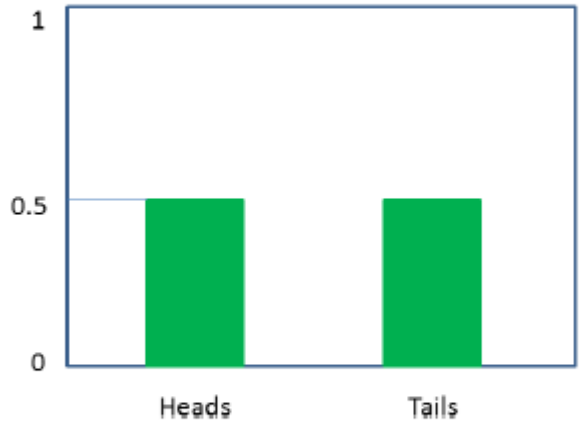
Analyzing attributes

PROBABILITY DISTRIBUTIONS

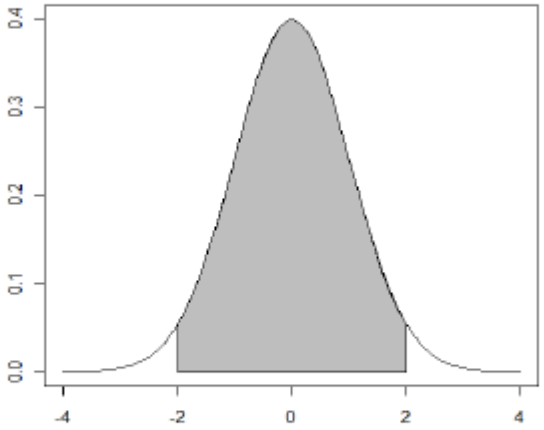
Random variable

- A variable that can take multiple values with different probabilities.
- The mathematical function describing these possible values along with their associated probabilities is called a probability distribution.

Discrete and Continuous



Countable



Measurable

Can any function be a probability distribution?

Discrete Distributions	Continuous Distributions
Probability that X can take a specific value x is $P(X = x) = p(x)$.	Probability that X is between two points a and b is $P(a \leq X \leq b) = \int_a^b f(x)dx$.
It is non-negative for all real x.	It is non-negative for all real x.
The sum of $p(x)$ over all possible values of x is 1, i.e., $\sum p(x) = 1$.	$\int_{-\infty}^{\infty} f(x)dx = 1$
Probability Mass Function	**Probability Density Function**

Histogram

A series of contiguous rectangles that represent the frequency of data in given class intervals.

How many class intervals?

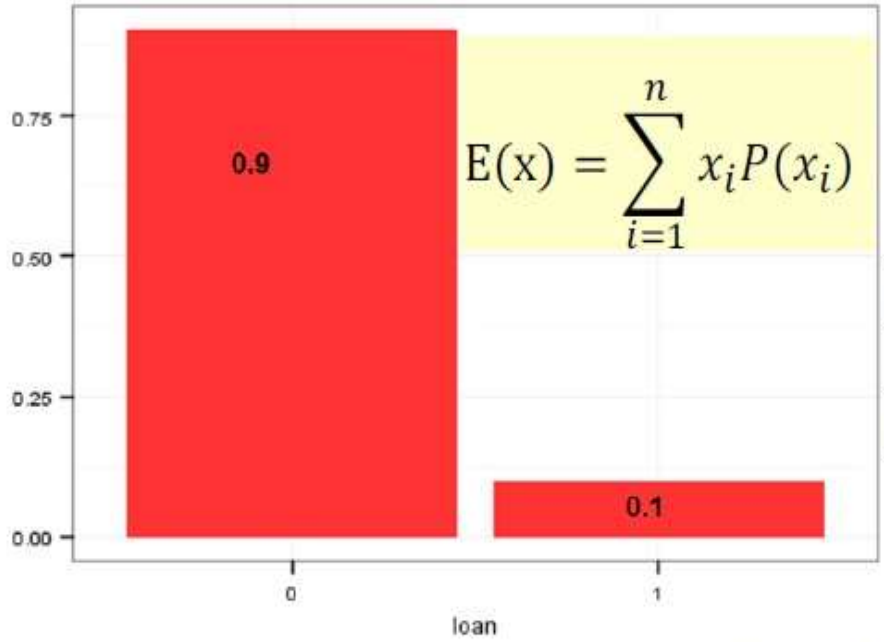
Rule of thumb: 5-15 (not too many and not too few)

Freedman-Diaconis rule:

$$No.of\ bins = \frac{(max - min)}{2 * IQR * n^{\frac{-1}{3}}},$$

$$where\ the\ denominator\ is\ the\ bin - width$$

Expectation: Discrete



$$\mathrm{E(x)} = \sum_{i=1}^{n} x_i P(x_i)$$

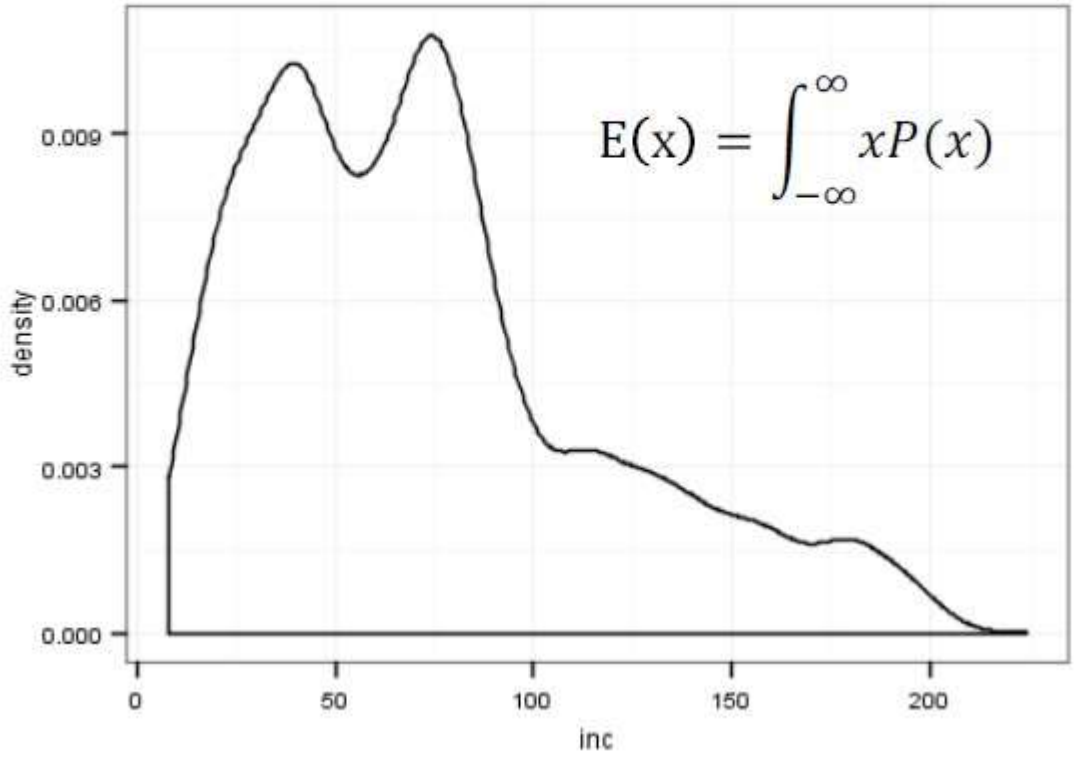
Recall anything like this?

Salary (BHD)	100	345	1000	9833
Frequency, f	10	1	10	2
Probability	0.43	0.04	0.43	0.09

$$\text{Mean, } \mu = \frac{\Sigma x}{n} = \frac{\Sigma f x}{\Sigma f} = \frac{100X10+345X1+1000X10+9833X2}{10+1+10+2} = 1348$$

$$\text{Expectation, } E(X) = 100 * 0.43 + 345 * 0.04 + 1000 * 0.43 + 9833 * 0.09 = 1348$$

Expectation: Continuous



Describing a Distribution – Summary of Moments

Measure	Formula	Description
Mean (μ)	$E(X)$	Measures the centre of the distribution of X
Variance (σ^2)	$E[(X-\mu)^2]$	Measures the spread of the distribution of X about the mean

Simplifying the Formula

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

$$= E[X^2] - 2\mu E[X] + \mu^2$$ (we get this from previous formula as m is just a number)

$$= E[X^2] - 2\mu^2 + \mu^2$$

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

Expectation Properties

$E(X+Y) = E(X) + E(Y)$ e.g., Playing a game each on 2 slot machines with different probabilities of winning. This is called Independent Observation.

$E(aX+b) = aE(X)+E(b) = aE(X) + b$ e.g., values x have been changed. This is called Linear Transformation.

If I have a portfolio of 30% TCS, 50% Wipro and 20% Ranbaxy stocks, the expected return of my portfolio is

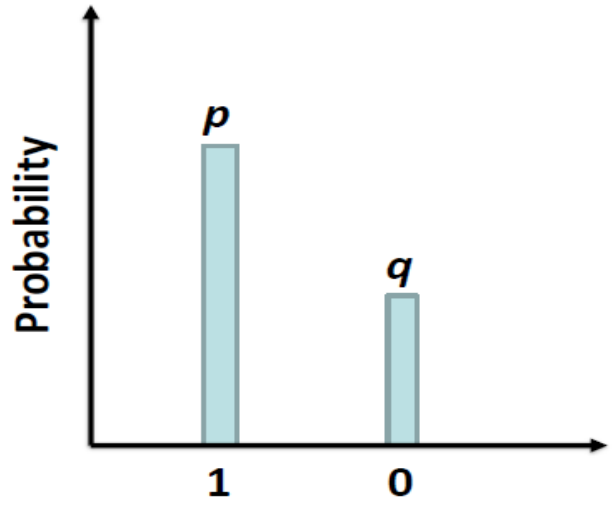
$E(\text{Portfolio}) = 0.3\ E(\text{TCS}) + 0.5\ E(\text{Wipro}) + 0.2\ E\ (\text{Ranbaxy})$

SOME COMMON DISTRIBUTIONS

Bernoulli

There are two possibilities (loan taker or non-taker) with probability *p* of success and *1-p* of failure

- Expectation: *p*
- Variance: *p(1-p)* or *pq*, where *q=1-p*



Geometric Distribution

Number of independent and identical Bernoulli trials needed to get ONE success, e.g., number of people I need to call for the first person to accept the loan.

Geometric Distribution

PMF*, $P(X = r) = q^{r-1}p$ — *(r-1)* failures followed by ONE success.

$P(X > r) = q^r$ — Probability you will need more than *r* trials to get the first success.

CDF**, $P(X \leq r) = 1 - q^r$ — Probability you will need *r* trials or less to get your first success.

$$E(X) = \frac{1}{p} \qquad Var(X) = \frac{q}{p^2}$$

* Probability Mass Function ** Cumulative Distribution Function

Geometric Distribution

- You run a series of independent trials.
- There can be either a success or a failure for each trial, and the probability of success is the same for each trial.
- The main thing you are interested in is how many trials are needed in order to get the first successful outcome.

Binomial Distribution

If there are two possibilities with probability *p* for success and *q* for failure, and if we perform *n* trials, the probability that we see *r* successes is

PMF, $\mathrm{P}(\mathrm{X} = \mathrm{r}) = C_r^n p^r q^{n-r}$

CDF, $\mathrm{P}(\mathrm{X} \leq \mathrm{r}) = \sum_{i=0}^{r} C_i^n p^i q^{n-i}$

Binomial Distribution

$E(X) = np$ $Var(X) = npq$

When to use?

- You run a series of independent trials.
- There can be either a success or a failure for each trial, and the probability of success is the same for each trial.
- There are a finite number of trials, and you are interested in the number of successes or failures.

Poisson Distribution

Probability of getting 15 customers requesting for loans in a given day given on average we see 10 customers

$\lambda = 10 \; and \; r = 15$

PMF, $\mathrm{P(X = r)} = \frac{e^{-\lambda}\lambda^{r}}{r!}$

CDF, $\mathrm{P(X \leq r)} = e^{-\lambda} \sum_{i=0}^{r} \frac{\lambda^{i}}{i!}$

Poisson Distribution

$E(X) = \lambda$ Can be equated to *np* of Binomial if *n* is large (>50) and *p* is small (<0.1)

$Var(X) = \lambda$ Can be equated to *npq* of Binomial in the above situation.

When to use?

- Individual events occur at random and independently in a given interval (time or space).
- You know the mean number of occurrences, λ, in the interval or the rate of occurrences, and it is finite.

Poisson Distribution

The probability that no customer will visit the store in one day

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

Probability that she will not have a customer for *n* days

$$e^{-n\lambda}$$

Exponential Distribution

Probability that a customer will visit in *n* days: $1 - e^{-n\lambda}$

$$CDF = 1 - e^{-n\lambda}, n \geq 0$$
$$PDF = \lambda e^{-n\lambda}, n \geq 0$$

Distributions

Geometric:	For estimating number of attempts before first success
Binomial:	For estimating number of successes in *n* attempts
Poisson:	For estimating *n* number of events in a given time period when on average we see *m* events
Exponential:	Time between events

Probability Distributions

Here are a few scenarios. Identify the distribution and calculate expectation, variance and the required probabilities.

Q1. A man is bowling. The probability of him knocking all the pins over is 0.3. If he has 10 shots, what is the probability he will knock all the pins over less than 3 times?

Q2. On average, 1 bus stops at a certain point every 15 minutes. What is the probability that no buses will turn up in a single 15 minute interval?

Q3. 20% of cereal packets contain a free toy. What is the probability you will need to open fewer than 4 cereal packets before finding your first toy?

Probability Distributions

Solutions

A man is bowling. The probability of him knocking all the pins over is 0.3. If he has 10 shots, what is the probability he will knock all the pins over less than 3 times?

X ~ B(10,0.3); n=10, p=0.3, q=1-0.3=0.7, r=0, 1, 2 (< 3)

E(X) = np = 3

Var(X) = npq = 2.1

$P(X = r) = {}^{n}C_{r}p^{r}q^{n-r}$

P(X=0) = 0.028; P(X=1) = 0.121; P(X=2) = 0.233

∴ P(X<3) = 0.028 + 0.121 + 0.233 = 0.382

Probability Distributions

Solutions

On average, 1 bus stops at a certain point every 15 minutes. What is the probability that no buses will turn up in a single 15 minute interval?

X ~ Po(1); λ=1, r=0

E(X) = λ = 1

Var(X) = λ = 1

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

P(X=0) = 0.368