Conditional probability

- Given that event A of probability P(A) > 0 occurred, the conditional probability of B given A is denoted by P(B|A) and defined as
$$P(B|A) = \frac{P(AB)}{P(A)}$$
- P(B|A) can be larger than, smaller than, or the same as P(B)
- Conditional probabilities satisfy the axioms of probability theory

The chain rule or product rule

- P(B|A) = P(AB)/P(A)
- P(AB) = P(B|A)P(A)

  P(ABCD…)
  = P(A)P(B|A)P(C|AB)P(D|ABC)…
- The chain rule also applies to conditional probabilities given an event H (say)

  P(ABCD… |H)
  = P(A|H)P(B|AH)P(C|ABH)P(D|ABCH)…

The theorem of total probability I

- The theorem of total probability allows us to compute unconditional probabilities from conditional probabilities

  $P(A) = P(A|B)P(B) + P(A|B^c)P(B^c)$

  $P(B) = P(B|A)P(A) + P(B|A^c)P(A^c)$
- The theorem also applies to conditional probabilities

  $P(A|C) = P(A|BC)P(B|C)+P(A|B^c C)P(B^c|C)$

The theorem of total probability II

Given a countable partition $A_1, A_2, \ldots A_n, \ldots$ of the sample space,

$$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + \ldots + P(B|A_n)P(A_n) + \ldots$$

P(B|C)

$$= P(B|A_1C)P(A_1|C) + P(B|A_2C)P(A_2|C) + \ldots + P(B|A_nC)P(A_n|C) + \ldots$$

Checking the answers

- $\min_i P(B|A_i) \le P(B) \le \max_j P(B|A_j)$
- In particular,

  $P(B) \ge \min\{P(B|A), P(B|A^c)\}$

  $P(B) \le \max\{P(B|A), P(B|A^c)\}$
- Equality holds if and only if P(A) = 0 or P(A) = 1 or $P(B|A) = P(B|A^c)$
- If P(A) = 1/2, $P(B) = [P(B|A) + P(B|A^c)]/2$

Why is all this stuff important?

- The chain rule or product rule allows us to compute a joint probability (probability of an intersection) as the product of various conditional probabilities
- The theorem of total probability allows us to find an unconditional probability from conditional probabilities
- Results are very important and very useful tools in probabilistic analyses

Conditional pmf of X

- The pmf of a discrete random variable **X** tells us the probabilities with which we observe **X** taking on various values
- When partial knowledge is available about the outcome, we should update the pmf probabilities from their original values to the corresponding conditional probabilities
- This updated or modified pmf is called the conditional pmf of **X**

Definition of conditional pmf of X

- The pmf of a discrete random variable **X** taking on values $u_1, u_2, \ldots$ is given by

  $p_{\mathbf{X}}(u_i) = P\{\mathbf{X} = u_i\}, i = 1, 2, \ldots$
- Given that event A (with P(A) > 0) has occurred, the conditional pmf of **X** is

  $p_{\mathbf{X}|A}(u_i|A) = P\{\mathbf{X} = u_i|A\}, i = 1, 2, \ldots$

  where the right side now has conditional probabilities given that A occurred
- $P\{\mathbf{X} = u_i|A\} = P[\{\mathbf{X} = u_i\} \cap A]/P(A)$

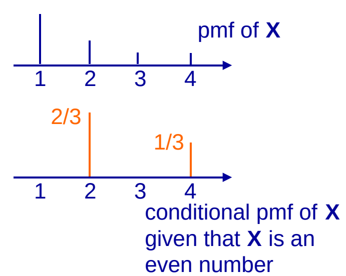
Some thoughts on conditional pmfs

- In defining the pmf of **X**, it was assumed that values $u_1, u_2, \ldots$ occur with nonzero probabilities, i.e. $p_{\mathbf{X}}(u_i) > 0$ for each u_i
- Given that the event A occurred, it might be that certain values u_i of **X** can never occur under these conditions
- $\{\mathbf{X} = u_i\} \cap A =$
- In the conditional pmf of **X**, $p_{\mathbf{X}|A}(u_i|A) = 0$ for such values u_i

Example: conditional pmf

- **X** has value 1, 2, 3, and 4 respectively with probabilities 1/2, 1/4, 1/8, and 1/8
- Let A = {**X** is an even number}; P(A) = 3/8
- The conditional pmf of **X** is given by
  - $p_{\mathbf{X}|A}(1|A) = 0$
  - $p_{\mathbf{X}|A}(3|A) = 0$
  - $p_{\mathbf{X}|A}(2|A) = 2/3$
  - $p_{\mathbf{X}|A}(4|A) = 1/3$
- Note that the conditional probabilities have the same ratios (1/4):(1/8) as (2/3):(1/3)

Graphical representation



Conditional mean and variance

- Just as conditional probabilities are a probability measure, conditional pmfs are valid pmfs
- The conditional expectation of **X** is the expectation of **X** computed using the conditional pmf of **X** instead of the pmf
- $E[\mathbf{X}|A] = \sum u_i \cdot p_{\mathbf{X}|A}(u_i|A)$
- Similarly, the conditional variance uses the conditional pmf instead of the pmf

Unconditional pmf of X

- The unconditional pmf of **X** is related to conditional pmfs via the theorem of total probability
- $p_X(u_i) = p_{X|A}(u_i|A) \bullet P(A) + p_{X|A}(u_i|A^c) \bullet P(A^c)$
- This extends to larger partitions $A_1, A_2, \ldots$ in the obvious way
- $p_X(u_i) = \sum p_{X|A_k}(u_i|A_k) \bullet P(A_k)$; sum is on k
- Don't get flustered by fancy notations; this is just the theorem of total probability

Unconditional expected value of X

- E[**X**], the unconditional expected value of **X**, is the weighted sum of the conditional expected values of **X**
- $$E[X] = \sum_i u_i \bullet p_X(u_i) = \sum_i u_i \bullet \sum_k p_{X|A_k}(u_i|A_k) \bullet P(A_k) = \sum_k P(A_k) \bullet \sum_i u_i \bullet p_{X|A_k}(u_i|A_k) = \sum_k P(A_k) \bullet E[X|A_k]$$
  (where $\sum_i u_i \bullet p_{X|A_k}(u_i|A_k) = E[X|A_k]$)

Unconditional variance of X

- The unconditional variance of **X** is found via the result $E[(X - a)^2] = var(X) + (\mu - a)^2$
- $$var(X) = \sum_i (u_i-\mu)^2 \bullet p_X(u_i) = \sum_i (u_i-\mu)^2 \bullet \sum_k p_{X|A_k}(u_i|A_k) \bullet P(A_k) = \sum_k P(A_k) \bullet \sum_i (u_i-\mu)^2 \bullet p_{X|A_k}(u_i|A_k) = \sum_k P(A_k) \bullet [var(X|A_k) + (\mu_k - \mu)^2]$$
  where μ_k denotes $E[X|A_k]$

Another way of looking at it

- $\mu = E[X] = \sum P(A_k) \bullet E[X|A_k] = \sum \mu_k \bullet P(A_k)$
- $var(X) = \sum P(A_k) \bullet [var(X|A_k) + (\mu_k - \mu)^2]$
  $= \sum var(X|A_k) \bullet P(A_k) + \sum (\mu_k - \mu)^2 \bullet P(A_k)$
- The first sum is the weighted sum of the conditional variances of **X**
- Let random variable **Y** take on values $\mu_1, \mu_2, \ldots$ with probabilities $P(A_1), P(A_2), \ldots$
- $E[Y] = \sum \mu_k \bullet P(A_k) = \mu$

Weird, weird, weird…

- **Y** take on values μ_k with probs $P(A_k)$
- $E[Y] = \sum \mu_k \bullet P(A_k) = \mu$
- $var(Y) = \sum (\mu_k - \mu)^2 \bullet P(A_k)$
- $var(X) = \sum var(X|A_k) \bullet P(A_k) + \sum (\mu_k - \mu)^2 \bullet P(A_k)$
- First sum is the mean of the conditional variances of **X**
- Second sum is the variance of the conditional means!

Really slick formulas…

- The mean of **X** is the weighted sum of the conditional means, that is, the mean of the conditional means of **X**
- $\mu = E[X] = \sum P(A_k) \bullet E[X|A_k] = \sum \mu_k \bullet P(A_k)$
- The variance of **X** is the mean of the conditional variances plus the variance of the conditional means
- $var(X) = \sum var(X|A_k) \bullet P(A_k) + \sum (\mu_k - \mu)^2 \bullet P(A_k)$

Example

- Given A, **X** is binomial; parameters (n, p)
- Given A^c, **X** is binomial; parameters (n, q)
- P(A) = 1/2
- $q \neq 1 - p$
- $P\{\mathbf{X} = k\} = P\{\mathbf{X}=k|A\}P(A) + P\{\mathbf{X}=k|A^c\}P(A^c)$

  = average of two binomial probabilities
- The unconditional pmf of **X** is a mess
- E[**X**] = ? $E[\mathbf{X}|A] = np$ and $E[\mathbf{X}|A^c] = nq$
- E[**X**] = (np + nq)/2

Example (continued)

- Given A, **X** is binomial; parameters (n, p)
- Given A^c, **X** is binomial; parameters (n, q)
- P(A) = 1/2
- $q \neq 1 - p$
- $\text{var}(\mathbf{X}|A) = np(1-p)$
- $\text{var}(\mathbf{X}|A^c) = nq(1-q)$
- The conditional mean is np or nq with equal probability
- $\text{var(conditional mean)} = [n(p-q)/2]^2$
- $\text{var}(\mathbf{X}) = [np(1-p)+nq(1-q)]/2 + [n(p-q)/2]^2$

Comment

- The example is a very simple illustration of the use of conditional expectations
- In many important instances, conditional expectations are much easier to calculate than the unconditional expectations
- The theorem of total probability is the tool that we use to obtain unconditional expectations from conditional expectations

Bayes' Formula … at last!

- Given that event A of probability P(A) > 0 occurred, the conditional probability of B given A is denoted by P(B|A) and defined as

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

- What is P(A|B)?

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

- Simplest version of Bayes' formula

Many names; some wrong 'uns too

- Bayes' formula

$$P(A|B) = P(B|A)P(A)/P(B)$$

  is also called Bayes' theorem, or Bayes' lemma, or often (mistakenly) Bayes' rule
- Bayes' rule (discussed later) refers to a methodology for decision-making that is an extremely controversial topic among statisticians
- No controversy about Bayes' formula

Is that all there is to it?

- Bayes' formula

$$P(A|B) = P(B|A)P(A)/P(B)$$

  is a very simple consequence of the definition of conditional probability
- In applications, P(B) is usually replaced by its equivalent expression as given by the theorem of total probability
- $P(A|B) = \dfrac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$

Look closely at the formula …

- $P(A|B) = \dfrac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)}$
- The numerator is always one of the terms in the denominator
- When is $P(B|A)P(A) \neq P(B|A)P(A)$?
- When it is calculated twice by an ECE 313 student in a hurry on homework or exam!
- Save and re-use the appropriate term!

General version of Bayes' formula

- When P(B) is obtained from $P(B|A_k)$'s via the more general version of the theorem of total probability, the more general total probability appears in the denominator
- The numerator is still one of the terms in the denominator
- $P(A_k|B) = \dfrac{P(B|A_k)P(A_k)}{\sum_i P(B|A_i)P(A_i)}$

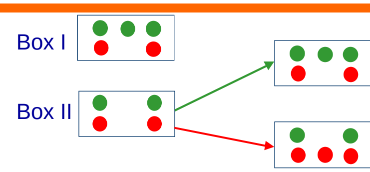
Example of use of Bayes' formula

- Example: Box I has 3 green and 2 red balls, while Box II has 2 green and 2 red balls. A ball is drawn at random from Box I and transferred to Box II. Then, a ball is drawn at random from Box II.
- G = event ball drawn from Box II is green
- A = event ball transferred is red
- $P(G|A) = 2/5$ • $P(G|A^c) = 3/5$ • $P(A) = 2/5$

Example (continued)

- G = event ball drawn from Box II is green
- A = event ball transferred is red
- $P(G|A) = 2/5$ • $P(G|A^c) = 3/5$ • $P(A) = 2/5$
- $P(G) = P(G|A)P(A) + P(G|A^c)P(A^c)$
  $= (2/5)(2/5) + (3/5)(3/5) = 13/25$
- $P(A|G) = P(G|A)P(A)/P(G) = 4/13$
- $P(A^c|G) = P(G|A^c)P(A^c)/P(G) = 9/13$
- Check: $P(A^c|G) = 1 – P(A|G)$

Does this make any sense?



- $P(A|G) = 4/13$, $P(A^c|G) = 9/13 > P(A|G)$
- If a green ball is drawn from Box II, it is reasonable to assume (more likely) that a green ball was transferred over!

Another example

- Whenever there is a fire, a fire alarm rings with probability $1 – 10^{–8}$. If there is no fire, the fire alarm does occasionally ring. The probability of such false alarms is $10^{–5}$
- Both these probabilities can be measured experimentally by the manufacturer (and touted in its advertising!)
- What the user is more concerned about is: When the fire alarm is ringing, is there actually a fire? or is it just a false alarm?

Where's the fire?

- R = event that the alarm is ringing
- F = event that there is a fire
- $P(R|F) = 1 - 10^{-8}$ • $P(R|F^c) = 10^{-5}$
- $P(R) = P(R|F)P(F) + P(R|F^c)P(F^c)$ depends on P(F)
- $P(F|R) = P(R|F)P(F)/P(R)$ depends on P(F)
- The manufacturer cannot know what P(F) is going to be; it depends on the location!

Living in a firetrap

- $P(R|F) = 1 - 10^{-8}$ • $P(R|F^c) = 10^{-5}$
- Suppose that P(F), the probability that there is a fire, is 10^{-7}
- $P(R) = P(R|F)P(F) + P(R|F^c)P(F^c)$

  $= (1 - 10^{-8})\bullet 10^{-7} + 10^{-5}\bullet(1 - 10^{-7})$

  $\approx 10^{-5} + 10^{-7} = 0.000010100\ldots$
- $P(F|R) = P(R|F)P(F)/P(R) \approx 10^{-7}/1.01\times 10^{-5}$

  $= 1/101$ $P(F^c|R) = 100/101$

Sleeping in & ignoring the ringing!

- $P(R|F) = 1 - 10^{-8}$ • $P(R|F^c) = 10^{-5}$
- Suppose that P(F), the probability that there is a fire, is 10^{-7}
- $P(F|R) = 1/101$ and $P(F^c|R) = 100/101$
- Out of 101 times that the alarm is ringing, only once will there be an actual fire!
- Should we just continue sleeping and ignore the alarm?

False alarms or false positives

- In many kinds of tests (including medical tests), a false alarm or a false positive is said to occur when the test indicates the existence of a condition when in fact the condition does not exist
- This is especially problematical when the condition itself is quite rare
- The false positives from the large number of tests in the absence of the condition mask the true positives

False negatives or false dismissal

- The opposite kind of test failure occurs when the test indicates the condition is absent when the condition is in fact present
- We are now looking at $P(F|R^c)$, the probability of being burnt alive in our beds as opposed to $P(F^c|R)$, the probability of freezing on the sidewalk while waiting for the fire trucks to show up!

False negatives

- $P(R|F) = 1 - 10^{-8}$ • $P(R|F^c) = 10^{-5}$
- $P(F) = 10^{-7}$ • $P(R) \approx 10^{-5} + 10^{-7}$
- $P(F|R^c) = P(R^c|F)P(F)/P(R^c)$

  $\approx 10^{-8}\times 10^{-7}/(1 - 10^{-5}) \approx 10^{-15}$
- In most tests, the probability of a false positive can be traded off against the probability of a false negative
- Paranoia versus ostrichism

What comes next?

- In the next two lectures, we shall discuss the general notion of decision making in the face of uncertainty
  - How to make decisions?
  - What are the probabilities that our decisions are incorrect?
  - What is the decision rule that minimizes the error probability?
  - What decision rule minimizes the costs?

Summary

- We discussed the notion of the conditional pmf of a random variable and related it to the pmf via the theorem of total probability
- We discussed conditional means and variances and their relation to the mean and variance
- We discussed Bayes' formula and some simple applications