14. Randomness and probability

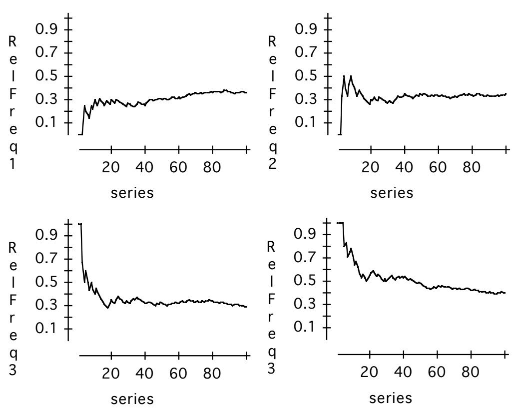
- Random phenomena, independent trials, outcomes, events.
- Law of Large Numbers,
  not to be confused with the "law of averages" fallacy.
- Empirical probability as long run relative frequency.
- Personal probability.
- Formal probability, basic rules.

- Random phenomenon. E.g., source of randomness in traffic light when you reach an intersection.
- Trials, outcomes, events. E.g., green, amber, red.
- Independent trials: outcome on next trial does not depend on outcomes of previous trials.
- Law of Large Numbers: Consider a sequence of independent trials. For any event $\boldsymbol{A}$, define

$$\text{relative frequency of } \boldsymbol{A} = \frac{\text{number of times } \boldsymbol{A} \text{ occurs}}{\text{number of trials}}.$$

  In the long run (as the number of trials grows), the relative frequency converges on a single value. This value is denoted $P(\boldsymbol{A})$ and is called the *empirical probability* of $\boldsymbol{A}$.

Relative frequencies of "Green" for 4 commuters. At 100 trips, relative frequencies are 0.36, 0.35, 0.29, and 0.40 .



- Coin tossing.
  - Buffon (1780): 2048/4040 = 0.5069
  - K. Pearson (1900): 12102/24000 = 0.5005
  - John Kerrich (1944): 5067/10000 = 0.5067
  - There is evidence (sciencenews.org) that a coin will land the same way it started about 51 percent of the time.
- Coin spinning: According to P. Diaconis, a spinning US penny will land as tails about 80 percent of the time because the extra material on the head side shifts the center of mass slightly.
- Murphy's Law.

- The "law of averages" fallacy.
  - Dear Abby (see text): Parents with seven children, all girls, were convinced their eighth child had to be a boy.
  - Seven bad poker hands in a row. I'm due.
  - Lightning never strikes in the same place twice.
  - Unconditional versus conditional probability.
  - The coin has no memory.
  - "In the long run, we are all dead", John Maynard Keynes.
  - Why did card-counting work in Black Jack?

- We can never observe an empirical probability, but we can sometimes calculate a *theoretical probability* based on a relevant model; e.g., equally likely outcomes, Normal models.

- Personal (subjective) probability: measure of personal belief.

Formal probability. Outcomes are elements in a *sample space* $\boldsymbol{S}$; i.e., the set of all possible outcomes. Events are subsets of $\boldsymbol{S}$. Probabilities are numbers assigned to events according to basic rules. Use Venn diagrams.

1. For any event $\boldsymbol{A}$, $0 \leq P(\boldsymbol{A}) \leq 1$.

2. Probability Assignment Rule: $P(\boldsymbol{S}) = 1$.

3. Complement Rule: $P(\boldsymbol{A}) = 1 - P(\boldsymbol{A}^C)$.

4. Addition Rule: $P(\boldsymbol{A} \textit{ or } \boldsymbol{B}) = P(\boldsymbol{A}) + P(\boldsymbol{B})$ provided $\boldsymbol{A}$ and $\boldsymbol{B}$ are disjoint (i.e., mutually exclusive, incompatible).

5. Multiplication Rule: $P(\boldsymbol{A} \; and \; \boldsymbol{B}) = P(\boldsymbol{A}) \times P(\boldsymbol{B})$ provided $\boldsymbol{A}$ and $\boldsymbol{B}$ are independent.

- The addition rule and multiplication rule generalize to more than two events. We study other generalizations in the next chapter.

- If $\boldsymbol{S}$ is finite (or countable), then $P(\boldsymbol{A})$ equals the sum of probabilities of all outcomes in $\boldsymbol{A}$. To be legitimate, the probabilities assigned to all outcomes in $\boldsymbol{S}$ must sum to one.

- If all outcomes in $\boldsymbol{S}$ are equally probable, then

$$P(\boldsymbol{A}) = \frac{\#\text{ of outcomes in } \boldsymbol{A}}{\#\text{ of outcomes in } \boldsymbol{S}}$$

  E.g., rolling a double in dice.

- In "continuous" probability models, all outcomes have zero probability. E.g., randomly select a person and measure their height. $\boldsymbol{S}$ = all possible heights. $\boldsymbol{A}$ = "height = 170 cm". $P(\boldsymbol{A})$ = ?

- Simple lottery: Pay one dollar for a ticket with a chance to win one of three possible prizes.
  - Chance of winning $\$100$ is 0.001
  - Chance of winning $\$10$ is 0.010
  - Chance of winning $\$2$ is 0.100

- Buy one lottery ticket. Probability you win a prize? Probability you win nothing?

- Buy two lottery tickets. Probability that you win nothing? Assumptions? Probability you win exactly one prize?

- Buy five lottery tickets. Probability that you win at least one prize? At most one prize?

Genetics example.

- Genes occur in pairs, one on each of two matching chromosomes. A child inherits one gene (in each pair) from each of its parents. Consider a gene controlling the colour of a flower.
  - WW produces white flower.
  - RR produce red flower.
  - WR and RW produce pink flower.
- Distribution of colours when we
  - cross white flower with white flower.
  - cross white flower with pink flower.
  - cross pink flower with pink flower.

15. Probability rules

- General addition rule.
- General multiplication rule.
- Independence and conditional probability.
- Tables and tree diagrams.
- Reversing the conditioning: Bayes' Rule.

- General Addition Rule. Motivate with Venn diagram.

  $\boldsymbol{A}$ *or* $\boldsymbol{B}$ = union of events = outcome is in $\boldsymbol{A}$ or $\boldsymbol{B}$, possibly both.

  $\boldsymbol{A}$ *and* $\boldsymbol{B}$ = intersection of events = outcome is in both $\boldsymbol{A}$ and $\boldsymbol{B}$.

$$P(\boldsymbol{A} \text{ } or \text{ } \boldsymbol{B}) = P(\boldsymbol{A}) + P(\boldsymbol{B}) - P(\boldsymbol{A} \text{ } and \text{ } \boldsymbol{B})$$

- Example: Asymmetric sea otters (Sanders & Schmidt).
  - While the skulls of most mammals are symmetric, there are exceptions. Sea otters have been noted to have naturally occurring asymmetry, where one side of the skull is larger than the other.
  - K. A. Shirley reported results from measuring cranial asymmetry in 387 sea otter skulls. Each skull was categorized by age (pup, juvenile, and adult) and by saggital crest deflection (to the left side, right side, and equal sides).

	Left	Right	Equal	Total
Pups	11	3	1	15
Juveniles	61	14	21	96
Adults	86	82	108	276
Total	158	99	130	387

- Randomly select one skull from the 387. Probability that sea otter was

1. a pup
2. left-sided
3. a pup with left sided deflection
4. a pup or left-sided
5. either a pup or left-sided, but not both (exclusive or)
6. a left-sided pup, right-sided juvenile, or equal-sided adult
7. left or right sided, but not an adult.
8. a left-sided pup or juvenile

- Conditional probability that $\boldsymbol{B}$ occurs given that $\boldsymbol{A}$ occurs. Motivate with Venn diagram.

$$P(\boldsymbol{B} \mid \boldsymbol{A}) = \frac{P(\boldsymbol{A} \; and \; \boldsymbol{B})}{P(\boldsymbol{A})}.$$

- Probability that sea otter is a pup given
  1. left-sided deflection
  2. right-sided deflection
  3. equal deflection
  4. it is not an adult
  5. it is an adult

- $\boldsymbol{A}$ and $\boldsymbol{B}$ are said to be independent if $P(\boldsymbol{B} \mid \boldsymbol{A}) = P(\boldsymbol{B})$.

- General Multiplication Rule: $P(\boldsymbol{A} \; and \; \boldsymbol{B}) = P(\boldsymbol{A}) \times P(\boldsymbol{B} \mid \boldsymbol{A})$.

- Independence $\neq$ disjoint. E.g., Deal one card.
  - $\boldsymbol{A}$ = face card, $\boldsymbol{B}$ = Ace
  - $\boldsymbol{A}$ = face card, $\boldsymbol{B}$ = Heart.

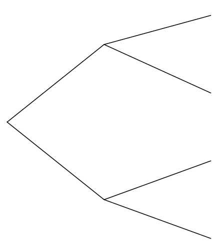
- Use tables to organize unconditional probabilities.

  Suppose 0.60 of a population support gun registration, 0.30 own guns, and 0.15 are gun owners who oppose gun registration. What proportion of the opposition do not own guns?

	Support	Oppose	Total
Own gun		0.15	0.30
Not own gun			
Total	0.60		1.00

- Use tree diagrams to organize conditional probabilities.

  Suppose 0.30 of a population own guns, 0.50 of gun owners support gun registration, and 0.643 of those without guns support gun registration. What is the probability that a randomly selected person is a gun-owning supporter? If you learn that the person supports gun registration, what is the conditional probability that the person owns a gun?



- Reverse conditioning. Suppose a team has probability 0.6 of winning the first game in a best-of-seven series. If the team wins the first game, the probability of winning the series is 0.821. If the team loses the first game, the probability of winning the series is 0.544. Given that the team wins the series, what is the probability that the team won the first game?

- Bayes' Rule is a formula for the preceding calculation. No need to learn this — just use the diagram approach.

$$P(\boldsymbol{A} \mid \boldsymbol{B}) = \frac{P(\boldsymbol{A})P(\boldsymbol{B} \mid \boldsymbol{A})}{P(\boldsymbol{A})P(\boldsymbol{B} \mid \boldsymbol{A}) + P(\boldsymbol{A}^C)P(\boldsymbol{B} \mid \boldsymbol{A}^C)}$$

- Roll two dice, one red and one green. Which pairs of the following events are disjoint? Which pairs are independent?
  - $A =$ "total is 7",
  - $B =$ "red die comes up 4",
  - $C =$ "at least one of the dice comes up 4".

- Three drinks. Three people go into a bar. Ann orders anisette, Bob orders beer, and Cathy orders coffee. The server brings the correct drinks but hands them out completely at random.
  Consider three events: $\boldsymbol{A} =$ "Ann gets her order", $\boldsymbol{B}$ and $\boldsymbol{C}$ similar.

  1. Which pairs of events are independent? Disjoint?
  2. Probability that Ann or Bob get their order.
  3. Probability that Ann gets her order given that Bob and Cathy get theirs.

- Example (Moore & McCabe). A six-sided die has four green and two red faces, and is balanced so that each face is equally likely to come up. The die will be rolled several times. You must choose one of the following three sequences of colours; you will win $25 if the first rolls of the die give the sequence you have chosen. Which of the following sequences should you choose? Why?
  - RGRRR
  - RGRRRG
  - GRRRRR

- Birthday problem: Suppose you have a random sample of n persons? How large does n have to be so that there is at least a 50-50 chance that two persons share the same birthday?

- Example. Toss a dime, nickel, and penny. You win the coins that come up heads. What is the probability that you win at least six cents? What are your assumptions? What would you be willing to pay to play this game?

- Example. Bob is a bit forgetful. If he doesn't make a "to do" list, the probability that he forgets something he is supposed to do is 0.1. Tomorrow Bob intends to run three errands, but forgets to write them on his list.

1. What is the probability that Bob forgets all three errands? What assumption did you make in order to calculate this probability?
2. What is the probability that Bob remembers at least one of the three errands?
3. What is the probability that Bob remembers the first errand but not the second or third?

- Mad Cow Example.
  - In 2002, 3.45 million cows were slaughtered in Canada. Among these, 3,377 were tested for BSE (Bovine Spongiform Encephalopathy).
  - Assume that the 3,377 cows tested represent a simple random sample from the 3.45 million cows slaughtered, even though this assumption is highly unrealistic.
  - How many of the slaughtered cows would need to have been infected to have an even chance (probability 1/2) of detecting at least one infected cow?

16. Random variables

- Probability distribution of a random variable. Notation and concepts.
- Mean = expected value, variance, standard deviation.
- Formulae for discrete random variables.
- Rules for means and variances.
- Combining Normal random variables.
- Correlation and covariance.

- Greedy pig.
  - All players stand, moderator rolls a die until a 5 comes up.
  - After each roll:
    If result $\neq 5$, those standing add points to their total.
    If result = 5, those standing lose all their points and game ends.
  - Players can sit down at any time to "lock in" points.
- When would you sit down?
- X = one player's point count at end of game.
- Distribution of X depends on strategy.
- Is there a "best" strategy?
- How many points can a good player expect to get on average?

- Random variable: numeric value based on outcome of a random event.
- Discrete if finite or countable number of values. Otherwise continuous.
- X denotes random variable, x denotes a particular value, $P(x) = P(X = x)$ denotes probability that $X = x$.
- Probability distribution of a discrete random variable X: the function $P(\cdot)$ or, equivalently, a table of values x and $P(x)$.
- Center of distribution: expected value of X = mean of X:

$$\mu = E(X) = \sum xP(x)$$

- $E(X)$ is a long-run average, not what we expect in a single trial.

- Aces and hearts.
  - Player pays $5 to draw one card from deck.
  - Player is paid $100 if card is ace of hearts.
  - Otherwise player is paid $10 if card is another ace.
  - Otherwise player is paid $5 if card is a heart.
  - Otherwise player is paid nothing.
- Would you play this game?
- X = net value received by player after one game, possibly negative.
- Distribution and expected value? Motivate with "ideal" sample.

- Spread: variance and standard deviation of a (discrete) random variable.

$$\sigma^2 = Var(X) = \sum (x - \mu)^2 P(x) = E\left\{(X - \mu)^2\right\}$$

$$\sigma = SD(X) = \sqrt{Var(X)}$$

- $Var(X)$ is the long run average squared deviation of X from μ.
- Aces and hearts.

- Another dice game: Roll die once. The player's payment X is \$0 for a 1, 2, or 3, \$5 for a 4 or 5, and \$50 for a 6. Find the distribution, mean, variance, and standard deviation.

- What happens if we increase each payment by \$10 ?

- What happens if we multiply each payment by 3 ?

- Now suppose the dice game is played twice, with independent payments X and Y. Find the distribution, mean, variance, and standard deviation for the combined payment $X + Y$.

- Find the distribution, mean, variance, and standard deviation for the difference in payments $X - Y$.

$$
\begin{aligned}
E(X \pm c) &= E(X) \pm c \\
Var(X \pm c) &= Var(X) \\
E(aX) &= aE(X) \\
Var(aX) &= a^2 Var(X) \\
SD(aX) &= |a| SD(X) \\
E(X \pm Y) &= E(X) \pm E(Y)
\end{aligned}
$$

If X and Y are independent, then

$$Var(X + Y) = Var(X - Y) = Var(X) + Var(Y)$$

Do *not* subtract variances! Do *not* add or subtract standard deviations! Compare formula with Pythagorean Theorem.

- Generalization of mean and variance rules: $X_1 + \ldots + X_n$.

- Note why subscripts are needed. Compare with $X + \ldots + X = nX$.

- Dice game: combined payments in ten games.

- Combining random variables (the bad news): Formulae for mean and variance are simple but expressions for the probability distribution usually become more complex. Consider tree diagrams.

- Continuous random variables: mean and variance formulae obtained by

  – replacing the probability function $P(x)$ by a "density" function,
  – replacing summation with integration (requires calculus).

  Other concepts and results are the same.

- Combining random variables (the good news): sums and differences of independent Normal random variables have Normal distributions.

- Auto prep. A used car dealer runs autos through a two-stage process to get them ready to sell. The mechanical checkup costs \$50 per hour and takes an average of 90 minutes with $SD = 15$ minutes. The appearance prep (wash, polish, etc.) costs \$6 per hour and takes an average of 60 minutes with $SD = 5$ minutes.

  a) What are the mean and standard deviation of the total time spent preparing a car? Assumptions?
  b) What are the mean and standard deviation of the total expense to prepare a car?
  c) What is the probability that it will take longer too do the appearance prep than the mechanical checkup? Assumptions?

Correlation and covariance.

- Suppose X and Y are random variables with means and variances $\mu_X, \mu_Y, \sigma_X^2, \sigma_Y^2$. The covariance between X and Y is

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

- If X and Y are independent, then $Cov(X,Y) = 0$. The converse need not be true. See the text for other properties.

- $Var(X \pm Y) = Var(X) + Var(Y) \pm 2Cov(X,Y)$.

- The covariance is related to the correlation:

$$\rho = Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = E\left\{\left(\frac{X - \mu_X}{\sigma_X}\right)\left(\frac{Y - \mu_Y}{\sigma_Y}\right)\right\}$$

- Bernoulli distribution: $P(X = 1) = p$ and $P(X = 0) = q = 1 - p$. Find the mean and variance of X.

- Binomial distribution: Let $X_1, X_2, \ldots, X_n$ be independent Bernoulli rvs, with same p. Find the mean and variance of $Y = X_1 + \ldots + X_n$.

- Find the mean and variance of $\hat{p} = Y/n$.

Tail probabilities for Binomial(n, p) distribution with p = 1/4

n=15		n=30		n=60			
x	P(X >= x)	x	P(X >= x)	x	P(X >= x)	x	P(X >= x)
0	1	0	1	0	1	20	0.0925
1	0.9866	1	0.9998	1	1.000	21	0.0541
2	0.9198	2	0.9980	2	1.000	22	0.0298
3	0.7639	3	0.9894	3	1.000	23	0.0154
4	0.5387	4	0.9626	4	1.000	24	0.0075
5	0.3135	5	0.9021	5	0.9998	25	0.0034
6	0.1484	6	0.7974	6	0.9990	26	0.0015
7	0.0566	7	0.6519	7	0.9969	27	0.0006
8	0.0173	8	0.4857	8	0.9912	28	0.0002
9	0.0042	9	0.3264	9	0.9788	29	0.0001
10	0.0008	10	0.1966	10	0.9548	30	0.0000
11	0.0001	11	0.1057	11	0.9141		
12	0.0000	12	0.0507	12	0.8524		
13	0.0000	13	0.0216	13	0.7684		
14	0.0000	14	0.0082	14	0.6651		
15	0.0000	15	0.0027	15	0.5494		
		16	0.0008	16	0.4312		
		17	0.0002	17	0.3204		
		18	0.0001	18	0.2247		
		19	0.0000	19	0.1486		

What Can Go Wrong?

- "Essentially, all models are wrong but some are useful", George Box

- If the model is wrong, so is everything else.

- Don't assume everything's Normal.
  You must Think about whether the Normality Assumption is justified.

- Watch out for variables that aren't independent:
  - You can add expected values of any two random variables.
  - You can only add variances of independent random variables.
  - Never add standard deviations or subtract variances.

Greedy pig, revisited. Optimal solution?