

# CMPT 210: Probability and Computation

## Lecture 22

---

Sharan Vaswani

July 29, 2022

## Two-sided Chernoff Bound

**Chernoff Bound:** Let  $T_1, T_2, \dots, T_n$  be mutually independent r.v.'s such that  $0 \leq T_i \leq 1$  for all  $i$ . If  $T := \sum_{i=1}^n T_i$ , for all  $c \geq 1$  and  $\beta(c) := c \ln(c) - c + 1$ ,

$$\Pr[T \geq c\mathbb{E}[T]] \leq \exp(-\beta(c) \mathbb{E}[T])$$

**Two-sided Chernoff Bound:** Bounds the probability that the r.v.  $T$  takes values (much) smaller or larger than the mean  $\mathbb{E}[T]$ :

$$\Pr[|T - \mathbb{E}[T]| \geq c\mathbb{E}[T]] \leq 2 \exp\left(\frac{-c^2 \mathbb{E}[T]}{3}\right)$$

Fussbook is redesigning their website to try to make it more appealing to users. In order to see if the new redesigned website actually helps, Fussbook decides to do an experiment.

Given  $m$  users of the website, let  $\mathcal{A}$  be the set of users who engage with the Fussbook posts (liking, sharing, etc) if *they are shown the old website*. Similarly,  $\mathcal{B}$  is the set of users that engage if *they are shown the new website*.

Define  $f_A$  (and  $f_B$ ) to be the fraction of users that engage when shown the old (or new website respectively), i.e.  $f_A := \frac{|\mathcal{A}|}{m}$  and  $f_B := \frac{|\mathcal{B}|}{m}$ . We assume that these fractions correspond to the proportion of users that prefer the old website vs the new.

The improvement on switching to the new website is defined in terms of the **lift** which is equal to  $f_B - f_A$ . If  $f_B - f_A > 0$ , then it makes sense to switch to the new website.

# A/B Testing

In order to estimate  $f_A$  and  $f_B$  and see if the new website actually helps, Fussbook decides to run an **A/B** test on  $m$  users.

---

**Algorithm** A/B Testing

---

```
1: function ABTest( $m$ )
2:  $X_A = 0, X_B = 0$  {Initialize number of users that engaged when shown the old/new website}
3: for  $i = 1, 2, \dots, m$  do
4:    $X_i \sim \text{Ber}(1/2)$ 
5:   if  $X_i = 1$  then
6:     Show user  $i$  old website.   If user engages,  $X_A = X_A + 1$ 
7:   else
8:     Show user  $i$  new website.   If user engages,  $X_B = X_B + 1$ 
9:   end if
10: end for
11: return  $\hat{f}_A = \frac{2X_A}{m}$  and  $\hat{f}_B = \frac{2X_B}{m}$ .
```

---

**Claim:**  $\hat{f}_A$  and  $\hat{f}_B$  are unbiased estimators of  $f_A$  and  $f_B$  respectively, i.e  $\mathbb{E}[\hat{f}_A] = f_A$  and  $\mathbb{E}[\hat{f}_B] = f_B$ . Note that  $X_A = \sum_{i \in \mathcal{A}} X_i$ .

$$\mathbb{E}[\hat{f}_A] = \mathbb{E}\left[\frac{2X_A}{m}\right] = \frac{2}{m}\mathbb{E}\left[\sum_{i \in \mathcal{A}} X_i\right] = \frac{2}{m}\left[\sum_{i \in \mathcal{A}} \mathbb{E}[X_i]\right] = \frac{2}{m}\left[\sum_{i \in \mathcal{A}} \Pr[X_i = 1]\right] = \frac{2}{m} \frac{|\mathcal{A}|}{2} = f_A.$$

Similarly,  $X_B = \sum_{i \in \mathcal{B}} (1 - X_i)$ .

$$\mathbb{E}[\hat{f}_B] = \mathbb{E}\left[\frac{2X_B}{m}\right] = \frac{2}{m}\mathbb{E}\left[\sum_{i \in \mathcal{B}} (1 - X_i)\right] = \frac{2}{m}\left[\sum_{i \in \mathcal{B}} \mathbb{E}[1 - X_i]\right] = \frac{2}{m}\left[\sum_{i \in \mathcal{B}} \Pr[X_i = 0]\right] = \frac{2}{m} \frac{|\mathcal{B}|}{2} = f_B.$$

## A/B Testing

**Claim:** With probability 0.95, the algorithm ABTest( $m$ ) estimates the lift to an error equal to 0.1 if  $m \approx 10517$ .

$$\begin{aligned}\Pr \left[ |\hat{f}_A - f_A| \geq \epsilon \right] &= \Pr \left[ \left| \frac{2X_A}{m} - f_A \right| \geq \epsilon \right] = \Pr \left[ \left| X_A - \frac{m f_A}{2} \right| \geq \frac{m\epsilon}{2} \right] \\&= \Pr \left[ \left| X_A - \frac{m f_A}{2} \right| \geq \underbrace{\frac{\epsilon}{f_A}}_c \frac{m f_A}{2} \right] \\&\leq 2 \exp \left( -\frac{1}{3} \left( \frac{\epsilon}{f_A} \right)^2 \frac{m f_A}{2} \right) \quad \text{(Two-sided Chernoff Bound)} \\&= 2 \exp \left( -\frac{\epsilon^2}{6} \frac{m}{f_A} \right) \leq 2 \exp \left( -\frac{\epsilon^2 m}{6} \right) \\&\quad \text{(Since } f_A < 1 \text{ and } \exp \text{ is a monotonically increasing function.)}\end{aligned}$$

Similarly, we can prove that,  $\Pr \left[ |\hat{f}_B - f_B| \geq \epsilon \right] \leq 2 \exp \left( -\frac{\epsilon^2 m}{6} \right)$ .

By the union-bound for two events  $A$  and  $B$ ,  $\Pr[A \cup B] \leq \Pr[A] + \Pr[B]$ . Hence,

$$\begin{aligned}\Pr \left[ |\hat{f}_A - f_A| \geq \epsilon \cup |\hat{f}_B - f_B| \geq \epsilon \right] &\leq 4 \exp \left( -\frac{\epsilon^2 m}{6} \right) \\ \implies \Pr \left[ |\hat{f}_A - f_A| < \epsilon \cap |\hat{f}_B - f_B| < \epsilon \right] &\geq 1 - 4 \exp \left( -\frac{\epsilon^2 m}{6} \right)\end{aligned}$$

$$\begin{aligned}|\text{estimated lift} - \text{true lift}| &= |[\hat{f}_B - \hat{f}_A] - [f_B - f_A]| = |[\hat{f}_B - f_B] + [\hat{f}_A - f_A]| \\ &\leq |[\hat{f}_B - f_B]| + |[\hat{f}_A - f_A]| \\ &\quad (\text{Triangle Inequality: For any constants } a, b, |a + b| < |a| + |b|)\end{aligned}$$

$$\implies \Pr[|\text{estimated lift} - \text{true lift}| \leq 2\epsilon] \geq 1 - 4 \exp \left( -\frac{\epsilon^2 m}{6} \right).$$

**Proof of Triangle inequality:** For any  $a, b$ ,  $ab \leq |a||b|$ .

$$\implies a^2 + b^2 + 2ab \leq a^2 + b^2 + 2|a||b| = |a|^2 + |b|^2 + 2|a||b| \text{ (since } \forall x, x^2 = |x|^2 \text{)}.$$

$$\implies (a + b)^2 = (|(a + b)|)^2 \leq (|a| + |b|)^2 \implies |a + b| \leq |a| + |b|.$$

In the claim, we want the RHS to be equal to 0.95 and  $2\epsilon = 0.1$ . Hence,

$$4 \exp\left(-\frac{\epsilon^2 m}{6}\right) = 0.05 \implies m = \frac{6 \ln(80)}{\epsilon^2} = \frac{6 \ln(80)}{(0.05)^2} \approx 10517.$$

Hence, Fussbook can estimate the lift upto an error of 0.1 accurately with probability 0.95 by running the A/B test on  $m \approx 10517$  users.



Questions?

**Sample (outcome) space  $\mathcal{S}$ :** Nonempty (countable) set of possible outcomes.

**Outcome  $\omega \in \mathcal{S}$ :** Possible “thing” that can happen.

**Event  $E$ :** Any subset of the sample space.

**Probability function** on a sample space  $\mathcal{S}$  is a total function  $\Pr : \mathcal{S} \rightarrow [0, 1]$ . For any  $\omega \in \mathcal{S}$ ,

$$0 \leq \Pr[\omega] \leq 1 \quad ; \quad \sum_{\omega \in \mathcal{S}} \Pr[\omega] = 1 \quad ; \quad \Pr[E] = \sum_{\omega \in E} \Pr[\omega]$$

**Union:** For mutually exclusive events  $E_1, E_2, \dots, E_n$ ,  
 $\Pr[E_1 \cup E_2 \cup \dots \cup E_n] = \Pr[E_1] + \Pr[E_2] + \dots + \Pr[E_n]$ .

**Complement rule:**  $\Pr[E] = 1 - \Pr[E^c]$

**Inclusion-Exclusion rule:** For any two events  $E, F$ ,  $\Pr[E \cup F] = \Pr[E] + \Pr[F] - \Pr[E \cap F]$ .

**Union Bound:** For any events  $E_1, E_2, E_3, \dots, E_n$ ,  $\Pr[E_1 \cup E_2 \cup E_3 \dots \cup E_n] \leq \sum_{i=1}^n \Pr[E_i]$ .

**Uniform probability space:** A probability space is said to be uniform if  $\Pr[\omega]$  is the same for every outcome  $\omega \in \mathcal{S}$ . In this case,  $\Pr[E] = \frac{|E|}{|\mathcal{S}|}$ .

**Conditional Probability:** For events  $E$  and  $F$ , probability of event  $E$  conditioned on  $F$  is given by  $\Pr[E|F]$  and can be computed as  $\Pr[E|F] = \frac{\Pr[E \cap F]}{\Pr[F]}$ .

**Probability rules with conditioning:** For the complement  $E^c$ ,  $\Pr[E^c|F] = 1 - \Pr[E|F]$ .

**Conditional Probability for multiple events:**

$$\Pr[E_1 \cap E_2 \cap E_3] = \Pr[E_1] \Pr[E_2|E_1] \Pr[E_3|E_1 \cap E_2].$$

**Bayes rule:** For events  $E$  and  $F$  if  $\Pr[E] \neq 0$ ,  $\Pr[F|E] = \frac{\Pr[E|F] \Pr[F]}{\Pr[E]}$ .

**Law of Total Probability:** For events  $E$  and  $F$ ,  $\Pr[E] = \Pr[E|F] \Pr[F] + \Pr[E|F^c] \Pr[F^c]$ .

**Independent Events:** Events  $E$  and  $F$  are said to be independent, if knowledge that  $F$  has occurred does not change the probability that  $E$  occurs, i.e.  $\Pr[E|F] = \Pr[E]$  and  $\Pr[E \cap F] = \Pr[E] \Pr[F]$ .

**Pairwise Independence:** Events  $E_1, E_2, \dots, E_n$  are pairwise independent, if for every pair of events  $E_i$  and  $E_j$  ( $i \neq j$ ),  $\Pr[E_i|E_j] = \Pr[E_i]$  and  $\Pr[E_i \cap E_j] = \Pr[E_i] \Pr[E_j]$ .

**Mutual Independence:** Events  $E_1, E_2, \dots, E_n$  are mutually independent, if for every subset of events, the probability that all the selected events occur equals the product of the probabilities of the selected events. Formally, for every subset  $S \subseteq \{1, 2, \dots, n\}$ ,  $\Pr[\cap_{i \in S} E_i] = \prod_{i \in S} \Pr[E_i]$ .

**Random variable:** A random “variable”  $R$  on a probability space is a total function whose domain is the sample space  $\mathcal{S}$ . The codomain is denoted by  $V$  (usually a subset of the real numbers), meaning that  $R : \mathcal{S} \rightarrow V$ .

**Indicator Random Variables:** An indicator random variable corresponding to an event  $E$  is denoted as  $\mathcal{I}_E$  and is defined such that for  $\omega \in E$ ,  $\mathcal{I}_E[\omega] = 1$  and for  $\omega \notin E$ ,  $\mathcal{I}_E[\omega] = 0$ .

**Probability density function (PDF):** Let  $R$  be a random variable with codomain  $V$ . The probability density function of  $R$  is the function  $\text{PDF}_R : V \rightarrow [0, 1]$ , such that  $\text{PDF}_R[x] = \Pr[R = x]$  if  $x \in \text{Range}(R)$  and equal to zero if  $x \notin \text{Range}(R)$ .

$$\sum_{x \in V} \text{PDF}_R[x] = \sum_{x \in \text{Range}(R)} \Pr[R = x] = 1.$$

**Cumulative distribution function (CDF):** The cumulative distribution function of  $R$  is the function  $\text{CDF}_R : \mathbb{R} \rightarrow [0, 1]$ , such that  $\text{CDF}_R[x] = \Pr[R \leq x]$ .

**Distribution** over a random variable can be fully specified using the cumulative distribution function (CDF) (usually denoted by  $F$ ). The corresponding probability density function (PDF) is denoted by  $f$ .

## Wrapping up

**Bernoulli Distribution:**  $f_p(0) = 1 - p$ ,  $f_p(1) = p$ . *Example:* When tossing a coin such that  $\Pr[\text{heads}] = p$ , random variable  $R$  is equal to 1 if we get a heads (and equal to 0 otherwise). In this case,  $R$  follows the Bernoulli distribution i.e.  $R \sim \text{Ber}(p)$ .

**Uniform Distribution:** If  $R : \mathcal{S} \rightarrow V$ , then for all  $v \in V$ ,  $f(v) = 1/|V|$ . *Example:* When throwing an  $n$ -sided die, random variable  $R$  is the number that comes up on the die.  $V = \{1, 2, \dots, n\}$ . In this case,  $R$  follows the Uniform distribution i.e.  $R \sim \text{Uniform}(1, n)$ .

**Binomial Distribution:**  $f_{n,p}(k) = \binom{n}{k} p^k (1 - p)^{n-k}$ . *Example:* When tossing  $n$  independent coins such that  $\Pr[\text{heads}] = p$ , random variable  $R$  is the number of heads in  $n$  coin tosses. In this case,  $R$  follows the Binomial distribution i.e.  $R \sim \text{Bin}(n, p)$ .

**Geometric Distribution:**  $f_p(k) = (1 - p)^{k-1} p$ . *Example:* When repeatedly tossing a coin such that  $\Pr[\text{heads}] = p$ , random variable  $R$  is the number of tosses needed to get the first heads. In this case,  $R$  follows the Geometric distribution i.e.  $R \sim \text{Geo}(p)$ .

## Wrapping up

**Expectation**/mean of a random variable  $R$  is denoted by  $\mathbb{E}[R]$  and “summarizes” its distribution. Formally,  $\mathbb{E}[R] := \sum_{\omega \in \mathcal{S}} \Pr[\omega] R[\omega]$

**Alternate definition of expectation:**  $\mathbb{E}[R] = \sum_{x \in \text{Range}(R)} x \Pr[R = x]$ .

**Expectation of transformed r.v's:** For a random variable  $X : \mathcal{S} \rightarrow V$  and a function  $g : V \rightarrow \mathbb{R}$ , we define  $\mathbb{E}[g(X)]$  as follows:  $\mathbb{E}[g(X)] := \sum_{x \in \text{Range}(X)} g(x) \Pr[X = x]$

**Linearity of Expectation:** For  $n$  random variables  $R_1, R_2, \dots, R_n$  and constants  $a_1, a_2, \dots, a_n, b_1, b_2, \dots, b_n$ ,  $\mathbb{E}[\sum_{i=1}^n a_i R_i + b_i] = \sum_{i=1}^n a_i \mathbb{E}[R_i] + b_i$ .

**Conditional Expectation:** For random variable  $R$ , the expected value of  $R$  conditioned on an event  $A$  is given by  $\mathbb{E}[R|A] = \sum_{x \in \text{Range}(R)} x \Pr[R = x|A]$

**Law of Total Expectation:** If  $R$  is a random variable  $\mathcal{S} \rightarrow V$  and events  $A_1, A_2, \dots, A_n$  form a partition of the sample space, then,  $\mathbb{E}[R] = \sum_i \mathbb{E}[R|A_i] \Pr[A_i]$ .



## Wrapping up

**Independent random variables:** We define two random variables  $R_1$  and  $R_2$  to be independent if for *all*  $x_1 \in \text{Range}(R_1)$  and  $x_2 \in \text{Range}(R_2)$ , events  $[R_1 = x_1]$  and  $[R_2 = x_2]$  are independent. More formally,

$$\Pr[(R_1 = x_1) \cap (R_2 = x_2)] = \Pr[(R_1 = x_1)] \Pr[(R_2 = x_2)]$$

**Independent random variables:** Two random variables  $R_1$  and  $R_2$  are independent if for *all*  $x_1 \in \text{Range}(R_1)$  and  $x_2 \in \text{Range}(R_2)$ ,

$$\Pr[(R_1 = x_1) | (R_2 = x_2)] = \Pr[(R_1 = x_1)]$$

$$\Pr[(R_2 = x_2) | (R_1 = x_1)] = \Pr[(R_2 = x_2)]$$

**Expectation of product of r.v's:** For two r.v's  $R_1$  and  $R_2$ ,

$$\mathbb{E}[R_1 R_2] = \sum_{x \in \text{Range}(R_1 R_2)} x \Pr[R_1 R_2 = x].$$

**Expectation of product of independent r.v's:** For independent r.v's  $R_1$  and  $R_2$ ,

$$\mathbb{E}[R_1 R_2] = \mathbb{E}[R_1] \mathbb{E}[R_2].$$

## Wrapping up

**Joint distribution:** between r.v.'s  $X$  and  $Y$  can be specified by its joint PDF as follows:

$$\text{PDF}_{X,Y}[x,y] = \Pr[X = x \cap Y = y].$$

If  $X$  and  $Y$  are independent random variables,  $\text{PDF}_{X,Y}[x,y] = \text{PDF}_X[x] \text{PDF}_Y[y]$ .

**Marginalization:** We can obtain the distribution for each r.v. from the joint distribution by marginalizing over the other r.v.'s i.e.  $\text{PDF}_X[x] = \sum_i \text{PDF}_{X,Y}[x, y_i]$ .

**Variance:** Standard way to measure the deviation from the mean. For r.v.  $X$ ,  $\text{Var}[X] = \mathbb{E}[(X - \mathbb{E}[X])^2] = \sum_{x \in \text{Range}(X)} (x - \mu)^2 \Pr[X = x]$  where  $\mu := \mathbb{E}[X]$ .

**Alternate definition of variance:**  $\text{Var}[X] = \mathbb{E}[X^2] - \mu^2 = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$ .

**Standard Deviation:** For r.v.  $X$ , the standard deviation of  $X$  is defined as  $\sigma_X := \sqrt{\text{Var}[X]} = \sqrt{\mathbb{E}[X^2] - (\mathbb{E}[X])^2}$ .

**Properties of variance:** For constants  $a, b$  and r.v.  $R$ ,  $\text{Var}[aR + b] = a^2\text{Var}[R]$ .

**Pairwise Independence of r.v's:** Random variables  $R_1, R_2, R_3, \dots, R_n$  are pairwise independent if for any pair  $R_i$  and  $R_j$ , for  $x \in \text{Range}(R_i)$  and  $y \in \text{Range}(R_j)$ ,  
 $\Pr[(R_i = x) \cap (R_j = y)] = \Pr[R_i = x] \Pr[R_j = y]$ .

**Linearity of variance for pairwise independent r.v's:** If  $R_1, \dots, R_n$  are pairwise independent,  
 $\text{Var}[R_1 + R_2 + \dots + R_n] = \sum_{i=1}^n \text{Var}[R_i]$ .

**Properties of variance:** If  $R_1, \dots, R_n$  are pairwise independent, for constants  $a_1, a_2, \dots, a_n$  and  $b_1, b_2, \dots, b_n$ ,  $\text{Var}[\sum_{i=1}^n a_i R_i + b_i] = \sum_{i=1}^n a_i^2 \text{Var}[R_i]$ .

## Wrapping up

**Covariance:** For two random variables  $R$  and  $S$ , the covariance between  $R$  and  $S$  is defined as:

$$\text{Cov}[R, S] = \mathbb{E}[(R - \mathbb{E}[R])(S - \mathbb{E}[S])] = \mathbb{E}[RS] - \mathbb{E}[R]\mathbb{E}[S].$$

**Properties of covariance:** If  $R$  and  $S$  are independent r.v.'s,  $\mathbb{E}[RS] = \mathbb{E}[R]\mathbb{E}[S]$  and  $\text{Cov}[R, S] = 0$ .  $\text{Cov}[R, R] = \text{Var}[R]$ .  $\text{Cov}[R, S] = \text{Cov}[S, R]$ .

**Covariance of sum of r.v.'s:** For r.v.'s  $R_1, R_2, \dots, R_n$ ,  
$$\text{Var} \left[ \sum_{i=1}^n R_i \right] = \sum_{i=1}^n \text{Var}[R_i] + 2 \sum_{1 \leq i < j \leq n} \text{Cov}[R_i, R_j].$$

If  $R_i$  and  $R_j$  are pairwise independent,  $\text{Cov}[R_i, R_j] = 0$  and  $\text{Var} \left[ \sum_{i=1}^n R_i \right] = \sum_{i=1}^n \text{Var}[R_i]$ .

The correlation between two r.v.'s  $R_1$  and  $R_2$  is defined as  $\text{Corr}[R_1, R_2] = \frac{\text{Cov}[R_1, R_2]}{\sqrt{\text{Var}[R_1] \text{Var}[R_2]}}$ .

$\text{Corr}[R_1, R_2] \in [-1, 1]$  and indicates the strength of the relationship between  $R_1$  and  $R_2$ .

## Wrapping up

**Bernoulli:** If  $R \sim \text{Bernoulli}(p)$ ,  $\mathbb{E}[R] = p$  and  $\text{Var}[R] = p(1 - p)$ .

**Uniform:** If  $R \sim \text{Uniform}(\{v_1, \dots, v_n\})$ ,  $\mathbb{E}[R] = \frac{v_1 + v_2 + \dots + v_n}{n}$  and  $\text{Var}[R] = \frac{[v_1^2 + v_2^2 + \dots + v_n^2]}{n} - \left( \frac{[v_1 + v_2 + \dots + v_n]}{n} \right)^2$ .

**Binomial:** If  $R \sim \text{Bin}(n, p)$ ,  $\mathbb{E}[R] = np$  and  $\text{Var}[R] = np(1 - p)$ .

**Geometric:** If  $R \sim \text{Geo}(p)$ ,  $\mathbb{E}[R] = \frac{1}{p}$  and  $\text{Var}[R] = \frac{1-p}{p^2}$ .

## Wrapping up

**Tail inequalities** bound the probability that the r.v. takes a value much different from its mean.

**Markov's Theorem:** If  $X$  is a non-negative random variable, then for all  $x > 0$ ,  
 $\Pr[X \geq x] \leq \frac{\mathbb{E}[X]}{x}$ .

**Chebyshev's Theorem:** For a r.v.  $X$  and all  $x > 0$ ,  $\Pr[|X - \mathbb{E}[X]| \geq x] \leq \frac{\text{Var}[X]}{x^2}$ .

**Weak Law of Large Numbers:** Let  $G_1, G_2, \dots, G_n$  be pairwise independent variables with the same mean  $\mu$  and (finite) standard deviation  $\sigma$ . Define  $T_n := \frac{\sum_{i=1}^n G_i}{n}$ , then for every  $\epsilon > 0$ ,  
 $\lim_{n \rightarrow \infty} \Pr[|T_n - \mu| \leq \epsilon] = 1$ .

**Chernoff Bound:** If  $T_1, T_2, \dots, T_n$  are mutually independent r.v.'s such that  $0 \leq T_i \leq 1$  for all  $i$ . If  $T := \sum_{i=1}^n T_i$ , for all  $c \geq 1$  and  $\beta(c) := c \ln(c) - c + 1$ , then,  
 $\Pr[T \geq c\mathbb{E}[T]] \leq \exp(-\beta(c) \mathbb{E}[T])$ .

**Two-sided Chernoff Bound:**  $\Pr[|T - \mathbb{E}[T]| \geq c\mathbb{E}[T]] \leq 2 \exp\left(-\frac{c^2 \mathbb{E}[T]}{3}\right)$

Questions?