CSE 312

Foundations of Computing II

Lecture 17: CLT & Polling

Review CDF of normal distribution

Fact. If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $Y = aX + b \sim \mathcal{N}(a\mu + b, a^2\sigma^2)$

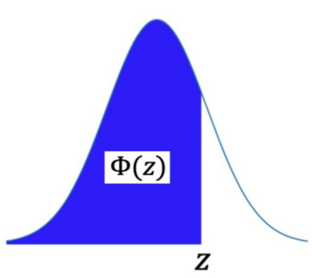
Standard (unit) normal = $\mathcal{N}(0, 1)$

CDF. $\Phi(z) = P(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} \mathrm{d}x$ for $Z \sim \mathcal{N}(0, 1)$

Note: $\Phi(z)$ has no closed form – generally given via tables

Review

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



Φ Table: $\mathbb{P}(Z \le z)$ 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

Review Analyzing non-standard normal in terms of $\mathcal{N}(0, 1)$

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then $\frac{X-\mu}{\sigma} \sim \mathcal{N}(0, 1)$

Therefore,

$$F_X(z) = P(X \leq z) = P\left(\frac{X-\mu}{\sigma} \leq \frac{z-\mu}{\sigma}\right) = \Phi\left(\frac{z-\mu}{\sigma}\right)$$

Review How Many Standard Deviations Away?

within k standard deviation of mean

Let $X \sim \mathcal{N}(\mu, \sigma^2)$.

$$P(|X-\mu| < k\sigma) = P\left(\frac{|X-\mu|}{\sigma} < k\right) =$$

$$= P\left(-k < \frac{X-\mu}{\sigma} < k\right) = \Phi(k) - \Phi(-k)$$

e.g. $k = 1$: 68%

$k = 2$: 95%

$k = 3$: 99%

Review Central Limit Theorem

$X_1, \dots, X_n$ i.i.d., each with expectation μ and variance σ^2

Define $S_n = X_1 + \cdots + X_n$ and

$$Y_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

$$\mathbb{E}[Y_n] = \frac{1}{\sigma\sqrt{n}}(\mathbb{E}[S_n] - n\mu) = \frac{1}{\sigma\sqrt{n}}(n\mu - n\mu) = 0$$

$$\mathrm{Var}(Y_n) = \frac{1}{\sigma^2 n}\big(\mathrm{Var}(S_n - n\mu)\big) = \frac{\mathrm{Var}(S_n)}{\sigma^2 n} = \frac{\sigma^2 n}{\sigma^2 n} = 1$$

Review Central Limit Theorem

$\mathrm{Var}(aZ) = a^2\,\mathrm{Var}(Z)$

$\bar{X} = \frac{X_1 + \cdots + X_n}{n}$

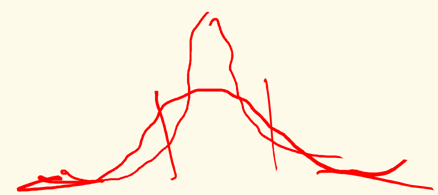
Total Variance $n\sigma^2$

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

$\frac{1}{n^2}$

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

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



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 $\sigma^2 = \mathrm{Var}(X_i)$

Agenda

- Central Limit Theorem (CLT) Review
- Polling

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 $i = 1, \dots, n$:

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

What type of r.v. is X_i?

Poll: pollev.com/paulbeame028

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

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

Problem: We don't know p

Polling Procedure

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

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: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

Random Variables

What type of r.v. is X_i?

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

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \underbrace{\sum_{i=1}^{n} \text{Var}(X_i)}_{\text{indep}} = p(1-p)\,n$$

$$\text{Var}(\bar{X}) = \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{p(1-p)\,n}{n^2} =$$

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

$$\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = np$$

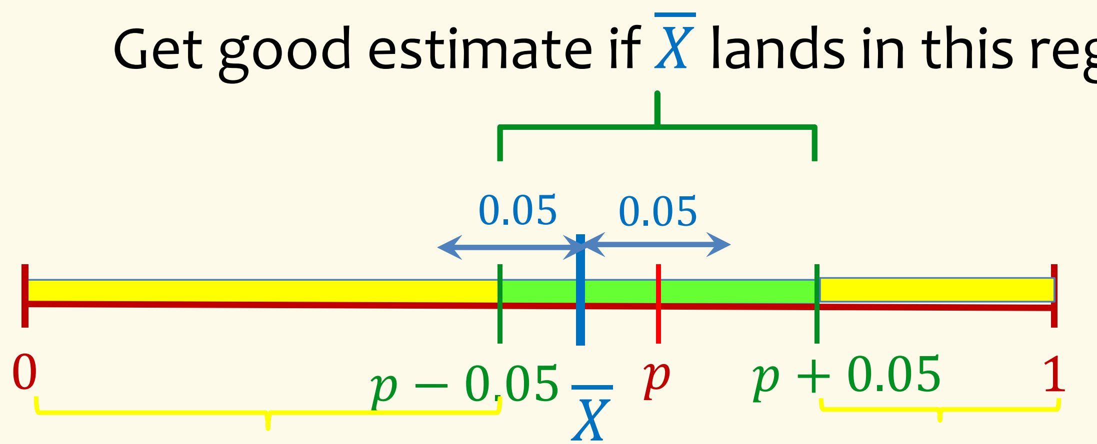
$$\mathbb{E}[\bar{X}] = \frac{\mathbb{E}\left(\sum_{i=1}^{n} X_i\right)}{n} = \frac{np}{n} = p$$

Poll: pollev.com/paulbeame028

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

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



Want $\underbrace{P\left(|\bar{X} - p| > 0.05\right)}$ ≤ 0.02 Bad Case

Central Limit Theorem

pollev.com/paulbeame028

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 $X_1, X_2, \dots, X_n$ where $\mathbb{E}[X_i] = \mu$ and $\mathrm{Var}(X_i) = \sigma^2$

As $n \to \infty$,

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

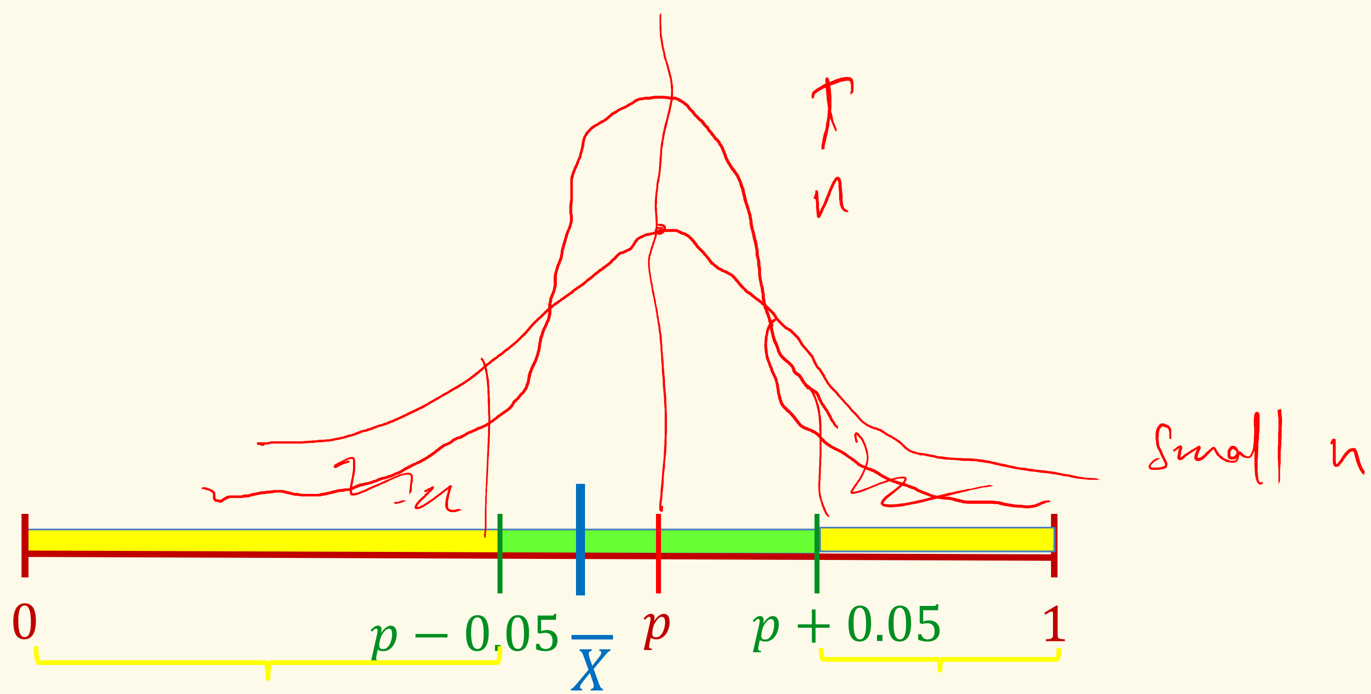
Restated: As $n \to \infty$,

p

p(1-p)

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

Roadmap: Bounding Error



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

Roadmap: Bounding Error

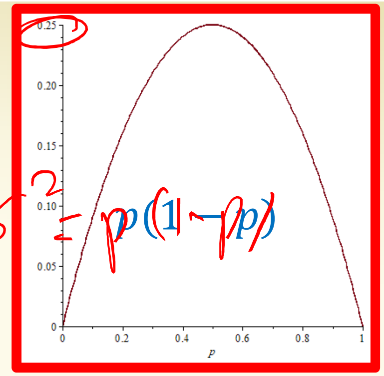
Goal: Find the value of n such that 98% of the time, the estimate $\bar{X}$ is within 5% of the true p

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

Following the Road Map

$0.25 = \frac{1}{4}$

$\sqrt{p(1-p)} \le \frac{1}{2}$ σ



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

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 ≥ 2

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

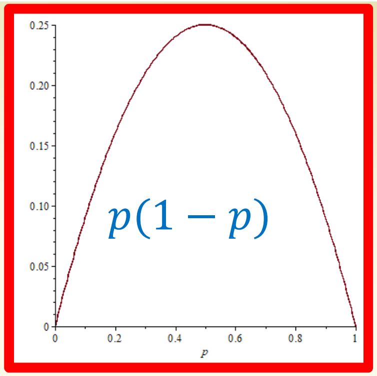
$$= P(|Z| > 0.05/\sigma) = P\left(|Z| > 0.05 \frac{\sqrt{n}}{\sqrt{p(1-p)}}\right)$$

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

$\ge 0.1\sqrt{n}$

Q: Why "≤"?
A: This condition on Z
is easier to satisfy

Following the Road Map



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

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 $\geq \mathbf{2}$

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

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

Want to choose n so that this is at most 0.02

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

4. Solve for n

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

- 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

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

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

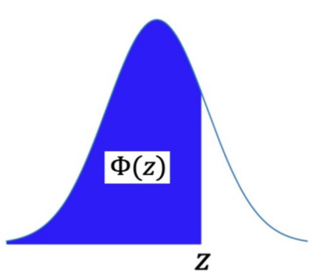
Table of $\Phi(z)$ CDF of Standard Normal Distribution

Choose n so

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

$\Phi(z) \geq 0.99$

From table $z = 2.33$ works



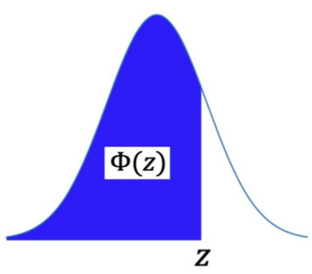
Φ Table: $\mathbb{P}(Z \leq z)$ when $Z \sim \mathcal{N}(0,1)$

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
$\Phi(z) \geq 0.99$

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 $Z \to \mathcal{N}(0,1)$ 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!