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.** $\Phi(z) = P(Z \leq z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2} dx$ 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
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?

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

$X_1, \ldots, X_n$ i.i.d., each with expectation $\mu$ and variance $\sigma^2$

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

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

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

$$\text{Var}(Y_n) = \frac{1}{\sigma^2 n} (\text{Var}(S_n - n\mu)) = \frac{\text{Var}(S_n)}{\sigma^2 n} = \frac{\sigma^2 n}{\sigma^2 n} = 1$$
Theorem. (Central Limit Theorem) The CDF of $Y_n$ converges to the CDF of the standard normal $\mathcal{N}(0,1)$, i.e.,

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

Also stated as:

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

“We have 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 $p$

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 $p$:

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

What type of r.v. is $X_i$?

<table>
<thead>
<tr>
<th>Type</th>
<th>$\mathbb{E}[X_i]$</th>
<th>$\text{Var}(X_i)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. Bernoulli</td>
<td>$p$</td>
<td>$p(1 - p)$</td>
</tr>
<tr>
<td>b. Bernoulli</td>
<td>$p$</td>
<td>$p^2$</td>
</tr>
<tr>
<td>c. Geometric</td>
<td>$p$</td>
<td>$\frac{1-p}{p^2}$</td>
</tr>
<tr>
<td>d. Binomial</td>
<td>$np$</td>
<td>$np(1 - p)$</td>
</tr>
</tbody>
</table>

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

Poll: [pollev.com/paulbeame028](http://pollev.com/paulbeame028)

<table>
<thead>
<tr>
<th>$\mathbb{E}[\bar{X}]$</th>
<th>$\text{Var}(\bar{X})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>a. $np$</td>
<td>$np(1 - p)$</td>
</tr>
<tr>
<td>b. $p$</td>
<td>$p(1 - p)$</td>
</tr>
<tr>
<td>c. $p$</td>
<td>$p(1-p)/n$</td>
</tr>
<tr>
<td>d. $p/n$</td>
<td>$p(1-p)/n$</td>
</tr>
</tbody>
</table>
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 $P(|\bar{X} - p| > 0.05) \leq 0.02$
Central Limit Theorem

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

As $n \to \infty$,

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

Restated: As $n \to \infty$,

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

Poll: In the limit $\bar{X}$ 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
Roadmap: Bounding Error

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

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

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

\[
P(|\bar{X} - p| > 0.05) = P(|Z| \cdot \sigma > 0.05) = P(|Z| > 0.05/\sigma) = P(|Z| > 0.05 \sqrt{n} / p(1-p)) \\
\leq P(|Z| > 0.1\sqrt{n})
\]

Q: Why “\( \leq \)”?
A: This condition on \( Z \) is easier to satisfy
Following the Road Map

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

2. By CLT $\overline{X} \rightarrow \mathcal{N}(\mu, \sigma^2)$ 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 $Z \rightarrow \mathcal{N}(0, 1)$

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

Want to choose $n$ so that this is at most 0.02

\[
\leq P\left( |Z| > 0.1\sqrt{n} \right)
\]

\[
= \frac{1}{\sqrt{p(1-p)}} \text{ is always } \geq 2
\]
4. Solve for $n$

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

• 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
4. Solve for $n$

Choose $n$ so

$$0.1\sqrt{n} \geq z \quad \text{where} \quad \Phi(z) \geq 0.99$$

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!