RANDOM VARIABLES AND PROBABILITY DISTRIBUTIONS

Lecture 1C

See Book Chapter/Sections 4.1 and 4.2

Probabilistic/Statistical View of Data

- As we saw earlier, we can view the data in Table 1 as a collection of column *random* variables using the following mapping:
  - {Age. Gender, Education, Credit Score, Income, Net Worth, Sales} → {X_1, X_2, X_3, X_4, X_5, X_6, X_7}.
- This permits us to work on understanding the marginal, joint and conditional distribution properties of the variables such as means, standard deviations, correlations etc.
- **Note:** *For the examples in this section, we are going to view this table as all of the data, i.e., as a population*

Table 1

ID	Name	Age	Gender	Education	Credit Score	Income	Net Worth	Sales
001	Adams, John	36	M	HS	350	38,900	65,924	1,535
002	Ramesh, Jyoti	23	F	Bachelors	600	172,000	178,154	2,196
003	Mendez, Nick	67	M	Bachelors	700	218,000	265,209	1,287
004	Mendez, Joan	38	F	PhD	550	182,000	85,277	2,143
005	Ritter, Jake	24	M	Masters	625	434,000	193,760	707
006	Rao, Eric	61	M	PhD	770	82,000	314,953	2,170
007	Blake, Ann	26	F	HS	490	112,000	192,946	1,229
008	Bishop, Marge	44	F	Masters	540	242,000	339,705	520
009	Ahmed, Mo	31	M	Masters	680	111,000	185,767	2,326
010	Shultz, Dante	44	M	Bachelors	280	66,000	97,778	588

Random Variables – Converting Events to Random Variables

- Working with events and their probabilities is difficult in a lot of situations.
- One solution is to associate algebraic variables with outcomes and events and obtain probabilities for values of variables. This gives us *random variables* that are easier to work with algebraically
- A random variable *assigns a numerical value to an experimental outcome*.
- Example:
  - Let X represent the number on the face of the red die, Y the number on the face of the yellow die.

	1	2	3	4	5	6
1	1,1	1,2	1,3	1,4	1,5	1,6
2	2,1	2,2	2,3	2,4	2,5	2,6
3	3,1	3,2	3,3	3,4	3,5	3,6
4	4,1	4,2	4,3	4,4	4,5	4,6
5	5,1	5,2	5,3	5,4	5,5	5,6
6	6,1	6,2	6,3	6,4	6,5	6,6

	1	2	3	4	5	6
1	1/36	1/36	1/36	1/36	1/36	1/36
2	1/36	1/36	1/36	1/36	1/36	1/36
3	1/36	1/36	1/36	1/36	1/36	1/36
4	1/36	1/36	1/36	1/36	1/36	1/36
5	1/36	1/36	1/36	1/36	1/36	1/36
6	1/36	1/36	1/36	1/36	1/36	1/36

  - Then *Outcome* (3, 4) becomes (X=3, Y=4), for example;
  - Then, the probability P(*3 on the red die AND 4 on the yellow die*) is P(X=3, Y=4) = 1/36
  - The axioms of probability will also hold for random variables.

Random Variables – Converting Events to Random Variables

- We can create many random variables depending on the underlying events.

	1	2	3	4	5	6
1	1/36	1/36	1/36	1/36	1/36	1/36
2	1/36	1/36	1/36	1/36	1/36	1/36
3	1/36	1/36	1/36	1/36	1/36	1/36
4	1/36	1/36	1/36	1/36	1/36	1/36
5	1/36	1/36	1/36	1/36	1/36	1/36
6	1/36	1/36	1/36	1/36	1/36	1/36

- Example:
  - Create a new random variable and obtain its probabilities from the outcomes of the toss of two dice.
  - Let W be the random variable that represents the event "*sum of the faces of the two dice*". Then, *Event* (3, 4) = (X=3, Y=4) becomes W=7.

  - Develop a table showing the random variable and its probabilities (i.e., the probability distribution):

W	2	3	4	5	6	7	8	9	10	11	12
P(W)	1/36	2/36	3/26	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

Random Variables

- One of the advantages of random variables is that sometimes we can get *closed-form formulas* such that when we plug in the value of the random variable, we get back probabilities!
- This **avoids the need to develop tables for the random variable values and its probabilities**.
- Let us take as a simple case, the probability that we will cast the **number 3** five times, when we cast one die 5 times.
- Let event **Success** =“number 3 on the die” and “**Failure**” = otherwise (i.e., the die is 1 or 2 or 4 or 5 or 6)
- In a single toss, the probability of Success = p = 1/6 and the probability of Failure = 5/6.
- Let the random variable X be the *number of “successes”* in n tosses of the die.
- The probability of getting X=5 times from (n=) 5 tosses or trials is given by the *closed-form formula*:
  - $\frac{n!}{x!n-x!}(p)^x(1-p)^{n-x}$ = $\frac{5!}{5!0!}(\frac{1}{6})^5(\frac{5}{6})^0$ = 0.000129
- One can immediately see the advantage of casting event and outcomes as random variables. Closed-form algebraic functions and operations replace set theoretic operations on exhaustive listing of outcomes.

Types of random variables

- We can divide random variables broadly into two types:
  - **Discrete** random variables have probabilities defined for discrete numerical values of the random variable
  - **Continuous** random variables have probabilities defined on ranges of values of the random variable

- Examples:
  - Gender can be made a (discrete) random variable by recasting it as: M = 0 and F = 1.
  - Income is a continuous random variable, though probabilities are only defined for ranges of its values

Table 1

ID	Name	Age	Gender	Education	Credit Score	Income	Net Worth	Sales
001	Adams, John	36	M	HS	350	38,900	65,924	1,535
002	Ramesh, Jyoti	23	F	Bachelors	600	172,000	178,154	2,196
003	Mendez, Nick	67	M	Bachelors	700	218,000	265,209	1,287
004	Mendez, Joan	38	F	PhD	550	182,000	85,277	2,143
005	Ritter, Jake	24	M	Masters	625	434,000	193,760	707
006	Rao, Eric	61	M	PhD	770	82,000	314,953	2,170
007	Blake, Ann	26	F	HS	490	112,000	192,946	1,229
008	Bishop, Marge	44	F	Masters	540	242,000	339,705	520
009	Ahmed, Mo	31	M	Masters	680	111,000	185,767	2,326
010	Shultz, Dante	44	M	Bachelors	280	66,000	97,778	588

Probability Distributions

- A **probability distribution** is a mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment.

- A discrete random variable is said to have a *probability mass function* (pmf) that supplies probabilities for *each value* of the random variable. The probabilities are non-negative and always sum up to 1 for the sample space (or for all values of the random variable).

  - pmf for X_1 (Gender): $p^{x1}(1-p)^{(1-x1)}$, where p is the probability of a Female and (1-p) is the probability of a male.
  - The pmf directly gives you the probabilities based on the values of the random variable. i.e., it is often expressed as a function of the random variable.

- A continuous random variable is said to have a *probability density function (pdf)* that (when integrated) supplies probabilities for *ranges of values* of the random variable. The probabilities are non-negative and the probability is 1 when the *pdf* is integrated over the entire range of the random variable.

  - Pdf for X_5 (Income): $1/\sqrt{2\pi\sigma^2}\, e^{-(x-\mu)^2/2\sigma^2}$
  - To find the Probability (X_5 <= 0.25), you will *integrate the above function* from –∞ to 0.25.

General Discrete Probability Distributions

	1	2	3	4	5	6
1	1/36	1/36	1/36	1/36	1/36	1/36
2	1/36	1/36	1/36	1/36	1/36	1/36
3	1/36	1/36	1/36	1/36	1/36	1/36
4	1/36	1/36	1/36	1/36	1/36	1/36
5	1/36	1/36	1/36	1/36	1/36	1/36
6	1/36	1/36	1/36	1/36	1/36	1/36

- Let **W** be the random variable representing the **sum of the faces on the two dice**.
- The *probability distribution* for this discrete random variable **W** is:

W =	2	3	4	5	6	7	8	9	10	11	12	Total
P	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36	36/36

- Note that **we do not have a formula for this general discrete distribution.** We had to explicitly enumerate X and the probability. In the next lecture we will obtain probabilities from a probability mass function (pmf) with known formulas for different distributions.

Practice with calculator

Cumulative Probability Distribution

W =	2	3	4	5	6	7	8	9	10	11	12	Total
P	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36	36/36

- The *cumulative probability distribution* gives the probability than a random variable is *less than or equal to* a certain quantity

- In the case of W (the sum on the faces if two dice), the *cumulative probabilities* are shown in the table below:

W ≤	2	3	4	5	6	7	8	9	10	11	12	Total
P	1/36	3/36	6/36	10/36	15/36	21/36	26/36	30/36	33/36	35/36	36/36	36/36

- In this case, for example, 4 corresponds to the 6/36 or the 16.67th percentile. That is, $P(W \leq 4) = 0.1667$

- Similarly, $P(W \leq 7) = 21/36 = 7/12$ which means that 7 is the 58.33th percentile.

Practice with calculator

Mean (Expected Value) and Measures of Central Tendency

- The two most commonly used properties of a distribution are its mean or expected value and its variance.

- *Expected value of W*:
  - is the long-run average value of W, as we repeat the experiment indefinitely.
  - It is **not** "the value we expect to occur". In fact, the expected value may never occur or not even exist.
    - The expected value of the roll of a *single die* is actually 21/6 or 7/2 or 3.5, which is never observed even though this will be the average numbers from tossing the die repeatedly for a long time.

- Expected Value of a discrete random variable;
  - $E(W) = \sum_{c \in A} cP(W = c)$. This also happens to be the *average* of the values of the random variable, but may not always be true. The Expected Value is the correct definition of the mean.

- The symbol for expected value for a random variable is **μ** and is also known as the *mean*.

- Mean value of W (sum of faces on two dice) = **7**

W	2	3	4	5	6	7	8	9	10	11	12	Total
P	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36	36/36
X*P	2/36	6/36	12/36	20/36	30/36	42/36	40/36	36/36	30/36	22/36	12/36	7

Practice with calculator

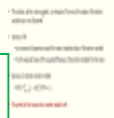
Mean (Expected Value) and Measures of Central Tendency

- The *mean* is an example of a measure of **central tendency**.
- The unit of the mean is the same as that of the random variable
- Other measures of central tendency include
  - the *mode* (the most common value in the distribution) and
  - the *median* (the observation which corresponds to a cumulative probability of 0.5 (or the 50^{th} percentile).

W	2	3	4	5	6	7	8	9	10	11	12	Total
P	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36	36/36

- The **mode** of the random variable W = sum of faces of two dice is **7**, because 7 occurs the most frequently (6/36)
- The **median** of the random variable W = sum of faces of two dice is **7**, because the probability $W \leq 7$ = 21/36 = probability $W \geq 7$.
  - **Note**: Sometimes you will find the median defined as the “middle value” when the observations are ranked. That applies to situations where all the outcomes are *equally likely*. (such as in a random sample). In this case, both values coincide.

Practice with calculator



Variance, Standard Deviation and Measures of Dispersion

- The variance, as the name suggests, is a measure of how much the values of the random variable vary or are "dispersed".
- *Variance of W:*
  - is a measure of dispersion around the mean or expected value of the random variable.
  - It is the expected value of the squared difference of the random variable from the mean
- Variance of a discrete random variable;
  - $V(W) = \sum_{c \in A}\left(c - E(W)\right)^2 P(W = c)$
- The symbol for the variance of a random variable is $\boldsymbol{\sigma^2}$ and its unit is the square of the unit of the random variable.

W	2	3	4	5	6	7	8	9	10	11	12	Total
$(W - E(W))^2$	25	16	9	4	1	0	1	4	9	16	25	
P	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36	36/36
$(W - E(W))^2 * P$	0.69	0.89	0.75	0.44	0.14	0.00	0.14	0.44	0.75	0.89	0.69	5.8333

Practice with calculator

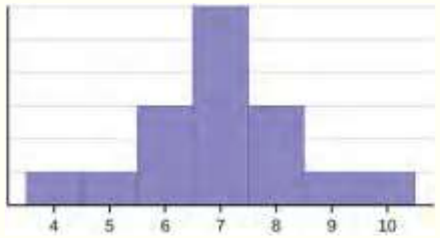
Variance, Standard Deviation and Measures of Dispersion

- Since the unit of variance is the **squared unit of the random variable**, in many cases the *standard deviation is preferred as a measure of dispersion because it has the same unit as the random variable*.
- The standard deviation (sd) is simply the square root of the variance. The symbol for the variance of a random variable is **σ**.
- Other measures of dispersion include the *range* (the difference between the largest and smallest value of the random variable), and
- The *coefficient of variation* = $\frac{\sqrt{\mathrm{Var}\ W}}{\mathrm{E(W)}} = \frac{\mathrm{SD(W)}}{\mathrm{E(W)}}$ that gives a scale-free way to assess the variance of the distribution of a random variable
- For our example:
  - Variance of the sum of the faces on two dice = 5 5/6 = 5.83333
  - **Note:** Sometimes you will find the formula for variance = $E(W^2) - (E(W))^2$. This is only true if all the values of the random variable are *equally likely* (i.e., have the same probability). In this case this formula will give 10, which is incorrect.
  - Standard Deviation of the sum of the faces on two dice = sqrt(5.83333) = 2.415
  - Range = 12 – 2 = 10. Range does not consider the probability distribution of the values.

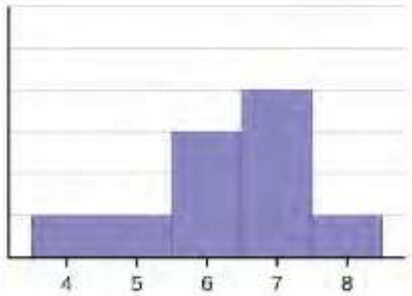
Skewness

- **Skewness** of a probability distribution is a measure of deviation from symmetry
- A **symmetric** distribution has zero skewness. Such a distribution has the **mean = median** of the distribution. If it has only one mode, the mode would also equal the mean and the median
- A “**left skewed**” distribution will appear chopped off on the right compared to the left. That is its left “tail” is longer. Such a distribution will have its **mean less that the median**, and both will be less than the mode
- A “**right skewed**” distribution will appear chopped off on the left compared to the right. That is its right “tail” is longer. Such a distribution will have its **mean greater that the median**, and both will be greater than the mode

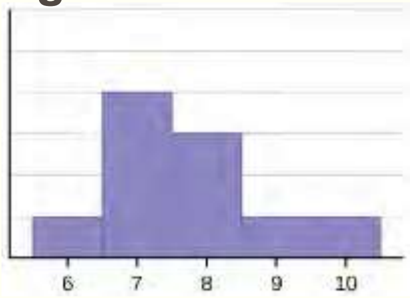
symmetric



“left skewed”



“right skewed”



Practice with calculator

Chebychev’s Inequality Theorem

- Relates the mean and variance of a distribution to the probability for **<u>any</u>** random variable:

$$P\,(|x - \mu| > c\sigma) < \frac{1}{c^2}$$

- In other words, X deviates or strays more than, say, 3 standard deviations from its mean *at most* only 1/9 of the time. This gives some concrete meaning to the concept of variance/standard deviation, regardless of the distribution of X.

- Another way to look at this is to say $P\,(x > \mu + c\sigma) + P\,(x < \mu - c\sigma) < \frac{1}{c^2}$

  - In the example of the sum of the faces of two dice, mean = 7 and std. dev = 2.4; 2 standard deviations is 4.8 and we should find that the probability (sum < 2.2) + probability(sum > 11.8) will be less than 25%.
  - P(w < 2.2) + P(w > 11.8) = 1/36 + 1/36 = 1/18 which is less than 0.25 and Chebychev’s inequality holds.

W	2	3	4	5	6	7	8	9	10	11	12	Total
P	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36	36/36

Practice with calculator

Problem – Converting Events to Random Variables

- Suppose you roll two dice. Let X be the *absolute* value of the difference between the numbers on the faces of the two dice.
    a) Show the probability distribution of X
    b) Calculate the expected value of X = 70/36 = 1.944
    c) Calculate the standard deviation of X = sqrt(2.0525) = 1.4327
    d) Using Chebychev's theorem, what are the two values that X lies past less than 25% of the time? Check that this is true.
    By Chebychev's theorem $P\,(X > \mu + c\sigma) + P\,(X < \mu - c\sigma) < \frac{1}{c^2}$.

    Let c = 2, then $P\,(X > 1.944 + 2 * 1.4327) + P\,(X < 1.944 - 2 * 1.4327) < 0.25$.

    i.e., P(X < -0.9214) + P(X > 4.8094) < 0.25.

    From our table P(X < -0.9214) = 0 and P(X > 4.8094) = 2/36 so answer is 2/36, which is less than 25% of the time, consistent with Chebychev's inequality.

X	0	1	2	3	4	5	Total	
P	6/36	10/36	8/36	6/36	4/36	2/36	36/36	
xP(X)	0	10/36	16/36	18/36	16/36	10/36	=70/36	**Expected value**
$(x - E(x))^2 * P(X)$	0.6301	0.2478	0.0007	0.1857	0.4695	0.5187	2.0525	**Variance**

	1	2	3	4	5	6
1	1/36	1/36	1/36	1/36	1/36	1/36
2	1/36	1/36	1/36	1/36	1/36	1/36
3	1/36	1/36	1/36	1/36	1/36	1/36
4	1/36	1/36	1/36	1/36	1/36	1/36
5	1/36	1/36	1/36	1/36	1/36	1/36
6	1/36	1/36	1/36	1/36	1/36	1/36

Practice with calculator

Book Problem – Example 4.2 – Book Page 250

- Suppose Nancy has classes **three days** a week. She attends classes all three days a week **80%** of the time, **two days 15%** of the time, **one day 4%** of the time, and **no days (misses all three days) 1%** of the time. Suppose one week is randomly selected.
- What is the random variable? X is number of days per week she attends classes
- What values does X take? 0, 1, 2, 3
- Show the probability distribution of the random variable X:

x	*P(x)*
0	0.01
1	0.04
2	0.15
3	0.80

- How many days does she attend per week, on average? = (0*0.01 + 1*0.04 + 2*0.15 + 3*0.80) = 2.74 = mean
- What is the standard deviation of number of days attended per week?
  = sqrt($(0 - 2.74)^2$ * (0.01) + $(1 - 2.74)^2$ * (0.04) + $(2 - 2.74)^2$ * (0.15) + $(3 - 2.74)^2$ * (0.8)) = 0.5
- What percentage of the time does she attend no more than 2 days? $P(X \leq 2) = 0.2$.
- What is the mode of the distribution? X=3 because it has the highest probability
- What is the median of the distribution? We cannot say, since there is no value of X close to the 50^{th} percentile.
- What kind of skewness does the distribution have? It is left-skewed since it has a longer left tail. This means that its median is greater than 2.74 (the mean)

Practice with calculator

Converting Events to RVs - Book Problem 4.74 (page 287)

- Suppose that you are offered the following "deal." You roll a die. If you roll a six, you win $10. If you roll a four or five, you win $5. If you roll a one, two, or three, you pay $6.
  - Define the Random Variable X.
  - Construct the table showing the probability mass and cumulative probabilities.
  - Over the long run of playing this game, what are your expected average winnings per game?
  - Based on numerical values, should you take the deal? Explain your decision in complete sentences.

- Solution:
  - X is winnings per game in dollars.

X	-6	5	10	
P(X)	3/6	2/6	1/6	
XP(X)	-18/6	10/6	10/6	E(X) = 2/6

  - Since the expected winnings on the long run is positive, if you play the game long enough you will win money. You should take the deal.

Practice with calculator

Working directly with RVs - Example: Book page 281

- Sometimes, we can work directly with random variables without the need to reference events, even though probabilities are only defined on events.
- Javier volunteers in community events each month. He does not do more than five events in a month. He attends exactly five events 35% of the time, four events 25% of the time, three events 20% of the time, two events 10% of the time, one event 5% of the time, and no events 5% of the time.
- X (random variable) = Number of events volunteered each month
- The probability mass and the *cumulative probabilities* are shown in the table. Since we don't have a closed form function relating X to its probability mass or cumulative probability, we don't use the term function. But taken together, the table represents the "function".

X	Probability P(X)	Cumulative Probability	x*P(X)	X^2	$x^2P(X)$
0	P(X=0) = 0.05	P(X≤0) = 0.05	0	0	0
1	P(X=1) = 0.05	P(X≤1) = 0.10	0.05	1	0.05
2	P(X=2) = 0.10	P(X≤2) = 0.20	0.20	4	0.40
3	P(X=3) = 0.20	P(X≤3) = 0.40	0.60	9	1.80
4	P(X=4) = 0.25	P(X≤4) = 0.65	1.00	16	4.00
5	P(X=5) = 0.35	P(X≤5) = 1.00	1.75	25	8.75
			3.60		15.00

The mean or expected value E(X) = $\Sigma xP(X)$ = 3.60.

Note that this is different from the average of X = 15/6 = 2.5. Arithmetic average gives the same probability to all values. So you should use mean or expected value for random variables, not arithmetic average.

Variance = $\Sigma x^2P(X) – (\Sigma xP(X))^2$ = 15.00 – 12.96 = 2.04

Standard Deviation = $\sqrt{2.04}$ = 1.4282.

Median = Value of random variable X with cumulative probability = 0.5. This lies between 3 and 4.

Mode = Value of random variable X with highest probability = 5

Probability that Javier volunteers for more than three events each month = P(X >=4) = 0.60 = 1 – P(X <= 3)