Slightly abbreviated section summaries for POLSCI 599: Statistical Methods for Political Science Research

Samuel Baltz
Fall 2018

Written to complement lectures by Yuki Shiraito

Contents

1 Section 1: 2018 September 13

Topics:

- Probability space $(\Omega, \mathcal{F}, P)$ [1]
- Kolmogorov's Axioms [2]
- Two proofs: [3]

  $P(A^c) = 1 - P(A)$

  $A_1 \subset A_2 \implies P(A_1) \leq P(A_2)$

After this class, I expect every student to be able to:

- define in words what a sample space Ω, set of events $\mathcal{F}$, and probability measure P are, and identify each in any simple experiment
- explain why we need $\mathcal{F}$, and cannot always define P directly in terms of Ω
- picture Ω and $\mathcal{F}$ as sets
- know and remember Kolmogorov's Axioms
- explain briefly the difference between a definition, axiom, and conclusion, and identify each in a simple statistical proof

BACK TO TABLE OF CONTENTS

(1)**Probability space,** $(\Omega, \mathcal{F}, P)$:

$\hookrightarrow \Omega$ is the **"sample space"**, which is the set of all outcomes of an experiment.

Definition: An **outcome** is the full result of one experiment

Example: Flip two coins. Then $\Omega = \{\{H, H\}, \{H, T\}, \{T, H\}, \{T, T\}\}$

$\hookrightarrow \mathcal{F}$ is the **"set of events"** of an experiment. This typically contains all possible combinations of the elements of Ω, together with $\emptyset$.

Definition: An **event** is any combination of 0 or more outcomes.

Example: Flip two coins. Then as an example,
$\mathcal{F} = \{\{\}, \{H, H\}, \cdots, \{T, T\}, \{\{H, H\}, \{H, T\}\}, \cdots, \{\{H, H\}, \{H, T\}, \{T, H\}, \{T, T\}\}\}$

Question: Can you interpret the meaning of the event $\mathcal{F}_i = \{\{H, H\}, \{H, T\}, \{T, H\}\}$? What is this event's substantive meaning? If we knew the probability of this event, what would that mean?

Remark: When we encounter $\mathcal{F}$, we should picture a set of nested sets. Think of $\mathcal{F}$ as the interface between Ω and our ability to assign probabilities to experimental results, which is formalized below:

$\hookrightarrow P$ is the **"probability measure"**, which is a function mapping from events onto numbers that we call probabilities.

Example: Using the above $\mathcal{F}_i$, we can calculate the number that P takes that event onto. In this case, just by counting the frequencies of outcomes, we find $P(\mathcal{F}_i) = \frac{3}{4}$.

(2)**Kolmogorov's Axioms**:

1. $P(A) \in \mathbb{R}$ so $P(A) \geq 0 \; \forall A \in \mathcal{F}$

2. $P(\Omega) = 1$

3. For $A_i, i \in \mathbb{N}$ mutually exclusive, $P(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$

(3)**Proof one:** We seek to show that $P(A^c) = 1 - P(A)$

Consider the definition of a complement,

$$A_1^c \equiv \Omega \setminus A_1$$

I take it to be equivalent that

$$A_1^c \cap A_1 = \emptyset$$

Then we can apply Kolmogorov's third axiom to A_1 and A_1^c, yielding

$$P(A_1 \cup A_1^c) = P(A_1) + P(A_1^c)$$

Now consider that, by the definition of a complement, $A_1 \cup A_1^c = \Omega$. So, by Kolmogorov's second axiom, we have $P(A_1 \cup A_1^c) = 1$. Then,

$$1 = P(A_1) + P(A_1^c)$$

$$P(A_1^c) = 1 - P(A_1)$$

□

Proof two: We seek to show that $A_1 \subset A_2 \implies P(A_1) \leq P(A_2)$

Consider $A_2 \cap A_1^c$. Notice that $A_1 \cap (A_2 \cap A_1^c) = \emptyset$, so by Kolmogorov's third axiom,

$$P(A_1 \cap (A_2 \cap A_1^c)) = P(A_1) + P(A_2 \cap A_1^c))$$

Because $A_1 \subset A_2$, we have $A_1 \cap (A_2 \cap A_1^c) = A_2$. So,

$$P(A_2) = P(A_1) + P(A_2 \cap A_1^c)$$

By Kolmogorov's first axiom, $P(A_2 \cap A_1) \geq 0$, so

$$P(A_2) \geq P(A_1)$$

□

2 Section 2: 2018 September 20

Topics:

- Conditional independence (conditioning and controlling) [1]
- Law of total probability [2]

After this class, I expect every student to be able to:

- define conditional independence
- give an example of conditionally independent events
- describe the law of total probability in intuitive terms

BACK TO TABLE OF CONTENTS

(1)**Conditional independence**:

Definition: In lecture we said that two events A_1, A_2 are **independent** if $P(A_1 \cap A_2) = P(A_1)P(A_2)$.

Definition: We define the **conditional probability** $P(A|B)$ of event A conditional on event B as $P(A|B) \equiv \frac{P(A \cap B)}{P(B)}$.

Definition: Two events A_1, A_2 are **conditionally independent** given another event A_3 if $P(A_1 \cap A_2|A_3) = P(A_1|A_3)P(A_2|A_3)$.

Interpretation: If we account for the relationships between A_1 and A_3 and between A_2 and A_3, then and A_1 and A_2 are conditionally independent given A_3, then knowing the probability A_1 will give us no information about the probability of A_2, and vice versa.

Tangible Example: $P(\text{taking 598} \cap \text{taking 599}) \neq P(\text{taking 598})P(\text{taking 599})$, but I think that, with s representing that the student is a political science student, it is nearly true that $P(\text{taking 598} \cap \text{taking 599}|s) = P(\text{taking 598}|s)P(\text{taking 599}|s)$

Political Science Example: If you have heard about "controlling" for a relationship in OLS regression, we are looking for exactly the idea of conditional independence. For example, in Canada, regionalism R heavily predicts vote choice V, but R barely predicts V if you control for Party ID, call it I. So $P(R \cap V) \neq P(R)P(V)$, because these two are closely related, but $P(R \cap V|I) \approx P(R|I)P(V|I)$.

Example of The Opposite: This was the day after my cat tried to kill me causing me to set fire to some rice, so I gave the example that whether or not I set fire to my house and whether or not I eat rice is typically independent, but my cat introduced a dependency between those two variables by distracting me.

(2)**Law of total probability**:

Definition: A **partition** S of a set X is a set of non-empty subsets of X so that every element in X is in exactly one subset in S.

Theorem (the Law of Total Probability): Consider a partition $\{B_1, \cdots, B_n\}$ of Ω, with $P(B_i > 0)$ $\forall i \in \mathbb{N}$ so $1 \leq i \leq n$. Then for any event A, $P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$.

Proof: By the definition of a set intersection,

$$P(A) = P(A \cap \Omega)$$

By the definition of a partition,

$$P(A) = P(A \cap \{B_1 \cup B_2 \cdots \cup B_n\})$$

Distributing the intersections over the unions, which is an elementary property of set intersections,

$$P(A) = P(\bigcup_{i=1}^{n}(A \cap B_i))$$

By Kolmogorov's third axiom, which applies by the definition of a partition,

$$P(A) = \sum_{i=1}^{n} P(A \cap B_i)$$

By the definition of conditional probability,

$$P(A) = \sum_{i=1}^{n} P(A|B_i)P(B_i)$$

□

3 Section 3: 2018 September 27

The material from this class was displaced for an R review. I do not summarize that review here because it is not curriculum material, so there's no need to remember it for its own sake.

BACK TO TABLE OF CONTENTS

4 Section 4: 2018 October 4

Topics:

- Relationship between: [1]

  Past concepts

  Random variables

  PFs

  CDFs

- An extended example: fivethirtyeight [2]

After this class, I expect every student to be able to:

- describe what a random variable is
- formally define a CDF
- identify whether or not a simple function is a CDF

BACK TO TABLE OF CONTENTS

(1) **From Ω to CDFs**:

Definition: A **random variable** X is a function $X : \Omega \to E$, where Ω is a sample space and E is any measurable space, for example $\mathbb{R}$.

Note: This is similar to the role that I said $\mathcal{F}$ plays: it is another interface that allows us to make meaningful statements about the outcomes of an experiment.

Note: Don't get tripped up by the fact that X maps from outcomes onto E, not from events onto E. Just notice that if X can map from one outcome, then it can map from a combination of outcomes.

Example: Flip 2 coins. Then $\Omega = \{\{HH\}, \{HT\}, \{TH\}, \{TT\}\}$. Say X represents the number of heads. Then we can talk about the probability of X taking on different values, which we write $P(X = x)$. In this example, just by counting, we can see that $P(X = 0) = \frac{1}{4}$, $P(X = 1) = \frac{1}{2}$, and $P(X = 2) = \frac{1}{4}$. Note what X is doing here: X takes the abstract idea of flipping coins and it attaches a number to it. The fact that X outputs 0, 1, or 2 allows us to discuss the frequency of different combinations of coin faces appearing.

Excercise: Draw the PMF and CDF of X in this example.

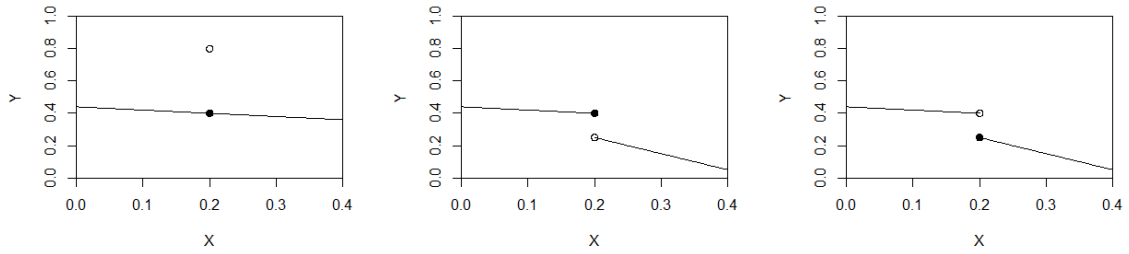
Definition: A **cumulative distribution function (CDF)**, denoted $F(\cdot)$, is any function that is

i) normalized beween 0 and 1
ii) monotonically increasing
iii) right-continuous

Interpretation: The cumulative distribution function for some realization x of a random variable X represents $F_X(x) = P(X \leq x)$.

Definition: The probability mass function $f_X(x)$ of a discrete random variable represents $f_X = P(X = x)$. The probability density function $f_X(x)$ of a continuous random variable represents a probability density, which we can integrate over a region to obtain a probability value.

Exercise: One of the following three plots represents a valid CDF. Which is it? What condition rules out the other 2?



Exercise: Practice drawing the CDF corresponding to some example PDFs. I also asked everyone to draw one curve that violates all three requirements of a CDF.

(2) **Extended example**:

↪ Here we looked at a plot on fivethirtyeight.com, called their "House Forecast", which showed the projected number of seats that Democrats and Republicans will win in the 2018 midterm congressional elections. We then answered the following questions:

Question: Is this a PDF, PMF, or CDF? Why?

Question: (After we had decided that it is almost a PMF) What CDF does this PMF correspond to? Draw that CDF. What question does that CDF answer?

Question: What random variable does this PMF correspond to?

Note: This is a Bayesian model. Nate Silver's problem (in its most crudely simplified form) is to find $P(\text{win}|\text{polls}) = \frac{P(\text{poll}|\text{win})P(\text{win})}{P(\text{poll})}$. The reason that this PMF is a distribution and not just a point estimate is that, when we don't know the prior for sure, we draw it from a reasonable-sounding distribution to simulate our uncertainty.

5 Section 5: 2018 October 11

Topics:

- Countability and uncountability [1]
- Inverses and quantiles [2]
- Two CDF properties with quantiles [3]

After this class, I expect every student to be able to:

- summarize the difference between "discrete" and "continuous"
- describe what a quantile function is

BACK TO TABLE OF CONTENTS

(1) **Countability and uncountability**:

$\hookrightarrow$ We have relied heavily on the ideas of "discrete" and "continuous" variables, but we haven't properly defined these terms. Luckily, there is an extremely accessible and famous proof that really shows the difference between these two ideas.

Remark: Discrete variables are defined on some (maybe improper) subset of the set of all integers ($\mathbb{Z}$, and recalling that $\mathbb{N} \subset \mathbb{Z}$), and continuous variables are defined on some interval of the set of real numbers ($\mathbb{R}$). We call the numbers that a discrete variable is defined on "countable" (simply because natural numbers are the numbers that humans count with), and the numbers that a continuous variable is defined on "uncountable". This naming convention is explained by the following theorem:

Theorem: It is not possible to associate every natural number with a real number, without leaving any real numbers out.

Note: This theorem says that there are more real numbers than there are integers or natural numbers, even though these sets of numbers are infinite.

Proof (Cantor's diagonal argument): First construct a countably infinite list of real numbers:

$$\begin{array}{cccccccc}
1 & . & 0 & 0 & 0 & 0 & 0 & \cdots \\
1 & . & 0 & 0 & 0 & 1 & 0 & \cdots \\
3 & . & 1 & 4 & 1 & 5 & 9 & \cdots \\
5 & . & 9 & 8 & 5 & 9 & 9 & \cdots \\
2 & . & 1 & 0 & 1 & 0 & 1 & \cdots \\
2 & . & 7 & 8 & 1 & 8 & 2 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array}$$

When I call this list "countably infinite", I mean that we can associate every number in it with a natural number, which means that we can count the numbers in the list. I can refer to the first number in the list, the second number in the list, the hundredth number in the list, and so on. So, this list is no larger than the set of natural numbers.

Now, consider the ith digit of each of number i in the list, so the 1st digit of the 1st number, the 2nd digit of the second number, and so on:

$$\begin{array}{ccccccccc}
\textcircled{1} & . & 0 & 0 & 0 & 0 & 0 & \cdots \\
1 & . & \textcircled{0} & 0 & 0 & 1 & 0 & \cdots \\
3 & . & 1 & \textcircled{4} & 1 & 5 & 9 & \cdots \\
5 & . & 9 & 8 & \textcircled{5} & 9 & 9 & \cdots \\
2 & . & 1 & 0 & 1 & \textcircled{0} & 1 & \cdots \\
2 & . & 7 & 8 & 1 & 8 & \textcircled{2} & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array}$$

Let's consider the number, call it r, that is formed by concatenating each of those digits together:

$$r = 1.04502\cdots$$

Now let's modify every digit of this number, say by adding 1 to every non-9 digit, and sending every 9 to 0. So we obtain a new real number, call it r':

$$r' = 2.15613\cdots$$

r' is real, but it is not in our countably infinite list, because its first digit differs from the first digit of the first number, its second digit differs from the second digit of the second number, and so on.

Therefore, given any countably infinite list of real numbers, this procedure can always produce a real number which was not in our countably infinite list. No matter how many countably infinite numbers we list, we can always produce a number which is not in that list. So any mapping from natural numbers to real numbers will never cover all of the real numbers.

□

(2) **Inverses and quantiles**:

Definition: The **inverse** f^{-1} of a function f returns the input of f when applied to the output of f. You can think of f^{-1} as "undoing" f. So, $f^{-1}(f(x)) = x\ \forall x$ in the domain of f.

Remark: We want to define something called the quantile function to answer the very useful question: "for what value of x do we have a cumulative probability p?" So, state a cumulative probability, like $p = \frac{1}{2}$, and we want to say what value of a random variable X corresponds to that cumulative probability.

Definition: A **quantile** function Q is given by $Q(p) = \inf\{x \in \mathbb{R} : F(x) \geq p\}$.

Note: We sketched some pictures to show that Q has the desired property: $Q(F(x)) = Q(P(X \leq x)) = x$. The definition is as messy as it is because we need to be able to handle the several types of discontinuities that are allowed in CDFs.

Examples: You have probably heard of probit and logit in empirical social science research. Probit is the quantile function of the normal distribution, and logit can be formulated as the quantile of a logistic distribution.

(3) **Two CDF properties with quantiles**:

We will need two lemmas:

Lemma 1: By the definition of monotonicity and the definition of a CDF, we can apply $F(\cdot)$ to both sides of an inequality without altering the inequality.

Lemma 2: For $U \sim \text{unif}(0; 1)$, $P(U \leq x) = x \ \forall x \in \mathbb{R}, 0 \leq x \leq 1$.

Property 1: For $U \sim \text{unif}(0; 1)$, $Y \equiv Q_X(U) \implies F_Y(y) = F_X(y)$.

Proof: Begin by writing the definition of the CDF of Y:

$$F_Y(y) = P(Y \leq y)$$

By the definition of Y in this question,

$$F_Y(y) = P(Q_X(U) \leq y)$$

By lemma 1,

$$F_Y(y) = P(F_X(Q_X(U)) \leq F_X(y))$$

Then since the quantile function is the inverse of the CDF,

$$F_Y(y) = P(U \leq F_X(y))$$

Invoking lemma 2,

$$F_Y(y) = F_X(y)$$

□

Property 2: For $U \sim \text{unif}(0; 1)$, $\inf\{x|F(X) > U\} \leq y \iff U \leq F_X(y)$.

Proof: To show the first direction, suppose

$$\inf\{x|F(X) > U\} \leq y$$

Then by lemma 2 and the definition of the quantile function,

$$Q_x(U) \leq y$$

By lemma 1,

$$F_x(Q_x(U)) \leq F_X(y)$$

Then since the quantile function is the inverse of the CDF,

$$U \leq F_X(y)$$

The other direction proceeds identically.

□

6 Section 6: 2018 October 18

Topics:

- Showing that a PDF is valid [1]
- Specifying conditions in which a PDF is valid [2]

After this class, I expect every student to be able to:

- show that simple PDFs integrate to 1

BACK TO TABLE OF CONTENTS

On this day, in addition to the exposition below, we had two extended conversations that I won't summarize here: one about why we claim that grades aren't important in a PhD program, and another about fraudulent research.

(1) Showing that a PDF is valid:

Exercise: Is the following PDF valid?

$$f(x,y) = \begin{cases} x+y & 0 \leq x,y \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

To check this, we need to show that it integrates up to 1 in the given interval. So,

$$\int_0^1 \int_0^1 f(x,y)dxdy = \int_0^1 \int_0^1 (x+y)dxdy$$

$$\int_0^1 \int_0^1 f(x,y)dxdy = \int_0^1 \left[\tfrac{1}{2}x^2 + xy\right]_0^1 dy$$

$$\int_0^1 \int_0^1 f(x,y)dxdy = \int_0^1 \left(\tfrac{1}{2} + y\right)dy$$

$$\int_0^1 \int_0^1 f(x,y)dxdy = \left[\tfrac{1}{2}y + \tfrac{1}{2}y^2\right]_0^1$$

$$\int_0^1 \int_0^1 f(x,y)dxdy = \tfrac{1}{2} + \tfrac{1}{2}$$

$$\int_0^1 \int_0^1 f(x,y)dxdy = 1$$

So, this is a valid PDF.

Note: To be truly careful, we should also check that the corresponding CDF is normalized between 0 and 1, monotonically increasing, and right-continuous. However, life is short and (with any luck) sections are even shorter, so I left this as an exercise.

(2) Specifying conditions in which a PDF is valid:

For what value of the real constant c is the following PDF valid?

$$f(x,y) = \begin{cases} cx^2y & -1 \leq x \leq 1, x^2 \leq y \leq 1 \\ 0 & \text{elsewhere} \end{cases}$$

To check this, we need to find the value of c that allows $f(x,y)$ to integrate up to 1 in the given interval. So,

$$\int_{-1}^1 \int_{x^2}^1 cx^2ydydx = \int_{-1}^1 \left[\tfrac{1}{2}xc^2y^2\right]_{y=x^2}^1 dx$$

$$\int_{-1}^1 \int_{x^2}^1 cx^2ydydx = c\int_{-1}^1 x^2\left(\tfrac{1}{2} - \tfrac{1}{2}x^4\right)dx$$

$$\int_{-1}^{1}\int_{x^2}^{1} cx^2 y dy dx = c\int_{-1}^{1} \tfrac{1}{2}x^2 - \tfrac{1}{2}x^6\big) dx$$

$$\int_{-1}^{1}\int_{x^2}^{1} cx^2 y dy dx = c\big[\tfrac{1}{6}x^3 - \tfrac{1}{14}x^7\big]_{-1}^{1}$$

$$\int_{-1}^{1}\int_{x^2}^{1} cx^2 y dy dx = c\big(\big(\tfrac{1}{6} - \tfrac{1}{14}\big) - \big(-\tfrac{1}{6} + \tfrac{1}{14}\big)\big)$$

$$\int_{-1}^{1}\int_{x^2}^{1} cx^2 y dy dx = c\tfrac{4}{21}$$

So $f(x, y)$ is a valid pdf if and only if $c = \frac{21}{4}$.

7 Section 7: 2018 October 25

Topics:

- An extended review of random variables and joint, conditional, and marginal distributions [1]

After this class, I expect every student to be able to:

- be able to explain in words how conditional, joint, and marginal distributions relate to an experiment through random variables
- be able to perform simple calculations to connect conditional, joint, and marginal distributions

BACK TO TABLE OF CONTENTS

① **An extended example of different types of distributions**:

Consider a cat, Pando, who likes to stand on the highest surface he can find and yell.

Pando, right, seeks out high ground so that people can hear him complain from farther away. Abyzou, left, goes on top of cupboards so that she can open them from above and steal the food inside.

When Pando is on a cabinet screaming, his behaviour is designed to influence Abyzou's; he wants somebody to play with him. But, like all cats, Pando behaves quite randomly. We could reasonably define a random variable, call it X, on what Pando is doing at a given moment, and another random variable Z which specifies what Abyzou is doing at a given moment. So, we can model Pando's behaviour as a probability distribution which is a function of Abyzou's behaviour. When we write the distribution of a random variable in terms of values of another random variable, we call that a **conditional probability distribution**; we represent the PMF of X conditional on Z as $f_{X|Z}(x|z)$. So, let's define the conditional probability distribution of X on Z using the following table, where the columns describe what Abyzou is doing and the rows describe what Pando is doing:

	Sleeping (0)	Looking out the window (1)	Running around (2)
Yelling (1)	$\frac{3}{5}$	$\frac{4}{5}$	$\frac{1}{5}$
¬ Yelling (0)	$\frac{2}{5}$	$\frac{1}{5}$	$\frac{4}{5}$

Note: This table should be interpreted as specifying the conditional probability that Pando is yelling given that Abyzou is sleeping, looking out the window, or running around. In order for this interpretation to make sense, we need the columns of the table to add up to one, since Pando's possible actions are mutually exclusive and span the entire sample space, and we want those numbers to legally represent the probability that Pando takes each

of those actions. The same is not true for the rows of the table, which **do not** represent a simple probability of the corresponding actions; the whole point of this exercise will be to figure out how to infer the probability of each of Pando's possible actions! The table says that if Abyzou is sleeping, then there is a $\frac{3}{5}$ probability that Pando is yelling and a $\frac{2}{5}$ probability that Pando is not yelling, but it **does not** say that, given that Pando is yelling, there is a $\frac{3}{5}$ probability that Abyzou is sleeping, a $\frac{4}{5}$ probability that she is looking out the window, and a $\frac{1}{5}$ probability that she is running around.

So, let's now assign a distribution of probabilities that Abyzou takes each of her possible actions. Previously we've just been calling this a "distribution", but when we're talking about the distribution of one variable out of a collection of multiple random variables, we call it a **marginal probability distribution**. Recall that a random variable assigns outcomes to some measurable space, so let's arbitrarily define the following correspondence:

Definition: Define the random variable $X : \Omega \to \mathbb{R}$ which maps Pando's behaviours as follows: $X(\text{Yelling}) = 1$, $X(\neg\text{Yelling}) = 0$.

Definition: Define the random variable $Z : \Omega \to \mathbb{R}$ which maps Abyzou's behaviours as follows: $Z(\text{Sleeping}) = 0$, $Z(\text{Looking out the window}) = 1$, $Z(\text{Running around}) = 2$.

Let's arbitrarily say that the probability that Abyzou is sleeping is given by $P(Z = 0) = \frac{3}{5}$, the probability that she is looking out the window is $P(Z = 1) = \frac{1}{5}$, and the probability that she is running around is $P(Z = 2) = \frac{1}{5}$. Note: since these are mutually exclusive outcomes which (let's say) cover all possibilities, they must sum to one.

Now, since $P(X|Z) = \frac{P(X \cap Z)}{P(Z)}$, we now have enough information to calculate the probability of any combination of two events happening. We call this the **joint probability distribution** of X and Z, and it tells us the probability that X takes some specific value while Z also takes some specific value. The logic here is pleasantly intuitive: if we want to know how likely it is that (for example) $X = 0$ and $Z = 0$, then we can simply figure out how likely it is that $Z = 0$, and then also figure out how likely it is that, given that $Z = 0$, X is also 0. In this example, this procedure looks like:

For $X = 0$ and $Z = 0$:

$$P(X = 0 \cap Z = 0) = P(X = 0|Z = 0) \cdot P(Z = 0)$$

$$P(X = 0 \cap Z = 0) = \tfrac{2}{5} \cdot \tfrac{3}{5}$$

$$P(X = 0 \cap Z = 0) = \tfrac{6}{25}$$

For $X = 0$ and $Z = 1$:

$$P(X = 0 \cap Z = 1) = P(X = 0|Z = 1) \cdot P(Z = 1)$$

$$P(X = 0 \cap Z = 1) = \frac{1}{5} \cdot \frac{1}{5}$$

$$P(X = 0 \cap Z = 1) = \frac{1}{25}$$

For $X = 0$ and $Z = 2$:

$$P(X = 0 \cap Z = 2) = P(X = 0|Z = 2) \cdot P(Z = 2)$$

$$P(X = 0 \cap Z = 2) = \frac{4}{5} \cdot \frac{1}{5}$$

$$P(X = 0 \cap Z = 2) = \frac{4}{25}$$

For $X = 1$ and $Z = 0$:

$$P(X = 1 \cap Z = 0) = P(X = 1|Z = 0) \cdot P(Z = 0)$$

$$P(X = 1 \cap Z = 0) = \frac{3}{5} \cdot \frac{3}{5}$$

$$P(X = 1 \cap Z = 0) = \frac{9}{25}$$

For $X = 1$ and $Z = 1$:

$$P(X = 1 \cap Z = 1) = P(X = 1|Z = 1) \cdot P(Z = 1)$$

$$P(X = 1 \cap Z = 1) = \frac{4}{5} \cdot \frac{1}{5}$$

$$P(X = 1 \cap Z = 1) = \frac{4}{25}$$

For $X = 1$ and $Z = 2$:

$$P(X = 1 \cap Z = 2) = P(X = 1|Z = 2) \cdot P(Z = 2)$$

$$P(X = 1 \cap Z = 2) = \frac{1}{5} \cdot \frac{1}{5}$$

$$P(X = 1 \cap Z = 2) = \frac{1}{25}$$

Now, we can tally up the joint distribution. Because we have now obtained valid mutually exclusive probabilities, we can derive the individual probability of each behaviour by simply summing up the rows and columns in the table of joint probabilities, as follows:

	Sleeping (0)	Looking out the window (1)	Running around (2)	Sum
Yelling (1)	$\frac{9}{25}$	$\frac{4}{25}$	$\frac{1}{25}$	$\frac{14}{25}$
$\neg$ Yelling (0)	$\frac{6}{25}$	$\frac{1}{25}$	$\frac{4}{25}$	$\frac{11}{25}$
Sum	$\frac{15}{25}$	$\frac{5}{25}$	$\frac{5}{25}$	1

The "sum" row shows how likely it is that each of Abyzou's actions happens on its own (and notice that it does match the probabilities we originally assigned to Abyzou's actions; if it didn't, then we would know that we had made a mistake). The "sums" column shows how likely it is that each of Pando's actions happens on its own. It must be true that entries of both the "sums" row and the "sums" column individually add up to 1.

8 Section 8: 2018 November 1

Topics:

- Random sampling [1]

  In-class example, identifying RVs

  Finite- and superpopulation assumption

After this class, I expect every student to be able to:

- describe, in words, the idea of sampling
- describe, mathematically and in words, what it means for variables to be iid, and how that relates to the idea of sampling
- define the finite population assumption and the superpopulation assumption
- identify the random variables in a sampling procedure

BACK TO TABLE OF CONTENTS

(1) **An extended example of different types of distributions**:

Definition: Random variables X, Y are **independent and identically distributed (iid)** if $F_{X,Y} = F_X(x)F_Y(y)$ and $F_X(x) = F_Y(x)$.

Example: n sequential coin tosses, $n \in \mathbb{N}$, will always be iid. With $X_i, i \in \mathbb{N}$ representing the probability of the coin showing a certain face after the i^{th} flip, we have $P(X_1 = 1) = P(X_2 = 1) = \cdots = P(X_n = 1)$, and similarly for all $P(X_i = 0)$, and also we know that the result in one flip does not affect the result in subsequent flips, so the $P(X_i = x)$ are all independent. Hence, this process is iid.

Example: In section I improvised an extended verbal example to do with sampling. Suppose we call n people and ask them m questions with q options each. I asked the class to take a few minutes and individually decide what the random variables are, what the outcomes of the random variables are, what numbers the random variables might map the outcomes onto, and whether or not those random variables are iid. Then, I asked: in social science, do we usually survey populations with or without replacement?

Definition: If we are sampling information from a population and we will eventually sample the value of every element of the population in finite time, then we are making the **finite population assumption**. This property holds whenever the population has a finite number of elements and we are sampling it without replacement.

Definition: If we are sampling information from a population and we will never sample the value of every element of the population in finite time, then we are making the **superpopulation assumption**. This property holds whenever the population has an infinite number of elements, or when the population has a finite number of elements and we are sampling it with replacement.

Remark: If we sample under the finite population assumption, then the sampled random variables are not identically distributed. This is because, every time we sample, we remove one individual from the population. So, the distribution of values among members of the population changes as we sample! If instead we sample under the superpopulation assumption, then the sampled random variables are identically distributed.

Next, to demonstrate all of this information, I conducted a poll in class. First, I asked every student two questions:

↪ Did you dress up for Halloween?
↪ How many cups of coffee do you drink per day on average?

I tallied every student's response, and, defining our class to be the population of interest, I drew the population values on the board. Next, generating random numbers on my phone, I conducted a few samples of these random variables from the class. Because we knew the population values (a luxury we never have in social science), we could answer the following

questions, where x_i and x_j, $i, j \in \mathbb{N}, i \neq j$ are two different realizations (students' sampled answers) of a random variable:

↪ What are the random variables in this example?

↪ What was the probability that $x_i = x_j$ under the finite population assumption?

↪ What was the probability that $x_i = x_j$ under the superpopulation assumption?

↪ On your own, calculate the probability that our next sample will have each possible value under each assumption.

↪ (If you finished quickly) can you write the general equation for the probability of two realisations being equal under each assumption?

Finally, having already described when the "identically distributed" part of iid can break down for sampled variables, I spoke briefly about dependencies in samples. The example I used is the idea of panel studies, which choose a group of people and then continue asking them questions over time to see how those exact peoples' opinions changed; this is an extreme example of dependent random variables in between two samples, because you're sampling values of the random variables from exactly the same people each time.

9 Section 9: 2018 November 8

Topics:

- Law of the Unconscious Statistician

  Theorem [1]

  Lemmas [2]

  Proof [3]

  Example [4]

After this class, I expect every student to be able to:

- find the expectation of a simple function of a random variable

BACK TO TABLE OF CONTENTS

① **The Law of the Unconscious Statistician**:

Theorem (LOTUS): Consider a function $g(X)$ of a random variable X. If X is discrete, then the expectation of $g(X)$ is given by

$$E[g(X)] = \sum_x g(x) f_X(x)$$

If X is continuous, then

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

The same exact idea applies to finding the expectation of a function of multiple random variables:

Theorem (Multivariate LOTUS): Consider a function $g(X, Y)$ of the random variables X and Y. If X and Y are discrete, then the expectation of $g(X, Y)$ is given by

$$E[g(X, Y)] = \sum_y \sum_x g(x, y) f_{X,Y}(x, y)$$

If X and Y are continuous, then

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{X,Y}(, yx) dx dy$$

Note: Multivariate LOTUS has an identical form for any number of at least 2 random variables; the statement above is WLOG because identical logic applies if you want to find the expectation of n random variables using their joint distribution.

Remark: LOTUS is truly remarkable for being the most obvious possible method for finding the expectation of a function of a random variable. All we have to do is take the usual expression for expecation of a random variable, and swap out x for $g(x)$.

Limitation: LOTUS requires that we know the distribution of X (or, in the case of multivariate LOTUS, the joint distribution of some $X_1, X_2, \cdots, X_n$).

② **Lemmas**:

Lemma (Integration by substitution, or "u-substitution"): For some function $\varphi(x)$ defined on the interval $[a; b]$ and a continuous function $f(\cdot)$,

$$\int_{\varphi(a)}^{\varphi(b)} f(u) du = \int_a^b f(\varphi(x)) \varphi'(x) dx$$

Proof of integration by substitution: By the chain rule, with F the antiderivative of f,

$$(f \circ \varphi)'(x) = F'(\varphi(x))\varphi'(x)$$

By the Fundamental Theorem of Calculus,

$$(f \circ \varphi)'(x) = f(\varphi(x))\varphi'(x)$$

So,

$$\int_a^b f(\varphi(x))\varphi'(x)dx = \int_a^b (F \circ \varphi)'(x)dx$$

$$\int_a^b f(\varphi(x))\varphi'(x)dx = (F\varphi(b)) - (F\varphi(a))$$

$$\int_a^b f(\varphi(x))\varphi'(x)dx = \int_{\varphi(a)}^{\varphi(b)} f(u)du$$

□

Lemma (Derivative of inverse): We need the following property

$$(f^{-1})'(a) = \frac{1}{f'(f^{-1}(a))}$$

I exclude the proof, which is much uglier than you might guess; we would have had to carefully define f, then use the Intermediate Value Theorem to show that its inverse is also continuous, then probably cry and go to lunch.

③ **Proof of LOTUS**:

I prove the continuous case of the single-variable version of LOTUS. I ran out of time to prove the discrete case, which isn't much of a problem, because it is tremendously easier to prove than the continuous case. The multivariate case is almost identical to the single-variable case.

Unfortunately, we have to impose some restrictions to make the proof tractable. First, we say that X is continuous, that g is a real-valued C^1 function, and that g has a monotonic inverse. This last assumption isn't pretty or reasonable, but it will make our job doable; the only proof I know of without this restriction requires a lot of measure theory that we don't discuss in this class.

First, define the random variable $Y \equiv g(X)$. Notice that this is a valid random variable because, by the definition of g, $X : \Omega \to \mathbb{R} \implies g(X) : \Omega \to \mathbb{R} \to \mathbb{R}$, so g is a function mapping from outcomes onto a measurable space; it satisfies our definition of a random variable. Now, proceed by application of the theorem of inverse function derivatives:

$$\frac{d}{dy}(g^{-1}(y)) = \frac{1}{g'(g^{-1}(y))}$$

Now notice that because $x = g^{-1}(y)$ by the definition of y,

$$dx = d[g^{-1}(y)]\frac{dy}{dy}$$

$$dx = \frac{d}{dy}[g^{-1}(y)]dy$$

$$dx = \frac{1}{g'(g^{-1}(y))}dy$$

Then, by the theorem of integration by substitution, and with $g(x) = y$ and $x = g^{-1}(y)$,

$$\int_{-\infty}^{\infty} g(x)f_X(x)dx = \int_{-\infty}^{\infty} yf_X(g^{-1}(y))\frac{1}{g'(g^{-1}(y))}$$

Call this Equation 1. Now, considering the definition of Y and its CDF, notice that

$$F_Y(y) = P(Y \leq y)$$

$$F_Y(y) = P(g(X) \leq y)$$

By the assumption that the inverse of g is monotonic, and using a very similar trick to the one we used when discussing quantiles,

$$F_Y(y) = P(X \leq g^{-1}(y))$$

$$F_Y(y) = F_X(g^{-1}(y))$$

And by the chain rule,

$$F_Y(y) = f_X(g^{-1}(y))\frac{1}{g'(g^{-1}(y))}$$

Plugging this expression into the line that we called Equation 1,

$$\int_{-\infty}^{\infty} yf_Y(y)dy = \int_{-\infty}^{\infty} g(x)f_X dx$$

Then, recalling the definition of Y,

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f_X dx$$

□

Note: The above proof has a close resemblence to the proof available on Wikipedia as of Fall 2018. There's a very good reason for that: I wrote nearly all of the Wikipedia page for LOTUS!

④**Example of LOTUS**:

To justify making everybody sit through the proof of LOTUS, I offered the following motivating example to illustrate the power of this theorem.

Example: Suppose that a random variable X has the following distribution:

$$f_X(x) = \begin{cases} 3x^2 & 0 < x < 1 \\ 0 & \text{else} \end{cases}$$

And let's say that we want to find $E[X^2]$.

Solution without using LOTUS: Let's pretend that you were taking a class with a very cruel GSI who didn't show you the beautiful tool that is LOTUS. Then you would have to proceed roughly as follows:

First, declare the random variable $Y \equiv X^2$. So, for $0 < y < 1$, let's begin by seeking the CDF of X in terms of a realisation of Y:

$$F_Y(y) = P(Y \leq y)$$

$$F_Y(y) = P(X^2 \leq y)$$

$$F_Y(y) = P(|X| \leq \sqrt{y})$$

$$F_Y(y) = P(-\sqrt{y} \leq X \leq \sqrt{y})$$

$$F_Y(\sqrt{y}) - F_X(-\sqrt{y})$$

By the range of Y,

$$F_Y(y) = F_X(\sqrt{y})$$

Next, we can find the PDF of Y by application of the chain rule:

$$f_Y(y) = F'_X(\sqrt{y})\tfrac{d}{dy}(\sqrt{y})$$

$$f_Y(y) = f_X(\sqrt{y})\tfrac{1}{2\sqrt{y}}$$

Because we know the PDF of X, we can substitute:

$$f_Y(y) = 3(\sqrt{y})^2\tfrac{1}{2\sqrt{y}}$$

$$f_Y(y) = \tfrac{3}{2}(\sqrt{y})$$

So,

$$f_Y(y) = \begin{cases} \frac{3}{2}\sqrt{y} & 0 < y < 1 \\ 0 & \text{else} \end{cases}$$

Now, that we know the distribution of Y, we can seek its expected value, as:

$$E[X^2] = E[Y]$$

$$E[X^2] = \int_{-\infty}^{\infty} y f_Y(y) dy$$

$$E[X^2] = \int_0^1 y \frac{3}{2}\sqrt{y} dy$$

$$E[X^2] = \frac{3}{2}\int_0^1 y^{\frac{3}{2}} dy$$

$$E[X^2] = \frac{3}{2}[\frac{2}{5}y^{\frac{5}{2}}]_{y=0}^1$$

$$E[X^2] = \frac{3}{5}$$

At last, we have found the expectation of Y without using LOTUS! Now, let's figure out how we would do it with LOTUS.

Solution using LOTUS: By LOTUS,

$$E[X^2] = \int_{-\infty}^{\infty} x^2 f_X(x) dx$$

$$E[X^2] = \int_0^1 x^2 3x^2 dx$$

$$E[X^2] = [\frac{3}{5}x^5]_{x=0}^1$$

$$E[X^2] = \frac{3}{5}$$

This was bad enough. Now imagine if the PDF had been messier, or if instead of X^2 the function had been, say, $\sin(iX)^{\frac{83}{e}}$. Thank goodness we know LOTUS now!

10 Section 10: 2018 November 15

Topics:

- Law of total variance

  Theorem [1]

  Lemmas [2]

  Proof [3]

- Partial correlations [4]

After this class, I expect every student to be able to:

- Explain in words what the law of total variance is
- Explain what partial correlations are

BACK TO TABLE OF CONTENTS

①**The law of total variance**:

Theorem (The law of total variance): For two random variables X, Y,

$$V(X) = E[V(X|Y)] + V(E[X|Y])$$

②**Lemmas**:

Theorem (Law of iterated expectations): For random variables X, Y, $E[X] = E[E[X|Y]]$

Since we had discussed this result in detail during lecture, and because the proof takes some time, I omitted it.

Theorem: For some random variable X, $V(X) = E[X^2] - E^2[X]$

Proof of variance property:

Begin with the definition of variance:

$$V(X) = E[(X - \mu)^2]$$

$$V(X) = E[(X - E[X])^2]$$

$$V(X) = E[X^2 - 2XE[X] + E^2[X]]$$

By the linearity of expectations,

$$V(X) = E[X^2] - 2E[XE[X]] + E[E^2[X]]$$

$$V(X) = E[X^2] - 2E^2[X] + E^2[X]$$

$$V(X) = E[X^2] - E^2[X]$$

□

This is considered an elementary property of variance, and is a tremendously important property to know when working with variances.

③**Proof of the law of total variance**:

Begin with the variance property:

$$V(X) = E[X^2] - E^2[X]$$

By the law of iterated expectations,

$$V(X) = E[E[X^2|Y]] - E^2[E[X|Y]]$$

By another application of the variance property,

$$V(X) = E[V(X|Y) + E^2[X|Y]] - E^2[E[X|Y]]$$

Again by the linearity of expectations,

$$V(X) = E[V(X|Y)] + E[E^2[X|Y]] - E^2[E[X|Y]]$$

Then by yet another application of the variance property,

$$V(X) = E[V(X|Y)] + V(E[X|Y])$$

□

④ **Partial correlations**:

I ended the class with a baroque example of partial correlations, using the example of support for the Green Party of Canada, which was a bit too visual to usefully reproduce here. The core point was that correlations can give us a very nice intuition for what we mean when we talk about "controlling for" a variable, which will be dealt with in much more detail in POLSCI 699.

11 Section 11: 2018 November 29

Topics:

- Moments
  - List of moments [1]
  - Moment-generating functions (MGFs) [2]
  - Finding moments from MGFs [3]
- The Gaussian MGF
  - Completing the square [4]
  - Derivation [5]

After this class, I expect every student to be able to:

- describe in words what a statistical moment and a moment generating function are
- explain what we mean when we talk about the first moment, the second moment, and so on

BACK TO TABLE OF CONTENTS

(1) **Moments**:

Definition: A **moment** is a measure of the shape of a function. Glossing over some (important!) complications, for this class we will say for simplicity that the nth moment m_n is given by $m_n = E[X^n]$.

The moments are given, in this order, by:

1. Mean, which roughly measures the center of the distribution
2. Variance, which roughly measures the spread of the distribution
3. Skewness, which roughly measures which way the distribution leans
4. Kurtosis, which roughly measures how thick the tails of the distribution are

The next few moments (hyperskewness, hyperflatness, and so on) no longer have easy interpretations. But what do I mean by imposing this order on the moments? This is where the MGF comes in.

(2) **Moment generating functions**:

Definition: The **moment-generating function** M is an alternate expression for the distribution of a random variable X, like the probability function and cumulative density function. It is stated in terms of e^{tX}, for some $t \in \mathbb{R}$, as

$$M_X(x) = E[e^{tX}]$$

(3) **Finding moments from the MGF**:

Consider the series expansion (which we didn't prove because I'm not a sadist):

$$e^{tx} = 1 + tX + \frac{t^2X^2}{2!} + \frac{t^3X^3}{3!} + \cdots$$

So, by the linearity of expectations,

$$E[e^{tx}] = 1 + tE[X] + \frac{t^2E[X^2]}{2!} + \frac{t^3E[X^3]}{3!} + \cdots$$

Representing the nth moment as m_n, $n \in \mathbb{N}$, we can use the definition of moments to substitute

$$E[e^{tX}] = 1 + tm_1 + \frac{t^2m_2}{2!} + \frac{t^3m_3}{3!} + \cdots$$

So if we differentiate this expansion n times with respect to t and set $t = 0$, we will obtain exactly the nth moment!

$$m_n = \frac{d^n}{dt^n} M_X\Big|_{t=0}$$

Exercise: Verify that this is true when $n = 2$.

④ **Tool for Gaussian MGF**:

Tool: To **complete the square** is to convert a polynomial from the form ax^2+bx+c to the form $a(x-h)^2+k$ using the fact that $(x+n)^2 = x^2+2xn+n^2$, with $a, b, c, h, k, n \in \mathbb{R}$.

Example: Complete the square for the polynomial $2x^2 + 20x + 60$.

$$2x^2 + 20x + 60 = 2x^2 + 20x + 60$$

$$2x^2 + 20x + 60 = 2(x^2 + 10x + 30)$$

$$2x^2 + 20x + 60 = 2(x^2 + 10x + 25) + 10$$

$$2x^2 + 20x + 60 = 2(x + 5)^2 + 10$$

Which is the form that we wanted it in, with $a = 2, h = -5, k = 10$. Evidently, completing the square is like factoring but for people who are really hardcore.

⑤ **Deriving the Gaussian MGF**:

What is the moment generating function of a normally distributed random variable, $X \sim \mathcal{N}(\mu, \sigma^2)$? A natural place to start is with LOTUS:

$$E[e^{tX}] = \int_{-\infty}^{\infty} e^{tX} f_X(x) dx$$

Substituting according to the definition of the normal distribution,

$$M_X(x) = \int_{-\infty}^{\infty} e^{tX} \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}\right) dx$$

$$M_X(x) = \int_{-\infty}^{\infty} \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2+tx}\right) dx$$

$$M_X(x) = \int_{-\infty}^{\infty} \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x^2-2x\mu+\mu^2)+tx}\right) dx$$

$$M_X(x) = \int_{-\infty}^{\infty} \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x^2-2x\mu-2\sigma^2tx+\mu^2)}\right) dx$$

$$M_X(x) = \int_{-\infty}^{\infty} \left(\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(x^2-2x(\mu+\sigma^2t)+\mu^2)}\right) dx$$

Now to simplify the mess in that exponent, let's complete the square to show that the following identity holds:

$$x^2 - 2x(\mu + \sigma^2 t) + \mu^2 = (x - (\mu + \sigma^2 t))^2 - (\mu + \sigma^2 t)^2 + \mu^2$$

We have $X^2 - 2x(\mu + \sigma^2 t) + \mu^2$. Let us define that $a \equiv \mu + \sigma^2 t$ and $b \equiv \mu^2$. So,

$$x^2 - 2x(\mu + \sigma^2 t) + \mu^2 = x^2 - 2ax + b$$

Completing the square,

$$x^2 - 2x(\mu + \sigma^2 t) + \mu^2 = x^2 - 2ax + a^2 - a^2 + b$$

$$x^2 - 2x(\mu + \sigma^2 t) + \mu^2 = (x - a)^2 - a^2 + b$$

Plugging this back into the expression we had for the exponent,

$$x^2 - 2x(\mu + \sigma^2 t) + \mu^2 = (x - (\mu + \sigma^2 t))^2 - (\mu + \sigma^2 t)^2 + \mu^2$$

And then plugging this identity into the expression for M_X,

$$M_X(x) = \int_{-\infty}^{\infty} \left(\frac{1}{\sigma\sqrt{2\pi}} e^{(x-(\mu+\sigma^2 t))^2 - (\mu+\sigma^2 t)^2 + \mu^2} \right) dx$$

Separating the exponents and moving constants out of the integral,

$$M_X(x) = e^{-\frac{1}{2\sigma^2}(-\mu+\sigma^2 t)^2 + \mu^2} \int_{-\infty}^{\infty} \left(\frac{1}{\sigma\sqrt{2\pi}} e^{(x-(\mu+\sigma^2 t))^2} \right) dx$$

But notice that what is within the integral is simply the PDF of a random variable with distribution $\mathcal{N}(\mu + \sigma^2 t, \sigma)$. Since we know this to be a valid PDF, it must integrate to one over the entire real line. So the Gaussian MGF reduces to

$$M_X(x) = e^{-\frac{1}{2\sigma^2}(-\mu+\sigma^2 t)^2 + \mu^2}$$

After some straightforward algebra to simplify the exponent, we arrive at the statement of the Gaussian MGF that Yuki states in his lecture slides:

$$M_X(x) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$$

□

12 Section 12: 2018 December 6

In this class I talked through a proof of the central limit theorem using MGFs that Yuki didn't have time to cover in lecture, so I refer you to his slides where it's already typed up. Then, we did a review exercise in groups.

BACK TO TABLE OF CONTENTS