MICROENGINEERING 110:
Probability and Statistics for Engineers

Module 2: Probability

Prof. Vivek Subramanian
Microengineering
EPFL

Course Syllabus

Topic	Key material

Probability	• Sample spaces and events
	• Properties of probability
	• Discrete Random Variables and Probability
	  • Binomial and Poisson Distributions
	• Continuous Random Variables and Probability
	  • Normal and other continuous distributions
	• Joint Probability Distributions
	  • Covariance and Correlation
	• Point Estimation

Probability

- the study of randomness and uncertainty.
- In any situation in which one of a number of possible outcomes may occur, the discipline of probability provides methods for quantifying the chances, or likelihoods, associated with the various outcomes.

Sample Spaces

- The sample space of an experiment, denoted by *S*, is the set of all possible outcomes of that experiment.
- Example: Suppose I have a factory that produces screws, which may be:
  - Defective (Denoted here by D)
  - Free of defects (Denoted here by N)
- Then, the sample space for seeing whether a single screw is defective or not is:
  - $S = \{N, D\}$
- Similarly, the sample space for examining 3 successive screws for defects is:
  - $S = \{NNN, NND, NDN, DNN, NDD, DND, DDN, DDD\}$

See python code in module2_sample_spaces_permutations_combinations.ipynb

Permutations and combinations in Python

- As you saw, permutations and combinations are useful to determine sample spaces, which will soon be useful in determining probability
- Python makes generating lists of permutations and combinations easy

```
import itertools

elements = ["A", "B", "C"]
permutations = list(itertools.permutations(elements))
print(permutations)
```

```
[('A', 'B', 'C'), ('A', 'C', 'B'), ('B', 'A', 'C'), ('B', 'C', 'A'), ('C', 'A', 'B'), ('C', 'B', 'A')]
```

```
combinations = list(itertools.combinations(elements, 2))
print(combinations)
```

```
[('A', 'B'), ('A', 'C'), ('B', 'C')]
```

```
Inspection_outcome = ['N','D']
Sample_Space_3_screws=list(itertools.product(Inspection_outcome,repeat=3))
print(Sample_Space)
```

```
[('N', 'N', 'N'), ('N', 'N', 'D'), ('N', 'D', 'N'), ('N', 'D', 'D'), ('D', 'N', 'N'), ('D', 'N', 'D'), ('D', 'D', 'N'), ('D', 'D', 'D')]
```

https://docs.python.org/3/library/itertools.html

Events

- An event is any collection (subset) of outcomes contained in the sample space *S*
  - An event is simple if it consists of exactly one outcome
  - An event is compound if it consists of more than one outcome.
- When an experiment is performed, a particular event A is said to occur if the resulting experimental outcome is contained in A.
  - In general, exactly one simple event will occur,
  - but many compound events will occur simultaneously.

Example: leaving the CO Building

- Upon exiting the CO Building, you can go left or right. Suppose 3 of you leave the building, the possible outcomes are:

```
[('L', 'L', 'L'), ('L', 'L', 'R'), ('L', 'R', 'L'), ('L', 'R', 'R'), ('R', 'L', 'L'), ('R', 'L', 'R'), ('R', 'R', 'L'), ('R', 'R', 'R')]
```

See python code in module2_sample_spaces_permutations_combinations.ipynb

- There are 8 simple events
- Some example compound events are:
  - A = {RLL, LRL, LLR} ; the event that exactly one student turns right
  - B = {LLL, RLL, LRL, LLR} ; the event that at most student turns right
  - C = {LLL, RRR} : the event that all three students turn in the same direction
- Suppose that when the experiment is performed, the outcome is LLL. Then the simple event E1 has occurred and so also have the events B and C (but not A).
- Question: What is the "probability" that if 3 students exit the class:
  - They will all turn left?
  - They will all turn in the same direction?
- What assumptions did you make in your answer?

Example: Gas station pump usage

- Two gas stations are located at a certain intersection. Each one has six gas pumps. Consider the experiment in which the number of pumps in use at a particular time of day is determined for each of the stations. An experimental outcome specifies how many pumps are in use at the first station and how many are in use at the second one. One possible outcome is (2, 2), another is (4, 1), and yet another is (1, 4).
- How many outcomes exist in the sample space?

See python code in module2_sample_spaces_permutations_combinations.ipynb

Example: Gas station pump usage

		Second Station						
		0	1	2	3	4	5	6
First Station	0	(0, 0)	(0, 1)	(0, 2)	(0, 3)	(0, 4)	(0, 5)	(0, 6)
	1	(1, 0)	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
	2	(2, 0)	(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
	3	(3, 0)	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
	4	(4, 0)	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
	5	(5, 0)	(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
	6	(6, 0)	(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

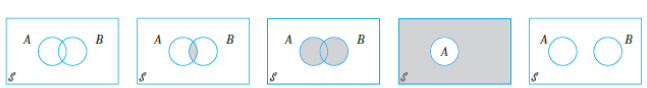
- What is the event that the number of pumps in use is the same for both stations?
- What is the event that the total number of pumps in use is 4?
- What is the event that at most one pump is in use at each station?

See python code in module2_sample_spaces_permutations_combinations.ipynb

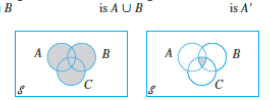
Some definitions

- The complement of an event A, denoted by A′, is the set of all outcomes in *S* that are not contained in A.
- The union of two events A and B, denoted by A $\cup$ B and read "A or B," is the event consisting of all outcomes that are either in A or in B or in both events (so that the union includes outcomes for which both A and B occur as well as outcomes for which exactly one occurs)—that is, all outcomes in at least one of the events.
- The intersection of two events A and B, denoted by A $\cap$ B and read "A and B," is the event consisting of all outcomes that are in both A and B.
- Let $\varnothing$ denote the null event (the event consisting of no outcomes whatsoever). When A $\cap$ B = $\varnothing$, A and B are said to be mutually exclusive or disjoint events.

Venn Diagrams

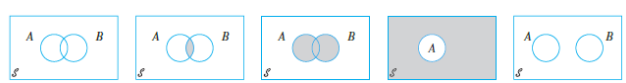


(a) Venn diagram of events A and B (b) Shaded region is $A \cap B$ (c) Shaded region is $A \cup B$ (d) Shaded region is A' (e) Mutually exclusive events



(f) Shaded region is $A \cup B \cup C$ (g) Shaded region is $A \cap B \cap C$

Axioms and Properties of Probability



(a) Venn diagram of events A and B (b) Shaded region is $A \cap B$ (c) Shaded region is $A \cup B$ (d) Shaded region is A' (e) Mutually exclusive events

AXIOM 1 For any event A, $P(A) \geq 0$.

AXIOM 2 $P(S) = 1$.

AXIOM 3 If $A_1, A_2, A_3, \ldots$ is an infinite collection of disjoint events, then

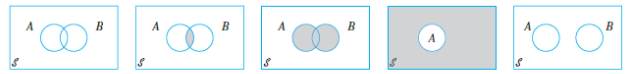
$$P(A_1 \cup A_2 \cup A_3 \cup \cdots) = \sum_{i=1}^{\infty} P(A_i)$$

$P(\varnothing) = 0$ where $\varnothing$ is the null event (the event containing no outcomes whatsoever). This in turn implies that the property contained in Axiom 3 is valid for a *finite* collection of disjoint events.

Example: Testing to find a working battery

- Consider testing batteries one by one until you find one having a voltage within a desired range. A success is called S and the failure is called F.
- The simple events are
  - E1 = {S},
  - E2 = {FS}
  - E3 = {FFS}
  - E4 = {FFFS},….
- Suppose the probability of any particular battery being satisfactory is .99.
- Then it can be shown that
  - P(E1) = .99
  - P(E2) = (.01)(.99)
  - P(E3) = $(.01)^2(.99)$,…
- In particular, because the Ei's are disjoint and S = E1 ∪ E2 ∪ E3 ∪…, it must be the case that
  - 1 = P(*S*) = P(E1) + P(E2) + P(E3) +…
    = .99[1 + .01 + $(.01)^2$ + $(.01)^3$ +…]
- Also provable by: $a + ar + ar^2 + ar^3 + \cdots = \frac{a}{1-r}$

More Properties of Probability



(a) Venn diagram of events *A* and *B* (b) Shaded region is $A \cap B$ (c) Shaded region is $A \cup B$ (d) Shaded region is A' (e) Mutually exclusive events

For any event *A*, $P(A) + P(A') = 1$, from which $P(A) = 1 - P(A')$.

For any event *A*, $P(A) \leq 1$.

For any two events *A* and *B*,
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

For any three events *A*, *B*, and *C*,
$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

Example: Sitting in a train

- During off-peak hours a commuter train has five cars. Suppose a commuter is twice as likely to select the middle car (Car3) as to select either adjacent car (Car2 or Car4), and is twice as likely to select either adjacent car as to select either end car (Car1 or Car5).
- Let p_i 5 P(car i is selected) = P(E_i).
- Then we have
  - $p_3 = 2p_2 = 2p_4$
  - $p_2 = 2p_1 = 2p_5 = p_4$.

$$1 = \sum P(E_i) = p_1 + 2p_1 + 4p_1 + 2p_1 + p_1 = 10p_1$$

- What is the probability of picking the first car?
- What is the probability of picking an end car (i.e., first or last)?
- What is the probability of picking one of the intermediate cars?

Equally Likely Outcomes

- In many experiments consisting of N outcomes, it is reasonable to assign equal probabilities to all N simple events. With p = P(E_i) for every i,

$$1 = \sum_{i=1}^{N} P(E_i) = \sum_{i=1}^{N} p = p \cdot N \quad \text{so } p = \frac{1}{N}$$

- Now consider an event A, with N(A) denoting the number of outcomes contained in A. Then

$$P(A) = \sum_{E_i \text{ in } A} P(E_i) = \sum_{E_i \text{ in } A} \frac{1}{N} = \frac{N(A)}{N}$$

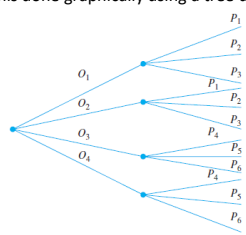
- Thus when outcomes are equally likely, computing probabilities reduces to counting: determine both the number of outcomes N(A) in A and the number of outcomes N in *S*, and determine their ratio.

The Product Rule: Useful while counting to determine probabilities

Product Rule for k-Tuples

Suppose a set consists of ordered collections of k elements (k-tuples) and that there are n_1 possible choices for the first element; for each choice of the first element, there are n_2 possible choices of the second element;…; for each possible choice of the first $k-1$ elements, there are n_k choices of the kth element. Then there are $n_1 n_2 \cdot \cdots \cdot n_k$ possible k-tuples.

- You may have seen this done graphically using a tree diagram



Permutations and Combinations

An ordered subset is called a **permutation.** The number of permutations of size k that can be formed from the n individuals or objects in a group will be denoted by $P_{k,n}$. An unordered subset is called a **combination.** One way to denote the number of combinations is $C_{k,n}$, but we shall instead use notation that is quite common in probability books: $\binom{n}{k}$, read "n choose k."

$$P_{k,n} = \frac{n!}{(n-k)!}$$

$$\binom{n}{k} = \frac{P_{k,n}}{k!} = \frac{n!}{k!(n-k)!}$$

Example: Shuffling a playlist

- Suppose I have a spotify playlist with 100 songs. Assume no song is repeated when shuffling the list, and suppose exactly 10 of the songs are by the Beatles.
- What is the probability that the first Beatles song played is the 5th song played while shuffling?
- Song 1 must be non-Beatles (NB) … 90/100
- Song 2, 3, and 4 must also be NB… what are the probabilities for each of these events?
- Song 5 must be Beatles (B). What is the probability for this event?
- $$P(1^{\text{st}}\text{ B is the }5^{\text{th}}\text{ song played}) = \frac{90 \cdot 89 \cdot 88 \cdot 87 \cdot 10}{100 \cdot 99 \cdot 98 \cdot 97 \cdot 96} = 0.0679$$

Example: Shuffling a playlist

Doing this using permutations:

- The number of permutations for the first 5 songs is simply $P_{5,100}$
- The number permutations by which we can select 4 NB songs out of the 90 NB songs is $P_{4,90}$
- The number of permutations by which we can select 1 B song out of the 10 B songs is 10
- Therefore, we can also write this as:

$$P(1^{\text{st}}\text{ B is the }5^{\text{th}}\text{ song played}) = \frac{90 \cdot 89 \cdot 88 \cdot 87 \cdot 10}{100 \cdot 99 \cdot 98 \cdot 97 \cdot 96} = \frac{P_{4,90} \cdot (10)}{P_{5,100}} = .0679$$

See python code in module2_sample_spaces_permutations_combinations.ipynb

Example: Shuffling a playlist

Doing this using combinations:

- Remember, combinations are about "choosing", without worrying about order.
- The number of ways to choose 10 of 100 songs is simply $C_{100,10}$. Effectively, we are choosing to give those slots to B songs.
- The number of ways to choose 9 of the last 95 songs to be B songs is simply $C_{95,9}$.
- If we do that, then then that means that the first 5 songs are *necessarily* 4 NBs and 1 B. There is only 1 way to order it so that we start with 4 NBs and then play a B, which means that

- $$P(1^{st}\text{ B is the }5^{th}\text{ song played}) = \frac{\binom{95}{9}}{\binom{100}{10}}$$ =0.0679

See python code in module2_sample_spaces_permutations_combinations.ipynb

Example: Shuffling a playlist

Interestingly, if we ask what is the probability of one of the first 5 songs played being a B, the probability is:

$$P(1^{st}\text{ B is the }1^{st}\text{ or }2^{nd}\text{ or }3^{rd}\text{ or }4^{th}\text{ or }5^{th}\text{ song played}) = \frac{\binom{99}{9}}{\binom{100}{10}} + \frac{\binom{98}{9}}{\binom{100}{10}} + \frac{\binom{97}{9}}{\binom{100}{10}} + \frac{\binom{96}{9}}{\binom{100}{10}} + \frac{\binom{95}{9}}{\binom{100}{10}} = .4162$$

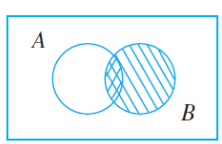
Which is pretty high, considering that only 10 out of 100 songs are B songs… why is this the case?

Conditional Probability

- We will use the notation P(A|B) to represent the conditional probability of A given that the event B has occurred. B is the "conditioning event."

For any two events A and B with $P(B) > 0$, the **conditional probability of A given that B has occurred** is defined by

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



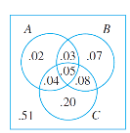
The Multiplication Rule

$$P(A \cap B) = P(A|B) \cdot P(B)$$

Example: Reading habits

- A news magazine publishes three columns entitled "Art" (A), "Books" (B), and "Cinema" (C). Reading habits of a randomly selected reader with respect to these columns are

Read regularly	A	B	C	$A \cap B$	$A \cap C$	$B \cap C$	$A \cap B \cap C$
Probability	.14	.23	.37	.08	.09	.13	.05



- What is P(A|B) and what does it mean

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{.08}{.23} = .348$$; this gives us the likelihood that a reader who reads "Books" also reads "Arts"

- The probability that the selected individual regularly reads the Art column given that he or she regularly reads at least one of the other two columns is

$$P(A|B \cup C) = \frac{P(A \cap (B \cup C))}{P(B \cup C)} = \frac{.04 + .05 + .03}{.47} = \frac{.12}{.47} = .255$$

- The probability that the selected individual reads at least one of the first two columns given that he or she reads the Cinema column is

$$P(A \cup B|C) = \frac{P((A \cup B) \cap C)}{P(C)} = \frac{.04 + .05 + .08}{.37} = .459$$

Example: Phone comparison

- An electronics store sells three different brands of phones. Of its phones sales, 50% are brand 1 (the least expensive), 30% are brand 2, and 20% are brand 3.
- Each manufacturer offers a 1-year warranty on parts and labor. It is known that 25% of brand 1's phones require warranty repair work, whereas the corresponding percentages for brands 2 and 3 are 20% and 10%, respectively.
- What is the probability that a randomly selected purchaser has bought a brand 1 Phone that will need repair while under warranty?
- What is the probability that a randomly selected purchaser has a phone that will need repair while under warranty?
- If a customer returns to the store with a phone that needs warranty repair work, what is the probability that it is a brand 1 phone ? A brand 2 phone? A brand 3 phone?

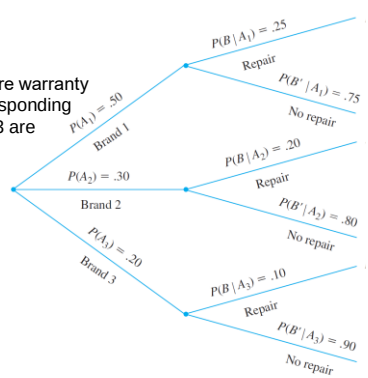
Example: Phone comparison

- An electronics store sells three different brands of phones. Of its phones sales, 50% are brand 1 (the least expensive), 30% are brand 2, and 20% are brand 3.
- Each manufacturer offers a 1-year warranty on parts and labor. It is known that 25% of brand 1's phones require warranty repair work, whereas the corresponding percentages for brands 2 and 3 are 20% and 10%, respectively
- Let A_i be the probability of purchasing a brand i, then:
  - $P(A_1) = 0.5$
  - $P(A_2) = 0.3$
  - $P(A_3) = 0.2$
- Let B be the probability that a phone needs repair. Then:
  - $P(B|A_1) = 0.25$
  - $P(B|A_2) = 0.2$
  - $P(B|A_3) = 0.1$

Example: Phone comparison

Of its phones sales, 50% are brand 1 30% are brand 2, and 20% are brand 3
$P(A_1) = 0.5$
$P(A_2) = 0.3$
$P(A_3) = 0.2$

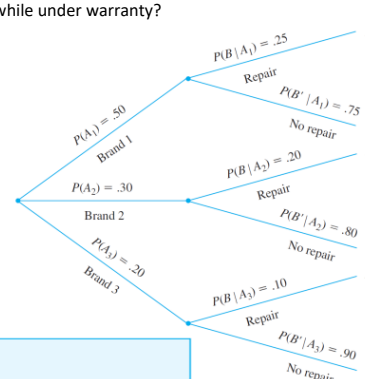
25% of brand 1's phones require warranty repair work, whereas the corresponding percentages for brands 2 and 3 are 20% and 10%, respectively.
$P(B|A_1) = 0.25$
$P(B|A_2) = 0.2$
$P(B|A_3) = 0.1$



Example: Phone comparison

- What is the probability that a randomly selected purchaser has bought a brand 1 Phone that will need repair while under warranty?

$$P(A_1 \cap B) = P(B|A_1) \cdot P(A_1) = .125.$$



Reminder:

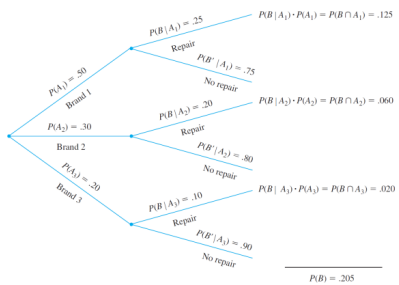
The Multiplication Rule

$$P(A \cap B) = P(A|B) \cdot P(B)$$

Example: Phone comparison

- What is the probability that a randomly selected purchaser has a phone that will need repair while under warranty?

$$P(B) = P[\text{(brand 1 and repair) or (brand 2 and repair) or (brand 3 and repair)}]$$
$$= P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B)$$
$$= .125 + .060 + .020 = .205$$



Example: Phone comparison

- If a customer returns to the store with a phone that needs warranty repair work, what is the probability that it is a brand 1 phone ? A brand 2 phone? A brand 3 phone?

$$P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{.125}{.205} = .61$$
$$P(A_2|B) = \frac{P(A_2 \cap B)}{P(B)} = \frac{.060}{.205} = .29$$
$$P(A_3|B) = 1 - P(A_1|B) - P(A_2|B) = .10$$

Reminder:

The Multiplication Rule

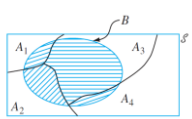
$$P(A \cap B) = P(A|B) \cdot P(B)$$

Some more theorems

The Law of Total Probability

Let $A_1, \ldots, A_k$ be mutually exclusive and exhaustive events. Then for any other event B,

$$P(B) = P(B|A_1)P(A_1) + \cdots + P(B|A_k)P(A_k)$$
$$= \sum_{i=1}^{k} P(B|A_i)P(A_i)$$



Some more theorems

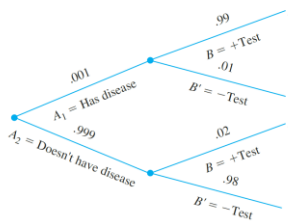
Bayes' Theorem

Let $A_1, A_2, \ldots, A_k$ be a collection of k mutually exclusive and exhaustive events with *prior* probabilities $P(A_i)$ $(i = 1, \ldots, k)$. Then for any other event B for which $P(B) > 0$, the *posterior* probability of A_j given that B has occurred is

$$P(A_j|B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{k} P(B|A_i) \cdot P(A_i)} \quad j = 1, \ldots, k$$

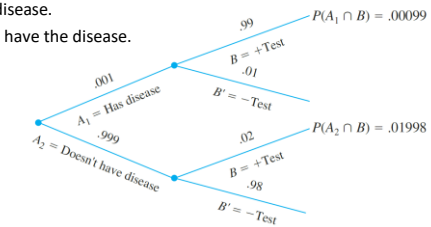
Example: Disease Testing

- Only 1 in 1000 adults is afflicted with a rare disease for which a diagnostic test has been developed. The test is such that when an individual actually has the disease, a positive result will occur 99% of the time, whereas an individual without the disease will show a positive test result only 2% of the time (i.e., the sensitivity of this test is 99% and the specificity is 98%). Let A_1 = individual has the disease.
- Let A_2 = individual does not have the disease.
- Let B = positive test result
- Then:
  - $P(A_1) = 0.001$
  - $P(A_2) = 0.999$
  - $P(B|A_1) = 0.99$
  - $P(B|A_2) = 0.02$



Example: Disease Testing

- If a randomly selected individual is tested and the result is positive, what is the probability that the individual has the disease?
- Let A_1 = individual has the disease.
- Let A_2 = individual does not have the disease.
- Let B = positive test result
- Then:
  - $P(A_1) = 0.001$
  - $P(A_2) = 0.999$
  - $P(B|A_1) = 0.99$
  - $P(B|A_2) = 0.02$



  - $P(B) = 0.00099+0.01998 = 0.02097$
  - Then, $P(A_1|B) = \dfrac{P(A_1 \cap B)}{P(B)} = \dfrac{.00099}{.02097} = .047$
  - In other words, the likelihood that a positive test actually is indicative of a disease is really low! Why? Because the rarity of the disease means that most positives are due to errors. To detect rare diseases, we need tests with REALLY low false positive rates

Independence

> Two events A and B are **independent** if $P(A|B) = P(A)$ and are **dependent** otherwise.

- Additionally, if A an B are independent, then so are
  - A′ and B
  - A and B′
  - A′ and B′

> A and B are independent if and only if (iff)
> $$P(A \cap B) = P(A) \cdot P(B)$$

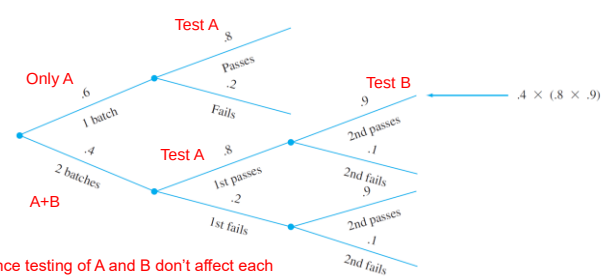
  - Proof:

$$P(A \cap B) = P(A|B) \cdot P(B) = P(A) \cdot P(B)$$

Since P(A|B) = P (A) if independent

Example exploiting independence: Testing of supplier batches

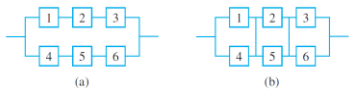
- Each weekday, a batch of components sent by a supplier A arrives at a certain inspection facility. Two days a week, a batch also arrives from a Supplier B. Eighty percent of all supplier A's batches pass inspection, and 90% of supplier B's do likewise. What is the probability that, on a randomly selected day, two batches pass inspection?



Since testing of A and B don't affect each other, we can chain them up by dealing with A and then B

Extending to multiple probabilities

- Consider 2 solar cell string configurations



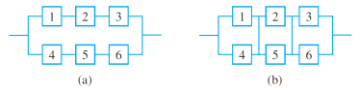
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cap B) = P(A) \cdot P(B)$$

- We want to know if the each system will work for 10 years. The system is assumed to "work" if power is available. Assuming the probability of any cell last 10 years is 0.9, and assuming cell failure is independent, which system has a higher probability of working for 10 years.
  - System (a): 1, 2, & 3 are in series, so all must work for the string to work. On the other hand, the stings are in parallel.
  - Let A_i be the probability that any cell i lasts 10 years. Then, the P of the system working for 10 years is:
  - $P(A_{system}) = P[(A_1 \cap A_2 \cap A_3) \cup (A_4 \cap A_5 \cap A_6)]$
    $= P(A_1 \cap A_2 \cap A_3) + P(A_4 \cap A_5 \cap A_6) - P(A_1 \cap A_2 \cap A_3) \cap P(A_4 \cap A_5 \cap A_6)$
    $= 0.9*0.9*0.9+0.9*0.9*0.9-0.9*0.9*0.9*0.9*0.9*0.9=0.927$

Extending to multiple probabilities

- Consider 2 solar cell string configurations



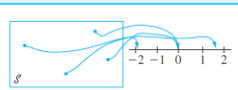
$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cap B) = P(A) \cdot P(B)$$

  - System (b): 1 and 4 are in parallel, which are in series with 2 and 5…
  - $P(A_{system}) = P[(A_1 \cup A_4) \cap (A_2 \cup A_5) \cap (A_3 \cup A_6)]$
    $=[P(A_1)+ P(A_4)-P(A_1 \cap A_4)][P(A_2)+ P(A_5)-P(A_2 \cap A_5)][P(A_3)+ P(A_6)-P(A_3 \cap A_6)]$
    $=[P(A_1)+ P(A_4)-P(A_1)P(A_4)][P(A_2)+ P(A_5)-P(A_2)P(A_5)][P(A_3)+ P(A_6)-P(A_3)P(A_6)]$
    $= (0.9+0.9-0.81)^3$
    $= 0.97$

Random variables

For a given sample space $\mathcal{S}$ of some experiment, a **random variable (rv)** is any rule that associates a number with each outcome in $\mathcal{S}$. In mathematical language, a random variable is a function whose domain is the sample space and whose range is the set of real numbers.



Any random variable whose only possible values are 0 and 1 is called a **Bernoulli random variable.**

Example:

- When a student calls a university help desk for technical support, he/she will either immediately be able to speak to someone (S, for success) or will be placed on hold (F, for failure). With $S = \{S, F\}$, define an rv X by $X(S) = 1$ $X(F) = 0$

Reminder: 2 gas pumps

- Two gas stations are located at a certain intersection. Each one has six gas pumps. Consider the experiment in which the number of pumps in use at a particular time of day is determined for each of the stations.

		Second Station						
		0	1	2	3	4	5	6
First Station	0	(0, 0)	(0, 1)	(0, 2)	(0, 3)	(0, 4)	(0, 5)	(0, 6)
	1	(1, 0)	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
	2	(2, 0)	(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
	3	(3, 0)	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
	4	(4, 0)	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
	5	(5, 0)	(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
	6	(6, 0)	(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

- Define rv's *X, Y,* and *U* by
  - *X* = the total number of pumps in use at the two stations
  - *Y* = the difference between the number of pumps in use at station 1 and the number in use at station 2
  - *U* = the maximum of the numbers of pumps in use at the two stations
  - If this experiment is performed and (2, 3) results, then $X((2, 3)) = 2 + 3 = 5$, so the observed value of *X* was $x = 5$.
  - Similarly, the observed value of *Y* would be $y = 2 - 3 = -1$,
  - and the observed value of *U* would be $u = \max(2, 3) = 3$.

Types of Random Variables

A **discrete** random variable is an rv whose possible values either constitute a finite set or else can be listed in an infinite sequence in which there is a first element, a second element, and so on ("countably" infinite).
A random variable is **continuous** if *both* of the following apply:

1. Its set of possible values consists either of all numbers in a single interval on the number line (possibly infinite in extent, e.g., from $-\infty$ to ∞) or all numbers in a disjoint union of such intervals (e.g., $[0, 10] \cup [20, 30]$).
2. No possible value of the variable has positive probability, that is, $P(X = c) = 0$ for any possible value c.

- This second point may seem counter-intuitive, but it makes sense when we consider that this references a single value, not an interval. Since we are continuous, intervals will have non-zero probability.

Probability Distribution Functions of a discrete RV

The **probability distribution** or **probability mass function** (pmf) of a discrete rv is defined for every number x by $p(x) = P(X = x) = P(\text{all } \omega \in \mathcal{S}: X(\omega) = x)$.

Example:

- Six boxes of components are ready to be shipped by a certain supplier. The number of defective components in each box is as follows:

Box	1	2	3	4	5	6
Number of defectives	0	2	0	1	2	0

- One of these boxes is to be randomly selected for shipment to a particular customer. Let *X* be the number of defectives in the selected box. The three possible *X* values are 0, 1, and 2. Then

$$p(0) = P(X = 0) = P(\text{box 1 or 3 or 6 is sent}) = \frac{3}{6} = .500$$

$$p(1) = P(X = 1) = P(\text{box 4 is sent}) = \frac{1}{6} = .167$$

$$p(2) = P(X = 2) = P(\text{box 2 or 5 is sent}) = \frac{2}{6} = .333$$

Plotting probability distribution functions

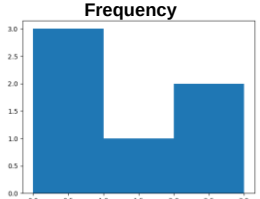
- We can generate histograms of such distribution functions, such as:

$$p(0) = P(X = 0) = P(\text{box 1 or 3 or 6 is sent}) = \frac{3}{6} = .500$$

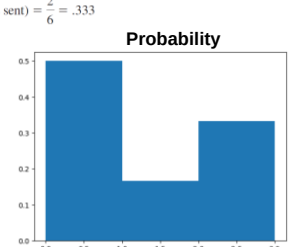
$$p(1) = P(X = 1) = P(\text{box 4 is sent}) = \frac{1}{6} = .167$$

$$p(2) = P(X = 2) = P(\text{box 2 or 5 is sent}) = \frac{2}{6} = .333$$

Frequency



Probability



- Similarly, we can generate cumulative distribution functions, etc.

See python code in module2_discrete_RVs.ipynb

Plotting probability distribution functions

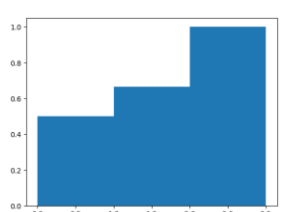
- We can generate histograms of such distribution functions, such as:

$$p(0) = P(X = 0) = P(\text{box 1 or 3 or 6 is sent}) = \frac{3}{6} = .500$$

$$p(1) = P(X = 1) = P(\text{box 4 is sent}) = \frac{1}{6} = .167$$

$$p(2) = P(X = 2) = P(\text{box 2 or 5 is sent}) = \frac{2}{6} = .333$$

Cumulative Probability



The **cumulative distribution function** (cdf) $F(x)$ of a discrete rv variable X with pmf $p(x)$ is defined for every number x by

$$F(x) = P(X \le x) = \sum_{y:\, y \le x} p(y)$$

For any number x, $F(x)$ is the probability that the observed value of X will be at most x.

Expected Value

Let X be a discrete rv with set of possible values D and pmf $p(x)$. The **expected value** or **mean value** of X, denoted by $E(X)$ or μ_X or just μ, is

$$E(X) = \mu_X = \sum_{x \in D} x \cdot p(x)$$

If the rv X has a set of possible values D and pmf $p(x)$, then the expected value of any function $h(X)$, denoted by $E[h(X)]$ or $\mu_{h(X)}$, is computed by

$$E[h(X)] = \sum_{D} h(x) \cdot p(x)$$

Example: For a linear function

$$E(aX + b) = a \cdot E(X) + b$$

(Or, using alternative notation, $\mu_{aX+b} = a \cdot \mu_X + b$)

Variance

Let X have pmf $p(x)$ and expected value μ. Then the **variance** of X, denoted by $V(X)$ or σ_X^2, or just σ^2, is

$$V(X) = \sum_{D} (x - \mu)^2 \cdot p(x) = E[(X - \mu)^2]$$

The **standard deviation** (SD) of X is

$$\sigma_X = \sqrt{\sigma_X^2}$$

Python Example:

```
from scipy.stats import rv_discrete

x = [10, 20, 30]
p = [0.2, 0.3, 0.5]
distribution = rv_discrete(values=(x, p))
print("Expected value: ", distribution.expect())
print("Variance: ", distribution.var())
print("Standard Deviation: ", distribution.std())
```

Expected value: 23.0
Variance: 61.0
Standard Deviation: 7.810249675906654

See python code in module2_discrete_RVs.ipynb

The binomial distribution

- There are many experiments that conform either exactly or approximately to the following list of requirements:

1. The experiment consists of a sequence of n smaller experiments called trials, where n is fixed in advance of the experiment.
2. Each trial can result in one of the same two possible outcomes (dichotomous trials), which we generically denote by success (S) and failure (F). The assignment of the S and F labels to the two sides of the dichotomy is arbitrary.
3. The trials are independent, so that the outcome on any particular trial does not influence the outcome on any other trial.
4. The probability of success P(S) is constant from trial to trial; we denote this probability by p.

An experiment for which Conditions 1–4 (a fixed number of dichotomous, independent, homogenous trials) are satisfied is called a **binomial experiment.**

The binomial random variable

The **binomial random variable** X associated with a binomial experiment consisting of n trials is defined as

X = the number of S's among the n trials

- Suppose, for example, that $n = 3$. Then there are eight possible outcomes for the experiment:
  - *SSS SSF SFS SFF FSS FSF FFS FFF*
- From the definition of X, $X(SSF) = 2$, $X(SFF) = 1$, and so on.
- Possible values for X in an n-trial experiment are $x = 0, 1, 2, \ldots, n$.
- We will often write X,Bin(n, p) to indicate that X is a binomial rv based on n trials with success probability p.

Probability of x in a binomial experiment

- **Consider an n=4 binomial experiment**

Outcome	x	Probability	Outcome	x	Probability
SSSS	4	p^4	*FSSS*	3	$p^3(1-p)$
SSSF	3	$p^3(1-p)$	*FSSF*	2	$p^2(1-p)^2$
SSFS	3	$p^3(1-p)$	*FSFS*	2	$p^2(1-p)^2$
SSFF	2	$p^2(1-p)^2$	*FSFF*	1	$p(1-p)^3$
SFSS	3	$p^3(1-p)$	*FFSS*	2	$p^2(1-p)^2$
SFSF	2	$p^2(1-p)^2$	*FFSF*	1	$p(1-p)^3$
SFFS	2	$p^2(1-p)^2$	*FFFS*	1	$p(1-p)^3$
SFFF	1	$p(1-p)^3$	*FFFF*	0	$(1-p)^4$

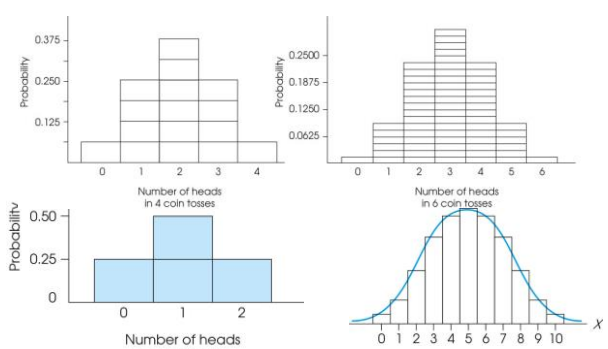
- If we wanted to determine the probability of a particular number of successes, i.e., a particular binomial random variable value, we would sum the probabilities of the individual occurences. For example, for x=3, we have:

$$b(3; 4, p) = P(FSSS) + P(SFSS) + P(SSFS) + P(SSSF) = 4p^3(1-p)$$

- More generally:

$$b(x; n, p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & x = 0, 1, 2, \ldots, n \\ 0 & \text{otherwise} \end{cases}$$

Binomial Distribution as a function of n



If $X \sim \text{Bin}(n, p)$, then $E(X) = np$, $V(X) = np(1-p) = npq$, and $\sigma_X = \sqrt{npq}$ (where $q = 1 - p$).

Binomial distributions in Python

```
from scipy.stats import binom

# Parameters
n = 10 # number of trials
x = 7 # number of successes
p = 0.2 # probability of success

print("Mean: ", binom.mean(n, p))
print("Variance: ", binom.var(n, p))
print("Probability mass function: ", binom.pmf(x, n, p))
print("Cumulative distribution function: ", binom.cdf(x,n,p))
```

Mean: 2.0
Variance: 1.6
Probability mass function: 0.000786432
Cumulative distribution function: 0.9999220736

See python code in module2_discrete_RVs.ipynb

The Dishonest Gambler Problem

- Consider a dishonest gambler, Denis Bloodnok, who makes money betting with a biased penny that he knows comes up heads, on the average, 8 times out of 10.
- For his penny the *probability p* of a head is 0.8 and the *probability q* = 1 - *p* of a tail is 0.2.
- Suppose he bets at even money that of *n* = *5* tosses of his penny at least four will come up heads. To make his bets appropriately, Bloodnok calculates the probability, with five tosses, of no heads, one head. two heads, and so on.
- With *y* representing the number of heads, what he needs are the *n* + 1 = six values: Pr(y=0), Pr(y=1),…
- Call the tossing of the penny five times a *trial* and denote the *outcome* by listing the heads (H) and tails (T) in the order in which they occur. The outcome *y* = O of getting no heads can occur in only one way, so
  - Pr(y = O) = Pr(T T T T T) = q x q x q x q x q = q^5 = 0.2^5 = 0.00032

The Dishonest Gambler Problem

- **Find Pr(y=1)**

  Pr(y = 1) = Pr(H T T T T) + Pr(T H T T T} + Pr(T T H T T) + Pr(T T T H T) + Pr(T T T T H)

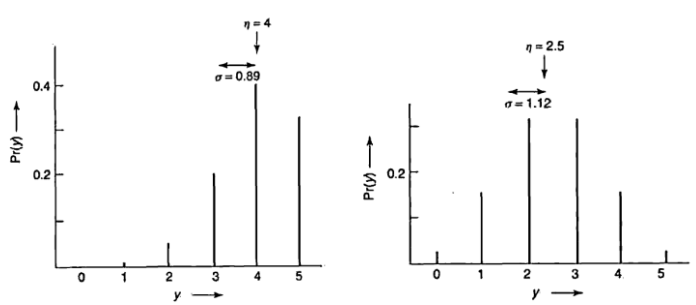
  Pr(y = 1) = $5pq^4$ = *5* x 0.8 x 0.2^4 = 5 x 0.000128 = 0.00640.
- **Find Pr(y=2)**
  - **Write out the possibilities**

    (H H T T T)
  - **Write out the equation**

The overall result

Number of Heads y	$\binom{5}{y} = \frac{5!}{y!(5-y)!}$	$p^y q^{n-y}$	Pr(y)
0	1	0.00032	0.00032
1	5	0.00128	0.00640
2	10	0.00512	0.05120
3	10	0.02048	0.20480
4	5	0.08192	0.40960
5	1	0.32768	0.32768
			1.00000

Graphically



Biased penny **Fair penny**

How does he make money?

- In this way Bloodnok calculates that using his biased penny the probability of *at least* four heads [given by Pr(4) + Pr(5)] is about 0.74. For the fair penny Pr(4) + Pr(5) yields a probability of only 0.19.
- For the fair penny a wager at even money that he can throw at least four heads appears unfavorable to him.
- By using his biased penny he can make an average of 48 cents for every dollar bet. (If he bets a single dollar on this basis 100 times, in 74 cases he will make a dollar and in 26 cases he will lose a dollar. His overall net gain is thus 74-26 = 48 dollars per 100 bet.)

The Poisson distribution

A discrete random variable X is said to have a **Poisson distribution** with parameter μ ($\mu > 0$) if the pmf of X is

$$p(x; \mu) = \frac{e^{-\mu} \cdot \mu^x}{x!} \qquad x = 0, 1, 2, 3, \ldots$$

Suppose that in the binomial pmf $b(x; n, p)$, we let $n \to \infty$ and $p \to 0$ in such a way that np approaches a value $\mu > 0$. Then $b(x; n, p) \to p(x; \mu)$.

- According to this result, *in any binomial experiment in which n is large and p is small, b(x; n, p) < p(x; μ), where μ = np*. As a rule of thumb, this approximation can safely be applied if $n > 50$ and $np < 5$.

Poisson distribution in Python

```
from scipy.stats import poisson

# Parameters
x = 1 # number of events
Lambda = 2/3 # lambda parameter

print("Mean: ", poisson.mean(Lambda))
print("Variance: ", poisson.var(Lambda))
print("Probability mass function: ", poisson.pmf(x, Lambda))
print("Cumulative distribution function: ", poisson.cdf(x, Lambda))
```

Mean: 0.6666666666666666
Variance: 0.6666666666666666
Probability mass function: 0.3422780793550613
Cumulative distribution function: 0.8556951983876534

See python code in module2_discrete_RVs.ipynb

PDF of Continuous random variables

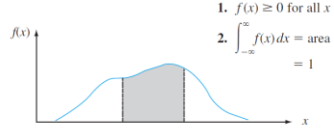
Let X be a continuous rv. Then a **probability distribution** or **probability density function** (pdf) of X is a function $f(x)$ such that for any two numbers a and b with $a \le b$,

$$P(a \le X \le b) = \int_a^b f(x)dx$$

That is, the probability that X takes on a value in the interval $[a, b]$ is the area above this interval and under the graph of the density function, as illustrated in Figure 4.2. The graph of $f(x)$ is often referred to as the *density curve*.

For $f(x)$ to be a legitimate pdf, it must satisfy the following two conditions:

1. $f(x) \ge 0$ for all x
2. $\int_{-\infty}^{\infty} f(x)dx$ = area under the entire graph of $f(x)$ $= 1$

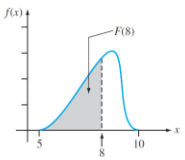


CDF of Continuous random variables

The **cumulative distribution function** $F(x)$ for a continuous rv X is defined for every number x by

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(y)\,dy$$

For each x, $F(x)$ is the area under the density curve to the left of x. This is illustrated in Figure 4.5, where $F(x)$ increases smoothly as x increases.



Computing probabilities for CDFs

Let X be a continuous rv with pdf $f(x)$ and cdf $F(x)$. Then for any number a,

$$P(X > a) = 1 - F(a)$$

and for any two numbers a and b with $a < b$,

$$P(a \le X \le b) = F(b) - F(a)$$

Expected value and variance

The **expected** or **mean value** of a continuous rv X with pdf $f(x)$ is

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

The **variance** of a continuous random variable X with pdf $f(x)$ and mean value μ is

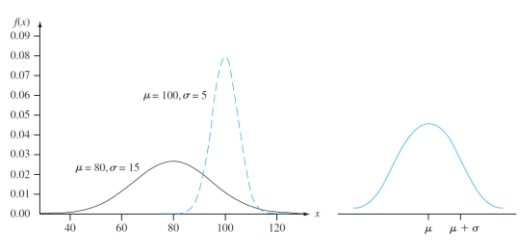
$$\sigma_X^2 = V(X) = \int_{-\infty}^{\infty} (x-\mu)^2 \cdot f(x)dx = E[(X-\mu)^2]$$

The **standard deviation** (SD) of X is $\sigma_X = \sqrt{V(X)}$.

Normal Distribution

A continuous rv X is said to have a **normal distribution** with parameters μ and σ (or μ and σ^2), where $-\infty < \mu < \infty$ and $0 < \sigma$, if the pdf of X is

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)} \quad -\infty < x < \infty \qquad (4.3)$$



Normal distributions in Python

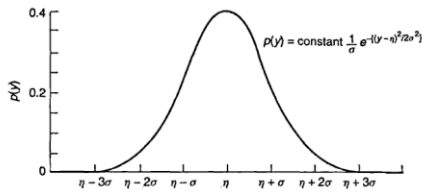
```
from scipy.stats import norm
# Parameters
x = 1.3 # value to look for
mu = 0 # mean
sigma = 1 # standard deviation

print("Mean: ", norm.mean(loc = mu, scale = sigma))
print("Variance: ", norm.var(loc = mu, scale = sigma))
print("Probability mass function: ", norm.pdf(x, loc = mu, scale = sigma))
print("Cumulative distribution function: ", norm.cdf(x, loc = mu, scale = sigma))
print("Survival function (1-cdf): ", norm.sf(x, loc = mu, scale = sigma))
```

```
Mean: 0.0
Variance: 1.0
Probability mass function: 0.17136859204780736
Cumulative distribution function: 0.9031995154143897
Survival function (1-cdf): 0.09680048458561036
```

See python code in module2_continuous_RVs.ipynb

The normal distribution

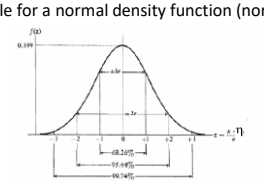
- **In the real-world, random variations often result in the "normal" distribution**



Characteristics of the normal distribution

- Symmetric, bell shaped
- Continuous for all values of X between -∞ and ∞ so that each conceivable interval of real numbers has a probability other than zero.

  $-\infty \le X \le \infty$
- • Two parameters, η and σ. Note that the normal distribution is actually a family of distributions, since η and σ determine the shape of the distribution.
- • The rule for a normal density function (normalized to have an area of 1) is



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

- About 2/3 of all cases fall within one standard deviation of the mean, that is $P(\eta - \sigma \le X \le \eta + \sigma) = .6826$.
- About 95% of cases lie within 2 standard deviations of the mean, that is $P(\eta - 2\sigma \le X \le \eta + 2\sigma) = .9544$

Why is the normal distribution important?

The central limit effect

- **Any measurement in a set of replicates is subject to numerous sources of error or "noise".**
- **Typically, in a well-designed experiment, the effect of these individual errors is small. As such, the overall error can then be approximated as a linear combination of the individual errors:**

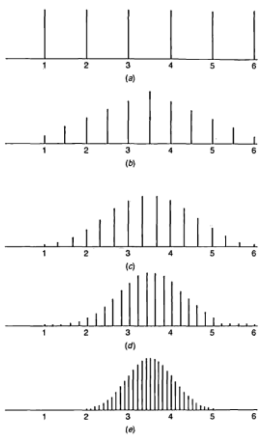
$$e = a_1 e_1 + a_2 e_2 + \cdots + a_n e_n$$

- **The central limit theorem states that the resulting distribution will tend to normality if:**
  - **There are several sources of error, such that the distribution of the individual error sources becomes unimportant**
  - **No individual error source dominates**

Robustness to assumption of normality

- **The techniques used here work well even if the fit to normality is only approximate.**

Example: Dice throwing



1 die thrown repeatedly

Average of 2 dice thrown repeatedly (i.e., sum of values / 2)

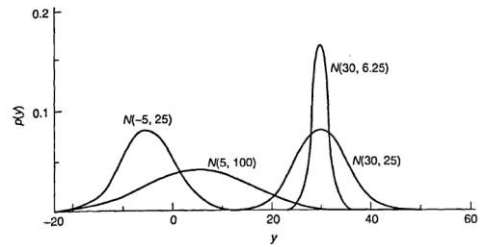
Average of 3 dice thrown repeatedly (i.e., sum of values / 3)

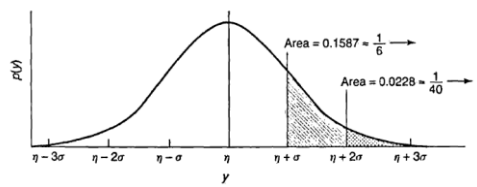
Average of 4 dice thrown repeatedly (i.e., sum of values / 4)

Average 5 dice thrown repeatedly (i.e., sum of values / 5)

Normal Distribution Characteristics

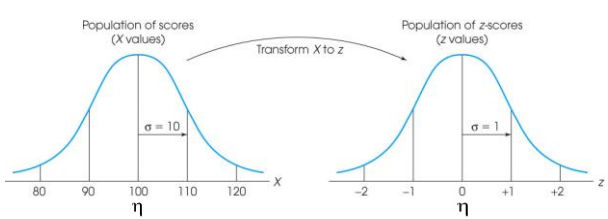
$N(\eta, \sigma^2)$





Z-scores

- **Populations are sometimes presented as z-scores to normalize with respect to the standard deviation**



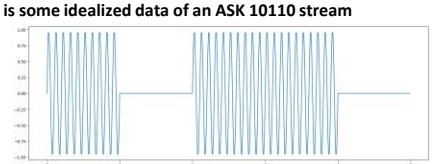
$$z = \frac{Y - \eta}{\sigma}$$

1. $\Pr(y > \eta + \sigma) = \Pr[(y - \eta) > \sigma] = \Pr\left[\left(\frac{y-\eta}{\sigma}\right) > 1\right] = \Pr(z > 1) = 0.1587$
2. $\Pr(z < -1) = 0.1587$
3. $\Pr(|z| > 1) = 0.3174$
4. $\Pr(z > 2) = 0.0228$
5. $\Pr(z < -2) = 0.0228$
6. $\Pr(|z| > 2) = 0.0455$

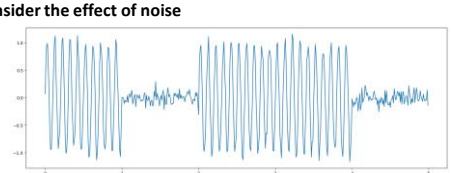
Back to our Camipro example

See python code in module2_continuous_RVs.ipynb

- **Here is some idealized data of an ASK 10110 stream**



- **Consider the effect of noise**



```
noise = np.random.normal(0,0.1,500)
modulated_signal_noisy=modulated_signal+noise
plot_signal(tc,modulated_signal_noisy)
```

Demodulation

- **To extract the 1's and 0's, a demodulator on the reader rectifies the input signal**

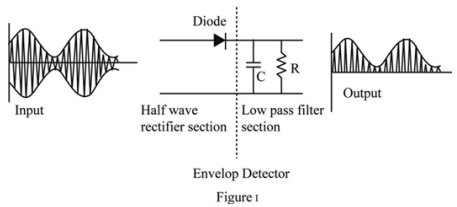
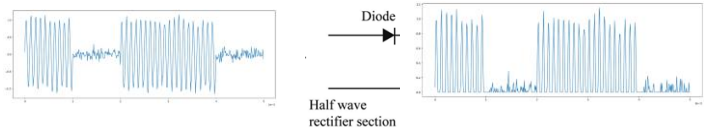


Figure 1

Rectification in python

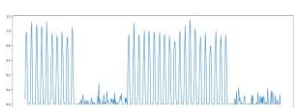
```
rectified_signal=np.zeros(500)
for counter in range(500):
    rectified_signal[counter]=modulated_signal_noisy[counter]
    if modulated_signal_noisy[counter] < 0:
        rectified_signal[counter]=0

plot_signal(tc,rectified_signal)
```



Data extraction

- **Let's assume the data extraction uses the peaks of the half-waves.**
- **Python allows us to easily extract peaks of waveforms**

```
from scipy.signal import find_peaks
peaksarray, _=find_peaks(rectified_signal)
peakvalues=np.zeros(len(peaksarray))
for counter in range(len(peaksarray)):
    peakvalues[counter]=rectified_signal[peaksarray[counter]]
```

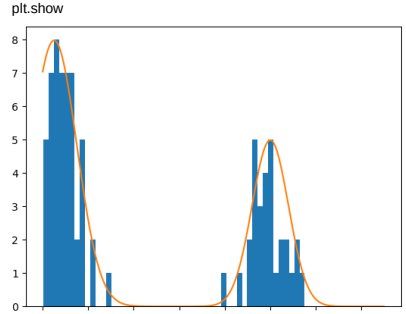


```
[0.9775125  1.11906718 1.08207663 1.05030563 1.12597284 0.96720406
 0.93322268 0.99352721 0.91196599 1.05140782 0.00278545 0.03121158
 0.06734725 0.13117669 0.05794183 0.05319874 0.04911876 0.08776536
 0.04323852 0.02650191 0.11461536 0.28906106 0.07803061 0.03735408
 0.13403531 0.0343706  0.17840012 0.05496864 0.08341722 0.07439708
 0.06892796 0.15936613 0.05021095 0.02298455 0.10826486 0.15002391
 0.09607602 0.08486772 0.94952447 1.10455066 0.94189156 1.01213539
 0.9939272  0.97959991 0.98274087 0.9438953  0.93246599 0.85503584
 0.99206036 1.07134819 1.15121215 1.02795644 0.91974752 0.968604
 0.79342024 0.99531486 0.92938227 0.9485021  0.00577397 0.07329029
 0.17116082 0.22104769 0.08531827 0.02427434 0.11967682 0.00510069
 0.1297011  0.11775567 0.03078041 0.16429437 0.11457015 0.09692613
 0.11108234 0.11180809 0.03362144 0.2167035  0.18201885 0.17533213
 0.12456165 0.0620643  0.13531726]
```

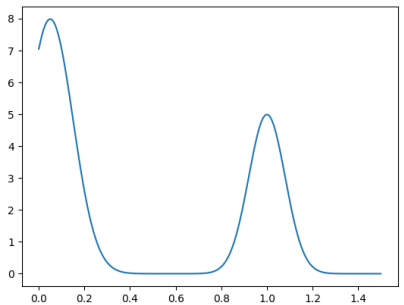

Data extraction

- **In fact, in reality, the 1's and 0's look like normal distibutions**

```
plt.hist(peakvalues,bins=50)
x_axis = np.arange(0,1.5,0.001)
plt.plot(x_axis, 2*norm.pdf(x_axis, 0.05, 0.1)+norm.pdf(x_axis,1,0.08))
plt.show
```



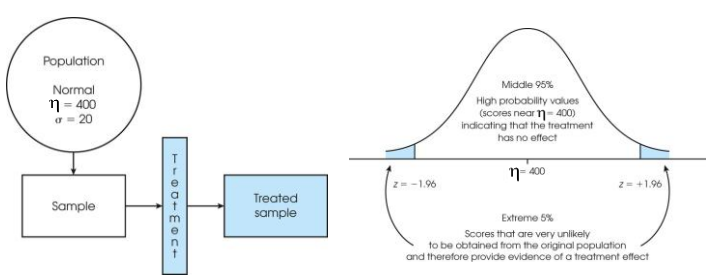
How would we estimate error rate probabilities?



Probability and Inferential Statistics

- **Probability is important because it establishes a link between samples and populations.**
- **For any known population it is possible to determine the probability of obtaining any specific sample.**
- **We will use this link as the foundation for inferential statistics.**
- **The general goal of inferential statistics is to use the information from a sample to reach a general conclusion (inference) about an unknown population.**
- **Typically a researcher begins with a sample.**
- **If the sample has a high probability of being obtained from a specific population, then the researcher can conclude that the sample is likely to have come from that population.**
- **If the sample has a very low probability of being obtained from a specific population, then it is reasonable for the researcher to conclude that the specific population is probably not the source for the sample.**

Checking if an experimental factor affects a measured response



The missing information in samples

- **We may know the mean of the population, but we rarely know the variance**

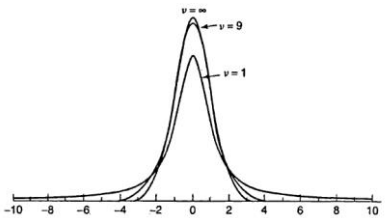
Definition	*Population*: a hypothetical set of N observations from which the sample of observations actually obtained can be imagined to come (typically N is very large)	*Sample*: a set of n available observations (typically n is small)
	Parameters	**Statistics**
Measure of location	Population mean $\eta = \sum y/N$	Sample average $\bar{y} = \sum y/n$
Measure of spread	Population variance $\sigma^2 = \sum(y-\eta)^2/N$	Sample variance $s^2 = \sum(y-\bar{y})^2/(n-1)$
	Population standard deviation $\sigma = +\sqrt{\sum(y-\eta)^2/N}$	Sample standard deviation $s = +\sqrt{\sum(y-\bar{y})^2/(n-1)}$

- **This means that we cannot easily determine the z score of a measurement. We define a new term, t, which is determined from the population mean and the sample variance**

$$z = \frac{y-\eta}{\sigma} \quad \text{vs} \quad t = \frac{y_0-\eta}{s}$$

Effect of sample size

- As we increase the number of samples, the distribution of the samples approaches the normal distribution *if the sampling is random*



- This sample distribution is called the student's t distribution.
  - The original discoverer was W. S. Gosset, who wrote under the pseudonym "student".

Repeated sampling

- In many experiments, we draw repeated independent samples from a general population
  - E.g.,
    - take a bucket full of numbered balls (the population)
    - Extract n random balls, calculate the sample mean $\bar{y}$, and put the balls back
    - Repeat multiple times
- This is called independent identically distributed observations.
- The distribution of the sample means have several interesting properties

$$E(\bar{y}) = \eta, \qquad V(\bar{y}) = \frac{\sigma^2}{n}$$

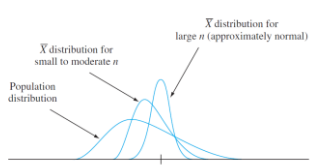
	Parent Distribution for Observations y	Sampling Distribution for Averages $\bar{y}$
Mean	η	η
Variance	σ^2	σ^2/n
Standard deviation	σ	$\sigma/\sqrt{n}$
Form of parent distribution	Any*	More nearly normal than the parent distribution

*This statement applies to all parent distributions commonly met in practice. It is not true for certain mathematical toys (e.g., the Cauchy distribution), which need not concern us here.

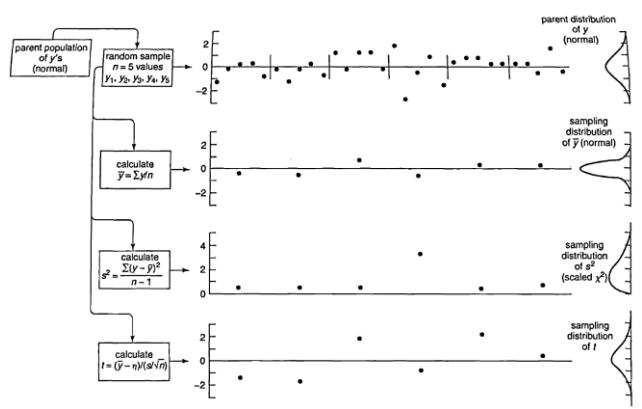
The central limit theorem and normality

The Central Limit Theorem (CLT)

Let $X_1, X_2, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Then if n is sufficiently large, $\bar{X}$ has approximately a normal distribution with $\mu_{\bar{X}} = \mu$ and $\sigma^2_{\bar{X}} = \sigma^2/n$, and T_o also has approximately a normal distribution with $\mu_{T_o} = n\mu$, $\sigma^2_{T_o} = n\sigma^2$. The larger the value of n, the better the approximation.



Special case: random sampling from a normal distribution

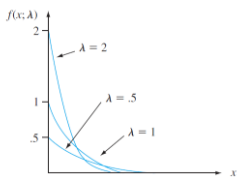


Notice that we could calculate a t-distribution of the $\bar{y}$ population purely from the s values of the individual samples. This is only possible because we are sampling from a normal distribution.. This is a useful fact since we often don't have access to the population σ, etc.

Other continuous distributions: The exponential distribution

X is said to have an **exponential distribution** with (scale) parameter λ ($\lambda > 0$) if the pdf of X is

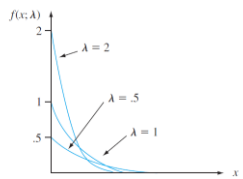
$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$



Other continuous distributions: The chi-squared distribution

X is said to have an **exponential distribution** with (scale) parameter λ ($\lambda > 0$) if the pdf of X is

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$



Others that are important: Chi-squared, Weibull, etc.

Joint Probability: Discrete Random Variables

- **Joint probability deals with determination of probabilities for 2 or more random variables**
- **Reminder: 2 gas station example**
  - Two gas stations are located at a certain intersection. Each one has six gas pumps. An experimental outcome specifies how many pumps are in use at the first station and how many are in use at the second one.
  - The sample space was

		Second Station						
		0	**1**	**2**	**3**	**4**	**5**	**6**
	0	(0, 0)	(0, 1)	(0, 2)	(0, 3)	(0, 4)	(0, 5)	(0, 6)
	1	(1, 0)	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
	2	(2, 0)	(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
First Station	**3**	(3, 0)	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
	4	(4, 0)	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
	5	(5, 0)	(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
	6	(6, 0)	(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)

  - These are discrete random variables, since we cannot have a fractional pump use, and we could calculate probabilities of outcomes.

Joint Probability: Discrete Random Variables

- **Joint probability deals with determination of probabilities for 2 or more random variables**

Let X and Y be two discrete rv's defined on the sample space $\mathcal{S}$ of an experiment. The **joint probability mass function** $p(x, y)$ is defined for each pair of numbers (x, y) by

$$p(x, y) = P(X = x \text{ and } Y = y)$$

It must be the case that $p(x, y) \geq 0$ and $\sum_x \sum_y p(x, y) = 1$.

Now let A be any particular set consisting of pairs of (x, y) values (e.g., $A = \{(x, y): x + y = 5\}$ or $\{(x, y): \max(x, y) \leq 3\}$). Then the probability $P[(X, Y) \in A]$ that the random pair (X, Y) lies in the set A is obtained by summing the joint pmf over pairs in A:

$$P[(X, Y) \in A] = \sum_{(x,y) \in A} \sum p(x, y)$$

Joint Probability: Discrete Random Variables

- **Reminder: 2 gas station example**
  - Assuming equal probability of use for all pumps, we have:

		Second Station						
		0	**1**	**2**	**3**	**4**	**5**	**6**
First Station	**0**	0.0204	0.0204	0.0204	0.0204	0.0204	0.0204	0.0204
	1	0.0204	0.0204	0.0204	0.0204	0.0204	0.0204	0.0204
	2	0.0204	0.0204	0.0204	0.0204	0.0204	0.0204	0.0204
	3	0.0204	0.0204	0.0204	0.0204	0.0204	0.0204	0.0204
	4	0.0204	0.0204	0.0204	0.0204	0.0204	0.0204	0.0204
	5	0.0204	0.0204	0.0204	0.0204	0.0204	0.0204	0.0204
	6	0.0204	0.0204	0.0204	0.0204	0.0204	0.0204	0.0204

  - What is the probability of 4 pumps being used in total?
  - What is the probability of no pumps being available?

Another example

- Consider the case of buying insurance. As you know, you can choose the amount of the franchise.
- Suppose we look at the deductibles chosen by a random customer for health (x) and automobile (y) insurance and we find:

	$p(x, y)$	y = 500	y = 1000	y = 5000
	100	.30	.05	0
x	500	.15	.20	.05
	1000	.10	.10	.05

- To proceed, we must (1) confirm this is a valid probability table:
  - All values are > 0
  - Sum is 1
- What is the probability that the auto insurance franchise is >= CHF500?

Marginal Probability

- The marginal probability is simply the probability function for one variable for the particular condition of another variable

The **marginal probability mass function of X,** denoted by $p_X(x)$, is given by

$$p_X(x) = \sum_{y:\, p(x,y)>0} p(x, y) \quad \text{for each possible value } x$$

Similarly, the **marginal probability mass function of Y** is

$$p_Y(y) = \sum_{x:\, p(x,y)>0} p(x, y) \quad \text{for each possible value } y.$$

- For example:

	$p(x, y)$	y = 500	y = 1000	y = 5000
	100	.30	.05	0
x	500	.15	.20	.05
	1000	.10	.10	.05

- $p_x(x)$ = 0.35 @ x=100, 0.4 @ x=500, 0.25 @ x=1000, 0 otherwise.

Joint Probability: Continuous Random Variables

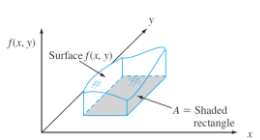
- **Joint probability deals with determination of probabilities for 2 or more continuous variables**

Let X and Y be continuous rv's. A **joint probability density function** $f(x, y)$ for these two variables is a function satisfying $f(x, y) \geq 0$ and $\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)\,dx\,dy = 1$. Then for any two-dimensional set A

$$P[(X, Y) \in A] = \iint_A f(x, y)\,dx\,dy$$

In particular, if A is the two-dimensional rectangle $\{(x, y): a \leq x \leq b, c \leq y \leq d\}$, then

$$P[(X,Y) \in A] = P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f(x, y)\,dy\,dx$$



Marginal Probability: Continuous Random Variables

The **marginal probability density functions** of X and Y, denoted by $f_X(x)$ and $f_Y(y)$, respectively, are given by

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y)\,dy \qquad \text{for } -\infty < x < \infty$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y)\,dx \qquad \text{for } -\infty < y < \infty$$

Example: Visiting McDonald's

- **Suppose you go to the McDonald's in Villeneuve, which has both a walk-in (y) and drive-through (x) service. Suppose the joint PDF that X,Y are in use is:**

$$f(x, y) = \begin{cases} \frac{6}{5}(x + y^2) & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

- **First, verify that this is a legitimate PDF:**
  - **Never negative**
  - **Total:**

$$= \int_0^1 \int_0^1 \frac{6}{5}(x + y^2)\,dx\,dy$$

$$= \int_0^1 \int_0^1 \frac{6}{5} x\,dx\,dy + \int_0^1 \int_0^1 \frac{6}{5} y^2\,dx\,dy$$

$$= \int_0^1 \frac{6}{5} x\,dx + \int_0^1 \frac{6}{5} y^2\,dy = \frac{6}{10} + \frac{6}{15} = 1$$

Example: Visiting McDonald's

$$f(x, y) = \begin{cases} \frac{6}{5}(x + y^2) & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

- **What is the probability that neither walk-up or drive-through is busy more than ¼ of the time?**

$$P\left(0 \le X \le \frac{1}{4}, 0 \le Y \le \frac{1}{4}\right) = \int_0^{1/4} \int_0^{1/4} \frac{6}{5}(x + y^2)\,dx\,dy$$

$$= \frac{6}{5} \int_0^{1/4} \int_0^{1/4} x\,dx\,dy + \frac{6}{5} \int_0^{1/4} \int_0^{1/4} y^2\,dx\,dy$$

$$= \frac{6}{20} \cdot \frac{x^2}{2}\Big|_{x=0}^{x=1/4} + \frac{6}{20} \cdot \frac{y^3}{3}\Big|_{y=0}^{y=1/4} = \frac{7}{640}$$

$$= .0109$$

Statistical dependence

In many experiments, we need to examine multiple responses and how they vary with each other. For example:

- **Suppose you were considering two characteristics, for example, the height y1 in cm and the weight y2 in kg of the population**
  - **distribution of heights p(y1); Distribution of weights p(y2)**
  - **the probability distribution of the weights of all people who were 150 cm tall. This distribution is written as p(y2|y1 = 150).**
  - **You would expect the conditional distribution p(y2 | y1 = 150) to be quite different from p(y2 | Y1 = 175)**
  - **Y1 and Y2 would be said therefore to be statistically dependent.**
  - **The likelihood of finding someone with a specific weight and height would be:**

$$p(y_1, y_2) = p(y_1) \times p(y_2 \mid y_1) = p(y_2) \times p(y_1 \mid y_2)$$

- **Now suppose that y3 was a measure of the IQ of the recruit. Y3 would be statistically independent of Y1, such that**

$$p(y_3|y_1) = p(y_3)$$

- **For statistically, independent variables, the probability of achieving some specific combination of y1 and y2 values is** $p(y_1, y_2) = p(y_1) \times p(y_2)$

Independence of RVs

Two random variables X and Y are said to be **independent** if for every pair of x and y values

$$p(x, y) = p_X(x) \cdot p_Y(y) \qquad \text{when } X \text{ and } Y \text{ are discrete}$$

or

$$f(x, y) = f_X(x) \cdot f_Y(y) \qquad \text{when } X \text{ and } Y \text{ are continuous}$$

If (5.1) is not satisfied for all (x, y), then X and Y are said to be **dependent.**

- Going back to the insurance example:
- Are x and y independent?

	$p(x, y)$	y = 500	1000	5000
x	100	.30	.05	0
	500	.15	.20	.05
	1000	.10	.10	.05

- If independent, then for all values of x and y, we should see that $p(x,y) = p_x(x) * p_x($ Check:
  - p(100,500) = 0.3, $p_x(100)$=0.35, $p_y(500)$ = 0.55
  - p(100,500) ≠ $p_x(100) * p_y(500)$, so therefore, not independent

Conditional Probability

Let X and Y be two continuous rv's with joint pdf $f(x, y)$ and marginal X pdf $f_X(x)$. Then for any X value x for which $f_X(x) > 0$, the **conditional probability density function of Y given that $X = x$** is

$$f_{Y|X}(y|x) = \frac{f(x, y)}{f_X(x)} \qquad -\infty < y < \infty$$

If X and Y are discrete, replacing pdf's by pmf's in this definition gives the **conditional probability mass function of Y when $X = x$.**

- Going back to the McDonald's example:

$$f(x, y) = \begin{cases} \frac{6}{5}(x + y^2) & 0 \le x \le 1, 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

- If we want to find the probability that the walk-up is busy at most half the time given that the drive-through is 0.8 is:

$$f_{Y|X}(y|.8) = \frac{f(.8, y)}{f_X(.8)} = \frac{1.2(.8 + y^2)}{1.2(.8) + .4} = \frac{1}{34}(24 + 30y^2) \quad 0 < y < 1$$

$$P(Y \le .5 | X = .8) = \int_{-\infty}^{.5} f_{Y|X}(y|.8)\,dy = \int_0^{.5} \frac{1}{34}(24 + 30y^2)\,dy = .390$$

Expected values and Covariance

Let X and Y be jointly distributed rv's with pmf $p(x, y)$ or pdf $f(x, y)$ according to whether the variables are discrete or continuous. Then the expected value of a function $h(X, Y)$, denoted by $E[h(X, Y)]$ or $\mu_{h(X,Y)}$, is given by

$$E[h(X, Y)] = \begin{cases} \sum_x \sum_y h(x, y) \cdot p(x, y) & \text{if } X \text{ and } Y \text{ are discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} h(x, y) \cdot f(x, y)\,dx\,dy & \text{if } X \text{ and } Y \text{ are continuous} \end{cases}$$

Covariance is used to estimate the degree of linear independence of two variables:

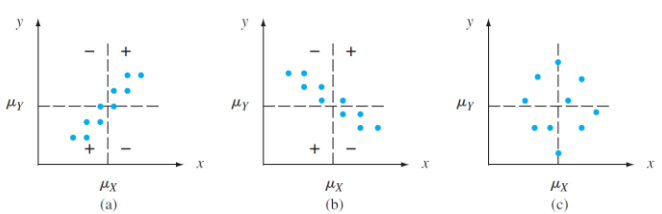
The **covariance** between two rv's X and Y is

$$\mathrm{Cov}(X, Y) = E[(X - \mu_X)(Y - \mu_Y)]$$

$$= \begin{cases} \sum_x \sum_y (x - \mu_X)(y - \mu_Y) p(x, y) & X, Y \text{ discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)(y - \mu_Y) f(x, y)\,dx\,dy & X, Y \text{ continuous} \end{cases}$$

$$\mathrm{Cov}(X, Y) = E(XY) - \mu_X \cdot \mu_Y$$

Covariance and correlation



Positive covariance | Negative Covariance | approx. zero covariance

The **correlation coefficient** of X and Y, denoted by $\mathrm{Corr}(X, Y)$, $\rho_{X,Y}$, or just ρ, is defined by

$$\rho_{X,Y} = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \cdot \sigma_Y}$$

Covariance and correlation in Python

- Python has built-in functions to calculate covariance and correlation.

```
np.cov(x,y))
np.corrcoef(x, y))
```

- The output is a matrix of form [i,j] where the i,i terms are variances or self-correlations and the i,j terms are covariances and correlations

See python code in module2_covariance.ipynb

Estimation

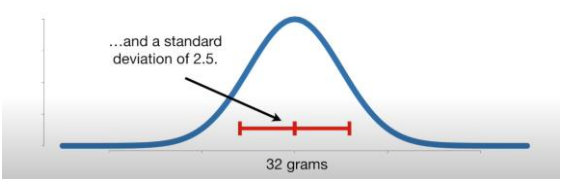
- **So far, we have discussed the probability of finding specific outcomes.**
- **Estimation is a complementary concept, where we have a specific outcome (e.g., a particular data set) and we want to find a distribution that from which that data set could have been obtained.**
- **Formally:**

> A **point estimate** of a parameter θ is a single number that can be regarded as a sensible value for θ. It is obtained by selecting a suitable statistic and computing its value from the given sample data. The selected statistic is called the **point estimator** of θ.

- **We will introduce 2 methods:**
  - **Method of moments**
  - **Maximum likelihood estimation**

Likelihood vs Probability

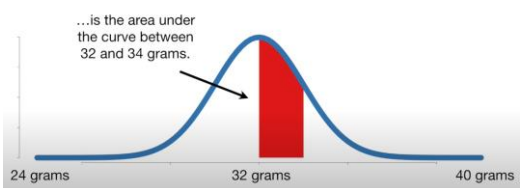
- **Consider a distribution of the weights of screws. The weights are normally distributed as below**



- **Why is this normally distributed?**

Review: Probability

- **The probability of finding a screw with a weight between 32 and 34 grams**



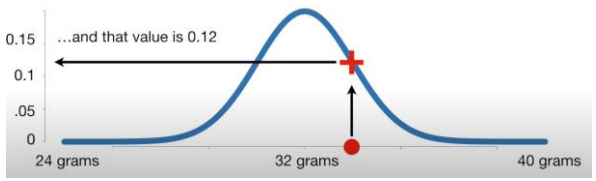
- **which is in fact, 0.29**
- **We can write this as:**

pr(weight between 32 and 34 grams | mean = 32 and standard deviation = 2.5)

This symbol means «given»

Likelihood

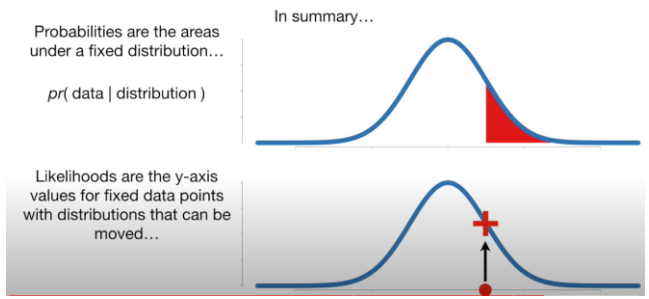
- **Suppose we measure a screw and find its weight to be 32 grams**
- **Then, the likelihood of measuring a 32 gram screw is the corresponding y-axis value for that date point**



- **We can write this as:**

$L(\text{mean} = 32 \text{ and standard deviation} = 2.5 \mid \text{screw weighs 34 grams})$

Likelihood vs Probability



In estimation, we know the data and want to find the distribution. Our goal then, is to maximize the likelihood that the data is from the distribution by picking the appropriate distribution parameters

Method of Moments

- **In the Method of Moments, we calculate "moments" from the known data and equate those to "moments" from the theoretical distribution.**
- **We then solve for the distribution parameters to obtain a distribution that is an estimate based on the data**
- **For example, consider a normal distribution:**

$$f(x_i; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

- **The parameters of this distribution are σ and μ. We will use the method of moments to find values of these so that we have an estimate for these values that correspond to a known data set**

Definition: Moments

- **Moments are defined for both the distribution and the sample data:**

1. $E(X^k)$ is the k^{th} (theoretical) moment of the distribution (**about the origin**), for $k = 1, 2, \ldots$
2. $E\left[(X-\mu)^k\right]$ is the k^{th} (theoretical) moment of the distribution (**about the mean**), for $k = 1, 2, \ldots$
3. $M_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$ is the k^{th} sample moment, for $k = 1, 2, \ldots$
4. $M_k^* = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^k$ is the k^{th} sample moment about the mean, for $k = 1, 2, \ldots$

- **In the method of moments, we just equate the theoretical moments to the sample moments. We need as many moments as there are paramteres in the distribution equation (for example, in the normal distribution we need 2 moments, since there are two parameters, σ and μ.**
- **This gives us a set of simultaneous equations, which we solve for the unknown parameters**
- **We denote these parameter estimates with the "hat" (^) symbol, to indicate that they are estimates, i.e., $\hat{\sigma}$ and $\hat{\mu}$**

MoM Example 1: Bernoulli random variables

- Consider a set of n data points taken from a Bernoulli distribution. As a reminder, these take the form:

$$f(k;p) = p^k(1-p)^{1-k} \quad \text{for } k \in \{0,1\}$$

  - Where p is the probability of obtaining 1.

- Suppose we have obtained the following data set from 10 samples:
  - 0, 1, 1, 0, 1, 1, 1, 0, 1, 0
- There is only 1 parameter to this distribution (p), so we only need the first theoretical and sample moment, which, for this distribution is: $E(X_i) = p$
- We can calculate the sample moment as: $p = \frac{1}{n}\sum_{i=1}^{n} X_i$

  = 0.6
- Therefore, the MoM estimate $\hat{p} = 0.6$, which we can use in the f equation above to identify the MoM-derived distribution from which the data likely came.

MoM Example 2: Normal random variables

- Consider a set of n data points taken from a Normal distribution:

$$f(x_i;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(x_i-\mu)^2}{2\sigma^2}\right]$$

- Suppose we have obtained the following data set from 3 samples:
  - x_1 = -2.321, x_2 = 1.112, and x_3 = -5.221
- For the normal distribution, the 1st and 2nd theoretical moments are:

$$E(X_i) = \mu \qquad E(X_i^2) = \sigma^2 + \mu^2$$

- The experimental moments are:

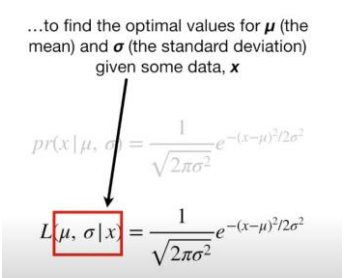
$$E(X) = \mu = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad E(X^2) = \sigma^2 + \mu^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

  = -2.143 = 11.29
- Which means, the estimators are: $\hat{\mu} = -2.143$ and $\hat{\sigma} = 2.588$
- Note that this standard deviation is biased, which is a limitation of this estimation method

See python code in module2_MoM.ipynb

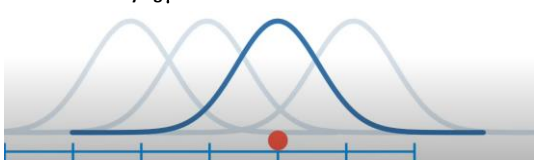
Maximum likelihood estimation (MLE)

- In MLE, we maximize the likelihood of the known data occurring, by adjusting the parameters of the distribution. For example, in a normal distribution, we would use MLE:

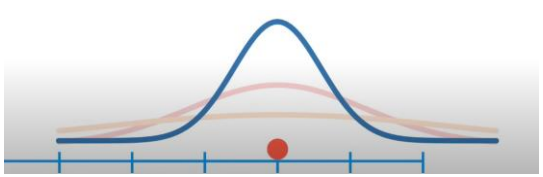


Maximum likelihood estimation (MLE) – single data point

- Given the known data point in red, which of the curves has the maximum likelihood for varying μ?



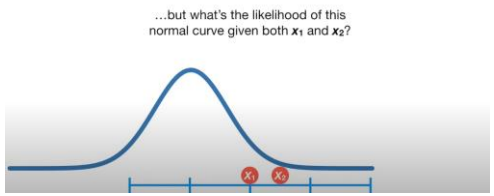
- And for varying σ?



Maximum likelihood estimation (MLE) – multiple data points

- **We can individually calculate the likelihood for each of the data points**

$$L(\mu = 28,\ \sigma = 2\,|\,x_1 = 32) \qquad L(\mu = 28,\ \sigma = 2\,|\,x_2 = 34)$$



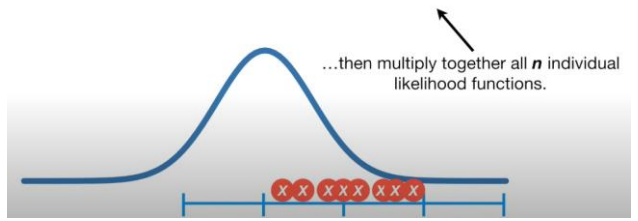
- **HINT: The samples are assumed to be independent.**

$$L(\mu = 28,\ \sigma = 2\,|\,x_1 = 32 \text{ and } x_2 = 34)$$
$$= L(\mu = 28,\ \sigma = 2\,|\,x_1 = 32) \times L(\mu = 28,\ \sigma = 2\,|\,x_2 = 34)$$

Maximum likelihood estimation (MLE) – multiple data points

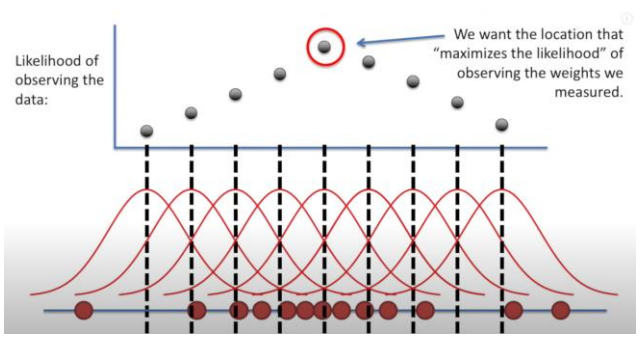
- **Extending to n data points, we:**

$$L(\mu,\ \sigma\,|\,x_1,\ x_2,\ \ldots,\ x_n) = L(\mu,\ \sigma\,|\,x_1) \times \ldots \times L(\mu,\ \sigma\,|\,x_n)$$
$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_1-\mu)^2/2\sigma^2} \times \ldots \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_n-\mu)^2/2\sigma^2}$$



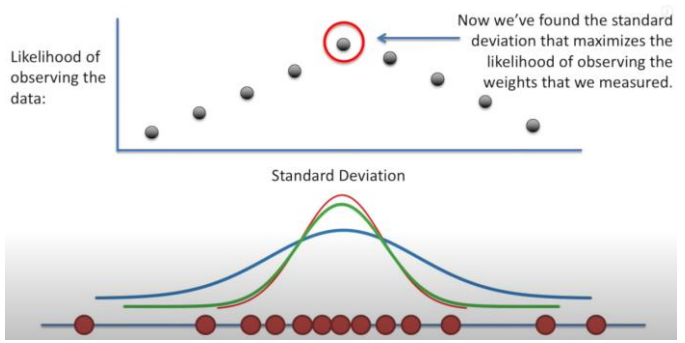
Maximum likelihood estimation (MLE)

- **More generally, if we several data points, we can calculate the overall likelihood and plot it**



Maximum likelihood estimation (MLE)

- **More generally, if we several data points, we can calculate the overall likelihood and plot it**



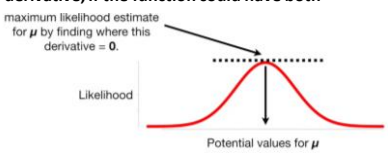
Maximum likelihood estimation (MLE) - methodology

- **Define the likelihood function (e.g., for a normal distribution)**

$$L(\mu, \sigma | x_1, x_2, \ldots, x_n) = L(\mu, \sigma | x_1) \times \ldots \times L(\mu, \sigma | x_n)$$

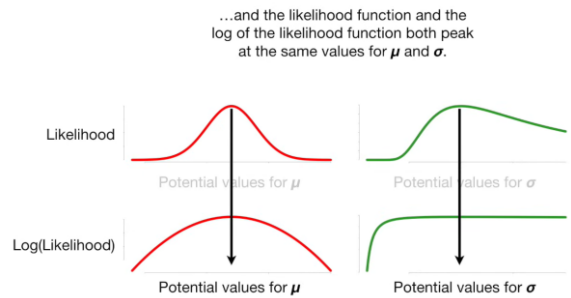
$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_1-\mu)^2/2\sigma^2} \times \ldots \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_n-\mu)^2/2\sigma^2}$$

- **And find derivatives of the function for each parameter (e.g, μ and σ), assuming other parameters are constant**
- **Then, we find the maximum by finding the point where the derivative goes to zero (technically, we should then verify it is a maximum and not a minimum by taking the 2nd derivative, if the function could have both maxima and minima)**



Maximum likelihood estimation (MLE) - methodology

- **Commonly, we actually take the log of the likelihood function since the math is often much easier**



Detailed example: Normal Distribution Log Transformation

- **Using the log converts the Π into Σ**

$$\ln[L(\mu, \sigma | x_1, \ldots, x_n)]$$

$$= \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_1-\mu)^2/2\sigma^2} \times \ldots \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_n-\mu)^2/2\sigma^2}\right)$$

$$= \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_1-\mu)^2/2\sigma^2}\right) + \ldots + \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_n-\mu)^2/2\sigma^2}\right)$$

...into addition.

Detailed example: Normal Distribution Log Transformation

- **We will go through the first term in detail, but every term is similar:**

$$\ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} \times e^{-(x_1-\mu)^2/2\sigma^2}\right)$$

$$= \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \ln\left(e^{-(x_1-\mu)^2/2\sigma^2}\right)$$

$$= \ln\left[(2\pi\sigma^2)^{-1/2}\right] - \frac{(x_1-\mu)^2}{2\sigma^2}\ln(e)$$

$$= -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(x_1-\mu)^2}{2\sigma^2}$$

$$= -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\ln(\sigma^2) - \frac{(x_1-\mu)^2}{2\sigma^2}$$

$$= -\frac{1}{2}\ln(2\pi) - \ln(\sigma) - \frac{(x_1-\mu)^2}{2\sigma^2}$$

Detailed example: Normal Distribution Log Transformation

- **So, we have:**

$$\ln\left[L(\mu, \sigma \mid x_1, \ldots, x_n)\right] = \ln\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_1-\mu)^2/2\sigma^2} \times \ldots \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x_n-\mu)^2/2\sigma^2}\right)$$

$$= -\frac{1}{2}\ln(2\pi) - \ln(\sigma) - \frac{(x_1-\mu)^2}{2\sigma^2} - \ldots - \frac{1}{2}\ln(2\pi) - \ln(\sigma) - \frac{(x_n-\mu)^2}{2\sigma^2}$$

$$= -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{(x_1-\mu)^2}{2\sigma^2} - \ldots - \frac{(x_n-\mu)^2}{2\sigma^2}$$

Detailed example: Normal Distribution Log Transformation

- **So, we have:**

$$\ln\left[L(\mu, \sigma \mid x_1, \ldots, x_n)\right] = -\frac{n}{2}\ln(2\pi) - n\ln(\sigma) - \frac{(x_1-\mu)^2}{2\sigma^2} - \ldots - \frac{(x_n-\mu)^2}{2\sigma^2}$$

- **We take derivatives with respect to μ and σ. The math is involved, but the answer is:**

$$\frac{\partial}{\partial\mu}\ln\left[L(\mu, \sigma \mid x_1, \ldots, x_n)\right] = \frac{1}{\sigma^2}\left[(x_1 + \ldots + x_n) - n\mu\right]$$

$$\frac{\partial}{\partial\sigma}\ln\left[L(\mu, \sigma \mid x_1, \ldots, x_n)\right] = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\left[(x_1-\mu)^2 + \ldots + (x_n-\mu)^2\right]$$

- **Set these to zero to find the maxima**
  - **In fact, for normal distributions, the answer is simply μ and**

Detailed example: Normal Distribution Log Transformation

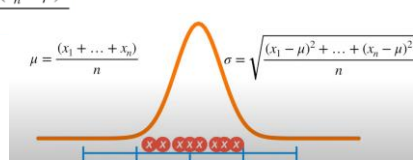
- **To find the maxima, we set to zero**

$$\frac{\partial}{\partial\mu}\ln\left[L(\mu, \sigma \mid x_1, \ldots, x_n)\right] = \frac{1}{\sigma^2}\left[(x_1 + \ldots + x_n) - n\mu\right]$$

$$\frac{\partial}{\partial\sigma}\ln\left[L(\mu, \sigma \mid x_1, \ldots, x_n)\right] = -\frac{n}{\sigma} + \frac{1}{\sigma^3}\left[(x_1-\mu)^2 + \ldots + (x_n-\mu)^2\right]$$

- $\hat{\mu} = \frac{(x_1 + \ldots + x_n)}{n}$
- $\hat{\sigma} = \sqrt{\frac{(x_1-\mu)^2 + \ldots + (x_n-\mu)^2}{n}}$

Again, the σ estimator is biased



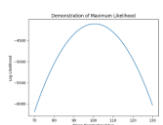
MLE in Python

- **We can generate some synthetic normal data:**

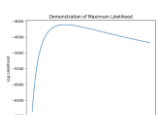
```
data = norm.rvs(loc=100,scale=15,size=1000,random_state=1)
```

- **We can use the log pdf of the normal distribution and plot for the mean and SD:**

```
def likelihood(params,data):
    return norm.logpdf(data,loc=params[0],scale=params[1]).sum()
x = np.linspace(70,130,1000)
y = [likelihood([val,15],data) for val in x]
plt.plot(x,y)
```



```
x = np.linspace(5,40,1000)
y = [likelihood([100,val],data) for val in x]
plt.plot(x,y)
```



- **We can use the "minimize" function (actually, negative of it to maximize) to find the MLE values**

```
def neglikelihood(params,data):
    return -1*likelihood(params,data)
result = minimize(neglikelihood,[90,10],args=(data))
print(result)
```

See python code in module2_MLE.ipynb

Other estimators

- **Exponential function:** $f_X(x_j) = \begin{cases} \lambda_0 \exp(-\lambda_0 x_j) & \text{if } x_j \in R_X \\ 0 & \text{otherwise} \end{cases}$
- **Estimator:** $\hat{\lambda}_n = \frac{n}{\sum_{j=1}^{n} x_j}$
- **Other examples can be found at:**

https://www.statlect.com/fundamentals-of-statistics/normal-distribution-maximum-likelihood