1 Recall the story about the two Duke University students and the flat tire.

As we worked out on the first quiz, assuming each guesses a tire with probability 1/4, the probability that they will both guess the same tire is 1/4.

The following question was asked of a class of students. "I was driving to school today, and one of my tires went flat. Which tire do you think it was?" The responses were as follows: right front, 40%, left front, 20%, right rear, 30%, left rear, 10%. Suppose that this distribution holds in the general population, and assume that the two Duke students are randomly chosen from the general population. What is the probability that they will give the same answer to the "which tire" question?

2 Some 10% (on average) of people who make reservations do not show up for their flights

- If the airline takes 30 reservations, the probability that 0,1,2, ... passengers will not show up is given by what probability distribution? and with what parameters?
- How well can this distribution be approximated by a Gaussian ("Normal") distribution? Explain.

3 Soft drink cans are labelled as containing 355 ml of product. In fact, the automated canning process fills cans with varying amounts of liquid. Suppose that the actual fills of these soft drink cans are normally distributed with a mean of 357 ml and a standard deviation of 2 ml.

- In the second quiz, we calculated that the probability that a randomly selected can will contain less than the promised 355 ml of soft drink is approximately 16%. Suppose a person purchases a 6-pack of this soft drink. What is the probability that at least one of the cans purchased will contain less than the promised 355 ml?

  Do you have any reservations about the assumptions you have made?
- If a person purchases a 6-pack of this soft drink, what is the probability that the average fill per can purchased will be less than the promised 355 ml?

4 Some 40% of 50 year old smokers have severe facial wrinkles[1]. Some 20% of 50 year old non-smokers have severe facial wrinkles. Some 30% of 50-year olds smoke.

- Person W is a 50 year old who has severe serious facial <u>w</u>rinkles. What is the probability that that person W is a smoker?
- Person N is a 50 year old who does <u>n</u>ot have severe facial wrinkles. What is the probability that person N is a smoker?
- The above 30% who smoke applies to Montreal. In Toronto, only 25% of persons smoke (but the 40% and 20% rates of wrinkles still apply). Is the probability that "Toronto person W" is a smoker higher / lower / the same as "Montreal person W"?

5 A person never puts money in a 25-cent parking meter, assuming that each time there is a probability of .05 of getting caught. The first offense costs nothing, the second costs 5 dollars, and subsequent offenses cost 20 dollars each. Under these assumptions, how does the

[1]* Altavista translates "wrinkles" as «rides».

expected cost of parking 100 times without paying the meter compare with the cost of paying the meter each time?

6. On the average, hotel guests weigh about 150 pounds with an SD of 25 pounds. An engineer is designing a large elevator for a convention hotel, to lift 100 people. If he designs it to lift 15,500 pounds, the chance it will be overloaded by a random group of 100 people is closest towhich of the following: 0.1 of 1%, 2%, 5%, 50%, 95%, 98%, 99.9% ? Explain. your reasoning.

7 A study of rush hour traffic in Montreal records the number Y of people in each car crossing the Champlain Bridge. Suppose that this number Y has mean 1.5 and standard deviation 0.75 in the population of all cars that cross the bridge during rush hours.

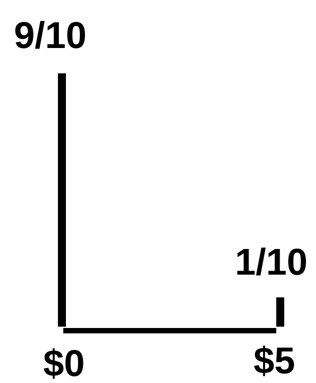
a Does the count Y have a binomial distribution? Why or why not?

b Could the exact distribution of Y be normal? Why or why not?

c Traffic engineers estimate that the capacity of the bridge is 700 cars per hour. According to the central limit theorem, what is the approximate distribution of the mean number of persons (y-bar) in 700 randomly selected cars crossing the bridge?

d The count of people in 700 cars is 700 times y-bar. Use your result from (c) to give an approximate distribution for the count. What is the probability that 700 cars will carry more than 1075 people?

8 • For the random variable shown below, calculate (i) the expected value (ii) the variance and (iii) the standard deviation.



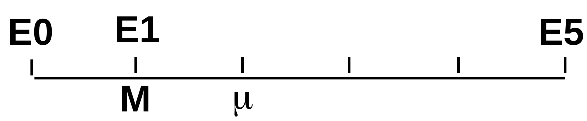
Suppose S is the sum of 2 iindependent random variables with the above distribution. Calculate (i) the distribution (ii) the expected value and (iii) the variance of the sum. Does its distribution look like any distribution we know?

9 If you try out the life expectancy calculator the link to which is given on the class web page, you will get the followning:

Age	Life Expectancy (Male)	Life Expectancy (Female)
0	76	81
20	77	82
40	78	83
60	80	84
80	87	89
90	94	95
99	101	101

- Explain this (to some at least) seeming paradox of increasing life expectancy for those with a higher completed age. Hint:Your answer should contain one of the following words "marginal", "conditional", "joint", "mutually exclusive", "Gaussian", "memoryless".

10 Three elevators are spaced unevenly along a wall, as illustrated below. They are named E0, E1, and E5 because of their distances from the leftmost one. It is equally that the next elevator to arrive is E0, E1, or E5 (and there is no indicator light!).



a If every day you arrive, you stand at M (i.e. the median E2 of the 3), what is the expected distance you will have to walk from M to the next elevator that comes? [count horizontal distances only i.e. ignore the fact that you will stand a little out from the wall.]

What is the expected value if you stand at the mean position (μ) of the 3 elevators, i.e., at a position (0+1+5)/3 = 2 units in from the leftmost elevator?

b If every day you arrive, you stand at M (i.e. the median E2 of the 3), what is the expected squared distance you will have to walk from M to the next elevator that comes?

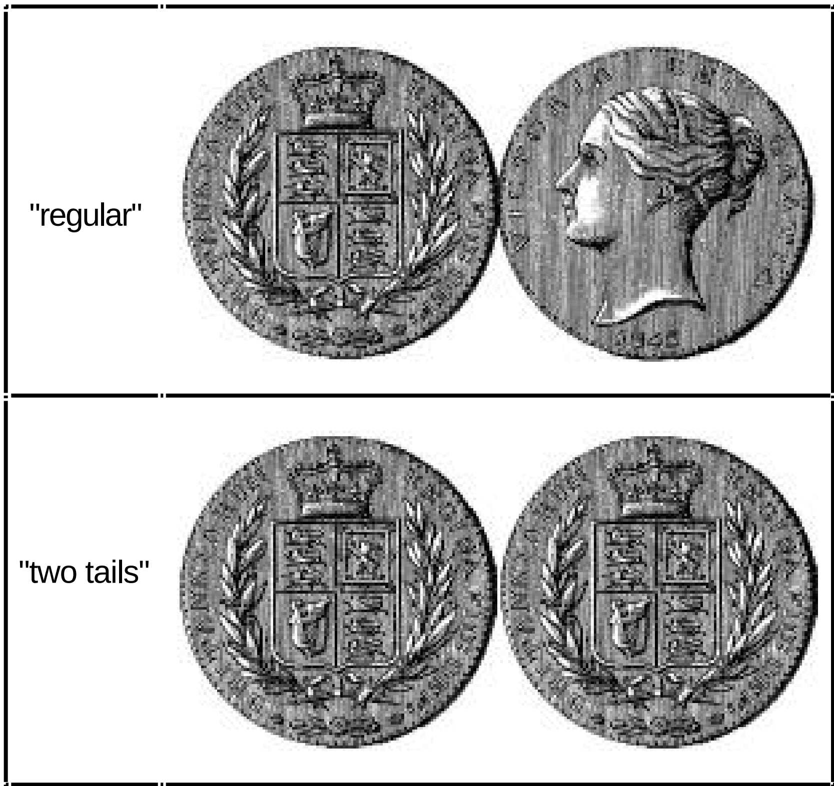
What is the expected squared distance if you stand at the mean position (μ) ?

Which measure of deviation would make you chose M over μ ?

c Say you measure the "loss" as the 4th power of the distance from each elevator to your waiting position. Which waiting position minimizes the average 4th power of distance?

11 Two coins, one "regular" and one with "two tails" (see below) are placed in a hat. One coin is selected at random and tossed twice. You are not allowed to see it until it lands. It lands "tails" both times.

• In light of this outcome, what is the probability that the side facing down is also "tails"? Explain your calculations fully/clearly. *Hint: be careful of intuition -- and draw a tree!*



12 Three coins, two "regular" and one with "two tails" are placed in a hat. One coin is selected at random and tossed. You are not allowed to see it until it lands. It lands "tails".

• In light of this outcome, what is the probability that the side facing down is also "tails"? Explain your calculations fully/clearly. *Hint: be careful of intuition -- and draw a tree!*

13 Sketch the probability distribution (pdf) for the following "<u>intervals</u> (time or distance) between events (locations etc.) " random variables, making thoughtful guesses as to their shape, location, (or, if they can be described by off-the-shelf distributions, the parameters of these, and their mean and variance. Make sure you put a clear scale on the horizontal axis.

the interval between...

- snowstorms in Montreal in January and February
- drips of a leaky faucet
- eruptions of the "Old Faithful" geyser in Yellowstone National Park

t• arrivals of an eastgoing Metro train at a specific station on a certain weekday (data collected over the part of the day when the Metro is in service)

- arrivals at a hospital emergency room between say 8 PM and 10 PM weekdays (it might be different at weekends or after the bars close!)
- road kill along a stretch of highway
- departures of minibuses that leave when all 6 seats have filled up (persons show up randomly)
- maternal deaths at a hospital that performs 5000 deliveries a year (assume that the maternal death rate is approximately 1 per 20,000)
- winning la Quotidienne 3 (if play the "3 different digits, any order" option every day; as we saw, the chance of winning with this option is 6/1000)
- having to replace ink cartridges for an printer (say you buy "2-packs" to save money --each individual cartridge lasts on average 1000 pages, with a Gaussian variation of 100 pages).
- expected total (and average) number of number of persons in 100 cars crossing the Champlain bridge, and SD thereof (60% of cars have 1 occupant, 30% have 2 and 10% have 3.

14 Consider the r.v. *year of publication of a randomly selected book in the McGill libraries*. Sketch what you think the (a) pdf (b) corresponding cdf look like. How do the variance and standard deviation of the random variable: *age of this randomly selected book*, relate to the corresponding variance and standard deviation of the random variable *year of publication of the book*?

The science librarian complains that the physical quality of the books published in computer science compares poorly with that in other disciplines, such as physics and engineering. The conclusion is based on the ages of the books that the library has to send for repair. Even though the average amount the damaged books have been used seems to be the same in the various disciplines, the average age of the damaged books in computer science is lower, i.e. they seem to fall apart sooner. Can you see flaws in the inference?

I'm sure our McGill librarians are too statistically savvy to reason this way. The above story is fictitious, but is based on the following true story. Every week, an older American physician used to read the obituaries (death notices) of physicians in the Journal of the American Medical Association. He wrote to the journal to say that his data analysis showed that the average age at death of the women physicians was a lot younger than that of the men physicians. He argued that these statistics prove his point that "women shouldn't be in medicine -- they can't handle the stress -- it kills them." The same argument was used to falsely conclude, from analysis of ages of death of left and right-handed baseball players, that left-handers die younger.

15 From the data on the length of time to diagnose and treat breast cancer, we saw that the initial mammogram was equally likely to be any one of the weekdays from Monday to Friday. Consider the mammograms performed the first week of October, say Monday October 1 to Friday October 5. Suppose that these women had surgery on one of the weekdays of the last full week of October, i.e., Monday October 22 to Friday October 26. Suppose further that the day of the surgery was also equally likely to be any one of these 5 weekdays, and unrelated to which day the mammogram was performed.

Calculate the probability distribution of the number of days from when a woman had the mammogram until she had the surgery.

Calculate the mean and the variance of this random variable.

Is this another déjà vu distribution? Did it arise in essentially the same way?

16 The place where a medical emergency happens is equally likely to be 1, 2, 3, .. 5 Km from where the nearest ambulance is located, and also equally likely to be 1, 2, ..., 10 Km from the nearest hospital. What is the probability distribution for the total distance the ambulance must travel to get to the injured the person and then transport the person to the hospital?

17 In the question about the total (and average) amount of soft drink in a 6-pack, it was -- unrealistically in my opinion -- assumed that the amount of soft drink in one can in the 6-pack was <u>independent</u> of that in one each of the others. In reality, since the cans in the same 6-pack came from the same "run" or "batch", where the machine setting may have drifted, the amount "over" or "under" μ in one can in the 6-pack is likely to be positively correlated with that in one of the others. i.e. if one can is above μ, the others are likely to be so as well. Say the degree of parities correlation is measured by a positive correlation coefficient of $\rho = +0.5$.

this related-ness can be modeled by saying that, for i,j = 1, 2, .. , 6, and $j \neq i$

Y_i = amount in can i $= \mu + e_i$;

$Var(e_i) = \sigma^2$;

$Correlation(e_i, e_j) = \rho$,

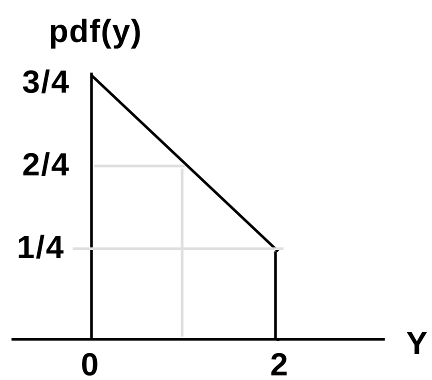
so $Covar(Y_i, Y_j) = Covar(e_i, e_j)$

$= \rho \, SD(Y_i) \, SD(Y_j)$

$= \rho \, \sigma^2$.

Calculate the variance of the sum (and thus the variance of the average [= sum/6]) of the amounts in the 6-pack

18 The random variable Y has pdf(y) = 1.5 – y if $0 \le Y < 1$, and = 0 elsewhere.



Find the pdf of the random variable $X = Y^{1/2}$

19 Some 15 of 20 CD burners are meant to be shipped to Toronto and 5 to Montreal.

If two are shipped to Vancouver by mistake and the "selection" is random, how does one calculate the probability that

a) both tape recorders should have gone to Toronto

b) one should have gone to Toronto and one to Montreal.

20 Seventy percent of the light aircraft that disappear while in flight in a certain country are subsequently discovered. Of the aircraft that are discovered, 60% have an emergency locator, whereas 90% of the aircraft not discovered do not have such a locator.

Suppose a light aircraft has disappeared.

(a) What is the probability that it has an emergency locator and it will not be discovered?

(b) What is the probability that it has an emergency locator?

(c) If it has an emergency locator, what is the probability that it will not be discovered?

21 Suppose a person's waiting time for the bus in the morning has mean 4 minutes and standard deviation 3 minutes, while the waiting time in the evening has mean 6 minutes and standard deviation 4 minutes. In a typical week, the person takes the bus 5 times in the morning and 3 times in the evening.

a Calculate the expected value and standard deviation of the total waiting time in a typical week. State your assumptions carefully.

b The shapes of the morning and afternoon waiting time distributions were not specified. Nevertheless, can you deduce the shape of the distribution of the total waiting time? Again, state your assumptions.

22 The most common form of colorblindness (dichromatism) is a sex-linked hereditary condition caused by a defect on the X chromosome. Thus, it is much more common in males than females; 7% of males are colorblind but only 0.5% of females are colorblind. In a certain population, 40% are male and 60% are female.

a Find the percentage of colorblind persons in the population.

b Find the percentage of colorblind persons that are male.

23 Suppose that there are **h** duck hunters, each a perfect shot. A flock of **d** ducks fly over, and each hunter selects one duck at random and shoots.

a Derive the probability distribution for the number of ducks that are killed.

b Calculate the mean and the standard deviation of the number of ducks that are killed.

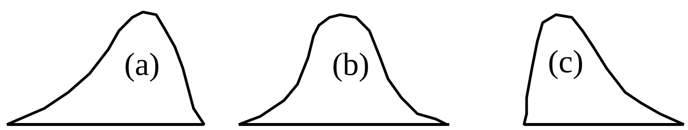
Do the above calculations first for the simplest case of **h** = 2 hunters and **d** = 2 ducks. Then move up to **h** = 2, **d**= 3; **h** = 2, **d**= 3; **h** = 3, **d** = 3; etc....

24 A metal rod is designed to fit into a circular hole on a certain assembly. The radius of the rod (R) is normally distributed with mean 1.000 cm and standard deviation 0.003 cm. The radius of the hole (H) is normally distributed with mean 1.010 cm and standard deviation 0.004 cm. The machining processes that produce the rod and the hole are independent. Find the probability that the rod is too big for the hole. Hint: work with the random variable H – R.

25 A doctor has successfully treated 20 patients for a certain skin condition. Of these, 9 received one drug (A) and 11 received another drug (B). Unpleasant side effects were reported in a total of 8 cases.

If the drugs are equally likely to produce side effects, what is the probability that y of those reporting side effects had received drug BA. Calculate the probabilities for y=0, 1, 2 .. , 8.

27 As part of a survey, a large company asked 1000 of its employees how far they commute to work each day (round trip). The average round trip commute distance was 18 Km, with an SD of 25 Km. Would a rough sketch of the histogram for the data look more like (a) or (b) or (c)? Or is there a mistake somewhere? Explain your answer.



28 Defects in the plating of large sheets of metal occur at random. On the average, there are 2.5 defects per 100 square feet.

find the probability that a sheet 5 feet by 8 feet will have no defects (b) at most one defect.

29 A sociologist decides that it would be useful, though not essential, to include in an experimental group at least 1 person whose annual income exceeds $1 million. The population from which the group is to be randomly selected contains 1% of such people.

How large a group must be selected in order to be 90% sure of including at least 1 such person?

Considerations of time and expense subsequently force the sociologist to limit the group size to 150. How sure can the sociologist now be about including at least one such person?

30 A large number, N, of people are subject to a blood test. This can be administered in two ways.

(i) Each person can be tested separately. In this case N tests are required.

(ii) The blood samples of k people can be pooled and analyzed together. If the test is negative, this one test suffices for the k people. If the test is positive, each of the k persons must be tested separately, and in all k + 1 tests are required for the k people

Assume the probability p that the test is positive is the same for all people and that people are stochastically independent.

a What is the probability that the test for a pooled sample of k people will be positive ?

b What is the expected value of the number, Y, of tests necessary under plan (ii) ?

31 The median of a continuous r. v. Y is a number m such that $P(Y \le m) = P(Y \ge m) = 0.5$.
Find the median in the case of the uniform, Gaussian(Normal) and exponential distributions.

32 Suppose that the distribution of trees in a forest is random, with an average of λ trees per unit area. Let R be the distance from a tree to its nearest neighbour. Show that R has a p.d.f

$$pdf_R(r) = 2\pi \lambda \exp[-\lambda \pi r^2]$$

Hint . take the "complement of the cdf" route.

33 Either a number 11 bus or a number 13 bus will take me to my destination. I always take the first bus that comes. I take the number 11 bus four times as often as the number 13 bus. Does this mean that buses run four times as often on the number 11 route as on the number 13 route?

34 The 3 Gaussian random variables whose pdf's and cdf' are shown below have SD's of 1/2, 1 and 2 respectively. Can you tell which pdf is which, and which cdf is which? Explain your reasoning.