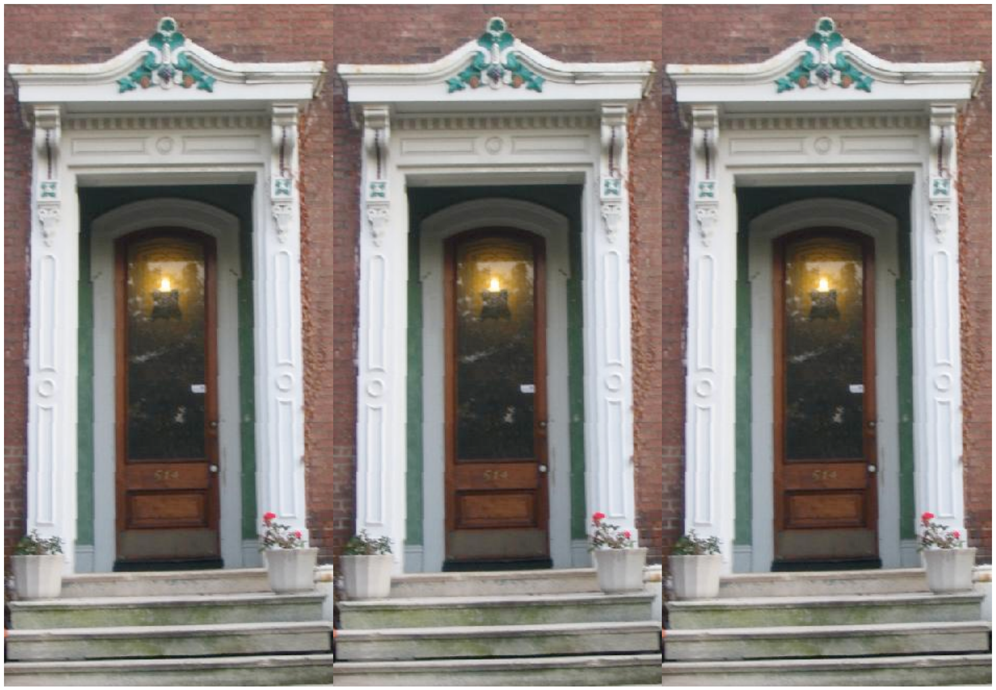
Review: probability

- RVs, events, sample space Ω
- Measures, distributions
  - disjoint union property (law of total probability—book calls this “sum rule”)
- Sample v. population
- Law of large numbers
- Marginals, conditionals

Monty Hall

keep

switch



has no prize

P1 ~~1/3~~ — O2 1/6, O3 ~~1/6~~ — 1/3

P2 ~~1/3~~ — O3

P3 1/3 — O2 — 2/3

Terminology

- Experiment = planned observations
- Prior = probability dist'n of parameter before experiment
  that we care about
- Posterior = dist'n after experiment

Example: model selection

- You're gambling to decide who has to clean the lab
- You are accused of using weighted dice!
- Two models:
  - fair dice: all 36 rolls equally likely $1/36$
  - weighted: rolls summing to 7 more likely → $1/12$ 7, $1/60$ not 7

prior: $1/10$ weighted $9/10$ not

observation: 2-5

posterior:

$3/360$

$\frac{1}{120}$	$\frac{9}{360}$
25%	75%
wt	-wt

	wt	-wt
1-6	$\frac{1}{10}\cdot\frac{1}{12}$	$\frac{9}{10}\cdot\frac{1}{36}$
2-5	$\frac{1}{10}\cdot\frac{1}{12}$	$9/10 \cdot 1/36$
⋮		
6-1		
1-1	$\frac{1}{10}\cdot\frac{1}{60}$	$\frac{9}{10}\cdot 1/36$
1-2	⋮	⋮
⋮		
6-5		
6-6		

Philosophy

- ***Frequentist*** v. ***Bayesian***
- Frequentist view: a probability is a property of the world (the coin has P(H) = 0.62)
- Bayesian view: a probability is a representation of our internal beliefs about the world (we think P(H) = 0.62)

Difference

- Bayesian is willing to assign P(E) to any E, even one which has happened already (although it will be 1 or 0 if E or ¬E has been observed)
- Frequentist will assign probabilities ***only*** to outcomes of future experiments
- Consider the question: what is the probability that coin #273 is fair?

Which is right?

- Both!
- Bayesians can ask more questions
- But for a question that makes sense to both, answer will agree
- Can often rephrase a Bayesian question in frequentist terms
  - answer may differ
  - either may see other's answer as a reasonable approximation

Independence

- X and Y are **independent** if, for all possible values of y, $P(X) = P(X \mid Y{=}y)$
    - equivalently, for all possible values of x, $P(Y) = P(Y \mid X{=}x)$

$P(x,y)/p(y) = P(x)$

$p(x|y) = P(x)$

    - equivalently, $P(X,Y) = P(X)\,P(Y)$
- Knowing X or Y gives us no information about the other

Independence: probability = product of marginals

AAPL price

Weather

	up	same	down	
sun	0.09	0.15	0.06	0.3
rain	0.21	0.35	0.14	0.7
	0.3	0.5	0.2	

$.5 \times .3 = .15$

Admin

- Slides and annotated slides

  http://www.cs.cmu.edu/~ggordon/10601/schedule.html

- Mailing list:

  10601-09f-announce@cs

- Recitation 6-8 PM GHC 8102 Matlab
  Wed (today!)

Readings

Bishop

- So far: p1–4, sec 1–1.2, sec 2–2.3
- We'll put them next to relevant lectures on schedule page
- They provide extra detail beyond what's in lecture—you are responsible for knowing it
- No specific due date

Expectations

- How much should we expect to earn from our AAPL stock?

$.09 \times \underline{1} + .21 \times \underline{1}$
$+ .06 \times \underline{(-1)} + .14 \times \underline{(-1)} = 0.1$

AAPL price

Weather	up	same	down
sun	0.09	0.15	0.06
rain	0.21	0.35	0.14

Weather	up	same	down
sun	+1	0	-1
rain	+1	0	-1

$E(X+Y) = E(X) + E(Y)$ $E(kX) = kE(X)$

Linearity of expectation

- Expectation is a linear function of numbers in bottom table
- E.g., change -1s to 0s or to -2s

$-2 \to -.1$
$-1 \to +.1$
$0 \to +.3$

AAPL price

Weather	up	same	down
sun	0.09	0.15	0.06
rain	0.21	0.35	0.14

Weather	up	same	down
sun	+1	0	-1
rain	+1	0	-1

Conditional expectation

- What if we know it's sunny?

	.3	.5	.2
X	+1	0	−1
	.3	0	−.2

sum = .1

AAPL price

Weather	up	same	down
sun	0.09	0.15	0.06
rain	0.21	0.35	0.14

Weather	up	same	down
sun	+1	0	-1
rain	+1	0	-1

Independence and expectation

- If X and Y are independent, then:
- Proof:

$E(X|Y) = E(X)$

$E(XY) = E(X)E(Y)$

$$E(XY) = \sum_{(x,y)\in\Omega} \underbrace{P(x,y)}_{= P(x)P(y)} xy$$

$$\left[\sum_x x P(x)\right]\left[\sum_y y P(y)\right]$$

$$= E(X)\ E(Y)$$

Sample means

- Sample mean = $\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$ not a R.V.
- Expectation of sample mean:

$$E\left(\frac{1}{N} \sum x_i\right) = \frac{1}{N} \sum E(x_i)$$

$$= \frac{1}{N} \sum \mu \rightarrow \text{pop'n mean}$$

$$= \mu$$

Estimators

- Common task: given a sample, infer something about the population
- An ***estimator*** is a function of a sample that we use to tell us something about the population
- E.g., sample mean is a good estimator of population mean
- E.g., linear regression

Law of large numbers (more general form)

- If we take a sample of size N from a distribution P with mean μ and compute sample mean $\bar{x}$
- Then $\bar{x} \to \mu$ as $N \to \infty$

Earlier version $[X = x_i]$

Bias

- Given an estimator T of a population quantity θ
- The ***bias*** of T is $E(T) - \theta$
- Sample mean is an unbiased estimator of population mean
- $T = (1 + \sum_{i=1}^{N} x_i) / (N+1)$ is biased

  but asymptotically unbiased as $N \to \infty$

Variance

- Two estimators of population mean: sample mean, mean of every 2nd sample
- Both unbiased, but one is more variable
- Measure of variability: variance

Variance

- If zero-mean: variance = $E(X^2)$
  - Ex: $\underbrace{\text{constant } 0}_{E(X^2)=0}$ v. $\underbrace{\text{coin-flip } \pm 1}_{E(X^2)=1}$
- In general: $E((X - E(X))^2)$

$E(X - E(X)) = E(X) - \underbrace{E(E(X))}_{E(X)} = 0$

Exercise: simplify the expression for variance

- $E((X - E(X))^2)$ $= E(X^2 - 2X\,E(X) + E(X)^2)$

$= E(X^2) - \underbrace{2\,E(X)\,E(X)}_{2\,E(X)^2} + E(X)^2$

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

$= \overline{X^2} - \bar{X}^2$

Exercise

- What is the variance of 3X?

assume $E(X) = 0$

$E((3X)^2) = E(9X^2) = 9E(X^2) = 9\,Var(X)$

Sample variance

$\frac{1}{N} \sum_{i=1}^{N} x_i$

- Sample variance = $\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2$

biased!

asyp. unbiased

- Expectation: $\frac{N-1}{N} Var(x)$
- Sample size correction:

$$\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

unbiased!

Bias-variance decomposition

- Estimator T of population quantity θ
- ***Mean squared error*** = E((T – θ)²) =

$$E((T - E(T) + E(T) - \theta)^2) =$$

$$E((T - E(T))^2 + 2(T - E(T))(E(T) - \theta) + (E(T) - \theta)^2)$$

$$= \underbrace{E((T - E(T))^2)}_{\text{variance}} + \cancel{2(E(T) - E(T))(E(T) - \theta)}$$

$$+ \underbrace{(E(T) - \theta)^2}_{\text{bias}^2}$$

Bias-variance tradeoff

- It's nice to have estimators w/ small MSE
- Typically there is a ***smallest possible*** MSE for a given amount of data
    - limited data provides limited information
- Estimator which achieves min is ***efficient*** (close for large N: ***asymptotically eff.***)
- Often can adjust estimator so MSE is due to bias or variance—the famed ***tradeoff***