Statistics 2R
Alexey Lindo

Joao Almeida-Domingues[*]

University of Glasgow

September 23rd, 2019 – December 4th, 2019

Contents

These lecture notes were collated by me from a mixture of sources , the two main sources being the lecture notes provided by the lecturer and the

[*]2334590D@student.gla.ac.uk

content presented in-lecture. All other referenced material (if used) can be found in the *Bibliography* and *References* sections.

The primary goal of these notes is to function as a succinct but comprehensive revision aid, hence if you came by them via a search engine , please note that they're not intended to be a reflection of the quality of the materials referenced or the content lectured.

Lastly, with regards to formatting, the pdf doc was typeset in LaTeX, using a modified version of Stefano Maggiolo's class

1 Introduction & Terminology

1.1 definition. Experiment any procedure that can be infinitely repeated and has a well-defined set of possible outcomes

1.2 definition. Trial a single performance of an experiment

1.3 definition. Outcome information obtained from one trial

1.4 definition. Stochastic Experiment an experiment which has more than one possible outcome, even when performed under identical conditions, where it is not known in advance which of the outcomes will occur when next performed. (a.k.a random experiment)

1.5 definition. Sample Space a set that contains all the possible outcomes of a random experiment

1.6 remark. usually denoted by S ; it can be finite, countable or uncountable

1.7 definition. Countable Set infinite set whose elements can be counted (e.g. natural numbers)

1.8 definition. Uncountable Set not countable (e.g. real numbers)

1.9 definition. Event collection of outcomes

1.10 remark. Any event E is necessarily a subset of S

1.11 definition. Simple Event an event that consists of a single outcome, i.e. $|E| = 1$

1.12 definition. Compound Event $|E| \geq 2$

1.1 *Sets*

This material has been covered at length at most other courses, hence it is crucial for higher level mathematics. See 2F for proofs on some of the results presented below

1.13 definition. Universal Set (S) set which includes all possible outcomes

1.14 definition. Empty Set ($\emptyset$) set which includes no elements

1.15 definition. Complement Set (E^c) set which includes everything which is not in E.

1.16 definition. Union ($E \cup F$) set which includes elements which are in at least one of E or F. *"E or F or both"*

1.17 definition. Intersection ($E \cap F$) set which includes elements which are both in E **and** F

1.18 definition. Disjoint Sets when $E \cap F = \emptyset$

1.19 definition. Subset ($E \subset F$) all elements of E which are also in F; $E \implies F$

1.20 definition. Relative Complement Set ($F \setminus E$) set which includes all elements in E except those which are also in E ; $F \setminus E = F \cap E^c$

1.21 definition. Disjoint Union of Sets $F = (F \cap E) \cup (F \cap E^c)$

All these results apply conversely to the intersection of events

Laws of Sets

Commutative : $A \cup B = B \cup A$

Associative : $A \cup (B \cup C) = (A \cup B) \cup C$

Distributive : $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

DeMorgan's : $(A \cup B)^c = A^c \cap B^c$

1.2 *Interpreting Probability*

1.22 definition. Probability of E ($P(E)$) informs us on how likely E is to occur in a given experiment

There are 3 main views, or ways to interpret the meaning of $P(E)$, nevertheless almost all of them use the same methods for calculating and manipulating it within a certain probability model. These will be covered in the next section

1. **Equally-Likely Outcome (ELO)** : Assumes that for k possible outcomes in S, all have the same likelihood of occurring, i.e.

$$P(E_1) = P(E_2) = \cdots = P(E_k) = \frac{1}{k} =$$

2. **Frequentist** : Assumes that for k possible outcomes in S, the probability of a given event occurring is weighted by considering how many times it occurs in a large number of trials of a given experiment. The concepts of limit and **relative frequency** are useful here, for practical reasons.

$$P(E) = \lim_{n \to \infty} \{rf_n(E)\} = \frac{n(E)}{n}$$

where $n(E)$ represents the number of outcomes in E

3. **Subjectivist** : The probability of an event is best treated as a statement of personal belief.

2 Counting ELOs

If E is an event in a sample space, S, with $|S| = N$ equally likely (simple) outcomes , then $P(E)$ is the sum of the probabilities of the outcomes in E,

$$P(E) = \frac{\text{number of favourable outcomes}}{\text{number of total possible outcomes}} = \frac{n(E)}{n(S)} = \frac{n(E)}{N}$$

Note that the sizes of both $n(E)$ and S can grow quite rapidly for relatively simple events, hence it is almost never desirable or feasible to exhaustively list them. Instead, we use combinatorics to count them.

2.1 *Multiplication Principle*

For a compound experiment λ with component experiments $\lambda_1, \ldots, \lambda_n$. If λ_1 has k_1 possible outcomes, then each of k_1 can pair with one of the k_2 possible outcomes of λ_2 and so on. So, no "path" needs to be necessarily taken, and therefore for each new experiment no new outcome needs to be excluded given the result of the previous. Hence, each event in k_2 could happen in any one of k_1 times, i.e the total possible number of outcomes for λ is given my the **cartesian product** of its component's sample sets

$$|S| = \prod_{i=1}^{n} k_i$$

Think of the way to create a one card as a compound experiment, where the first stage is picking the rank and the second is picking the suit. So our initial set is composed of 13 distinct possible outcomes. Our second set is composed of 4 distinct possible outcomes. Now note that choosing (1 out of the 4 possible suits) in the second stage can be achieved in 13 different ways, i.e our suit choice is not determined by the the rank chosen, each suit is available to each rank. Our unique card is however determined by the combination of (R,S), so for each card I can choose 13 different ranks and I can do this 4 times for each different suit $(13C + 13D + 13H + 13S) = 13 \times 4 = |S1| \times |S2|$

2.2 *Addition Principle*

Unlike before, there are cases where certain outcomes can occur simultaneously. In this cases a choice in a component experiment will affect the number of possible outcomes available in the next experiment. Say for example if you can only score one 6 in a two-diced throw. Then, in the first stage you could either (A) get a 6 or (B) not get a 6. Note that if A, then the second stage cannot also involve A. If B, then the second stage cannot also involve B. So, we have $|S_A| = 1 \times 5$ which include all possible outcomes given that a 6 is scored in the first dice , and $|S_B| = 5 \times 1$, which accounts for all cases where a 6 is scored in the second dice. In other words, for each stage we either exclude A or B given wether B or A happened before.

$$|S| = \sum_{i=1}^{n} k_n$$

2.1 remark. Note that the number of distinct outcomes can still be the same, in which case $|S| = k \times n$

2.2 remark. In general, if we want to count the outcomes for E_1 **and** B we use MP. E_1 **or** B involves using AP

2.3 *Counting using Combinatorics*

In order to count the possible outcomes, we use the concept of Combinations and Permutations. Which gives us the total possible number of arrangements from a given set. The main difference between the two being the fact that one takes into consideration the order on which the elements are chosen, while the other is not.

Permutation

2.3 definition. **Permutation** total possible ordered arrangements from a set of distinct objects

2.4 notation. ^nP_r

$$^nP_r = \frac{n!}{(n-r)!}$$

2.5 remark. Note from above that in the denominator we exclude all other distinct objects in the main set not chosen

2.6 remark. Note that if $n = r$, then $^nP_r = n!$

2.7 example. Say we need the total number of possible two distinct letter combination from the following set $\{A, B, C, D\}$. We could approach it as a multistage step problem, and use the multiplication principle in the following way:

1. From the $\{A, B, C, D\}$ choose 1 ; We know that there are 4 possible ways to do this. ; Say we choose A
2. Given that the objects need to be distinct we now have the set $\{B, C, D\}$ from which to choose, of which there are 3 possible ways to do so
3. Hence, we reach the conclusion that there are a total of $4 \times 3 = 12$ possible distinct arrangements

Note however that in a set of size 4 this is easy to see, but our initial goal was to find a method which works for large sets. Say, if you had 100 elements, and you needed a 50 element object then you would need $100 \times 99 \times 98 \cdots \times 51$. Alternatively, we can attempt to generalize our method.

1. We note that if we wanted to find the total number of arrangements for the original set, then that would just be $n!$ which is easy to compute with an aid of a calculator.
2. Then we note that from that total number if we only need r elements then there are $n - r$ elements for which we calculated possible arrangements which are not needed, and hence should not be accounted for in the total possible outcomes. In other words, there are $(n - r)!$ arrangements which we are not interested in
3. Finally, we exclude the arrangements in (2) by dividing by (1). Giving us

$$\frac{n!}{(n-r)!} =^n P_r$$

Combinations

2.8 definition. Combinations Subset of permutations, where the order does not matter. i.e $AB = BA$

Expanding on the notion of permutation above, all we need to do is to account for this repeated entries. We note that from the set of chosen objects, each object is composed of elements of size r, which can be arranged in $r!$ different ordered ways, i.e ${}^rP_r$. However, if the order doesn't matter, then this $r!$ count as 1 and need to be excluded.

$${}^nC_r = \frac{{}^nP_r}{{}^rP_r} = \frac{\frac{n!}{(n-r)!}}{\frac{r!}{(r-r)!}} = \frac{n!}{r!(n-r)!}$$

2.9 remark. The following useful results follow from above

$$\binom{n}{n} = \binom{n}{0} = 1 \qquad \text{and} \qquad \binom{n}{r} = \binom{n}{n-r}$$

3 Axioms of Probability

As discussed in 1.2 there is disagreement about the interpretation of $P(E)$ but there is general agreement about how valid deductions should be made within a probability model. There must be restrictions on the probabilities that can be assigned to events in S in order to make sure that the probabilities assigned to different events are consistent with each other. We say that a probability model is guaranteed to assign consistent probabilities to all events if it obeys the following Axioms of Probability laid down by *Kolmogorov (1933)*

Axioms

1. $0 \leq P(E) \leq 1$, for any event $E \subseteq S$
2. $P(S) = 1$
3. $P(E_1 \cup E_2) = P(E_1) + P(E_2)$ if $E_1 \cap E_2 = \emptyset$
4. If $\bigcap_{i=1}^{n} E_i = \emptyset$, then $P(\bigcup_{i=1}^{n} E_i) = \sum_{i=1}^{n} P(E_i)$

note that 3 is a special case of 4 where $E_{i>2} = \emptyset$

4 Rules of Probability

A common task within a given probability model is to calculate unknown probabilities from known ones. This could be calculated by successful application of the axioms above, however there are more key general results which can be derived from the axioms which make this process here. We call these results the *rules of probability*

4.1 remark. The axioms always refer to **disjoint events** , hence this is a crucial fact to be aware of when deriving the rules ; a common step when proving them is to split a compound event into its disjoint parts

4.2 proposition. $P(E^c) = 1 - P(E)$

Proof.

(1) $$P(E \cup E^c) = P(E) + P(E^c) \qquad (Axiom 3)$$

(2) $$P(E \cup E^c) = P(S) = 1 \qquad (Axiom 2)$$

(3) $$P(E) + P(E^c) = 1 \iff P(E^c) = 1 - P(E) \qquad (1), (2)$$

QED

4.3 proposition. $P(\emptyset) = 0$

Proof. We take $S^c = \emptyset$ then, from 4.2 , $P(\emptyset) = 1 - P(S) = 0$ QED

4.4 proposition. $P(E) \leq P(F)$, *for* $E \subseteq F$

Proof.

$$F = E \cup F \setminus E \qquad \text{(union of disjoint events)} \tag{4}$$
$$P(F) = P(E) + P(F \setminus E)(Axiom3) \tag{5}$$
$$\geq P(E), \text{ since } F \setminus E \geq 0 \qquad (Axiom1) \tag{6}$$

QED

4.5 example. See Workshop1.6

☛*Generalization of 3 , where we take into account the fact that for non-disjoint events $A \cap B$ is counted twice*

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

Proof.

1. Split the compound event $A \cup B$ into 3 disjoint events

$$A \setminus B = X = A \cap B^c \qquad B \setminus A = Y = B \cap A^c \qquad Z = A \cap B$$

$$P(A \cup B) = P(X) + P(Y) + P(Z) \qquad (Axiom4)$$

2. Now using Axiom 3, we can start reconstructing the original sets from the disjoint ones

Look ahead, $A \cap B$ is present in the RHS of the equality, so probably does not need to be simplified

$$P(A) = P(X \cup Z) = P(X) + P(Z); \qquad P(B) = P(Y \cup Z) = P(Y) + P(Z)$$

3. Finally, substituting the last line back into $P(A \cup B)$

$$\begin{aligned} P(A \cup B) &= P(X) + P(Y) + P(Z) \\ &= (P(A) - P(Z)) + (P(B) - \cancel{P(Z)}) + \cancel{P(Z)} \\ &= P(A) + P(B) - P(Z) \\ &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

QED

4.7 proposition. $P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$

Proof. Similar to above, but there are 7 disjoint events, hence writing out all the disjoint unions becomes slightly hairy, but the procedure is the same. See Workshop1.5 for full proof QED

4.8 proposition. ***Boole's Inequality***

$$P\left(\bigcup_{i=1}^{n} E_i\right) \leq \sum_{i=1}^{n} P\left(E_i\right)$$

Proof. By induction

1. **Base Case:** Show true for $n = 2$. From 4.6 we have that,

$$\begin{aligned} L.H.S = P(E_1 \cup E_2) &= P(E_1) + P(E_2) - P(E_1 \cap E_2) \\ &\leq P(E_1) + P(E_2) = R.H.S \qquad (Axiom1) \end{aligned}$$

Alternatively we can just note that $max(P(E_1 \cup E_2))$ happens when they're disjoint

2. **Induction Hypothesis:** Assume true for $n \geq 2$
3. **Induction Step:** Proving true for $n + 1$

$$P\left(\bigcup_{i=1}^{n+1} E_i\right) \leq \sum_{i=1}^{n+1} P\left(E_i\right)$$

$$\begin{aligned} L.H.S = P\left(\bigcup_{i=1}^{n+1} E_i\right) &= P\left(\bigcup_{i=1}^{n} E_i\right) \cup E_{n+1}) \\ &\leq P\left(\bigcup_{i=1}^{n} E_i\right) + P\left(E_{n+1}\right) \qquad from\ 4.6 \\ &\leq \sum_{i=1}^{n} P\left(E_i\right) + P(E_{n+1}) \qquad by\ the\ induction\ hypothesis \\ &\leq \sum_{i=1}^{n+1} P\left(E_i\right) = R.H.S \end{aligned}$$

QED

4.9 example. See Workshop 1.6 , 1.7

5 Random Variables

5.1 definition. **Random Variable** A function $X : S \to \mathbb{R}$, which associates a unique numerical value $X(s)$ with every outcome $s \in S$

When working with random variables, we can summarise each outcome of an experiment with a single numerical value. In general, R.V can be split into two broad categories , depending on the values they can assume, they can either be (i) Discrete (ii) Continuous .

5.1 *Discrete Random Variables*

5.2 definition. DRV a random variable with countable range, i.e where $X(s)$ is countable for all s

5.3 definition. Realisation Each $X(s)$

5.4 definition. Range Space All possible $X(s)$

5.5 notation. R_x

5.6 example. Likert scale , we have $S = \{\text{"Strongly disagree"}, \dots, \text{"Strongly Agree"}\}$, and we can map each response into a numerical value such that $X(s) = \{-2, \dots, 2\}$

Although usually we speak of probabilities associated with X , $P(X = 0), P(X < 2)$, these probabilities are induced from probabilities of equivalent events in the original sample space. For example $P(X = x)$ corresponds to the probability of the equivalent event $\{s \in S : X(S) = x\}$

5.7 definition. Probability Mass Function list of probabilities associated with each of its possible values. $p_x(x) = P(X = x), x \in \mathbb{R}$, such that $0 \leq p_x(X) \leq 1, x \in \mathbb{R}$ and $\sum_{x \in R_x} p_x(x) = 1$

In essence, *p.m.f* maps all the possible discrete values a random variable could take on, and maps them to their probabilities

5.8 remark. $p_x(x) = 0,\ \forall\ x \notin R_x$

weighted just means that one takes into consideration their likelihood/probability

5.9 definition. Expect Value / Mean the sum of the weighted possible values for X

$$\mu = E(X) = \sum_{x \in S}^{n} x p(x)$$

5.10 definition. Variance a measure of the spread of the possible values

$$\sigma^2 = E\left[(X - \mu)^2\right]$$

Where $E(X^2)$ corresponds to the sum of the weighted possible values squared, i.e.

$$E\left[X^2\right] = \sum_{x \in S} x^2 \cdot p(x)$$

Therefore,

$$\begin{aligned}
\sigma^2 &= \sum_{x \in S} \left[x^2 \cdot p(x)\right] - \mu^2 \\
&= \sum_{x \in S} \left[x^2 \cdot p(x)\right] - \left[\sum_{x \in S} x \cdot p(x)\right]^2
\end{aligned}$$

5.11 remark. It is common to represent p.m.f as the set of pairs $\{(x, p_x(x)), x \in R_x\}$

5.12 definition. **Cumulative Distribution Function** function giving the probability that $X \leq x : F_x(x) = P(X \leq x)$, such that

1. $0 \leq F_x(X) \leq 1$
2. $F_x(-\infty) = P(X \leq -\infty) = 0$ and $F_x(\infty) = P(X \leq \infty) = 1$
3. $x_1 \leq x_2$, then $P(X \leq x_1) \leq P(X \leq x_2) \equiv F_x(x_1) \leq F_x(x_2)$, i.e **increasing function**

5.13 remark. Given the discrete nature of the drv , then c.d.f for a given x is just the sum of the p.m.f for each discrete value smaller than x , i.e $F_x(x) = \sum_{r \in R_x \leq x} p_x(x), x \in \mathbb{R}$

Given the relation between p.m.f and c.d.f , we can always revert the relation and instead recover the first from the latter by subtracting consecutive terms : $p_x(1) = F_x(1) - F_x(0)$, in general

$$p_x(x) = F_x(x) - F_x(x-1)$$

To do (1)

6 Distributions

6.1 definition. **Gamma Function**

$$\Gamma(\alpha) = \int x^{\alpha-1} e^{-x} dx = (\alpha-1)!, \ \alpha = 1, 2, \ldots$$

6.1 *Continuous Random Variables*

For continuous variables it doesn't make sense to look at $P(X = x)$, instead we look at the probability for a given range of x

6.2 proposition. $P(X = x) = 0 \; \forall x$

6.3 remark. For continuous distributions, $P(X = x) = 0$, so the expressions $P(X \leq x) = P(X < x)$

6.4 remark. Note that by definition of μ and c.d.f , we have $\mu = F_X(x) = \frac{1}{2}$. We want the value for which $X \leq \frac{1}{2} \equiv X = \frac{1}{2}$

The c.d.f can be expressed as the integral of its p.d.f

$$F_X(x) = \int\limits_{-\infty}^{x} f_X(t)dt$$

For X within a range, we have

$$\mathrm{P}(a < \mathrm{X} \leq b) = F_X(b) - F_X(a) = \int\limits_{a}^{b} f_X(t)dt$$

6.5 example. Take a distribution with the p.d.f $f_X(x) = \frac{2}{a} - \frac{2}{a^2}x, \quad 0 < x < a$. Then,

6.6 proposition. ***Median*** *x such that $F_X(x) = \frac{1}{2}$*

More generally, given the definition of c.d.f, given $F_X(x)$ we can find the quartile of a given distribution by finding the x which gives the desired cut-off point. For more complex distributions, which cannot be integrated in closed form, x can be read from tables of numerical approximations

6.7 definition. Percentile $\varepsilon_k = F_X(x) \geq \frac{k}{100}$, as discussed above, given the continuous nature of the variable $F_X(x) \geq \frac{k}{100}$

6.8 remark. Note that is just the same as asking for the cut-off point for quartiles, the $\frac{1}{100}$ is just a convention in order for quantiles to be expressed as percentages. e.g , median = $\varepsilon_{50} = \frac{50}{100} = \frac{1}{2}$

6.9 definition. Quantile Function $F_X^{-1}(x)$

6.10 definition. Mean $E(X) = \int_a^b x f_X(x)$

6.11 definition. Variance $Var(X) = [E(X^2) - E(X)^2]$, where

$$E(X^2) = \int_a^b x^2 f_X(x)$$

Normal

Computation Standard Normal

1. Compute the mean $\mu = E(X)$, and standard deviation $\sigma = \sqrt{Var(X)}$
2. Input into formula $Z = \frac{X-\mu}{\sigma}$
3. Find appropriate Z in cdf tables ; Recall for x in table , we have $\phi(x) = Z \leq x$ and for $-x$ we have $1 - \phi(x) = Z \leq -x$

6.12 example. For $X \sim N(-1, 0.25)$ we have $Z = \frac{X-(-1)}{\sqrt{0.25}} \sim N(0,1)$. Hence,

1. $\mathbf{P(X < 0.5)}$

   $P(Z < \frac{0.5-(-1)}{\sqrt{0.25}}) = P(Z < 3)$, consulting the values of the $c.d.f$ we have $\phi(3) = P(X < 0.5) = 0.9987$

2. $\mathbf{P(X < -1.25)}$

   $P(Z < \frac{-1.25-(-1)}{\sqrt{0.25}}) = P(Z < -0.5)$, consulting the values of the $c.d.f$ we have $1 - \phi(0.5) = P(X < -1.25) = 0.3805$

3. $\mathbf{P(-2 < X < 0)}$

   $P(\frac{-2-(-1)}{\sqrt{0.25}} < Z < \frac{0-(-1)}{\sqrt{0.25}}) = P(-2 < Z < 2)$. Hence $Z > -2$ and $Z < 2$, which gives us the range $P(Z < 2) - P(Z < -2)$ consulting the values of the $c.d.f$ we have $\phi(2) = P(X < 2) = 0.9772$ and $1 - \phi(2) = P(X < -2)$ therefore, $P(Z < 2) - P(Z < -2) = P(-2 < X < 0) = 2\phi(2) - 1 = 0.9544$

4. **Expected value, variance , standard deviation , median of X**

   $E(X) = -1$; $Var(X) = 0.25$; $\sigma = 0.5$; $P(X < \varepsilon_{50}) = 0.5 \implies \varepsilon_{50} = -1$

To do (2)

6.2 *Gamma*

To do...

- ☐ 1 (p. 13): complete example 8.4 lecture notes ; p.45,56,48,49
- ☐ 2 (p. 15): Add to continuous subsec
- ☐ 3 (p. 16): https://en.wikipedia.org/wiki/Experiment_(probability_theory)
- ☐ 4 (p. 16): https://www3.nd.edu/~dgalvin1/10120/10120_S16/Topic09_7p2_Galvin.pdf
- ☐ 5 (p. 16): https://socratic.org/questions/how-do-you-use-a-probabili