Probability Review Outline (Revised)

Heidi Klumpe

Tuesday, September 27, 2016 15:29

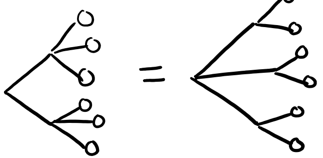
I. Naive Definition of Probability (Counting + Sampling)
II. Axioms and Notation
III. PMF v. PDF
IV. Common summary statistics
V. Review of freq. vs Bayes
VI. Questions from class

I. Naive Definition

$$P(x_i) = \frac{1}{\#\text{ of events}}$$ } assumes all events equally likely

★ Need to know how to count!

- multi-component experiment (multiplication rule)
  - 2 cone types, 3 ice cream flavors



$$\# = \prod_{i}^{N} n_i = n_1 \cdot n_2 \cdots n_N$$

# of outcomes for step i

  - experimental conditions

$\left(\begin{matrix}3\text{ treatments}\\2\text{ controls}\end{matrix}\right)\left(6\text{ timepoints}\right)\left(\begin{matrix}2\text{ tech}\\\text{replicates}\end{matrix}\right)$

- number of possible samples of size k from a population of size n

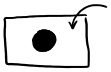
	(permutations) order matters	(combinations) order doesn't matter
with replacement	n^k	$\binom{n+k-1}{k}$ ← replacement adds options (nonrigorous intuition)
w/o replacement	$\frac{n!}{(n-k)!}$	$\frac{1}{k!}\frac{n!}{(n-k)!} = \binom{n}{k}$

multiplication rule

divide by # of identical permutations

II. Axioms and Notation

sample space



P(A)

$P(\bar{A})$

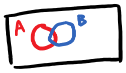
- sum rule: $P(A) + P(\bar{A}) = 1$
  → complement
- product rule: $P(A,B) = P(A|B)\,P(B) = P(B|A)\,P(A)$
  → conditional probability
- independence : $P(A|B) = P(A)$

can always add conditions

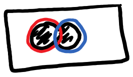
$P(A|C) + P(\bar{A}|C) = 1$

$P(A,B|C) = P(A|B,C)\,P(B|C)$

- union and intersection

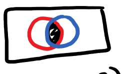


"OR"



$P(A \cup B)$

"AND"



$P(A \cap B)$

$P(A \cap B) = P(A|B)\,P(B)$

$= P(A)\,P(B)$ if A and B are independent

(also, $P(A \cap B) = P(B \cap A) = P(A,B)$)

$P(A \cup B)$ $\quad$ $P(A \cap B)$

$P(A \cap B) = P(A|B)\,P(B)$

$= P(A)$ … if A and B are independent

(also, $P(A \cap B) = P(\ldots$ $= P(A,B))$

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

n.b. $P(A|B) \neq P(A \cap B)$ though it might make sense as an English sentence

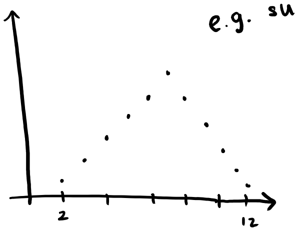
$= \dfrac{P(A \cap B)}{P(B)}$

we could restate most probabilities this way (whether you're using a Bayesian or a frequentist approach, it's best practice to include prior knowledge in your calculation)

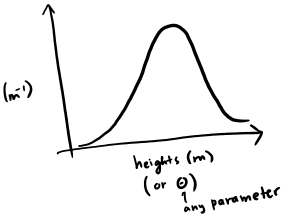
III. PMF v. PDF

- probability mass function
  - → discrete variables
  - → probabilities sum to 1
    - $\sum_{\text{all } i} P_i = 1$
    - must be unitless
    - $P(x_i)$ can have a finite value

e.g. sum of dice

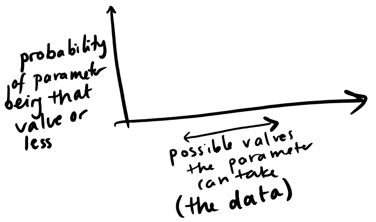


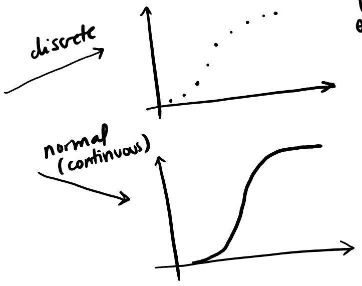
- probability density function
  - → continuous variables
  - → probabilities integrate to 1
    - $\int_{-\infty}^{\infty} d\theta\, P(\theta) = 1$
    - must have units $\frac{1}{\theta}$
    - $P(a \leq x \leq b) = \int_a^b d\theta\ P(\theta)$
    - $P(1.6542\ldots m) = 0$, $P(A) = \int_A^A d\theta\, P(\theta) = 0$



we can't compute the probability of being a specific height

- cumulative distribution function





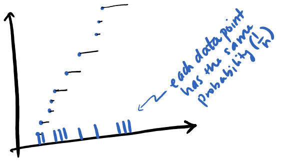
$\lim_{\theta \to -\infty} = 0$

$\lim_{\theta \to \infty} = 1$

eCDF (empirical)

or

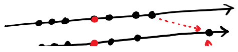
discrete data points for continuous parameter



IV. Common Summary Statistics

→ mean
- arithmetic
- geometric (log spaced data)

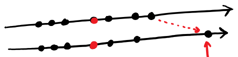
n.b. median is less sensitive to outliers



IV.

→ mean
- · arithmetic
- · geometric (log spaced data)

n.b. sensitive to outliers



median unchanged even if highest point increases many fold! (average changes a lot)

"value you most likely expect"

- expectation value - $\langle x \rangle$ weighted average of possible outcomes

$$\sum_{i=1}^{N} x_i P_i \quad \text{(discrete)}$$

n.b. if all x_i are equally likely $P_i = \frac{1}{N}$ and $\langle x \rangle = \frac{1}{N} \sum_i x_i$, which is the definition of the average!

$$\int_x dx \; x P(x) \quad \text{(continuous)}$$

- variance - $\sigma^2 = \langle (x - \langle x \rangle)^2 \rangle$ weighted average of spread

$$= \langle x^2 - 2x\langle x \rangle + \langle x \rangle^2 \rangle$$

(⟨ ⟩ is a linear operator, separate terms) and remove constants

$$= \langle x^2 \rangle - 2\langle x \langle x \rangle \rangle + \langle \langle x \rangle^2 \rangle$$

(⟨x⟩ is also a constant!

$$\langle \langle x \rangle^2 \rangle = \langle x \rangle^2 \longrightarrow \sum_j \underbrace{\Big(\sum_i x_i P_i\Big)^2}_{\text{no } j \text{ index}} P_j = \Big(\sum_i x_i P_i\Big)^2 \underbrace{\sum_j P_j}_{=1 \text{ by sum rule}}$$

)

$$= \langle x^2 \rangle - 2\langle x \rangle^2 + \langle x \rangle^2$$

$$= \langle x^2 \rangle - \langle x \rangle^2$$

- can be summarized as "moments" of moment generating functions

first moment: $\mu_1 = \langle x \rangle$

second moment: $\mu_2 = \langle x^2 \rangle$

→

mean $= \langle x \rangle$

var $= \sigma^2 = \mu_2 - \mu_1^2 = \langle x^2 \rangle - \langle x \rangle^2$

st. dev. $= \sigma$

coeff. of variance $= \dfrac{\sigma}{\langle x \rangle}$ (normalized stdev)

given $P(x)$ you can derive analytical expressions for mean and stdev (easier for discrete)

Example: Poisson Distribution

① We know the PMF : $P(k) = \dfrac{\lambda^k e^{-\lambda}}{k!}$

② We can find ⟨k⟩ b/c we know P(k)!

$$\langle k \rangle = \sum_{k=1}^{\infty} k \left(\frac{\lambda^k e^{-\lambda}}{k!} \right)$$

k=0 term is equal to 0

→ $\dfrac{k}{k!} = \dfrac{1}{(k-1)!}$

→ we can remove constants

→ $\displaystyle\sum_{x=0}^{\infty} \frac{\lambda^x}{x!} = e^{\lambda}$ (exponential series formula)

$$= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!}$$

let j= k-1

$$= \lambda e^{-\lambda} \sum_{j=0}^{\infty} \frac{\lambda^j}{j!}$$

$$= \lambda e^{-\lambda} e^{\lambda}$$

$$\langle k \rangle = \lambda$$

③ $\sigma_k^2 = \langle k^2 \rangle - \underbrace{\langle k \rangle^2}_{\lambda^2}$

$$\langle k^2 \rangle = \sum_{k=1}^{\infty} k^2 P(k)$$

$$= \sum_{k=1}^{\infty} k^2 \frac{\lambda^k e^{-\lambda}}{k!}$$

→ $\dfrac{k^2}{k!} = \dfrac{k}{(k-1)!}$

→ $\lambda e^{-\lambda}$ comes out as a constant

$$= \lambda e^{-\lambda} \sum_{k=1}^{\infty} k \frac{\lambda^{k-1}}{(k-1)!}$$

need this term to be 1 to use our exponential series formula

$$= \lambda e^{-\lambda} \left(\sum_{k=1}^{\infty} (k-1+1) \frac{\lambda^{k-1}}{(k-1)!} \right)$$

can split summation terms

our exponential

$$= \lambda e^{-\lambda} \left(\sum_{k=1}^{\infty} (k-1+1) \frac{\lambda^{k-1}}{(k-1)!} \right)$$

can split summation terms

$$= \lambda e^{-\lambda} \left[\sum_{\cancel{k=1}}^{\infty} (k-1) \frac{\lambda^{k-1}}{(k-1)!} + \sum_{k=1}^{\infty} (1) \frac{\lambda^{k-1}}{(k-1)!} \right]$$

k=2 (k=1 term is zero)

$$\downarrow \frac{(k-1)}{(k-1)!} = \frac{1}{(k-2)!}$$

$$\sum_{k=2}^{\infty} \frac{\lambda^{k-1}}{(k-2)!}$$

pull out a constant

$$\lambda \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!}$$

let $j = k-2$

$$\lambda \sum_{j=0}^{\infty} \frac{\lambda^{j}}{j!}$$

let $j = k-1$

$$\sum_{j=0}^{\infty} \frac{\lambda^{j}}{j!}$$

$$= \lambda e^{-\lambda} \left[\lambda e^{\lambda} + e^{\lambda} \right]$$

$$\langle k^2 \rangle = \lambda^2 + \lambda$$

Therefore ...

$$\sigma_k^2 = \langle k^2 \rangle - \langle k \rangle^2$$

$$= \lambda^2 + \lambda - (\lambda)^2$$

$$= \lambda$$

Cool! We just showed, precisely that the variance and mean of the Poisson distribution are the same!

V. Review of frequentist v. Bayesian approaches

long run probability over many repetitions

degree of plausibility

how would we estimate how far away Jupiter is?

want a way of quantifying how more plausible an idea is after observing something

→ Desiderata

① Probability represented by real #s

② Be rational: with more information, probability increases monotonically

③ Be consistent → more than 1 way to get the right answer
→ consider all relevant information (I)
→ equivalent states of knowledge represented by equivalent probability

→ Derive Bayes' Theorem

$$P(H_i, D \mid I) = P(D, H_i \mid I)$$

rewrite each side using product rule

$$\underline{P(H_i \mid D, I)}\, P(D \mid I) = P(D \mid H_i, I)\, P(H_i \mid I)$$

$$P(H_i \mid D, I) = \frac{P(D \mid H_i, I)\, P(H_i \mid I)}{P(D, I)}$$

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

how likely to observe data given hypothesis is true (need to model experimental variation)

plausibility of hypothesis before the expt

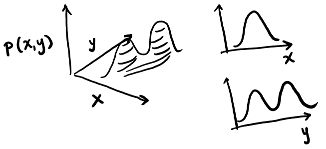
probability of observing the data, given what we knew before the experiment

also, note that there is nothing specifically Bayesian about this derivation - we invoked only axioms of probability (e.g. product rule)

the "Bayesian" component is that we use probability to describe our degree of belief, i.e. plausibility

→ Marginalization

- $P(X = x_i) = \sum_j P(X = x_i, Y = y_j)$



- By sum rule: $\sum_j P(H_j \mid D, I) = 1$

substitute using Bayes' Theorem

$$\sum_j \frac{P(D|H_j,I)\,P(H_j,I)}{P(D|I)} = 1$$

constant wrt j

$$P(D|I) = \sum_j P(D|H_j,I)\,P(H_j|I) \quad \text{(model)}$$

$$\text{or} \int_\theta d\theta \; P(D|\theta,I)\,P(\theta,I) \quad \text{(parameter)}$$

intuitively, this says the probability of observing the data (given only our prior knowledge) is related to sum of all model predictions for all possible models (think of <u>all</u> the ways the data can come about, weight them by their plausibility $(p(H_j|I))$, and you can assign a number to how likely you are to observe the data)

evidence is also "fully marginalized likelihood"

as above, when we marginalized over Y to convert $P(x,y) \rightarrow P(x)$ we marginalize over all hypotheses to convert $P(D,H_j) \rightarrow P(D)$

- Note that we can restate Bayes' theorem with this expression for the evidence...

$$P(H_i|D,I) = \frac{P(D|H_i,I)\,P(H_i|I)}{P(D|I)}$$

$$P(H_i|D,I) = \frac{P(D|H_i,I)\,P(H_i|I)}{\sum_j P(D|H_j,I)\,P(H_j|I)}$$

n.b. this looks a bit like the naïve definition of probability where we have all possiblites (here various models, weighted by plausibility) in the denominator, and the specific thing we're interested in is in the numerator