CSE 312

Foundations of Computing II

19: Recap polling + Law of Total Expectation

www.slido.com/2226110

Agenda

- Polling
- Odds and ends including Law of total expectation

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 $i = 1, \ldots, n$:

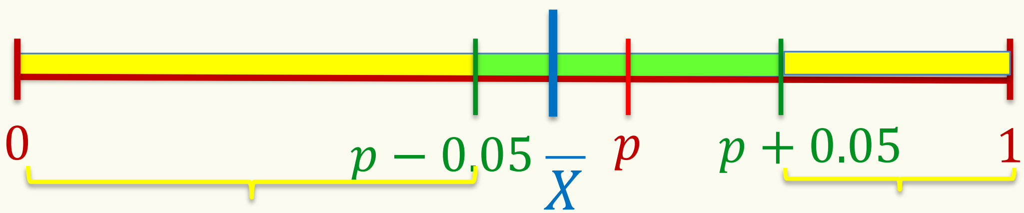
1. Pick uniformly random person to call (prob: $1/N$)
2. Ask them how they will vote

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

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

Roadmap: Bounding Error

Question: for what n is $P\left(\left|\overline{X} - p\right| > 0.05\right) \leq 0.02$



Crucial observation: the more samples we take, the more likely $\overline{X}$ is to be close to its expectation p since as $n \to \infty$,

By Central Limit Theorem $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$

Question: for what n is $P\left(\left|\overline{X}-p\right|>0.05\right)\leq 0.02$

By Central Limit Theorem $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i\rightarrow\mathcal{N}\left(p,\frac{p(1-p)}{n}\right)$

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

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

$\boxed{\text{Question: for what } n \text{ is } P\left(\left|\overline{X} - p\right| > 0.05\right) \le 0.02}$

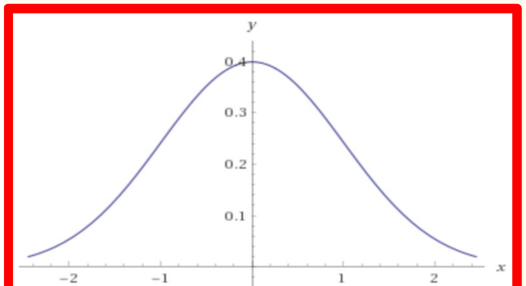
By Central Limit Theorem $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \to \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$

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

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

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

so $0.05\frac{\sqrt{n}}{\sqrt{p(1-p)}} \ge 2 \cdot 0.05\sqrt{n} \quad = 0.1\sqrt{n}$



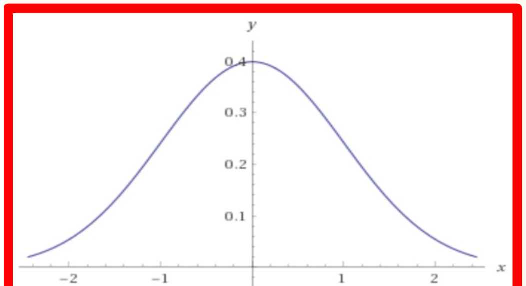
Question: for what n is $P\left(\left|\overline{X} - p\right| > 0.05\right) \leq 0.02$

By Central Limit Theorem $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \rightarrow \mathcal{N}\left(p, \frac{p(1-p)}{n}\right)$

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

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

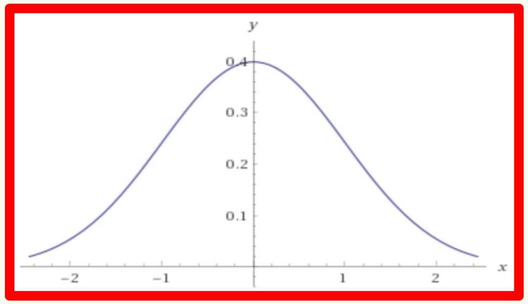
so $0.05\frac{\sqrt{n}}{\sqrt{p(1-p)}} \geq 2 \cdot 0.05\sqrt{n} = 0.1\sqrt{n}$



So $P\left(|Z| > 0.05\frac{\sqrt{n}}{\sqrt{p(1-p)}}\right) \leq P\left(|Z| > 0.1\sqrt{n}\right)$

Want to choose n so that this is at most 0.02

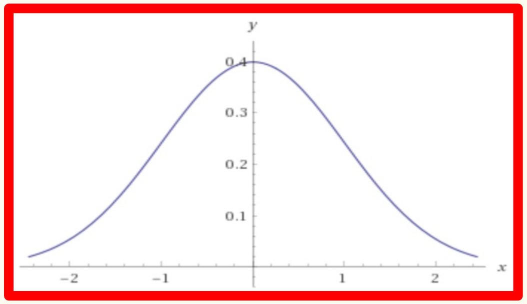
Solve for n such that $P(|Z| > 0.1\sqrt{n}) \leq 0.02$ *where* $Z \rightarrow \mathcal{N}(0,1)$



- This assumes n is large enough that $Z \sim \mathcal{N}(0,1)$

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

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



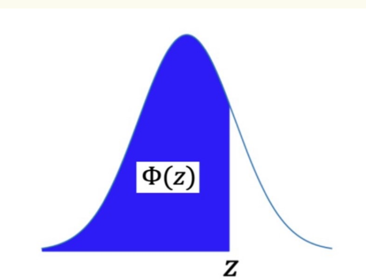
Or equivalently, choose n such that

$P(Z \leq 0.1\sqrt{n}) \geq 0.99$

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

Choose n so
$P(Z \leq 0.1\sqrt{n}) \geq 0.99.$
i.e.,
$\Phi(0.1\sqrt{n}) \geq 0.99$

From table $z = 2.33$ works



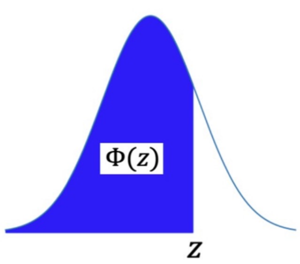
Φ Table: $\mathbb{P}(Z \leq z)$ when $Z \sim \mathcal{N}(0,1)$

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

Question: for what n is $P\left(\left|\overline{X}-p\right|>0.05\right)\leq 0.02$

Choose n so
$P(Z\leq 0.1\sqrt{n})\geq 0.99.$
i.e.,
$\Phi(0.1\sqrt{n})\geq 0.99$

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 $Z\sim\mathcal{N}(0,1)$
- Since we only have $Z\rightarrow\mathcal{N}(0,1)$ there is some loss due to approximation error (which can be dealt with).

Summary: We found an approximate``confidence interval”

We are trying to estimate some parameter (e.g. p). We output an estimator $\overline{X}$ such that $P\left(\left|\overline{X} - p\right| > \epsilon\right) \leq \delta$ for some (ϵ, δ).

- Often found using CLT, other approaches also important (especially when variance is unknown).
- We say that we are $(1 - \delta)$* 100% confident that the result of our poll ($\overline{X}$) is an accurate estimate of p to within ϵ* 100% percent.
- In our example, $(\epsilon = 0.05, \delta = 0.02)$.

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

- Polling
- Odds and ends, including Law of Total Expectation

Conditional Expectation

Definition. Let X be a discrete random variable then the **conditional expectation** of X given event A is

$$\mathbb{E}[X \mid A] = \sum_{x \in \Omega_X} x \cdot P(X = x \mid A)$$

Note:

- Linearity of expectation still applies here

$$\mathbb{E}[aX + bY + c \mid A] = a\, \mathbb{E}[X \mid A] + b\, \mathbb{E}[Y \mid A] + c$$

Law of Total Expectation

Law of Total Expectation (event version). Let X be a random variable and let events $A_1, \ldots, A_n$ partition the sample space. Then,

$$\mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}[X \mid A_i] \cdot P(A_i)$$

Law of Total Expectation (random variable version). Let X be a random variable and Y be a discrete random variable. Then,

$$\mathbb{E}[X] = \sum_{y \in \Omega_Y} \mathbb{E}[X \mid Y = y] \cdot P(Y = y)$$

Proof of Law of Total Expectation

Follows from Law of Total Probability and manipulating sums

$$\mathbb{E}[X] = \sum_{x \in \Omega_X} x \cdot P(X = x)$$

$$= \sum_{x \in \Omega_X} x \cdot \sum_{i=1}^{n} P(X = x \mid A_i) \cdot P(A_i)$$ (by LTP)

$$= \sum_{i=1}^{n} P(A_i) \sum_{x \in \Omega_X} x \cdot P(X = x \mid A_i)$$ (change order of sums)

$$= \sum_{i=1}^{n} P(A_i) \cdot \mathbb{E}[X \mid A_i]$$ (def of cond. expect.)

Example – Flipping a Random Number of Coins

Suppose someone gave us $Y \sim \mathrm{Poi}(5)$ fair coins and we wanted to compute the expected number of heads X from flipping those coins.

By the Law of Total Expectation

$$\mathbb{E}[X] = \sum_{i=0}^{\infty} \mathbb{E}[X \mid Y = i] \cdot P(Y = i) =$$

Example – Flipping a Random Number of Coins

Suppose someone gave us $Y \sim \mathrm{Poi}(5)$ fair coins and we wanted to compute the expected number of heads X from flipping those coins.

By the Law of Total Expectation

$$\mathbb{E}[X] = \sum_{i=0}^{\infty} \mathbb{E}[X \mid Y = i] \cdot P(Y = i) = \sum_{i=0}^{\infty} \frac{i}{2} \cdot P(Y = i)$$

$$= \frac{1}{2} \cdot \sum_{i=0}^{\infty} i \cdot P(Y = i)$$

$$= \frac{1}{2} \cdot \mathbb{E}[Y] = \frac{1}{2} \cdot 5 = 2.5$$

Example -- Elevator rides

The number X of people who enter an elevator on the ground floor is a Poisson random variable with mean 10. If there are N floors above the ground floor, and if each person is equally likely to get off at any one of the N floors, independently of where others get off, compute the expected number of stops the elevator will make before discharging all the passengers.

Law of total probability for continuous random variables.

Definition. Let A be an event and Y a continuous random variable. Then

$$P[A] = \int_{-\infty}^{\infty} P(A|Y = y) f_Y(y) \mathrm{d}y$$

Example use of law of total probability

Suppose that the time until server 1 crashes is $X \sim Exp\ (\lambda)$ and the time until server 2 crashes is independent, with $Y \sim Exp\ (\mu)$.

What is the probability that server 1 crashes before server 2?

Example use of law of total probability

$X \sim Exp\ (\lambda), Y \sim Exp\ (\mu)$.

What is the probability that $X < Y$?

$$P(X < Y) = \int_0^\infty \Pr(X < Y \mid X = x)\ f_X(x) dx$$

$$= \int_0^\infty \Pr(Y > X \mid X = x) \lambda e^{-\lambda x}\ \ dx$$

$$= \int_0^\infty \Pr(Y > x \mid X = x)\ \ \lambda e^{-\lambda x}\ \ dx$$

$$= \int_0^\infty \Pr(Y > x) \lambda e^{-\lambda x}\ \ dx$$

$$= \int_0^\infty e^{-\mu x}\ \lambda\ e^{-\lambda x}\ \ dx$$

$$= \frac{\lambda}{\lambda + \mu} \int_0^\infty (\lambda + \mu) \cdot e^{-\mu x}\ e^{-\lambda x}\ dx$$

$$= \frac{\lambda}{\lambda + \mu}$$

Alternative approach

$X \sim Exp\ (\lambda), Y \sim Exp\ (\mu)$.

What is the probability that $X < Y$?

$$P(X < Y) = \int_{x=0}^{\infty} \int_{y=x}^{\infty} f_{X,Y}(x, y) \mathrm{dy}\ \mathrm{d}x$$

$$= \int_{x=0}^{\infty} \int_{y=x}^{\infty} f_X(x) \cdot f_Y(y) \mathrm{dy}\ \mathrm{d}x$$

Covariance: How correlated are X and Y?

Recall that if X and Y are independent, $\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$

Definition: The **covariance** of random variables X and Y,

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X] \cdot \mathbb{E}[Y]$$

Unlike variance, covariance can be positive or negative. It has has value 0 if the random variables are independent.

$\mathrm{Cov}(X, X) = ?$

$$\mathrm{Cov}(X,Y) = \mathbb{E}[XY] - \mathbb{E}[X]\cdot\mathbb{E}[Y]$$

Two Covariance examples:

Suppose $X \sim \mathrm{Bernoulli}(p)$

If random variable $Y = X$ then

$$\mathrm{Cov}(X,Y) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \mathrm{Var}(X) = p(1-p)$$

If random variable $Z = -X$ then

$$\begin{aligned}\mathrm{Cov}(X,Z) &= \mathbb{E}[XZ] - \mathbb{E}[X]\cdot\mathbb{E}[Z]\\ &= \mathbb{E}[-X^2] - \mathbb{E}[X]\cdot\mathbb{E}[-X]\\ &= -\mathbb{E}[X^2] + \mathbb{E}[X]^2 = -\mathrm{Var}(X) = -p(1-p)\end{aligned}$$

Reference Sheet (with continuous RVs)

	Discrete	Continuous
Joint PMF/PDF	$p_{X,Y}(x,y) = P(X = x, Y = y)$	$f_{X,Y}(x,y) \neq P(X = x, Y = y)$
Joint CDF	$F_{X,Y}(x,y) = \sum_{t \leq x} \sum_{s \leq y} p_{X,Y}(t,s)$	$F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(t,s) ds dt$
Normalization	$\sum_{x} \sum_{y} p_{X,Y}(x,y) = 1$	$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx dy = 1$
Marginal PMF/PDF	$p_X(x) = \sum_{y} p_{X,Y}(x,y)$	$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy$
Expectation	$E[g(X,Y)] = \sum_{x} \sum_{y} g(x,y) p_{X,Y}(x,y)$	$E[g(X,Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y) dx dy$
Conditional PMF/PDF	$p_{X \mid Y}(x \mid y) = \frac{p_{X,Y}(x,y)}{p_Y(y)}$	$f_{X \mid Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}$
Conditional Expectation	$E[X \mid Y = y] = \sum_{x} x p_{X \mid Y}(x \mid y)$	$E[X \mid Y = y] = \int_{-\infty}^{\infty} x f_{X \mid Y}(x \mid y) dx$
Independence	$\forall x, y, p_{X,Y}(x,y) = p_X(x) p_Y(y)$	$\forall x, y, f_{X,Y}(x,y) = f_X(x) f_Y(y)$