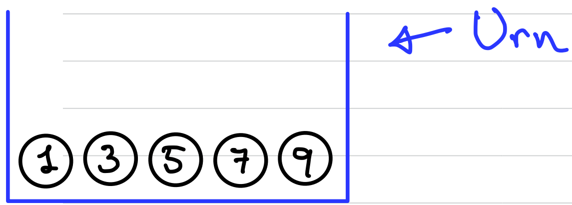
Lecture 3

Problem from Lecture 2



a) Draw a ball uniformly at random (u.a.r.)
b) record the number on the ball
c) return the ball to the urn
d) repeat n times.

Let S_n = sum over the observed numbers

$\boxed{Q}$ $\mathbb{P}[S_n \text{ is divisible by } 5]$?

X_i = # from the i^{th} draw

$S_n = X_1 + \dots + X_n$

Let $R = \{1, 3, 5, 7, 9\}$

Law of Total Probability $\Rightarrow$

$$\mathbb{P}[S_n \text{ is divisible by } 5] = \sum_{a \in R} \underbrace{\mathbb{P}[S_n \text{ is divisible by } 5 \mid X_n = a]}_{S_{n-1}+a \text{ is divisible } 5} \underbrace{\mathbb{P}[X_n = a]}_{\frac{1}{5}}$$

a	$S_{n-1} \bmod 5$
1	4
3	2
5	0
7	3
9	1

$$= \frac{1}{5} \sum_{k=0}^{4} \mathbb{P}[\underbrace{S_{n-1} = k \bmod 5}]$$

Call this event E_k.

$E_0, E_1, \dots, E_4$ partition $\Omega \Rightarrow$

$$= \frac{1}{5} \underbrace{\mathbb{P}[\Omega]}_{1} = \boxed{\frac{1}{5}}$$

Problem of the Day

Alice and Bob play the following guessing game.

1) Bob writes down two different numbers on two separate cards:

$\boxed{X}$ $\boxed{Y}$ Say $X, Y > 0$

2) Alice picks one of the cards uniformly at random and looks at the number.

3) Alice wins if she correctly guesses which of the two cards has a larger number.

$\boxed{Q}$ Can Alice do better than random guess?

Bayes' formula

Let $B_1, \dots, B_n \in \mathcal{F}$ be a partition of Ω. Then, for any $A \in \mathcal{F}$ with $\mathbb{P}(A) > 0$,

$$\mathbb{P}[B_i \mid A] = \frac{\mathbb{P}[B_i \cap A]}{\mathbb{P}[A]} = \frac{\mathbb{P}[A \mid B_i]\,\mathbb{P}[B_i]}{\mathbb{P}[A]} = \frac{\mathbb{P}[A \mid B_i]\,\mathbb{P}[B_i]}{\sum_{j=1}^{n} \mathbb{P}[A \mid B_j]\,\mathbb{P}[B_j]}$$

e.g.) Suppose you get tested for a disease and the test result comes back "+".

$\boxed{Q}$ Should you worry?

Suppose disease prevalence $= p$

"D" = event of having the disease

$\mathbb{P}[D] = p$ (prior)

$\mathbb{P}[D^c] = 1-p$

test result

$$\mathbb{P}(+ \mid D^c) = FPR$$
$$\mathbb{P}(- \mid D) = FNR$$

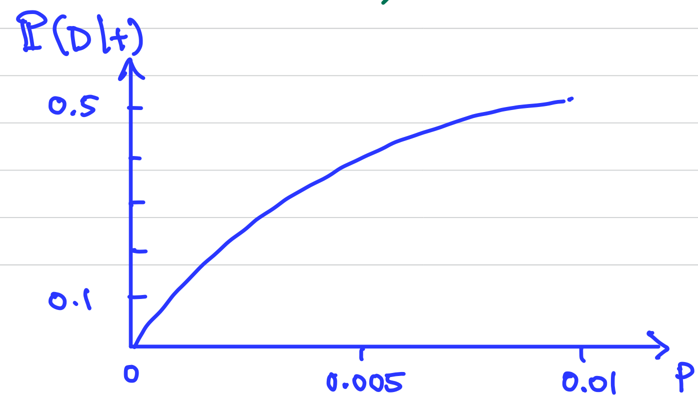
$$\mathbb{P}[D \mid +] = \frac{\mathbb{P}[+ \mid D]\,\mathbb{P}[D]}{\mathbb{P}[+ \mid D]\,\mathbb{P}[D] + \mathbb{P}[+ \mid D^c]\,\mathbb{P}[D^c]}$$

posterior

$$= \frac{(1-FNR)\,p}{(1-FNR)\,p + FPR\,(1-p)}$$

$$= \frac{(1-FNR)\,p}{FPR + p\,(1-FNR-FPR)}$$

If $FPR = FNR = 0.01$, then



$X \sim \text{Bernoulli}(p)$, $0 < p < 1$

Success — Fail

$\mathbb{P}[X=1] = p$, $\mathbb{P}[X=0] = 1-p$.

$\mathbb{E}[X] = 1\cdot p + 0(1-p) = p$

$\text{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$
$= p - p^2 = p(1-p)$

Bernoulli process

independent and identically distributed

$X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Bernoulli}(p)$

$$\underbrace{0\ 0\ 1\ 0\ 0\ 0\ 1\ 1\ 0\ 0\ 0\ \cdots\ 1\ 0\ 1\ 0\ 1}$$

S_n = total number of 1s
$= X_1 + \cdots + X_n$

$S_n \sim \text{Binomial}(n, p)$

$$\mathbb{P}[S_n = k] = \binom{n}{k} p^k (1-p)^{n-k}$$

$\mathbb{E}[S_n] = \mathbb{E}[X_1 + \cdots + X_n]$

by linearity of $\mathbb{E}$ $= \sum_{i=1}^{n} \mathbb{E}[X_i] = np$

$\text{Var}(S_n) = \sum_{i=1}^{n} \text{Var}(X_i)$ (by $\perp\!\!\!\perp$ of $X_1, \ldots, X_n$)
$= np(1-p)$

$$\underline{0\ 0\ 1}\ \underline{0\ 0\ 0\ 1}\ \underline{1}\ 0\ 0\ 0\ \cdots\ 1\ \underline{0\ 1}\ \underline{0\ 1}$$

$W_1 = 3$ $W_2 = 4$ $W_3 = 1$

W_i = waiting time between the $(i-1)^{th}$ & i^{th} successes

$W_1, W_2, W_3, \ldots$ are $\perp\!\!\!\perp$

$W_i \sim \text{Geometric}(p)$ $\forall i$

$\mathbb{P}[W = k] = (1-p)^{k-1} p$, $k = 1, 2, 3, \ldots$

$\mathbb{E}[W] = \sum_{k=1}^{\infty} k(1-p)^{k-1} p = \frac{1}{p}$

$\text{Var}[W] = \mathbb{E}[W^2] - \mathbb{E}[W]^2$
$= \left(\sum_{k=1}^{\infty} k^2 (1-p)^{k-1} p\right) - \left(\frac{1}{p}\right)^2$
$= \frac{1-p}{p^2}$

These moments can be computed more easily using generating functions (Later lectures)

$r \in \mathbb{N} = \{1, 2, 3, \dots\}$ fixed positive int.

T_r = Total waiting time to the r^{th} success

$$T_r = W_1 + W_2 + \cdots + W_r$$

where $W_1, \dots, W_r \overset{iid}{\sim} \text{Geometric}(p)$

$$\underbrace{0\,0\,1\,0\,1\,1\,0\,0\,0\,1\,\cdots\,0\,0}_{r-1 \text{ successes}}\,\overset{r^{th} \text{ success}}{1}$$

$$\mathbb{P}[T_r = n] = \binom{n-1}{r-1} p^{r-1} (1-p)^{n-r} \cdot p$$

$$= \binom{n-1}{r-1} p^r (1-p)^{n-r}$$

F_r = # failures before r^{th} success

$$F_r + r = T_r$$

$$\mathbb{P}[F_r = k] = \binom{r+k-1}{r-1} p^r (1-p)^k$$

$$= \binom{r+k-1}{k} p^r (1-p)^k$$

$F_r \sim \text{Negative Binomial}(r, p)$

We can compute $\mathbb{E}[F_r]$ & $\text{Var}[F_r]$ without using the prob. mass function directly.

$$\mathbb{E}[F_r] = \mathbb{E}[T_r] - r$$

$$= \frac{r}{p} - r = \frac{r(1-p)}{p}$$

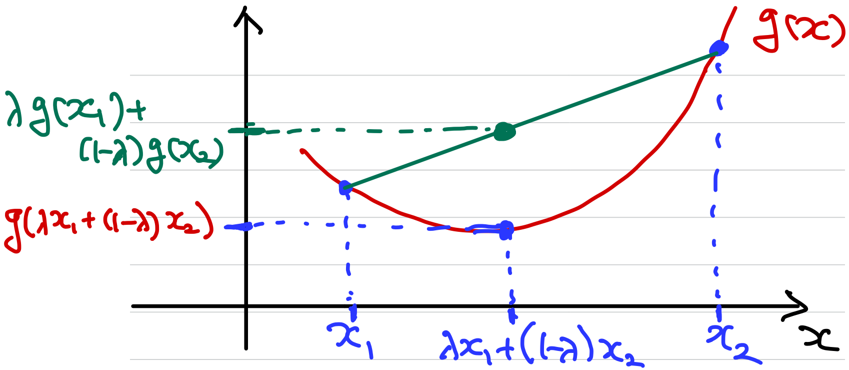
$$\text{Var}[F_r] = \text{Var}[T_r - r]$$

$$= \text{Var}[T_r]$$

$$= r\,\text{Var}[W_1]$$

$$= \frac{r(1-p)}{p^2}$$

NB distribution is widely used in single-cell genomics



A weighted sum $0 \le \lambda \le 1$

$\boxed{\text{Def}}$ (Convex function)

A function $g : (a, b) \to \mathbb{R}$ is said to be <u>convex</u> if

$$g(\lambda x_1 + (1-\lambda) x_2) \le \lambda g(x_1) + (1-\lambda) g(x_2)$$

$\forall\ x_1, x_2 \in (a, b)$ and $\forall\ 0 \le \lambda \le 1$.

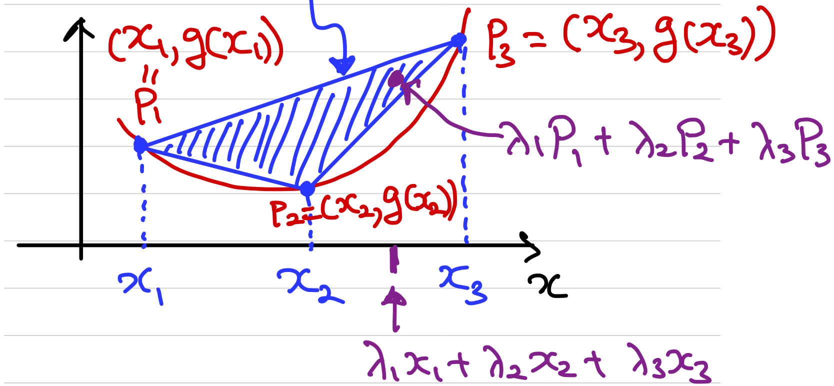
g is called strictly convex if the equality holds only for $\lambda = 0, \lambda = 1$ or $x_1 = x_2$.

(The graph of g over (a,b) contains no straight lines.)

$P_1, P_2, P_3 \in \mathbb{R}^2$

Convex hull of $\{P_1, P_2, P_3\}$

$$= \left\{ (x, y) \in \mathbb{R}^2 \;\middle|\; \begin{array}{l} (x,y) = \lambda_1 P_1 + \lambda_2 P_2 + \lambda_3 P_3 \\ \lambda_1 + \lambda_2 + \lambda_3 = 1, \quad \lambda_i \ge 0 \end{array} \right\}$$



$$g(\lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3)$$
$$\le \lambda_1 g(x_1) + \lambda_2 g(x_2) + \lambda_3 g(x_3)$$

$\forall\ x_1, x_2, x_3 \in \mathbb{R}$ and $\forall \sum_{i=1}^{3} \lambda_i = 1$, $\lambda_i \in [0, 1]$

g strictly convex $\Leftrightarrow$ equality holds only for 1) $\lambda_i = 1$ and $\lambda_j = 0$ for $j \ne i$

or 2) $x_1 = x_2 = x_3$.

Thm (Jensen's Inequality)
A convex function $g: (a,b) \to \mathbb{R}$ satisfies

$$g\left(\sum_{i=1}^{n} \lambda_i x_i\right) \le \sum_{i=1}^{n} \lambda_i g(x_i),$$

$\forall \lambda_1, \dots, \lambda_n$ satisfying $\sum_{i=1}^{n} \lambda_i = 1$, $\lambda_i \ge 0$ $\forall i$.

Corollary Let X be a $\mathbb{R}$-valued discrete RV and g a convex function. Then,

$$g(\mathbb{E}[X]) \le \mathbb{E}[g(X)]$$

If g is strictly cvx and X is not constant, then

$$g(\mathbb{E}[X]) < \mathbb{E}[g(X)].$$

Pf Let $\lambda_i = \mathbb{P}(X = x_i)$ in Jensen's inequality. □

(Holds for cts RVs also)

Pf of Jensen's inequality

Induction on n.

- Base case: $n=2$. True by cvx def.
- Induction Hyp: Assume true for all $n = 2, \dots, k$
- Show for $n = k+1$:

$$g\left(\sum_{i=1}^{k+1} \lambda_i x_i\right) = g\left(\sum_{i=1}^{k} \lambda_i x_i + \lambda_{k+1} x_{k+1}\right)$$

$$= g\left((1-\lambda_{k+1}) \frac{\sum_{i=1}^{k} \lambda_i x_i}{(1-\lambda_{k+1})} + \lambda_{k+1} x_{k+1}\right)$$

Def of convexity of $g \Rightarrow$

$$\le (1-\lambda_{k+1})\, g\left(\frac{\sum_{i=1}^{k} \lambda_i x_i}{1-\lambda_{k+1}}\right) + \lambda_{k+1} g(x_{k+1})$$

By induction hyp.

$$\le (1-\lambda_{k+1}) \left[\sum_{i=1}^{k} \frac{\lambda_i}{1-\lambda_{k+1}} g(x_i)\right] + \lambda_{k+1} g(x_{k+1})$$

$$= \sum_{i=1}^{k+1} \lambda_i g(x_i)$$

□

Suppose $\mathbb{P}$ & $\mathbb{Q}$ are two probability measures on $(\Omega, \mathcal{F})$, and let X be a discrete RV s.t.

$$\mathbb{P}[X = x] = p(x), \quad x \in \text{Range}(X)$$

$$\mathbb{Q}[X = x] = q(x), \quad x \in \text{Range}(X)$$

Def (Shannon entropy)

$$H(\mathbb{P}) = -\sum_x p(x) \log p(x) = -\mathbb{E}_p[\log p(X)]$$

Def (Cross entropy)

Cross entropy of $\mathbb{Q}$ relative to $\mathbb{P}$:

$$H(\mathbb{P}, \mathbb{Q}) = -\sum_x p(x) \log q(x)$$

$$= -\mathbb{E}_p[\log q(X)]$$

Def (KL Divergence)

The Kullback-Leibler divergence of $\mathbb{P}$ from $\mathbb{Q}$ is defined as

$$KL(\mathbb{P} \| \mathbb{Q}) = H(\mathbb{P}, \mathbb{Q}) - H(\mathbb{P})$$

$$= -\sum_x p(x) \log \left[\frac{q(x)}{p(x)}\right]$$

$$= -\mathbb{E}_p\left[\log \frac{q(X)}{p(X)}\right]$$

For continuous RV, replace

$p(x)$ with pdf and

$\sum_x$ with $\int dx$

Has lots of applications in machine learning (e.g., loss function, EM algorithm, VAE)