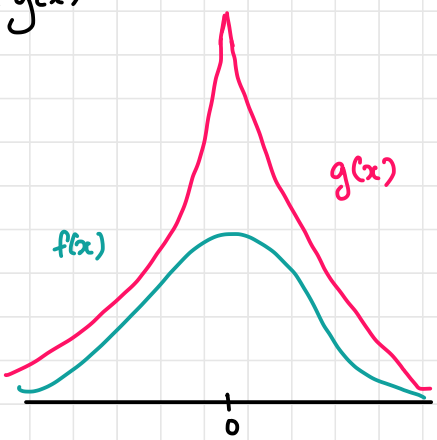
$$c = \int_{-\infty}^{\infty} t(x)\,dx \geq \int_{-\infty}^{\infty} f(x)\,dx = 1$$

- However, $\boxed{\frac{t(x)}{c}}$ is a density

- Let $g(x) = \frac{t(x)}{c} \Rightarrow t(x) = c\,g(x)$

- If $X \sim N(0,1^2)$, the pdf of $|X|$ is

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



- Let $g(x) = \sqrt{\frac{2e}{\pi}}\, e^{-x}$

Steps

1. Generate exponential Y with mean 1
2. Generate U from U(0,1)
3. If $U \leq e^{-(Y-1)^2/2}$, accept Y $\qquad \frac{f(Y)}{g(Y)} \leq c$

   Else, go to step 1
4. Return $X = Y$ or $X = -Y$ with probability 0.5

Box-Muller Method

- Box-Muller Transform transforms from a two-dimensional uniform distribution to a two-dimensional bivariate normal distribution (complex normal distribution)
- If U_1 and U_2 are independent RVs from $U(0,1)$

$$z_1 = \sqrt{-2 \ln U_1}\ \cos(2\pi U_2)$$ ← for single-variable normal distribution

$$z_2 = \sqrt{-2 \ln U_1}\ \sin(2\pi U_2)$$

No need to remember this:

- By solving for U_1 and U_2

$$z_1^2 + z_2^2 = -2 \ln U_1$$

$$-\ln U_1 = \frac{z_1^2 + z_2^2}{2}$$

$$U_1 = e^{-\left(\frac{z_1^2+z_2^2}{2}\right)}$$

$$\frac{z_2}{z_1} = \tan(2\pi U_2)$$

$$2\pi U_2 = \tan^{-1}\left(\frac{z_2}{z_1}\right)$$

$$U_2 = \frac{1}{2\pi} \tan^{-1}\left(\frac{z_2}{z_1}\right)$$

- Taking the Jacobian $\frac{\partial(U_1, U_2)}{\partial(z_1, z_2)}$

$$\begin{vmatrix} \frac{\partial U_1}{\partial z_1} & \frac{\partial U_1}{\partial z_2} \\ \frac{\partial U_2}{\partial z_1} & \frac{\partial U_2}{\partial z_2} \end{vmatrix} = -\left(\frac{1}{\sqrt{2\pi}} e^{-z_1^2/2}\right)\left(\frac{1}{\sqrt{2\pi}} e^{-z_2^2/2}\right)$$

Q27. Suppose the height of adult males in a certain area is normally distributed with a mean of 168 cm and a standard deviation of 8cm. Simulate the height of 4 adults.

$$X \sim N(168, 8^2) \qquad \mu = 168$$

$$\sigma = 8$$

$$Z = \frac{X - \mu}{\sigma} \qquad n = 4$$

- We first generate random Z values

- $Z = \sqrt{-2\ln(U_1)}\cos(2\pi U_2)$

X_1

Let $U_1 = 0.2432$, $U_2 = 0.5214$

$$Z_1 = \sqrt{-2\ln(0.2432)}\cos(2\pi \times 0.5214)$$

$$Z_1 = -1.666$$

$$X_1 = Z_1 \times \sigma + \mu = -1.666 \times 8 + 168$$

$X_1 = 154.67$ cm

X_2

Let $U_1 = 0.8921$, $U_2 = 0.6232$

$$Z_2 = \sqrt{-2\ln(0.8921)}\cos(2\pi \times 0.6232)$$

$$Z_2 = -0.3417$$

$X_2 = 165.27$ cm

X_3

Let $U_1 = 0.4421$, $U_2 = 0.0012$

$$z_3 = \sqrt{-2\ln(0.4421)}\cos(2\pi \times 0.0012)$$
$$z_3 = 1.2776$$

$X_3 = 178.22$ cm

X_4

Let $U_1 = 0.9921$, $U_2 = 0.7324$

$$z_4 = \sqrt{-2\ln(0.9921)}\cos(2\pi \times 0.7324)$$
$$z_4 = -0.0139$$

$X_4 = 167.89$ cm

Q25. Suppose the no. of shipments, x, on the loading dock of a company is either 0, 1 or 2. Generate RVs given $U = 0.23, 0.52, 0.81, 0.34$

(discrete RVs – inverse)

x	$P(x)$	$F(x)$
0	0.5	0.5
1	0.3	0.8
2	0.2	1.0

Let $U \sim U(0,1)$

$$x = \begin{cases} 0 & U \leq 0.5 \\ 1 & 0.5 < U \leq 0.8 \\ 2 & U > 0.8 \end{cases}$$

$U_1 = 0.23 \Rightarrow U < 0.5 \Rightarrow X_1 = 0$

$U_2 = 0.52 \Rightarrow 0.5 < U \leq 0.8 \Rightarrow X_2 = 1$

$U_3 = 0.81 \Rightarrow U > 0.8 \Rightarrow X_3 = 2$

$U_4 = 0.34 \Rightarrow U < 0.5 \Rightarrow X_4 = 0$