Introduction to Probability Theory – Part 4 [1]

Erhard Hinrichs

Seminar für Sprachwissenschaft
Eberhard-Karls Universität Tübingen

[1] Largely based on material from Sharon Goldwater's tutorial *Basics of Probability Theory* (henceforth abbreviated as SGT), available at: `https://homepages.inf.ed.ac.uk/sgwater/math_tutorials.html`

Random variables and discrete distributions

- Definition of a random variable
- Examples of RVs
- Probability mass functions and cumulative distribution functions
- Restating the probability rules using random variables
- The Bernoulli distribution
- Other discrete distributions

Random Variable

- Definition: A RANDOM VARIABLE (or RV) is a variable that represents all the possible events in some partition of the sample space.
- More formally: A random variable is a function from events to probabilities.
- The distribution over an RV simply tells us the probability of each value.
- We use the notation $P(X)$ as a shorthand meaning "the entire distribution over X", in contrast to P(X = x), which means "the probability that X takes the value x".

What are RVs good for?

- We are often interested in some function of the possible outcomes, rather than the individual outcomes themselves.
  - For example: We may only be interested in the sum of the outcomes of throwing two dice, rather than in the outcomes of each die.
  - Such functions may be expressed as a single or more than one random variable.
- We are sometimes only interested in the probability distribution of the random variable as a whole, and not in the probabilities of individual events.

Example 1: Tossing two Fair Dice simultaneously

(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

Let S be a random variable with domain $2 \ldots 12$ and with:
$P\{S = 2\} = \frac{1}{36}$, $P\{S = 3\} = \frac{1}{18}$, $P\{S = 4\} = \frac{1}{12}$, $P\{S = 5\} = \frac{1}{9}$,$P\{S = 6\} = \frac{5}{36}$,$P\{S = 7\} = \frac{1}{6}$,$P\{S = 8\} = \frac{5}{36}$,$P\{S = 9\} = \frac{1}{9}$, $P\{S10\} = \frac{1}{12}$, $P\{S = 11\} = \frac{1}{18}$,$P\{S = 12\} = \frac{1}{36}$

Example 2: Random Selection of 3 Balls from an Urn with 20 Balls[2]

Three balls are to be randomly drawn without replacement from an urn numbered 1 through 20. What is the probability that at least one of the balls selected has a number as large or larger than 19?

Let X denote the largest number selected. Then X is a random variable taking on one of the values 3, 4, . . . , 20.

[2] Example taken from Sheldon Ross (2002). A First Course in Probability, 6th Edition. Prentice Hall

Notation and Terminology

$$\binom{n}{r} = \frac{n!}{(n-r)!r!} \qquad \textit{for} \quad r \leq n \qquad (1)$$

$\binom{n}{r}$ (called "n choose r") represents the number of possible combinations of n objects taken r at a time. The numbers $\binom{n}{r}$ are called *binomial co-efficients* and the probabilities $p(r)$ are called *binomial probabilities*.

Example:

$$\binom{20}{3} = \frac{20 \times 19 \times 18}{1 \times 2 \times 3} = 1140 \qquad (2)$$

Solution to Example 2

Under the assumption that the $\binom{20}{3}$ possible selections are equally likely to occur, then

$$P\{X = i\} = \frac{\binom{i-1}{2}}{\binom{20}{3}} \qquad i = 3, \ldots, 20 \tag{3}$$

$$P\{X = 20\} = \frac{\binom{19}{2}}{\binom{20}{3}} = \frac{\frac{19\times 18}{2}}{\frac{20\times 19\times 18}{3\times 2\times 1}} = \frac{3}{20} = .150 \tag{4}$$

$$P\{X = 19\} = \frac{\binom{18}{2}}{\binom{20}{3}} = \frac{51}{380} \approx .134 \tag{5}$$

Hence:

$$P\{X \geq 19\} \approx .134 + .150 = .224$$

Example 3: Wifi Users on a Train

There are five passengers on a train with wifi service. Define a random variable X that maps the number of people using the train's wifi service to the associated probability distribution, for which we make the following modelling assumptions:

2/5 people always use it.
1/5 never uses it.
2/5 use it every other day at random and independently.

Calculate the following probabilities:

P{X=0}, P{X=1}, P{X=2}, P{X=3}, P{X=4}, P{X=5}

Example 3 – Solution: Wifi Users

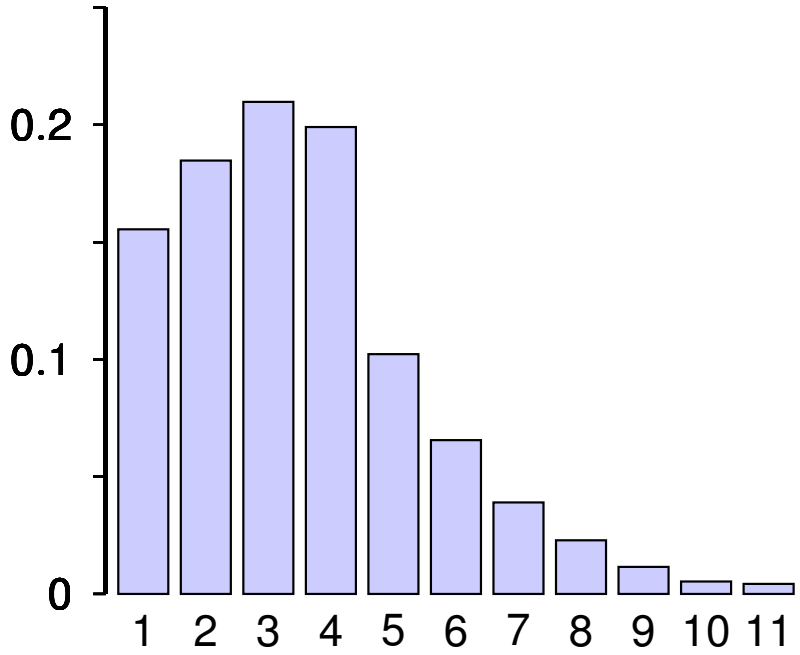
From these assumptions, it follows that:

$P\{X=0\} = 0$	$P\{X=1\} = 0$
$P\{X=2\} = 1/4$	$P\{X=3\} = 1/2$
$P\{X=4\} = 1/4$	$P\{X=5\} = 0$

Example due to Prof. Iain Collins, Macquarie University; see
`https://www.youtube.com/watch?v=MM6QM3y8pvI`

Probability mass function[3]

- *Probability mass function (PMF)* of a *discrete* random variable (X) maps every possible (x) value to its probability ($P(X = x)$).

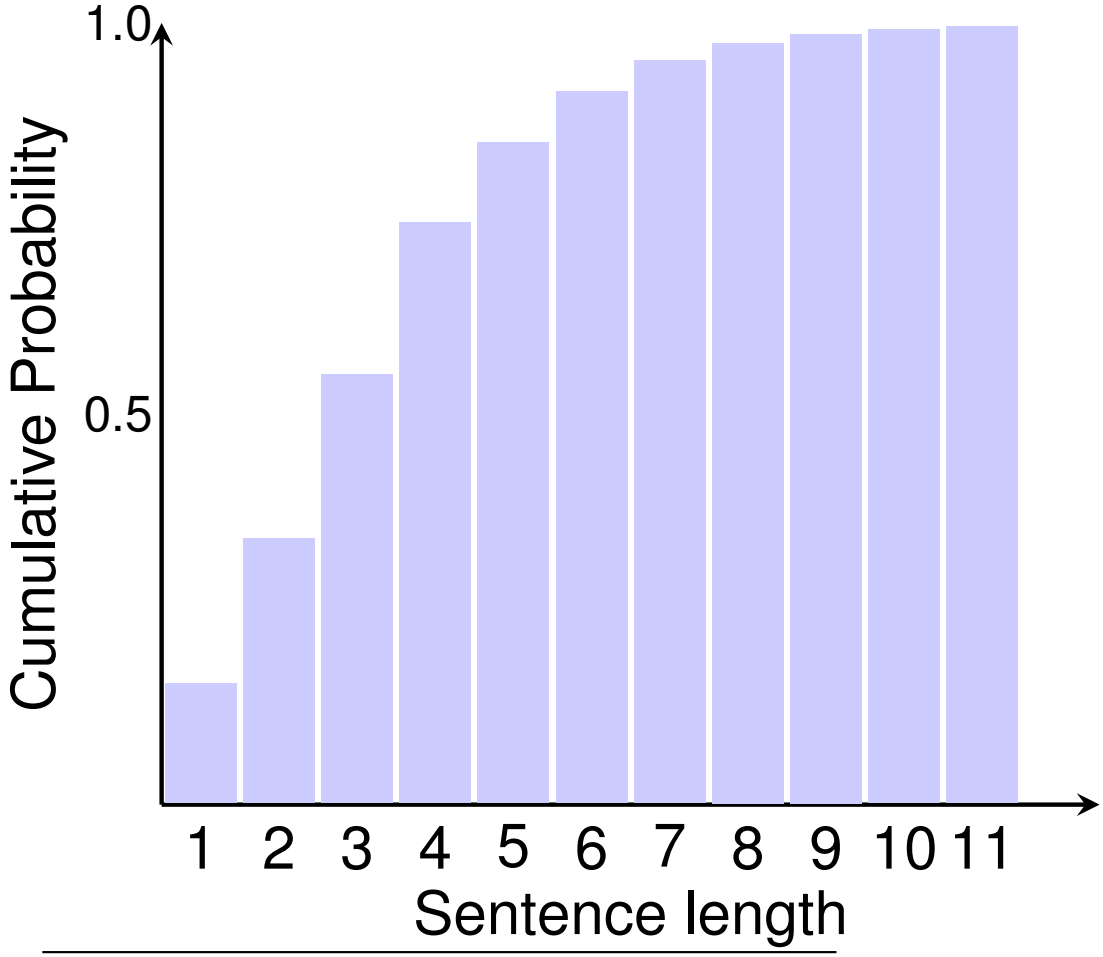


x	$P(X = x)$
1	0.155
2	0.185
3	0.210
4	0.194
5	0.102
6	0.066
7	0.039
8	0.023
9	0.012
10	0.005
11	0.004

[3] slide due to Cagri Cöltekin

Cumulative distribution function[4]

- $F_X(x) = P(X \leq x)$



Length	Prob.	C. Prob.
1	0.16	0.16
2	0.18	0.34
3	0.21	0.55
4	0.19	0.74
5	0.10	0.85
6	0.07	0.91
7	0.04	0.95
8	0.02	0.97
9	0.01	0.99
10	0.01	0.99
11	0.00	1.00

[4] slide due to Cagri Cöltekin

Bernoulli RVs and Distributions; Categorical RCs and Distributions

A random variable X with probability mass function

$$p(0) = P\{X = 0\} = 1 - p$$
$$p(1) = P\{X = 1\} = p$$

is called a **Bernoulli random variable**.
If an experiment has exactly two outcomes: 1 ("success") with probability p and 0 ("failure") with probability $1 - p$, then this discrete probability distribution is called a **Bernoulli distribution**.
The **categorical distribution** is a generalization of the Bernoulli distribution for a **categorical random variable**, i.e. for a discrete variable with more than two possible outcomes.

The Binomial Distribution

If we have a Bernoulli variable X and we repeatedly sample values from $P(X)$ (e.g., flip a coin many times), we will get some number of 1's (heads) and some number of 0's (tails). The binomial distribution with parameters p and n describes the probability of getting a certain number of 1's out of a total of n samples when the probability of a 1 on a single sample is p.

A random variable whose distribution follows the following formula is said to have a BINOMIAL DISTRIBUTION:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \qquad k = 0, 1, \ldots, n \qquad (6)$$

$$with : \binom{n}{k} = \frac{n!}{(n-k)! \cdot k!}$$

The Multinomial Distribution

This distribution is an extension of the binomial distribution for random variables with more than two possible values.

A random variable whose distribution follows the following formula is said to have a MULTINOMIAL DISTRIBUTION:

$$P(X_1 = x_1, X_2 = x_2 \ldots X_k = x_k) = \begin{cases} \frac{n!}{x_1!\ldots x_k!} p_1^{x_1} \times \cdots \times p_k^{x_k} & \text{when } \sum_{n=i}^{k} x_i = n \\ \text{otherwise} \quad 0 \end{cases} \tag{7}$$

with random variables $X_1, \ldots X_k$ and parameters $p_1, \ldots p_k$ (probabilities) and n (trials)

Example

A document contains 8 commas, 3 colons, and 5 exclamation marks. 6 punctuation marks are randomly chosen with replacement. What is the probability that 3 are commas, 1 as a colon, and 2 are exclamation marks?

$$P(X_1 = 3, X_2 = 1, X_3 = 2) = \frac{6!}{3!1!2!}\left(\frac{8}{16}\right)^3\left(\frac{3}{16}\right)^1\left(\frac{5}{16}\right)^2 \tag{8}$$

$$\approx 7.629395e-06 \tag{9}$$

The Geometric Distribution

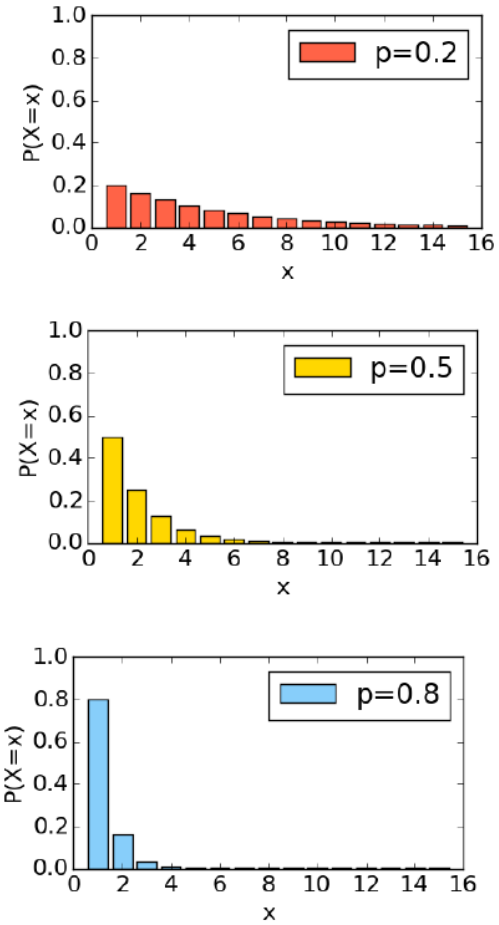
Example 5.2.2.
A random variable whose distribution follows the following formula is said to have a GEOMETRIC DISTRIBUTION:

$$P(X = n) = (1 - p)^{n-1} p \qquad \text{for positive integer } n \qquad (10)$$

The geometric distribution characterizes any situation in which:

- We repeat some random experiment until we succeed, and we are only interested in how long it will take to succeed.
- Each attempt is independent and has the same probability of success: the attempts are INDEPENDENT AND IDENTICALLY DISTRIBUTED or IID.

Geometric Distributions with Different Parameter Values



The Poisson Distribution

A random variable whose distribution follows the following formula is said to have a POISSON DISTRIBUTION:

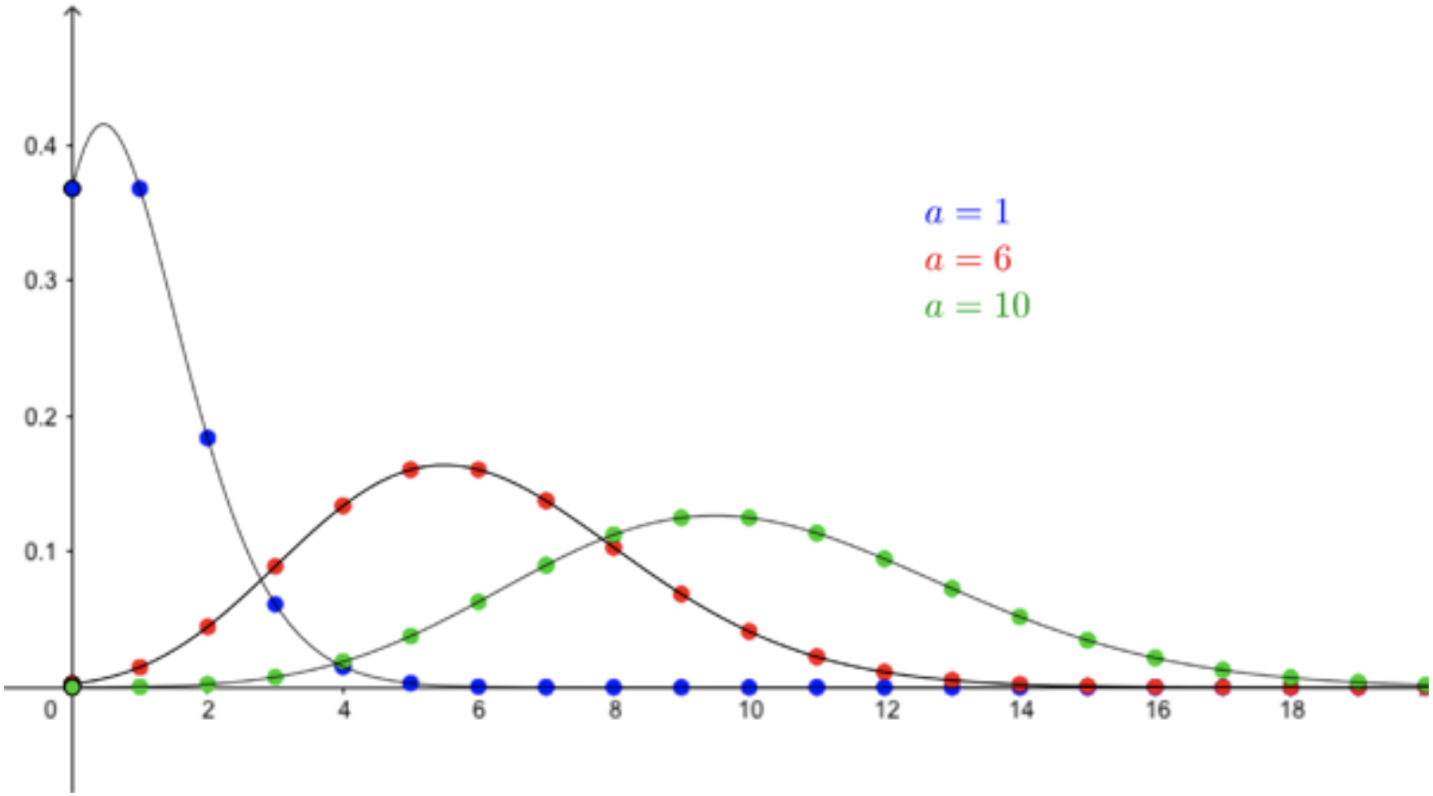
$$f(k;\lambda) = Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{11}$$

where:

- k is the number of occurrences ($k = 0, 1, 2 \ldots$)
- e is Euler's number 2.71828, the base of the natural logarithm: $\ln e^x = x$
- the parameter λ is the mean value of the random variable X

Leonhard Euler

Poisson PMF with different parameter values[5]



The Poisson distribution for *a*=1 (blue), *a*=6 (red), and *a*=10 (green).

[5]Graphics taken from
`https://plus.maths.org/content/poisson-distribution`

Examples - due to Ugarte et al. (2016)

The average number of goals in a World Cup soccer match is approximately 2.5 and the Poisson model is appropriate for this scenario, with λ= 2.5 as the average event rate per match.

P (k goals in a match) = $\frac{2.5^k e^{-2.5}}{k!}$

P (k = 0 goals in a match) = $\frac{2.5^0 e^{-2.5}}{0!} = \frac{e^{-2.5}}{1} \approx 0.082$

P (k = 1 goal in a match) = $\frac{2.5^1 e^{-2.5}}{1!} = \frac{2.5 e^{-2.5}}{1} \approx 0.205$

P (k = 2 goal goals in a match) = $\frac{2.5^2 e^{-2.5}}{2!} = \frac{6.25 e^{-2.5}}{2} \approx 0.257$

Examples - taken from Wikipedia

On a particular river, overflow floods occur once every 100 years on average. Calculate the probability of $k = 0, 1, 2, 3, 4, 5$, or 6 overflow floods in a 100-year interval, assuming the Poisson model is appropriate. Because the average event rate is one overflow flood per 100 years, $\lambda = 1$.

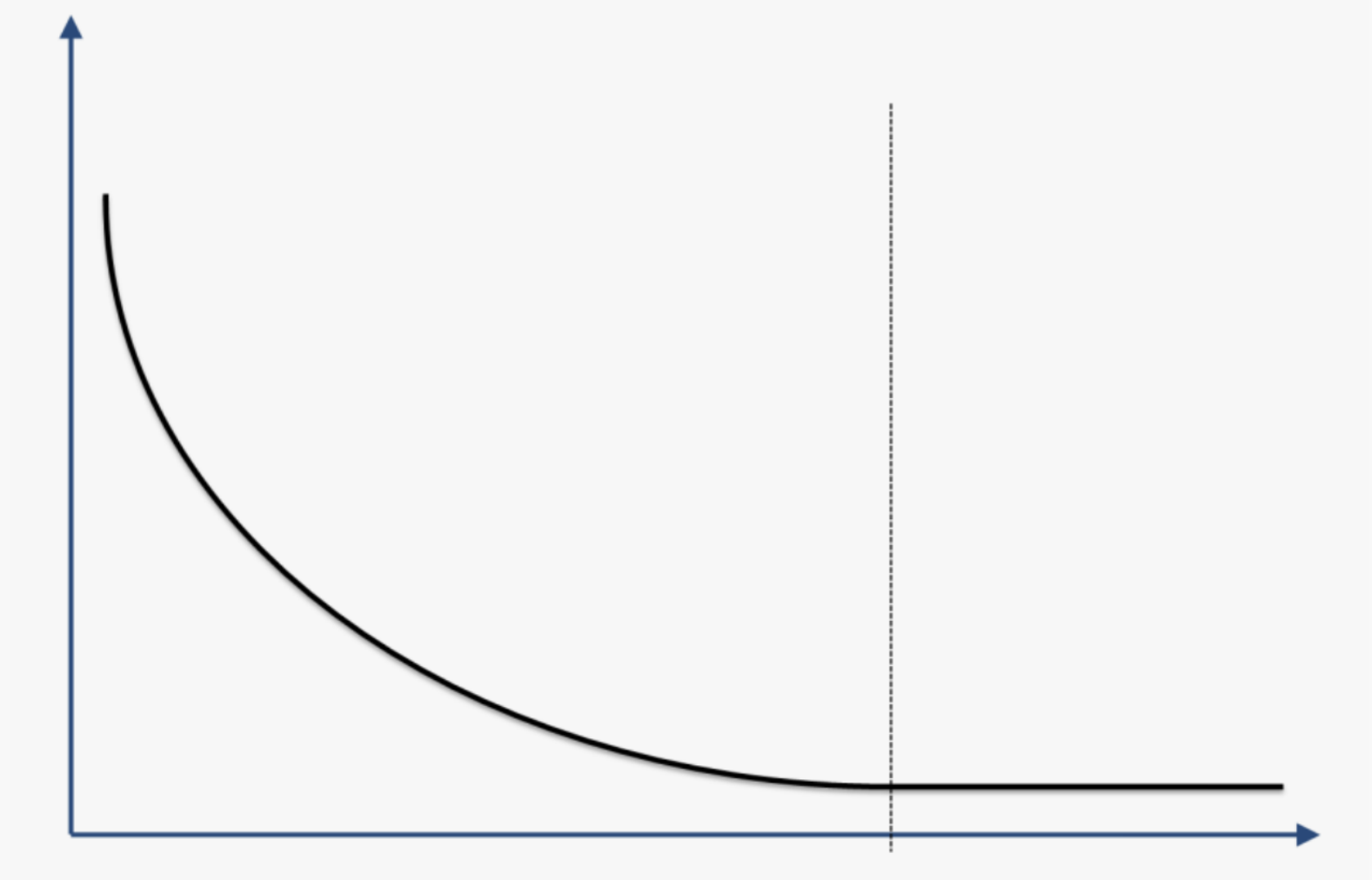
P (k overflow floods in 100 years) = $\frac{\lambda^k e^{-\lambda}}{k!} = \frac{1^k e^{-1}}{k!}$

P ($k = 0$ overflow floods in 100 years) = $\frac{1^0 e^{-1}}{0!} = \frac{e^{-1}}{1} \approx 0.368$

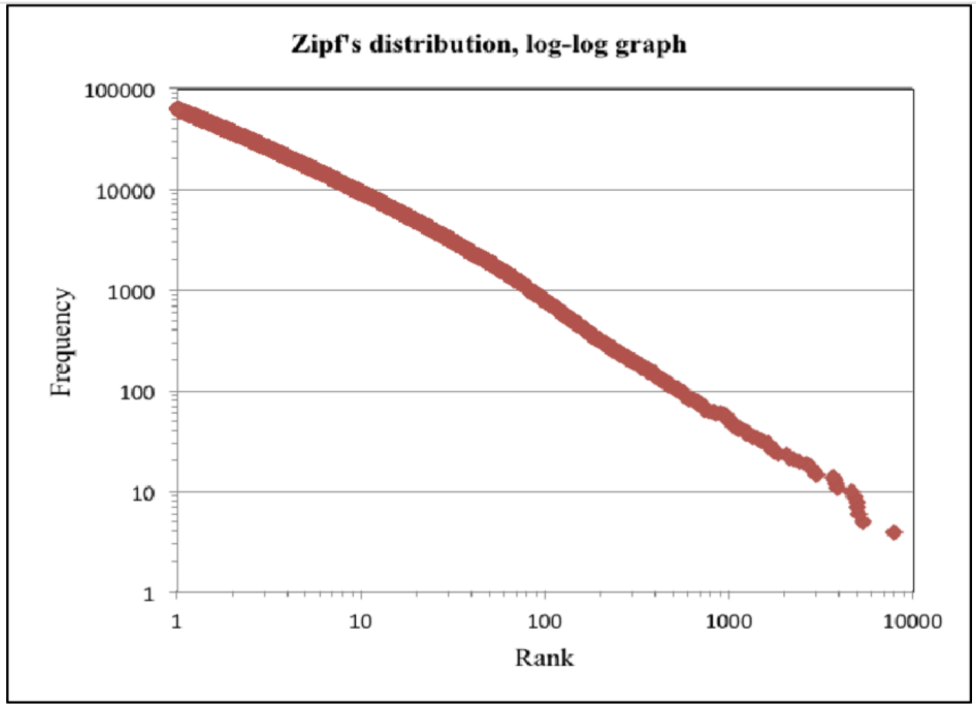
P ($k = 1$ overflow flood in 100 years) = $\frac{1^1 e^{-1}}{1!} = \frac{e^{-1}}{1} \approx 0.368$

P ($k = 2$ overflow floods in 100 years) = $\frac{1^2 e^{-1}}{2!} = \frac{e^{-1}}{2} \approx 0.184$

Power Law Graph



Visualizing Zipf's Law in log-log space



Zipf distribution log-log graph for types across the web-based corpus (Sharoff, 2006)

Word frequency and ranking in a log log graph follow a nice straight line, as is typical of a power-law.

Zipf's Law: General Formulation

$f(k; s, N)$, where:

N	is the number of elements
k	their rank
s	the value of the exponent characterizing the distribution (at least 1).

More specifically:

$$f(k; s, N) = \frac{\frac{1}{k^s}}{\sum_{n=1}^{N} \frac{1}{n^s}} = \frac{1}{\sum_{n=1}^{N} \frac{1}{n^s}} \times \frac{1}{k^s}$$

Remark: The term $\sum_{n=1}^{k} \frac{1}{n}$ is a generalized harmonic number $H_{N,s}$

Harmonic Numbers

The n-th harmonic number is the sum of the reciprocals of the first n natural numbers.

$H_n = 1 + \frac{1}{2} + \frac{1}{3} \cdots \frac{1}{n} = \sum_{n=1}^{k} \frac{1}{n}$

Starting with $n = 1$, the sequence of harmonic numbers begins as follows:

$1, \frac{3}{2}, \frac{11}{6}, \frac{25}{12}, \frac{137}{60}, \ldots$

Harmonic Numbers and Natural Logarithm

n	H_n	$ln(n)$	fraction
5	2.28333	1.60943	137/60
50	4.49921	3.91202	11703651829592112/2601270879967823
500	6.79282	6.21460	
5,000	9.09451	8.51719	
50,000	11.397	10.81977	
500,000	13.69958	13.12236	

Discrete Random Variable

Assumptions for a discrete random variable:

- X is a random variable with enumerably many possible values.
- This means that it is possible to create a one-to-one mapping between these values and the integers (think of it as assigning a unique integer to each x_i value of X).

Restating the Requirements for a Discrete Probability Distribution or Probability Mass Function

$$0 \leq P(X = x_i) \leq 1 \tag{16}$$

$$\sum_{i=1}^{n} P(X = x_i) = 1 \tag{17}$$

Example: Re-write the JPT from Example 4.4.2 using random variable notation

Table 4.4.2:

	R	G	B	
S	1/3	1/5	2/15	2/3
¬ S	2/15	2/15	1/15	1/3
	7/15	1/3	1/5	

Re-write in RV notation

	X = r	X = g	X = b	
Y = solid	1/3	1/5	2/15	2/3
Y = patchy	2/15	2/15	1/15	1/3
	7/15	1/3	1/5	

Restating the Laws of Probabilities in terms of Random Variables

verbose	concise	
$P(X = x) = 1 - P(X \neq x)$	$P(x) = 1 - P(\neg x)$	complement
$P(X = x) = \sum_y P(X = x, Y = y)$	$P(x) = \sum_y P(x, y)$	law of total prob
$P(X = x \mid Y = y) = \frac{P(X=x, Y=y)}{P(Y=y)}$	$P(x \mid y) = \frac{P(x,y)}{P(y)}$	defn of cond prob
$P(X = x, Y = y) = P(X = x \mid Y = y)P(Y = y)$	$P(x, y) = P(x \mid y)P(y)$	product rule
$P(X = x \mid Y = y) = \frac{P(Y=y \mid X=x)P(X=x)}{P(Y=y)}$	$P(x \mid y) = \frac{P(y \mid x)P(x)}{P(y)}$	Bayes' Rule

Conditional Probability Distribution

- We can refer to a (marginal) distribution $P(X)$ and a joint distribution (X, Y)
- But it is incorrect to write $P(X \mid Y)$ because a conditional distribution has to be conditioned on a particular event (and not a random variable).
- Valid ways to refer to a conditional distribution include $P(X|Y = y)$ or $P(X|y)$

Independent Random Variables

Two random variables X and Y are independent iff for *all* values x and y:

$$P(X = x, Y = y) = P(X = x) \times P(Y = Y) \tag{23}$$

We can say that the *events* X = x and Y = y are independent iff Eq (23) is true for those particular events, regardless of whether other outcomes of X and Y obey Eq. (23).

Law of Large Numbers[6]

For a random variable X with expected population mean $E(X)$, we define a random variable $\overline{E(X_n)}$ for the sample mean of n observations:

$$\overline{E(X_n)} = \frac{x_1 + \cdots + x_n}{n} \tag{24}$$

The Law of Large Numbers tell us that

$$\overline{E(X_n)} \to E(X_n) \text{ for } \; n \to \infty \tag{25}$$

Please note the difference between expected population mean $E(X)$ and sample mean $\overline{E(X_n)}$ and please beware of the gambler's paradox!

Example: If X is the number of heads after 100 tosses of a fair coin, then the population mean of $E(X) = 100 \times .5 = 50$

[6] Presentation based on `https://www.khanacademy.org/math/statistics-probability/random-variables-stats-library/expected-value-lib/v/law-of-large-numbers`

Expectation and variance

Expected Value (Mean) of a random variable X

$$E[X] = \mu = \sum_x x \cdot P(X = x) \tag{26}$$

Expectation of a function of a random variable

$$E[f(X)] = \sum_x f(x) \cdot P(X = x) \tag{27}$$

Variance and Standard Deviation

Variance

$$Var[X] = \sigma^2 = E[(X - E[X])^2] \tag{28}$$

$$Var[X] = \sigma^2 = E[X^2] - (E[X])^2 \tag{29}$$

Standard Deviation

$$SD[X] = \sigma = \sqrt{Var[X]} \tag{30}$$

Example 6.1.3; SGT, p. 29

I offer you to let you roll a single die, and will give you a number of pounds equal to the square of the number that comes up. How much would you expect to win by playing this game?

$$\begin{aligned} E[X^2] &= (1/6)(1^2) + (1/6)(2^2) + (1/6)(3^2) \\ &\quad + (1/6)(4^2) + (1/6)(5^2) + (1/6)(6^2) \\ &= 91/6 \approx 15.17 \end{aligned}$$

Another Instructive Example[7]

Let X denote a random variable that takes on any of the values -1,0,1 with respective probabilities

$$P\{X = -1\} = .2 \qquad P\{X = 0\} = .5 \qquad P\{X = 1\} = .3$$

Compute $E[X^2]$

[7] Example and solution taken from Sheldon Ross (2002). A First Course in Probability, 6th Edition. Prentice Hall

Solution

Letting $Y = X^2$, it follows that the probability mass function of Y is given by:

$$P\{Y = 1\} = P\{X = -1\} + P\{X = 1\} = .5 \tag{31}$$

$$P\{Y = 0\} = P\{X = 0\} = .5 \tag{32}$$

Hence: $E[X^2] = E[Y] = 1(.5) + 0(.5) = .5$

Readers should note that

.5 = $E[X^2] \neq (E[X])^2 = .01$!!

Proof of Equivalence

$$
\begin{aligned}
Var[X] &= E[(X - E[X])^2] \\
Var[X] &= E[(X - \mu)^2] \\
&= \sum_x (x - \mu)^2 p(x) \\
&= \sum_x (x^2 - 2\mu x + \mu^2) p(x) \\
&= \sum_x x^2 p(x) - 2\mu \sum_x x p(x) + \mu^2 \sum_x p(x) \\
&= E[X^2] - 2\mu^2 + \mu^2 \\
&= E[X^2] - \mu^2 \\
&= E[X^2] - (E[X])^2
\end{aligned}
$$

Measures of central tendency: mode and mean [8]

- The *mode* of a frequency distribution is the midpoint or class name of the most frequent measurement class.
  - If a case were drawn at random from the distribution, that case is more likely to fall in the *modal class*.
  - In the graph of any distribution, the modal class shows the highest "peak" or "bump" of the intervals presented.
- *mean* (aka: average, expectation)

[8] see William L. Hays (1994). *Statistics*. Fifth Edition. Harcourt Brace College Publishers, p. 165

Measures of central tendency: median [9]

- The *median* is the score corresponding to the middle class when all individual cases are arranged in order of scores.
  - The median reflects the score that divides the cases into two intervals having equal frequency or probability.
  - Thus, if you drew a case at random from any set of *N* observations and guessed that this score shows the median score, you are just as likely to be guessing too high or too low.

[9] see William L. Hays (1994). *Statistics*. Fifth Edition. Harcourt Brace College Publishers, p. 165

Median in a Grouped Frequency Distribution

- If the data have been arranged into a grouped frequency distribution, the median is defined as the point at or below which exactly 50% of the cases fall.
- The first step in finding the median of a grouped distribution is to construct the cumulative frequency distribution. Such a cumulative frequency distribution is illustrated in Table 1.1.

Median: Example for a Cumulative Frequency Distribution (Table 1.1.)

Class	f	cf
74 - 78	10	200
69 - 73	18	190
64 - 68	16	172
59 - 63	16	156
54 - 58	11	140
49 - 53	27	129
44 - 48	17	102
39 - 43	49	85
34 - 38	22	36
29 - 33	6	14
24 - 28	8	8
	200	

Where to place the Median

- The last column in Table 1.1 shows the cumulative frequencies for the class intervals. Because by definition, the median will be that point in the distribution at or below which 50% of the cases fall, the cumulative frequency at the median score should be .50N. Thus, the cumulative frequency is .50(200) = 100 at the median for this example.
- The median could not fall in any interval below the limit of 43.5, because 85 cases fall at or below that point.
- However the median does fall below the real limit 48.5, because 102 cases fall at or below that point. We have thus located the median as being in the class interval with real limits 43.5 and 48.5.

Median in a Cumulative Frequency Distribution

$$\textbf{median} = \text{ lower real limit } + i\left(\frac{.50N - cf \text{ below lower limit}}{f \text{ in interval}}\right) \tag{33}$$

where the *lower real limit* belongs to the interval containing the median, and the *cf* refers to the cumulative frequency up to the lower limit of the interval, *f* to the frequency of the interval containing the median, and *i* is the class interval size.

Median for Data in Table 1.1

$$\text{median} = 43.5 + 5\left[\frac{.50(200) - 85}{17}\right] \approx 47.9 \tag{34}$$

Median in a Discrete Probability Distribution

- The notions of mean, median and mode also apply to distributions of discrete random variables.
- They summarize the probability distributions of the RVs.
- They also may serve as constants or "parameters" entering into the mathematical rule giving the probability for any value or interval of values of the random variable.

Median: Example of a Discrete Probability Distribution (Table 1.2)

Class Interval	x	p(x in interval)	xp(x in interval)
74 - 78	76	.050	3.80
69 - 73	71	.090	6.39
64 - 68	66	.080	5.28
59 - 63	61	.080	4.88
54 - 58	56	.055	3.08
49 - 53	51	.135	6.89
44 - 48	46	.085	3.91
39 - 43	41	.245	10.05
34 - 38	36	.110	3.96
29 - 33	31	.030	.93
24 - 28	26	.040	1.04
			50.21

Computing the Median in a Discrete Probability Distribution

$$\textbf{median} = \text{ lower real limit } + i\left\{\frac{.50 - p(X \leq \text{ lower real limit}}{p(\text{ lower limit} \leq X \leq \text{ upper limit})}\right\} \tag{35}$$

where the *lower limit* and *upper limit* belong to the interval containing the median.

Median for Data in Table 1.2

$$\text{median} = 43.5 + \frac{5(.50 - .425)}{.085} \approx 47.91 (\text{where } \; p(x \leq 43.5) = .425) \tag{36}$$

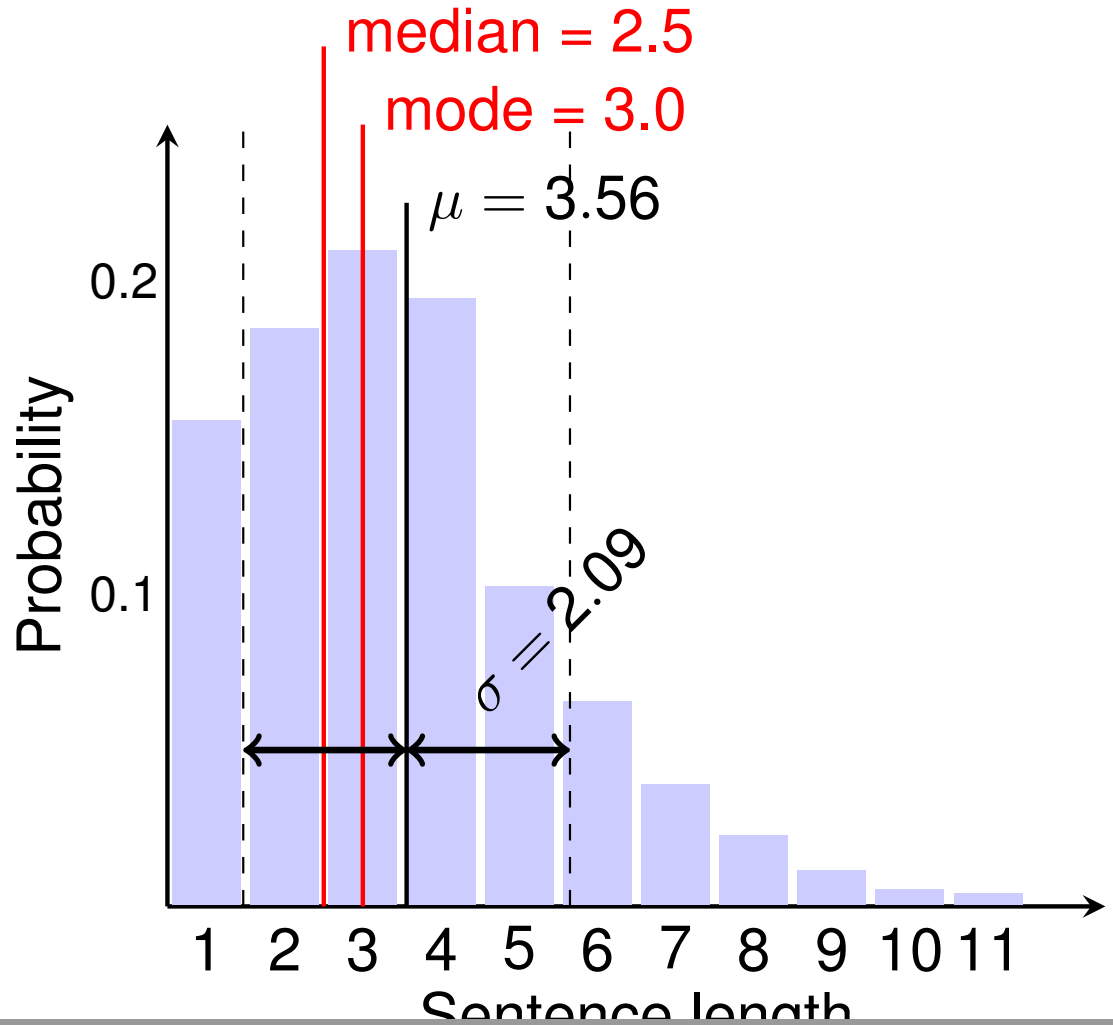
The Mean of a Probability Distribution as the Expectation of a Random Variable

$$E(X) = \sum_x = xp(x) = \text{ mean of } X \tag{37}$$

Example: Mean for Data in Table 1.2

$$\textbf{mean} = \sum_x = xp(x \text{ in interval}) = 50.21 \tag{38}$$

Measures of central tendency: mode, median, mean



Comparing Mean and Median

- The mean is used for normal distributions.
- The mean is not applicable to all distributions because it is highly sensitive to data outliers; i.e. values in the sample that are too small to too large.
- The median is more suitable for skewed distributions since it is more robust with respect to data outliers.