Probability & Statistics Intro/Refresher

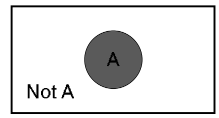
CPS 1/296

Ron Parr

A Simplistic View of Probability

- Probabilities defined over events
- Space of all possible events is "event space"

Event space:



- Think: Playing blindfolded darts with the Venn diagram..

Understanding Probabilities

- Relative frequencies work well for dice and coin flips
- For more complicated events, this is problematic

- Probability Clinton will win white house in 2008?
  - This event only happens once
  - We can't count frequencies
  - Still seems like a meaningful question

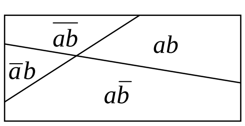
- In general, all events are unique
- "Reference Class" problem

Frequentism and Subjectivism

- Frequentists hold that probabilities must come from relative frequencies
- This is a purist viewpoint
- This is corrupted by the fact that relative frequencies are often unobtainable
- Often requires complicated and convoluted assumptions to come up with probabilities
- Subjectivists: probabilities are degrees of belief
  - Taints purity of probabilities
  - Often more practical

Event spaces for discrete RVs

- 2 variable case



- Important: Event space grows exponentially in number of random variables
- Components of event space = atomic events

Joint Distributions

- A joint distribution is an assignment of probabilities to every possible atomic event
- We can define all other probabilities in terms of the joint probabilities:

$$P(a) = \sum_{e_i \in e(a)} P(e_i)$$

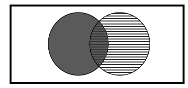
$$P(a) = P(a \wedge b) + P(a \wedge \neg b)$$

Why Probabilities Are Messy

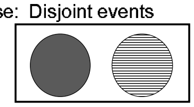
- Probabilities are not truth-functional
- To compute P(a and b), need joint distribution
  - sum out all of the other events from distribution
  - In general, it is not a function of P(a) and P(b)
  - In general, it is not a function of P(a) and P(b)
  - In general, it is not a function of P(a) and P(b)
  - This fact led to many approximations methods such as certainty factors and fuzzy logic (Why?)

Independence

- Convenient when it occurs, but don’t count on it
- When you have it:
  - P(A and B) = P(A)*P(B)
  - P(A or B) = P(A) + P(B) – P(A*B)



- Special case: Disjoint events



P(A or B) = ???

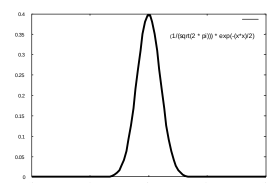
Kolmogorov’s axioms of probability

- 0<=P(a)<=1
- P(true) = 1; P(false)=0
- P(a or b) = P(a) + P(b) – P(a and b)
  - Subtract to correct for double counting
- This is sufficient to specify probability theory for discrete variables
- Continuous variables need density functions

Continuous Random Variables

- Domain is some interval, region, or union of regions
- Uniform case: Simplest to visualize
  (event probability is proportional to area)
- Non-uniform case visualized with extra dimension

Gaussian (normal/bell) distribution:



Updating Kolmogrov’s Axioms

- Use lower case for probability density
- Use end of the alphabet for continuous vars
- For discrete events: $0 \leq P(a) \leq 1$
- For densities: $0 \leq p(x)$
- Is p(x)>1 possible???

Requirements on Continuous Distributions

- p(x)>1 *is* possible so long as:

$$\int_x p(x)dx = 1$$

- Don’t confuse p(x) and P(X=x)
- P(X=x) for any x = 0!

$$P(x \in A) = \int_A p(x)dx$$

Cumulative Distributions

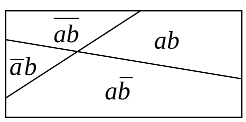
- When distribution is over numbers, we can ask:
  - P(X>=c) for some c
  - P(X<c) for some c
  - P(a<=X<=b) for some, a and b
- Solve by
  - Summation
  - Integration
- Cumulative sometimes called
  - CDF
  - Distribution function

Conditional Probabilities

- Ordinary probabilities = unconditional or prior probabilities
- P(a|b) = probability of a **given that we know *only* b**
- If we know c and d, we can't use P(a|b) directly
  (annoying, but important detail!)
- P(a|a)=1

Conditional Probability

- P(b|a) = Probability of event b given that event a is true



- Idea: In what fraction of a event space is b also true?

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

Conditioning for discrete events

- Suppose we know P(ABCDE) ← Joint
- Observe B=b, update our beliefs:

$$P(acde \mid b) = \frac{P(abcde)}{P(b)} = \frac{P(abcde)}{\sum_{ACDE} P(AbCDE)}$$

Condition with Bayes's Rule

$$P(A \wedge B) = P(B \wedge A)$$

$$P(A \mid B)P(B) = P(B \mid A)P(A)$$

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

Expectation

$$E(X) = \sum_X XP(X)$$

- Matches some colloquial notions of average
- "Mean"
- Arithmetic mean (uniform weights)
- For continuous random variables:

$$E(X) = \int_X Xp(X)dX$$

Nota bene: We will be assuming that E(X) is finite.

Properties of Expectation

$$E(f(X)) = \sum_X f(X)P(X)$$

$E(aX) = ???$ $aE(X)$

$E(aX+b) = ???$ $aE(X)+b$

$E(X+Y) = ???$ $E(X)+E(Y)$

$E(XY) = ???$ If X,Y are independent: $E(X)E(Y)$

(Proofs on board)

Sums of Expectations

- Suppose we have f(X) and g(Y).
- What is the expected value of f(X)+g(Y)?

$$\underset{XY}{E}\, f(X)+g(Y) = \sum_{XY} P(X \wedge Y)(f(X)+g(Y))$$
$$= \sum_{XY} P(X \wedge Y) f(X) + \sum_{XY} P(X \wedge Y)g(Y)$$
$$= \sum_X \sum_Y P(X \wedge Y) f(X) + \sum_Y \sum_X P(X \wedge Y)g(Y)$$
$$= \sum_X f(x) \sum_Y P(X \wedge Y) + \sum_Y g(Y) \sum_X P(X \wedge Y)$$
$$= \sum_X f(x)P(X) + \sum_Y g(Y) \sum_X P(X \wedge Y)$$
$$= \underset{X}{E}\, f(X) + \underset{Y}{E}\, f(Y)$$

Linearity of Expectation

Expectation Can Minimize Loss

- Suppose you need to bet on an outcome (e.g. die roll)
- Suppose loss is squared error, want:

$$\min_y E(X-y)^2$$

- Minimize and solve for y

(Proof on board)

Variance

- Hard to define in words
- "How much we trust the mean"

$$Var(X) = E(X-E(X))^2$$

- Compare with our loss function

Nota bene: We will be assuming that Var(X) is finite.

Properties of Variance

$$Var(X) = E(X-E(X))^2$$

$Var(aX) = ???$ $a^2 Var(X)$

$Var(aX+b) = ???$ $a^2 Var(X)$

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

$Var(X)+Var(Y)+2E[(X-E(X))(Y-E(Y))]$

If X,Y are independent: $Var(X)+Var(Y)$

(Proofs on board)

Covariance

$$Var(X+Y) = Var(X)+Var(Y)+2E[(X-E(X))(Y-E(Y))]$$

- Covariance captures the leftover:

$$Cov(X,Y) = Cov(Y,X) = E[(X-E(X))(Y-E(Y))]$$

$$Var(X+Y) = Var(X)+Var(Y)+2Cov(X,Y)$$

- If X,Y are independent, Cov(X,Y)=??? 0

Standard Deviation

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

- Even harder to define in English
- Sometimes more natural than variance:

$$SD(aX) = \quad aSD(X)$$

- Often not, for X,Y independent:

$$SD(X+Y) = \sqrt{SD^2(X) + SD^2(Y)}$$

Sample Mean

- Suppose we observe $X_1 \ldots X_n$
- What is our estimate for E(X)?

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

- Why?

$$E(\bar{X}) = E\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{nE(X)}{n} = E(X)$$

Also…

Chebyshev's Inequality

- Let X have finite mean and variance:

$$P(|X - E(X)| \geq c) \leq \frac{Var(X)}{c^2}$$

- Variance governs our chances of missing the mean

Convergence of Sample Mean

- Apply Chebyshev's inequality to sample mean

$$P(|\bar{X} - E(\bar{X})| \geq c) \leq \frac{Var(\bar{X})}{c^2}$$

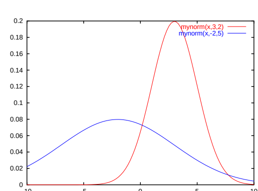
$$Var(\bar{X}) = \quad Var\left(\sum_{i=1}^{n} \frac{X_i}{n}\right) = \sum_{i=1}^{n} \frac{1}{n^2} Var(X_i) = \frac{Var(X)}{n}$$

$$\lim_{n \to \infty} P(|\bar{X} - E(\bar{X})| \geq c) \leq \frac{Var(X)}{nc^2} = 0$$

The Normal Distribution

- AKA:
  - Gaussian Distribution
  - Bell Curve
- Mean mu
- Standard Deviation sigma

$$p(x;\mu,\sigma) = \frac{1}{(2\pi)^{1/2}\sigma} \exp\left[-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right]$$



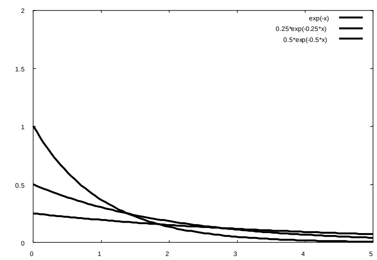
Why is the Normal so Popular???

- Some convenient properties:
  - Properties of expectation, variance well understood
  - Exponential drop off away from mean
  - Closed under linear combination
  - Central limit theorem
- Some inconvenient properties:
  - Annoying (but doable) to sample from
  - No nice closed form for the CDF

Exponential Distribution

- Good for modeling event durations where some small p of event terminating at any time.
- Defined for positive x: $p(x) = \frac{1}{\beta} e^{\frac{-x}{\beta}}$

Examples of Exponential Dist.



Poisson Process

- Suppose a sequence of independent random events occurs one after the other
- Delay between events is exponential
- For some time t, what is the distribution over the number of events that have occurred?

$$P(N(t) = n) = \frac{t^n}{\beta^n n!} e^{\frac{-t}{\beta}}$$

- mean = variance = t/β