CSE 312

Foundations of Computing II

Lecture 17: Polling

Continuity Correction & Distinct Elements

Review: Central Limit Theorem

$$Y_n = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

$\lim_{n\to\infty} (S_n = X_1 + \cdots + X_n) \to \mathcal{N}(n\cdot\mu, n\sigma^2)$ $\sim \mathcal{N}(0,1)$

Theorem. (Central Limit Theorem) The CDF of Y_n converges to the CDF of the standard normal $\mathcal{N}(0,1)$, i.e.,

$$\lim_{n\to\infty} P(Y_n \leq y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-x^2/2} \mathrm{d}x$$

Application: Use Normal Distribution to Approximate Y_n
No need to understand Y_n !!

Also stated as:

- $\lim_{n\to\infty} Y_n \to \mathcal{N}(0,1)$
- $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} X_i \to \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ for $\mu = \mathbb{E}[X_i]$ and $\sigma^2 = \mathrm{Var}(X_i)$

Magic Mushrooms

In Fall 2020, Oregonians voted on whether to legalize the therapeutic use of “magic mushrooms”.

Poll to determine the fraction p of the population expected to vote in favor.

- Call up a random sample of n people to ask their opinion
- Report the empirical fraction

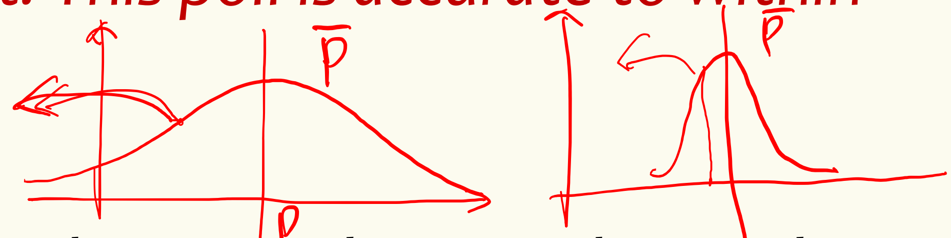
Questions

- Is this a good estimate?
- How to choose n?

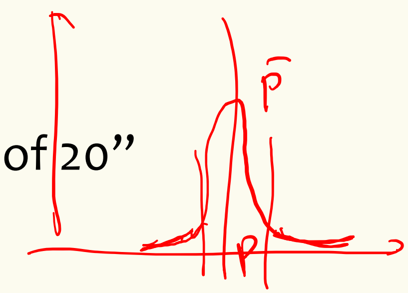
Polling Accuracy

Often see claims that say

"Our poll found 80% support. This poll is accurate to within 5% with 98% probability"*



Will unpack what this and how they sample enough people to know this is true.



* When it is 95% this is sometimes written as "19 times out of 20"

Formalizing Polls

Population size N, true fraction of voting in favor p, sample size n.

Problem: We don't know p, want to estimate it

Polling Procedure

for $i = 1, \ldots, n$:

1. Pick uniformly random person to call (prob: $1/N$)
2. Ask them how they will vote

$$X_i = \begin{cases} 1, & \text{voting in favor} \\ 0, & \text{otherwise} \end{cases}$$

Report our estimate of p: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Formalizing Polls

Population size N, true fraction of voting in favor p, sample size n.

Problem: We don't know p

What type of r.v. is X_i?

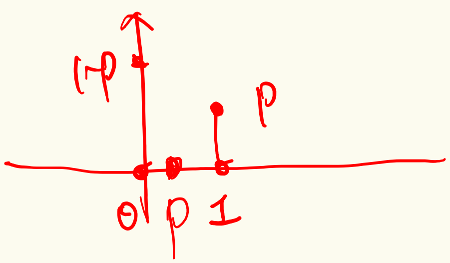
Poll: pollev.com/rachel312

	Type	$\mathbb{E}[X_i]$	$\mathrm{Var}(X_i)$
a.	Bernoulli	p	$p(1-p)$
b.	Bernoulli	p	p^2
c.	Geometric	p	$\frac{1-p}{p^2}$
d.	Binomial	np	$np(1-p)$

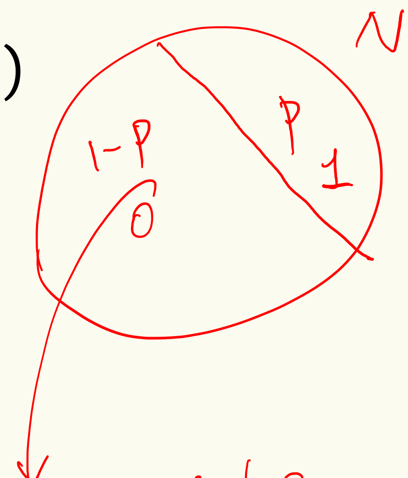
Polling Procedure

for $i = 1, \ldots, n$:

1. Pick uniformly random person to call (prob: $1/N$)
2. Ask them how they will vote



$$X_i = \begin{cases} 1, & \text{voting in favor} \\ 0, & \text{otherwise} \end{cases}$$



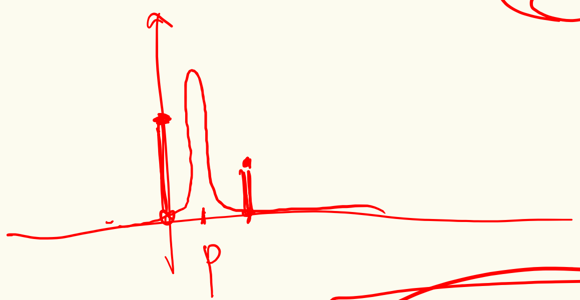
Report our estimate of p: $\hat{p} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$X_i = 1 \mid 0$

$P[X_i = 1] = p$

Random Variables

What type of r.v. is X_i?



	Type	$\mathbb{E}[X_i]$	$\mathrm{Var}(X_i)$
a.	Bernoulli	p	$p(1-p)$
b.	Bernoulli	p	p^2
c.	Geometric	p	$\frac{1-p}{p^2}$
d.	Binomial	np	$np(1-p)$

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

$E[\bar{X}] = \frac{1}{n} \sum E[X_i] = \frac{n \cdot p}{n}$

$Var[\sum X_i] = \sum Var[X_i] = n \cdot p(1-p)$

$Var[\frac{\sum X_i}{n}] = \frac{Var[\sum X_i]}{n^2} = \frac{p(1-p)}{n}$

Poll: pollev.com/rachel312

	$\mathbb{E}[\bar{X}]$	$\mathrm{Var}(\bar{X})$
a.	np	$np(1-p)$
b.	p	$p(1-p)$
c.	p	$p(1-p)/n$
d.	p/n	$p(1-p)/n$

Central Limit Theorem

pollev.com/rachel312

Poll: In the limit $\overline{X}$ is…?

a. $\mathcal{N}(0,1)$
b. $\mathcal{N}(p, p(1-p))$
c. $\mathcal{N}(p, p(1-p)/n)$
d. I don't know

With i.i.d random variables $X_1, X_2, \ldots, X_n$ where $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$

As $n \to \infty$,

$$Y_n = \frac{X_1 + X_2 + \cdots X_n - n\mu}{\sigma\sqrt{n}} \to \mathcal{N}(0,1)$$

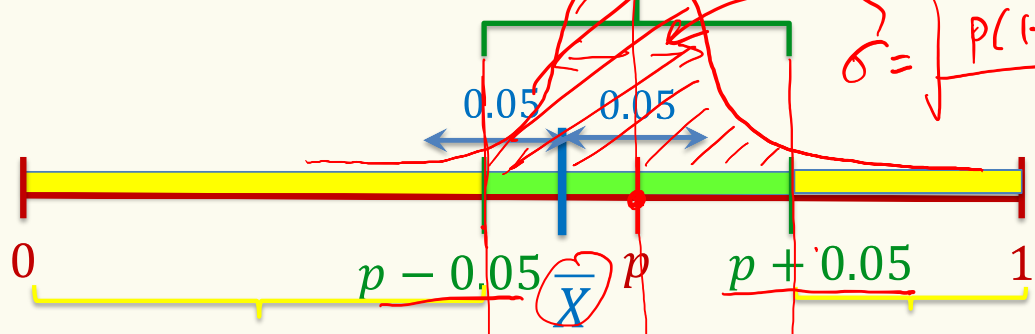
As $n \to \infty$,

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

Roadmap: Bounding Error

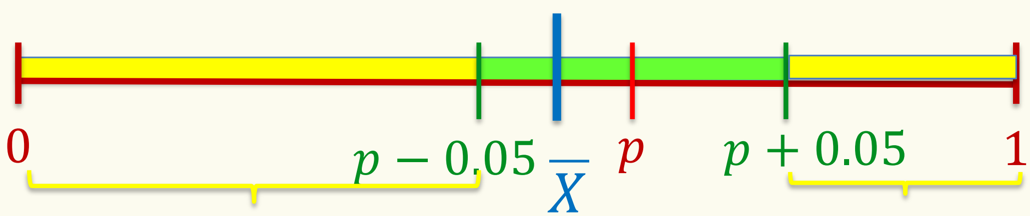
Goal: Find the value of n such that 98% of the time, the estimate $\bar{X}$ is within 5% of the true p

Get good estimate if $\bar{X}$ lands in this region



Want $P\left(\left|\bar{X} - p\right| > 0.05\right) \leq 0.02$

Roadmap: Bounding Error



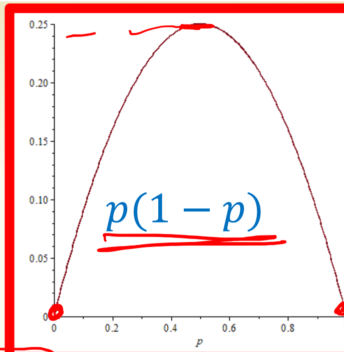
Want $P\left(\left|\overline{X} - p\right| > 0.05\right) \leq 0.02$

Roadmap: Bounding Error

Goal: Find the value of n such that 98% of the time, the estimate $\bar{X}$ is within 5% of the true p

1. Define probability of a "bad event" $P(|\bar{X} - p| > 0.05) \leq 0.02$
2. Apply CLT
3. Convert to a standard normal
4. Solve for n

Following the Road Map



1. Want $P(|\overline{X} - p| > 0.05) \le 0.02$
2. By CLT $\overline{X} \to \mathcal{N}(\mu, \sigma^2)$ where $\mu = p$ and $\sigma^2 = p(1-p)/n$
3. Define $Z = \frac{\overline{X}-\mu}{\sigma} = \frac{\overline{X}-p}{\sigma}$. Then, by the CLT $Z \to \mathcal{N}(0,1)$

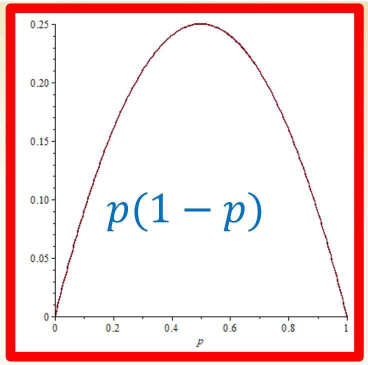
$\frac{\mathbf{1}}{\sqrt{\boldsymbol{p(1-p)}}}$ is always $\ge \mathbf{2}$

$$P(|\overline{X} - p| > 0.05) = P(|Z|\sigma > 0.05)$$

$$= P(|Z| > 0.05/\sigma) = P\left(|Z| > 0.05 \frac{\sqrt{n}}{\sqrt{p(1-p)}}\right)$$

$$\le P(|Z| > 0.1\sqrt{n})$$

Following the Road Map



1. Want $P\left(\left|\overline{X}-p\right| > 0.05\right) \le 0.02$

2. By CLT $\overline{X} \to \mathcal{N}(\mu, \sigma^2)$ where $\mu = p$ and $\sigma^2 = p(1-p)/n$

3. Define $Z = \frac{\overline{X}-\mu}{\sigma} = \frac{\overline{X}-p}{\sigma}$. Then, by the CLT $Z \to \mathcal{N}(0,1)$

$\frac{1}{\sqrt{p(1-p)}}$ is always $\ge \mathbf{2}$

$$P\left(\left|\overline{X}-p\right| > 0.05\right) = P(|Z| \cdot \sigma > 0.05)$$

$$= P(|Z| > 0.05/\sigma) = P\left(|Z| > 0.05\frac{\sqrt{n}}{\sqrt{p(1-p)}}\right)$$

Want to choose n so that this is at most 0.02

$$\le P(|Z| > 0.1\sqrt{n}) \le 0.02$$

4. Solve for n

We want $P(|Z| > 0.1\sqrt{n}) \leq 0.02$ *where* $Z \rightarrow \mathcal{N}(0,1)$

- If we actually had $Z \sim \mathcal{N}(0,1)$ then enough to show that $P(Z > 0.1\sqrt{n}) \leq 0.01$ since $\mathcal{N}(0,1)$ is symmetric about 0

≤ 0.01

- Now $P(Z > z) = 1 - \Phi(z)$ where $\Phi(z)$ is the CDF of the Standard Normal Distribution

$P(Z \leq z) \geq 0.99$

- So, want to choose n so that $0.1\sqrt{n} \geq z$ where $\Phi(z) \geq 0.99$

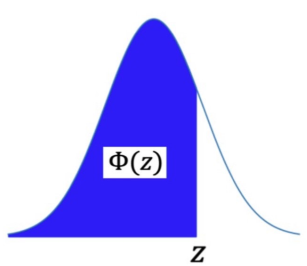
Table of $\Phi(z)$ CDF of Standard Normal Distribution

Choose n so

$0.1\sqrt{n} \geq z$ where

$\Phi(z) \geq 0.99$

From table $z = 2.33$ works



Φ Table: $\mathbb{P}(Z \leq z)$ when $Z \sim \mathcal{N}(0,1)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5	0.50399	0.50798	0.51197	0.51595	0.51994	0.52392	0.5279	0.53188	0.53586
0.1	0.53983	0.5438	0.54776	0.55172	0.55567	0.55962	0.56356	0.56749	0.57142	0.57535
0.2	0.57926	0.58317	0.58706	0.59095	0.59483	0.59871	0.60257	0.60642	0.61026	0.61409
0.3	0.61791	0.62172	0.62552	0.6293	0.63307	0.63683	0.64058	0.64431	0.64803	0.65173
0.4	0.65542	0.6591	0.66276	0.6664	0.67003	0.67364	0.67724	0.68082	0.68439	0.68793
0.5	0.69146	0.69497	0.69847	0.70194	0.7054	0.70884	0.71226	0.71566	0.71904	0.7224
0.6	0.72575	0.72907	0.73237	0.73565	0.73891	0.74215	0.74537	0.74857	0.75175	0.7549
0.7	0.75804	0.76115	0.76424	0.7673	0.77035	0.77337	0.77637	0.77935	0.7823	0.78524
0.8	0.78814	0.79103	0.79389	0.79673	0.79955	0.80234	0.80511	0.80785	0.81057	0.81327
0.9	0.81594	0.81859	0.82121	0.82381	0.82639	0.82894	0.83147	0.83398	0.83646	0.83891
1.0	0.84134	0.84375	0.84614	0.84849	0.85083	0.85314	0.85543	0.85769	0.85993	0.86214
1.1	0.86433	0.8665	0.86864	0.87076	0.87286	0.87493	0.87698	0.879	0.881	0.88298
1.2	0.88493	0.88686	0.88877	0.89065	0.89251	0.89435	0.89617	0.89796	0.89973	0.90147
1.3	0.9032	0.9049	0.90658	0.90824	0.90988	0.91149	0.91309	0.91466	0.91621	0.91774
1.4	0.91924	0.92073	0.9222	0.92364	0.92507	0.92647	0.92785	0.92922	0.93056	0.93189
1.5	0.93319	0.93448	0.93574	0.93699	0.93822	0.93943	0.94062	0.94179	0.94295	0.94408
1.6	0.9452	0.9463	0.94738	0.94845	0.9495	0.95053	0.95154	0.95254	0.95352	0.95449
1.7	0.95543	0.95637	0.95728	0.95818	0.95907	0.95994	0.9608	0.96164	0.96246	0.96327
1.8	0.96407	0.96485	0.96562	0.96638	0.96712	0.96784	0.96856	0.96926	0.96995	0.97062
1.9	0.97128	0.97193	0.97257	0.9732	0.97381	0.97441	0.975	0.97558	0.97615	0.9767
2.0	0.97725	0.97778	0.97831	0.97882	0.97932	0.97982	0.9803	0.98077	0.98124	0.98169
2.1	0.98214	0.98257	0.983	0.98341	0.98382	0.98422	0.98461	0.985	0.98537	0.98574
2.2	0.9861	0.98645	0.98679	0.98713	0.98745	0.98778	0.98809	0.9884	0.9887	0.98899
2.3	0.98928	0.98956	0.98983	0.9901	0.99036	0.99061	0.99086	0.99111	0.99134	0.99158
2.4	0.9918	0.99202	0.99224	0.99245	0.99266	0.99286	0.99305	0.99324	0.99343	0.99361
2.5	0.99379	0.99396	0.99413	0.9943	0.99446	0.99461	0.99477	0.99492	0.99506	0.9952
2.6	0.99534	0.99547	0.9956	0.99573	0.99585	0.99598	0.99609	0.99621	0.99632	0.99643
2.7	0.99653	0.99664	0.99674	0.99683	0.99693	0.99702	0.99711	0.9972	0.99728	0.99736
2.8	0.99744	0.99752	0.9976	0.99767	0.99774	0.99781	0.99788	0.99795	0.99801	0.99807
2.9	0.99813	0.99819	0.99825	0.99831	0.99836	0.99841	0.99846	0.99851	0.99856	0.99861
3.0	0.99865	0.99869	0.99874	0.99878	0.99882	0.99886	0.99889	0.99893	0.99896	0.999

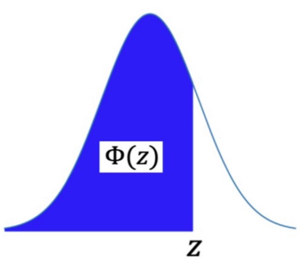
4. Solve for n

Choose n so
$0.1\sqrt{n} \geq z$ where
$\Phi(z) \geq 0.99$

$\Pr[|\bar{X} - p| \geq 0.05] \leq 0.02$

From table $z = 2.33$ works

- So we can choose $0.1\sqrt{n} \geq 2.33$
  or $\sqrt{n} \geq 23.3$
- Then $n \geq 543 \geq (23.3)^2$ would be good enough … if we had $Z \sim \mathcal{N}(0,1)$
- We only have $Z \rightarrow \mathcal{N}(0,1)$ so there is some loss due to approximation error.
- Maybe instead consider $z = 3.0$ with $\Phi(z) \geq 0.99865$ and $n \geq 30^2 = 900$ to cover any loss.



Idealized Polling

So far, we have been discussing "idealized polling". Real life is normally not so nice ☹

Assumed we can sample people uniformly at random, not really possible in practice

- Not everyone responds
- Response rates might differ in different groups
- Will people respond truthfully?

Makes polling in real life much more complex than this idealized model!

Agenda

- Continuity correction
- Application: Counting distinct elements

Example – Y_n is binomial

We understand binomial, so we can see how well approximation works

We flip n independent coins, heads with probability $p = 0.75$.

X = # heads $\quad \mu = \mathbb{E}(X) = 0.75n \quad \sigma^2 = \text{Var}(X) = p(1-p)n = 0.1875n$

$\mathbb{P}(X \leq 0.7n)$

n	exact	$\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$ approx
10	0.4744072	0.357500327
20	0.38282735	0.302788308
50	0.25191886	0.207108089
100	0.14954105	0.124106539
200	0.06247223	0.051235217
1000	0.00019359	0.000130365

Example – Naive Approximation

Fair coin flipped (independently) **40** times. Probability of **20** or **21** heads?

Exact. $\mathbb{P}(X \in \{20,21\}) = \left[\binom{40}{20} + \binom{40}{21}\right]\left(\frac{1}{2}\right)^{40} \approx \boxed{0.2448}$

Approx. X = # heads $\quad \mu = \mathbb{E}(X) = 0.5n = 20 \quad \sigma^2 = \mathrm{Var}(X) = 0.25n = 10$

$$\mathbb{P}(20 \le X \le 21) = \Phi\left(\frac{20-20}{\sqrt{10}} \le \frac{X-20}{\sqrt{10}} \le \frac{21-20}{\sqrt{10}}\right)$$

$$\approx \Phi\left(0 \le \frac{X-20}{\sqrt{10}} \le 0.32\right)$$

😢

$$= \Phi(0.32) - \Phi(0) \approx \boxed{0.1241}$$

Example – Even Worse Approximation

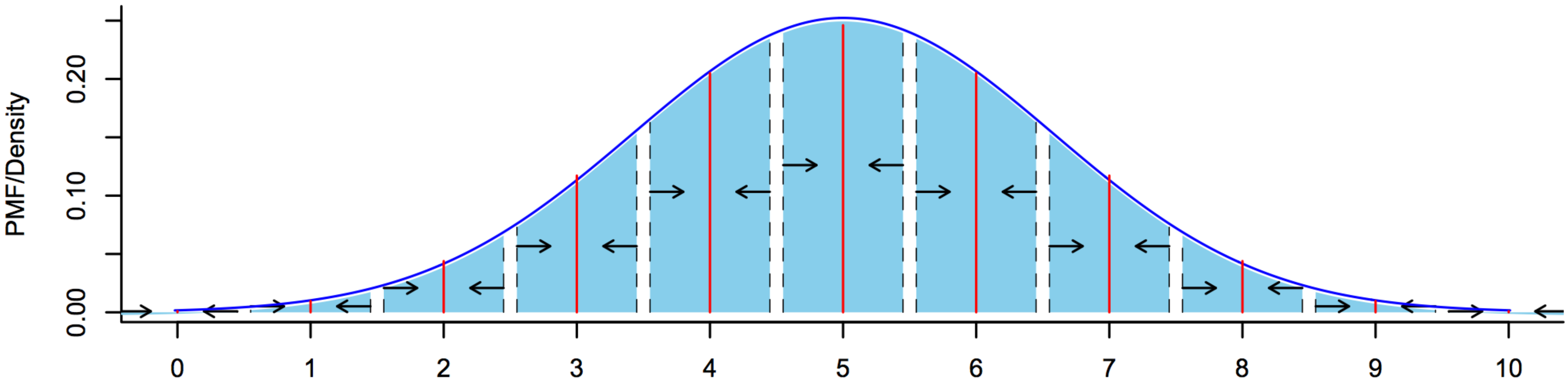
Fair coin flipped (independently) **40** times. Probability of **20** heads?

Exact. $\mathbb{P}(X = 20) = \binom{40}{20}\left(\frac{1}{2}\right)^{40} \approx \boxed{0.1254}$

Approx. $\mathbb{P}(20 \leq X \leq 20) = 0$ 😢

Solution – Continuity Correction

Probability estimate for i: Probability for all x that round to i!



To estimate probability that discrete RV lands in (integer) interval $\{a, \dots, b\}$, compute probability continuous approximation lands in interval $[a - \frac{1}{2}, b + \frac{1}{2}]$

Example – Continuity Correction

Fair coin flipped (independently) **40** times. Probability of **20** or **21** heads?

Exact. $\mathbb{P}(X \in \{20,21\}) = \left[\binom{40}{20} + \binom{40}{21}\right]\left(\frac{1}{2}\right)^{40} \approx \boxed{0.2448}$

Approx. X = # heads $\quad \mu = \mathbb{E}(X) = 0.5n = 20 \quad \sigma^2 = \mathrm{Var}(X) = 0.25n = 10$

$$\mathbb{P}(19.5 \le X \le 21.5) = \Phi\left(\frac{19.5 - 20}{\sqrt{10}} \le \frac{X - 20}{\sqrt{10}} \le \frac{21.5 - 20}{\sqrt{10}}\right)$$

$$\approx \Phi\left(-0.16 \le \frac{X - 20}{\sqrt{10}} \le 0.47\right)$$

$$= \Phi(0.47) - \Phi(-0.16) \approx \boxed{0.2452}$$

👍

Example – Continuity Correction

Fair coin flipped (independently) **40** times. Probability of **20** heads?

Exact. $\mathbb{P}(X = 20) = \binom{40}{20}\left(\frac{1}{2}\right)^{40} \approx \boxed{0.1254}$

Approx. $\mathbb{P}(19.5 \le X \le 20.5) = \Phi\left(\frac{19.5-20}{\sqrt{10}} \le \frac{X-20}{\sqrt{10}} \le \frac{20.5-20}{\sqrt{10}}\right)$

$$\approx \Phi\left(-0.16 \le \frac{X-20}{\sqrt{10}} \le 0.16\right)$$

$$= \Phi(0.16) - \Phi(-0.16) \approx \boxed{0.1272}$$

Agenda

- Continuity correction
- Application: Counting distinct elements

Data mining – Stream Model

- In many data mining situations, data often not known ahead of time.
  - Examples: Google queries, Twitter or Facebook status updates, YouTube video views
- Think of the data as an infinite stream
- Input elements (e.g. Google queries) enter/arrive one at a time.
  - We cannot possibly store the stream.

Question: How do we make critical calculations about the data stream using a limited amount of memory?

Stream Model – Problem Setup

Input: sequence (aka. "stream") of N elements $x_1, x_2, \ldots, x_N$ from a known universe U (e.g., 8-byte integers).

Goal: perform a computation on the input, in a single left to right pass, where:

- Elements processed in real time
- Can't store the full data ⇒ use minimal amount of storage while maintaining working "summary"

What can we compute?

32, 12, 14, 32, 7, 12, 32, 7, 32, 12, 4

Some functions are easy:

- Min
- Max
- Sum
- Average

Today: Counting distinct elements

32, 12, 14, 32, 7, 12, 32, 7, 32, 12, 4

Application

You are the content manager at YouTube, and you are trying to figure out the **distinct** view count for a video. How do we do that?

Note: A person can view their favorite videos several times, but they only count as 1 **distinct** view!

Other applications

- IP packet streams: How many distinct IP addresses or IP flows (source+destination IP, port, protocol)
  - Anomaly detection, traffic monitoring
- Search: How many distinct search queries on Google on a certain topic yesterday
- Web services: how many distinct users (cookies) searched/browsed a certain term/item
  - Advertising, marketing trends, etc.

Counting distinct elements

32, 12, 14, 32, 7, 12, 32, 7, 32, 12, 4

N = # of IDs in the stream = 11, m = # of distinct IDs in the stream = 5

Want to compute number of **distinct** IDs in the stream.

- *Naïve solution: As the data stream comes in, store all distinct IDs in a hash table.*
- *Space requirement:* $\Omega(m)$

YouTube Scenario: m is huge!

Counting distinct elements

32, 12, 14, 32, 7, 12, 32, 7, 32, 12, 4

N = # of IDs in the stream = 11, m = # of distinct IDs in the stream = 5

Want to compute number of **distinct** IDs in the stream.

How to do this <u>without</u> storing all the elements?

Detour – I.I.D. Uniforms

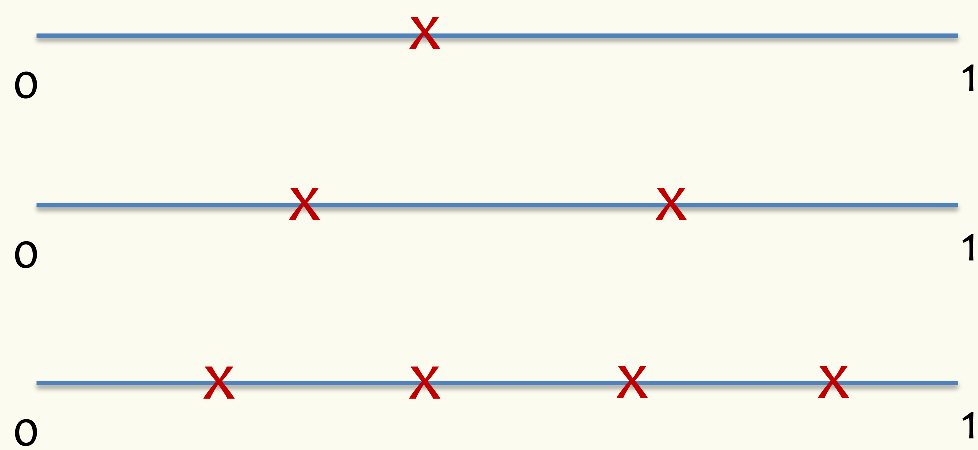
If $Y_1, \cdots, Y_m \sim \text{Unif}(0,1)$ (i.i.d.) where do we expect the points to end up?

"Evenly spread out"

$m = 1$

$m = 2$

$m = 4$



What is some intuition for this?

Detour – I.I.D. Uniforms

If $Y_1, \cdots, Y_m \sim \text{Unif}(0,1)$ (i.i.d.) where do we expect the points to end up?

$m = 1$

Y_1 has expected value $1/2$

… but probably isn't very close to the middle

… and Y_2 is more likely to be in the bigger gap

$m = 2$

Detour – Min of I.I.D. Uniforms

If $Y_1, \cdots, Y_m \sim \mathrm{Unif}(0,1)$ (i.i.d.) where do we expect the points to end up?

e.g., what is $\mathbb{E}[\min\{Y_1, \cdots, Y_m\}]$?

CDF: Observe that $\min\{Y_1, \cdots, Y_m\} \geq y$ if and only if $Y_1 \geq y, \dots, Y_m \geq y$

(Similar to Section 6)

$$P(\min\{Y_1, \cdots, Y_m\} \geq y) = P(Y_1 \geq y, \dots, Y_m \geq y)$$

$y \in [0,1]$

$$= P(Y_1 \geq y) \cdots P(Y_m \geq y) \quad \text{(Independence)}$$

$$= (1-y)^m$$

$$\Rightarrow P(\min\{Y_1, \cdots, Y_m\} \leq y) = 1 - (1-y)^m$$

Detour – Min of I.I.D. Uniforms

Useful fact. For any random variable Y taking non-negative values

$$\mathbb{E}[Y] = \int_0^\infty P(Y \geq y)\mathrm{d}y$$

Proof (Not covered)

$$\mathbb{E}[Y] = \int_0^\infty x \cdot f_Y(x)\,\mathrm{d}x = \int_0^\infty \left(\int_0^x 1\,\mathrm{d}y\right) \cdot f_Y(x)\,\mathrm{d}x = \int_0^\infty \int_0^x f_Y(x)\,\mathrm{d}y\,\mathrm{d}x$$

$$= \iint_{0 \leq y \leq x \leq \infty} f_Y(x) = \int_0^\infty \int_y^\infty f_Y(x)\,\mathrm{d}x\,\mathrm{d}y = \int_0^\infty P(Y \geq y)\,\mathrm{d}y$$

Detour – Min of I.I.D. Uniforms

$Y_1, \cdots, Y_m \sim \mathrm{Unif}(0,1)$ (i.i.d.)

$Y = \min\{Y_1, \cdots, Y_m\}$

Useful fact. For any random variable Y taking non-negative values

$$\mathbb{E}[Y] = \int_0^\infty P(Y \geq y) \mathrm{d}y$$

$$\mathbb{E}[Y] = \int_0^\infty P(Y \geq y) \mathrm{d}y = \int_0^1 (1-y)^m \mathrm{d}y$$

$$= -\frac{1}{m+1}(1-y)^{m+1}\Big|_0^1 = 0 - \left(-\frac{1}{m+1}\right) = \frac{1}{m+1}$$

Detour – Min of I.I.D. Uniforms

If $Y_1, \cdots, Y_m \sim \text{Unif}(0,1)$ (iid) where do we expect the points to end up?

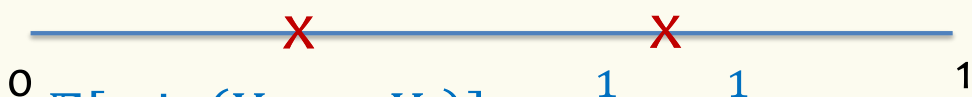
In general, $\mathbb{E}[\min(Y_1, \cdots, Y_m)] = \frac{1}{m+1}$

$$\mathbb{E}[\min(Y_1)] = \frac{1}{1+1} = \frac{1}{2}$$

$m = 1$

$$\mathbb{E}[\min(Y_1, Y_2)] = \frac{1}{2+1} = \frac{1}{3}$$

$m = 2$



$$\mathbb{E}[\min(Y_1, \cdots, Y_4)] = \frac{1}{4+1} = \frac{1}{5}$$

$m = 4$

Distinct Elements – Hashing into $[\mathbf{0}, \mathbf{1}]$

Hash function $h: U \to [0,1]$
Assumption: For all $x \in U$, $h(x) \sim \mathrm{Unif}(0,1)$ and mutually independent

$x_1 = 5$	$x_2 = 2$	$x_3 = 27$	$x_4 = 35$	$x_5 = 4$
$h(5)$	$h(2)$	$h(27)$	$h(35)$	$h(4)$

5 distinct elements

$\to$ 5 i.i.d. RVs $h(x_1), \ldots, h(x_5) \sim \mathrm{Unif}(0,1)$

$\to \mathbb{E}[\min\{h(x_1), \ldots, h(x_5)\}] = \frac{1}{5+1} = \frac{1}{6}$

Distinct Elements – Hashing into $[0, 1]$

Hash function $h: U \to [0,1]$
Assumption: For all $x \in U$, $h(x) \sim \text{Unif}(0,1)$ and mutually independent

$x_1 = 5$	$x_2 = 2$	$x_3 = 27$	$x_4 = 5$	$x_5 = 4$
$h(5)$	$h(2)$	$h(27)$	$h(5)$	$h(4)$

4 distinct elements

$\Rightarrow$ 4 i.i.d. RVs $h(x_1), h(x_2), h(x_3), h(x_5) \sim \text{Unif}(0,1)$ and $h(x_1) = h(x_4)$

$\Rightarrow \mathbb{E}[\min\{h(x_1), \dots, h(x_5)\}] = \mathbb{E}[\min\{h(x_1), h(x_2), h(x_3), h(x_5)\}] = \frac{1}{4+1}$

Distinct Elements – Hashing into $[0, 1]$

Hash function $h: U \to [0,1]$
Assumption: For all $x \in U$, $h(x) \sim \mathrm{Unif}(0,1)$ and mutually independent

$x_1, x_2, \dots, x_N$ contains m distinct elements

$h(x_1), h(x_2), \dots, h(x_N)$ contains m i.i.d. rvs $\sim \mathrm{Unif}(0,1)$

and $N - m$ repeats

$$\mathbb{E}[\min\{h(x_1), \dots, h(x_N)\}] = \frac{1}{m+1} \Longleftrightarrow m = \frac{1}{\mathbb{E}[\min\{h(x_1), \dots, h(x_N)\}]} - 1$$

The MinHash Algorithm – Idea

$$m = \frac{1}{\mathbb{E}[\min\{h(x_1), \ldots, h(x_N)\}]} - 1$$

1. Compute $\mathrm{val} = \min\{h(x_1), \ldots, h(x_N)\}$
2. Assume that $\mathrm{val} \approx \mathbb{E}[\min\{h(x_1), \ldots, h(x_N)\}]$
3. Output $\mathrm{round}\left(\frac{1}{\mathrm{val}} - 1\right)$

The MinHash Algorithm – Implementation

Algorithm **`MinHash`**$(x_1, x_2, \dots, x_N)$

$\mathrm{val} \leftarrow \infty$

for $i = 1$ **to** N **do**

$\quad \mathrm{val} \leftarrow \min\{\mathrm{val}, h(x_i)\}$

return $\mathrm{round}\left(\frac{1}{\mathrm{val}} - 1\right)$

Memory cost = just remember val (with sufficient precision)

MinHash Example

Stream: 13, 25, 19, 25, 19, 19

Hashes: 0.51, 0.26, 0.79, 0.26, 0.79, 0.79

What does MinHash return?

Poll: pollev.com/rachel312

a. 1
b. 3
c. 5
d. No idea

MinHash Example II

Stream: 11, 34, 89, 11, 89, 23

Hashes: 0.5, 0.21, 0.94, 0.5, 0.94, 0.1

Output is $\frac{1}{0.1} - 1 = 9$

Clearly, not a very good answer!

Not unlikely: $P(h(x) < 0.1) = 0.1$

The MinHash Algorithm – Problem

Algorithm **MinHash**$(x_1, x_2, \dots, x_N)$

$\text{val} \leftarrow \infty$

for $i = 1$ **to** N **do**

$\quad \text{val} \leftarrow \min\{\text{val}, h(x_i)\}$

return $\text{round}\left(\frac{1}{\text{val}} - 1\right)$

But, val is not $\mathbb{E}[\text{val}]$!
How far is val from $\mathbb{E}[\text{val}]$?

$$\text{Var}(\text{val}) \approx \frac{1}{(m+1)^2}$$

$\text{val} = \min\{h(x_1), \dots, h(x_N)\}$ $\quad \mathbb{E}[\text{val}] = \frac{1}{m+1}$

How can we reduce the variance?

Idea: Repetition to reduce variance!

Use k **independent** hash functions $h^1, h^2, \cdots h^k$

Algorithm **`MinHash`**$(x_1, x_2, \ldots, x_N)$

$\mathrm{val}_1, \ldots, \mathrm{val}_\mathrm{k} \leftarrow \infty$

for $i = 1$ **to** N **do**

$\quad \mathrm{val}_1 \leftarrow \min\{\mathrm{val}_1, h^1(x_i)\}, \ldots, \mathrm{val}_\mathrm{k} \leftarrow \min\{\mathrm{val}_k, h^k(x_i)\}$

$\mathrm{val} \leftarrow \frac{1}{k}\sum_{i=1}^{k} \mathrm{val}_\mathrm{i}$

return $\mathrm{round}\left(\frac{1}{\mathrm{val}} - 1\right)$

$$\mathrm{Var}(\mathrm{val}) = \frac{1}{k}\frac{1}{(m+1)^2}$$

MinHash and Estimating # of Distinct Elements in Practice

- MinHash in practice:
  - One also stores the element that has the minimum hash value for each of the k hash functions
    - Then, just given separate MinHashes for sets A and B, can also estimate
      - what fraction of $A \cup B$ is in $A \cap B$; i.e., how similar A and B are

- Another randomized data structure for distinct elements in practice:
  - HyperLoglog - even more space efficient but doesn't have the set combination properties of MinHash