intro

- foreshadowing/context: under all prob computations are *sample spaces*
- rarely work with sample spaces directly, unless they're simple (heads/tails)
  - (so, we start here)

what is prob?

- objective prob: long run freq of occurrence (eg heads in coin flip)
  - often called frequentist/classical methods.
  - used more often in undergrad CMU
- subjective prob: a possibly informed belief in rate of occurrence of event
  - can called bayesian

set notation

- $A \supset B, A \subset B, A \cup B, A \cap B, \overline{A}$ aka A^C
- let the set of all experimental outcomes $\Omega = A \cup \overline{A}$
- $A \cap B = \emptyset \Longrightarrow A$ and B are mutually exclusive aka disjoin
- distributive/associative laws
- de morgan's ($\overline{A \cup B} = \overline{A} \cap \overline{B}$, etc)

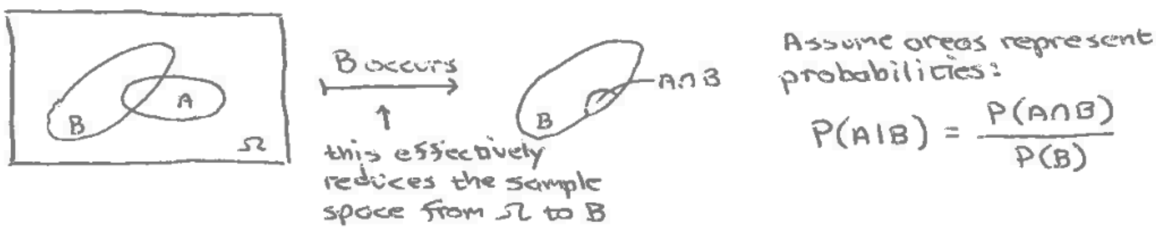
what are experiments?

- make observations
- passive: just collect data
- active: control setting
- sample space (Ω)
  - two coins tossed? $\Omega = \{\text{HH}, \text{HT}, \text{TH}, \text{TT}\}$
    - HH is simple event
    - TH is compound event ("at least one tail")
  - free throws until miss? $\Omega = \{\text{M, HM, HHM, ...}\}$
  - relative freqs of above? don't know! need more info

probability

sample space probabilities

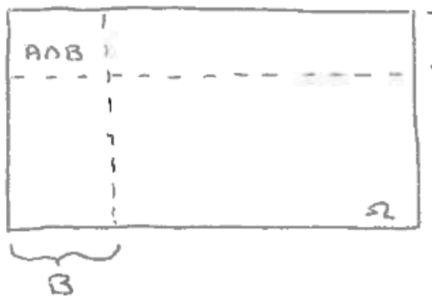
- conditional probability: A, when we know B



- A and B are independent
  - iff $P(A \mid B) = P(A)$
  - iff $P(B \mid A) = P(B)$
  - iff $P(A \cap B) = P(A)P(B)$
- if there are three events, A and B are *conditionally* independent

- if $P(A \cap B \mid C) = P(A \mid C)P(B \mid C)$

The following is an example of independence as rendered on a Venn diagram:



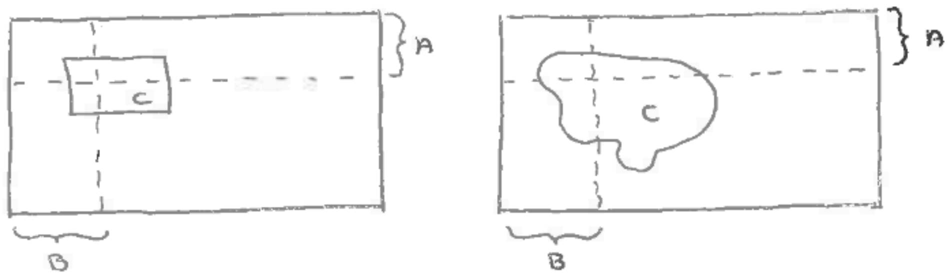
$$P(A \mid B) = \frac{P(A\cap B)}{P(B)} = \frac{P(A)}{P(\Omega)} = P(A)$$

area ratios are equal

Let A, B, and C be three separate events in Ω, all of which have non-zero probability of occurring. A and B are *conditionally independent* if

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

The following shows conditional independence and dependence as rendered on a Venn diagram:



- multiplicative law:

$$\begin{aligned} P(A \cap B) &= P(A)P(B \mid A) = P(B)P(A|B) \\ &= 0 \text{ if } A, B \text{ disjoint} \\ &= P(A)P(B) \text{ if } A, B \text{ independent} \end{aligned}$$

We can generalize this result to n events $\{A_1, \cdots, A_n\}$:

$$P(A_1\cap\cdots\cap A_n) = P(A_1\mid A_2\cap\cdots\cap A_n)\,P(A_2\cap\cdots\cap A_n)$$

$$P(A_2\mid A_3\cap\cdots\cap A_n)\,P(A_3\cap\cdots\cap A_n)$$

$$P(A_3\mid A_4\cap\cdots\cap A_n)\,P(A_4\cap\cdots\cap A_n)$$

etc.

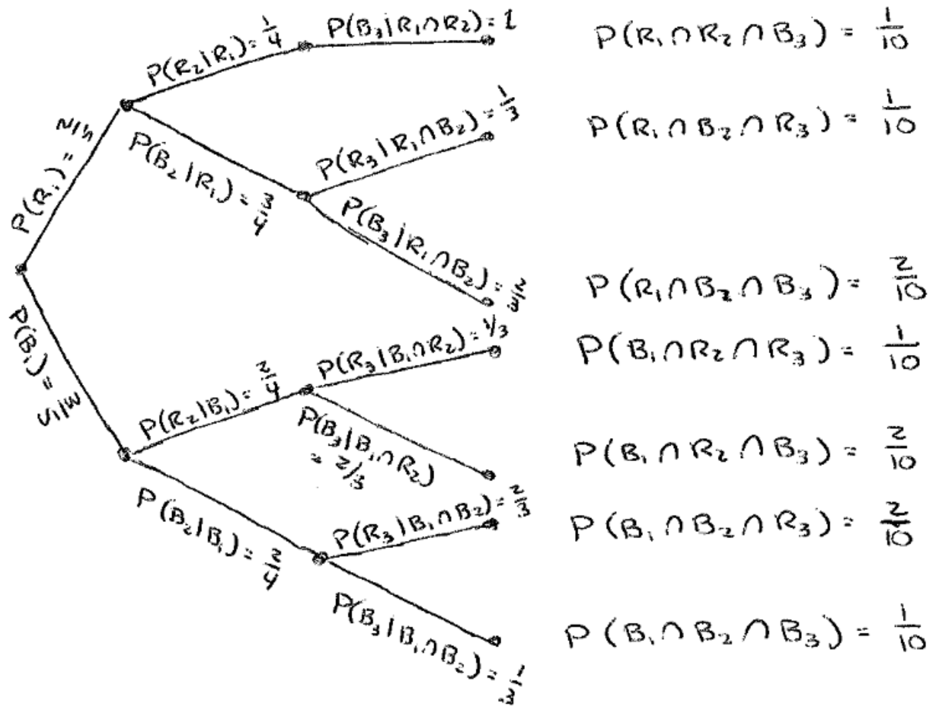
$$= P(A_n)\prod_{i=1}^{n-1} P(A_i\mid A_{i+1}\cap\cdots\cap A_n)$$

- additive law:

$$\begin{aligned} P(A \cup B) &= P(A) + P(B) - P(A \cap B) \\ &= P(A) + P(B) \text{ if } A, B \text{ disjoint} \\ &= P(A) + P(B) - P(A)P(B) \text{ if } A, B \text{ independent} \end{aligned}$$

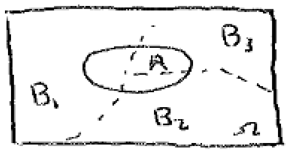
- decision trees: good for when probabilities change
  - eg picking colored balls without replacement, not like probability of heads of fair coin

→ **Example**: you pull three balls out from an urn which has two red and three black balls, without replacement. (a) What is the probability that the second ball drawn is red? (b) What is the probability that the second ball drawn is red, if the third ball drawn is black?



- for these next two, let $\{B_i\}$ be a partition of Ω.
- law of total probability.

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



  - "probability of B_1. then times the probability that i landed in A, in B_1, etc etc"
  - helpful when given conditional probs but not A itself

- Bayes' Rule:

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{\sum_{i=1}^{k} P(A|B_i)P(B_i)} = \frac{P(A|B_j)P(B_j)}{P(A)}$$

→ **Example: The Monty Hall Problem.** This problem is named for the long-time host of the game show *Let's Make a Deal*. The simple version goes as follows: you are shown three doors; behind two of the doors are goats and behind the other is a car. You choose a door (say Door #1). Monty Hall then opens, say, Door #3 to reveal a goat, and asks you if you want to switch to Door #2.

So: do you stick with Door #1 or switch to Door #2?

Assume Door #1 has been selected.

O_i = "Monty opens Door i" C_i = "car is behind Door i"

$P(C_i) = \frac{1}{3}$ for all i ← car could be behind any door

$$\Omega = \{O_2 \cap C_1, \cancel{O_2 \cap C_2}, O_2 \cap C_3, O_3 \cap C_1, O_3 \cap C_2, \cancel{O_3 \cap C_3}\}$$

Monty will <u>not</u> open the door the car is behind.

What is $P(C_2|O_3)$?

$$P(C_2|O_3) = \frac{P(O_3|C_2)P(C_2)}{P(O_3)} = \frac{P(O_3|C_2)P(C_2)}{P(O_3|C_2)P(C_2) + P(O_3|C_1)P(C_1)}$$

$P(O_3|C_2) = 1$ (no choice) $P(O_3|C_1) = \frac{1}{2}$

$$\Rightarrow P(C_2|O_3) = \frac{1 \cdot 1/3}{1 \cdot 1/3 + 1/2 \cdot 1/3} = \boxed{\frac{2}{3}}$$

⇒ change your pick to Door #2!

other probability paradigm??

- need better paradigm for continuous situations/nondiscrete outcomes: random variables, prob distributions
- random variable: like a function, maps events in Ω to, eg, real number line (the output. meters ran maybe)
  - but tbh just think of random variables as their outputs
  - can have diff random variables for same experiment: number of tails observed, number of heads, whether not heads observed
  - denoted with uppercase Latin eg X, $P(X = x)$
  - NO INHERENT NOTION OF PROBABILITY!
- functions of random variables are random variables! *statistics* are *functions of data* are *random variables*
- properties of discrete and continuous prob distributions

	Discrete	Continuous
Name	pmf: probability mass function	pdf: probability density function
Symbol	$p_X(x)$	$f_X(x)$
Properties	$0 \leq p_X(x) \leq 1$	$f_X(x) \geq 0$
	$\sum_{\text{all } x} p_X(x) = 1$	$\int_{\text{all } x} f_X(x) dx = 1$
	$P(a \leq X \leq b) = \sum_{x \in [a,b]} p_X(x)$	$P(a \leq X \leq b) = \int_a^b f_X(x) dx$
	$P(a < X < b) = \sum_{x \in (a,b)} p_X(x)$	$P(a < X < b) = \int_a^b f_X(x) dx$
	$E[X] = \sum_{\text{all } x} x p_X(x)$	$E[X] = \int_{\text{all } x} x f_X(x) dx$
(law of unconscious statistician)	$E[g(x)] = \sum_{\text{all } x} g(x) p_X(x)$	$E[g(x)] = \int_{\text{all } x} g(x) f_X(x) dx$

- expected value operator $E[X] = \mu_X$, mean of the distribution X was sampled from
  - $E[cX] = cE[X]$
  - $E[c] = c$
  - $E[x + y] = E[x] + E[y]$
- variance operator $V[X] = \sigma^2$
  - $V[X] = E\left[(x - \mu)^2\right] = \text{(simplifies to) } E[X^2] - (E[X])^2$
  - apparently $V[x + y] = V[x] + V[y]$??? sep 18 class example 2
  - variance is not width, but the square of the width. think about units of $V[X]$ vs units of $E[X]$
- translation/scaling's effects on mean and variance: $(X \to X + b)$
  - $E[X + b] = E[X] + b$ (translation shifts mean)
  - $E[aX] = aE[X]$ (scaling shifts mean multiplicatively)
  - $V[X + b] = V[X]$ (translation doesn't effect width)
  - $V[aX] = a^2 V[X]$ (scaling widens exponentially. verify via shortcut formula)

- cumulative distribution function (cdf)
  - accumlated prob up to x, inclusive

	Discrete	Continuous
Symbol	$F_X(x)$	$F_X(x)$
Definition	$F_X(x) = \sum_{y \le x} p_Y(y)$	$F_X(x) = \int_{-\infty}^{x} f_Y(y)dy$
Limiting Properties	for both, $F_X(-\infty) = 0$ and $F_X(\infty) = 1$	
Reln to pmf/ pdf	$p_X(x)$ is magnitude of jump in $F_X(x)$ at coord x	$f_X(x) = \frac{d}{dx} F_X(x)$
Reln to Quantile q	$X = \min\{x : F_X(x) \ge q\}$	inverse cdf, $X = F_X^{-1}(q)$
Reln to Prob. Over Range	It's complicated.	$P(a \le X \le b)$ $= P(a < X < b)$ $= F_X(b) - F_X(a)$

  - inverse cdf: given total prob q to left of (and including) some x_0, what is x_0?

	input	output
cdf	x_0	q
inverse cdf	q	x_0

families of distributions

- up till now, practice problems have fixed θ in equations. we often have to find it. (use law of total prob)
- deriving from Law of Total Prob ($P(A) = \sum_{i=1}^{k} P(A|B_i)P(B_i)$) if we know
  - $p_{X|\theta}(x|\theta)$
  - with θ weights from $p_\Theta(\theta)$,

$$p_X(x) = \sum_\theta p_{X|\theta}(x|\theta) p_\Theta(\theta)$$

  - if θ is continuous, we may just use an integral
  - Now, what happens if X is a continuous random variable?

Discrete Θ: $f_X(x) = \sum_{\text{all } \Theta} f_{X|\Theta}(x|\Theta)\, p_\Theta(\Theta)$

Continuous Θ: $f_X(x) = \int_{\text{all } \Theta} f_{X|\Theta}(x|\Theta)\, f_\Theta(\Theta)\, d\Theta$

- plotting such a pdf: $f_X(x) = \int_0^\infty \theta x^{\theta-1} \cdot e^{-\theta} d\theta$

```
x   <- seq(0.001,1,by=0.001)
f.X <- rep(NA,length(x))

f <- function(theta,x) {
  return(theta*x^(theta-1)*exp(-theta))
}
for ( ii in 1:length(x) ) {  # we loop over x's indices (1, 2, ..., length(x))
  f.X[ii] <- integrate(f,0,Inf,x=x[ii])$value
}

# Now let's plot!
library(ggplot2)
df <- data.frame(x,f.X)
ggplot(data=df,mapping=aes(x=x,y=f.X)) +
  geom_line(col="firebrick") +
  ylim(c(0,3)) +
  ylab(expression(f[X]*"(x)"))
```

data sampling code

- inverse-transform sampling
  1. sample a $q \in (0, 1)$. (e.g. `runif()`, random uniform)
  2. plug q into $x = F_X^{-1}(q)$, record x
  3. repeat n times for a sample size of n
- rejection sampling
  1. choose finite domain $[a, b]$ for $f_X(x)$ that's good enough
  2. let $\max(f_X(x)) = m$
  3. repeat until n samples recorded:
     1. randomly sample $x' \in [a, b]$ and $y' \in [0, m]$.
     2. if $y' \leq f_X(x')$, keep the data point. otherwise, reject it and continue
- see `media/data_sampling.Rmd`

statistics

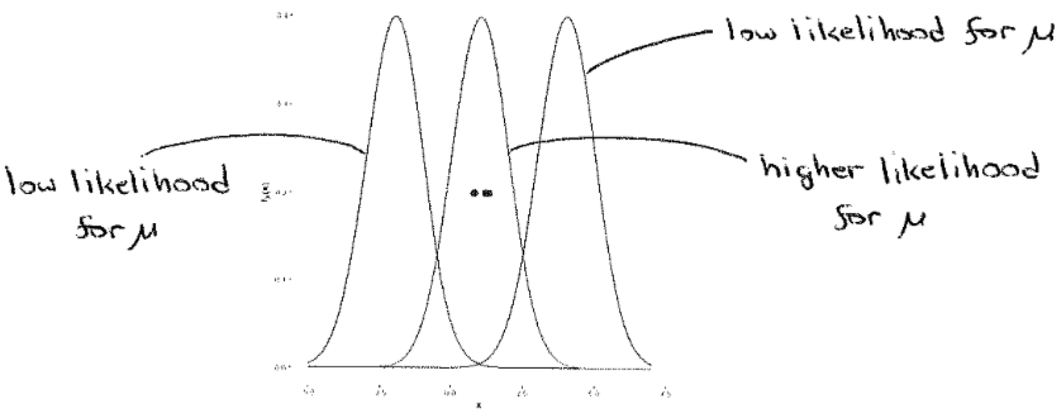
- iid: independant and identically distributed
- reiterating: these are just functions of observed data
  - $Y = X_1 + \cos(X_2) - \frac{X_{37}}{\pi}$ is a statistic, but not informative
  - $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_1$ is as well, and is informative of μ
  - $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_1 - \overline{X}\right)^2$ is informative of σ^2
- statistics are random variables!!!
  - drawn from sampling distributions (which are just pmfs/pdfs)file:
- sample mean, stddev, range, median
- $E\left[\overline{X}\right] = E[X]$, $V\left[\overline{X}\right] = \frac{V[X]}{n}$ hold for all distributions
- standard error is the width of the distribution; the standard deviation of a statistic; $\sqrt{V[Y]}$, where Y is a random variable/sampling distribution/pmf or pdf for a statistic
- expected value of the sample variance S^2 is σ^2 after LOTS of math
  - this is Wednesday: Statistics and Sampling Distributions class example 1
  - why is $\frac{1}{n-1}$ there? create an unbiased example of the population estimate??

likelihood function

- the likelihood function quantifies how likely a θ is given our data
- let it be defined:

$$\mathcal{L}\left(\theta \mid \vec{X}\right) = \prod_{i=1}^{n} f_X(x_i \mid \theta)$$

  for continuous data, simply using $p_X(x_i \mid \theta)$ for discrete
- notice that



- why? in inference, we want to estimate θ given data! so we maximize $\mathcal{L}$ for θ
- to make math easier, we use the log-likelihood $\ell\left(\theta \mid \vec{X}\right) = \log \mathcal{L}\left(\theta \mid \vec{X}\right)$

bias and variance

- bias: $B\left[\hat{\theta}\right] = E\left[\hat{\theta} - \theta\right] = E\left[\hat{\theta}\right] - \theta$
  - estimator biased: $B\left[\hat{\theta}\right] = 0$, unbiased $B\left[\hat{\theta}\right] \neq 0$
- variance: $V\left[\hat{\theta}\right]$ (recall defn above)
- mean-squared error (MSE): $\mathrm{MSE}\left[\hat{\theta}\right] = E\left[\left(\hat{\theta} - \theta\right)^2\right] \underset{\text{simpl to}}{=} V\left[\hat{\theta}\right] + \left(B\left[\hat{\theta}\right]\right)^2$
  - select "best" estimator by taking the one with lowest MSE

Maximum Likelihood Estimation

- procedure (it's literally just AP Calc maximization):
  1. find $\mathcal{L}\left(\theta \mid \vec{X}\right)$
  2. find $\ell\left(\theta \mid \vec{X}\right)$
  3. compute $\ell'\left(\theta \mid \vec{X}\right)$, partial or normal with respect to θ
  4. solve the above equal to 0 for θ, now called $\hat{\theta}_{\mathrm{MLE}}$ (also replace x_i with X_i)
- this does not work with domain-specifying parameters (e.g. $f_X(x) = \frac{1}{\theta}$ for $x \in [0, \theta]$)
- property of MLEs: invariance property
  - if $\theta' = g(\theta)$, then $\hat{\theta}'_{\mathrm{MLE}} = g\left(\hat{\theta}_{\mathrm{MLE}}\right)$
  - e.g., .0 what

more abt confidence intervals

$P\left(\hat{\theta}_L \leq \theta \leq \hat{\theta}_R\right)$ where the theta HATS are the r.v.s!

point estimates dont need sampling distributions! point estimates do!

when doing the math, finding lower and upper bound should be the same thing except having a swapped q

erm.

we skipped a bunch of notetaking. it's exam 3 time.