Normal Approximation of Binomial

Theorem $Y \sim \text{Binom}(n, p)$, then $\frac{Y - np}{\sqrt{np(1-p)}} \xrightarrow[n\to\infty]{} Z \sim N(0,1)$ in distribution.

(DeMoivre – Laplace Theorem)

Special case of Central Limit Theorem

Prof: $Y = X_1 + X_2 + \cdots + X_n$, X_i i.i.d Bernoulli(p) $X_i = \begin{cases} 1 & \text{w/ prob } p \\ 0 & \text{w/ prob } 1-p \end{cases}$

$\mathbb{E}Y = n\mathbb{E}X_1 = np$

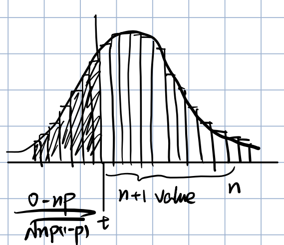
$\text{Var}\, Y = n \text{Var} X_1 = np(1-p)$

CLT says $\underbrace{\frac{Y - \mathbb{E}Y}{\sqrt{\text{Var}\, Y}}}_{\text{"standardized"}} \to Z \sim N(0,1)$ $\text{Var} X_1 = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \mathbb{E}X_1 - (\mathbb{E}X_1)^2 = p - p^2 = p(1-p)$

Y_n takes value $\{0, 1, 2, \ldots, n\}$

Z_n takes value $\left\{\frac{0-np}{\sqrt{np(1-p)}}, \frac{1-np}{\sqrt{np(1-p)}}, \ldots, \frac{n-np}{\sqrt{np(1-p)}}\right\}$ $n+1$ possible values.

$Z \sim N(0,1)$ takes any value in the set of real numbers.



convergence in distribution means

$F_{Z_n}(t) \to F_Z(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$

CDF converges.

$\frac{Y_n - np}{\sqrt{np(1-p)}} \approx N(0,1)$

$Y_n - np \approx \sqrt{np(1-p)} \cdot N(0,1)$

$Y_n \approx np + \sqrt{np(1-p)} \cdot N(0,1) = N(np, np(1-p))$

$F_{Y_n}(t) = \sum_{k \le t} \binom{n}{k} p^k (1-p)^{n-k} \approx \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi\, np(1-p)}} e^{-\frac{(t-np)^2}{2np(1-p)}}$

· Take away

Approximate Binom(n, p) by $N(np, np(1-p))$ for large n (e.g. $n \ge 20$)

10/31/2022

Confidence Interval for percentiles (distribution free)

X is continuous RV

$m(X) = \min\{t : F_X(t) = \frac{1}{2}\}$
median

$\pi_p(X) = \min\{t : F_X(t) = p\}$
pth quantile / $100p\%$ percentile.

$p = \frac{1}{2} = 0.5$ 50% percentile = median
$p = \frac{1}{4} = 0.25$ 25% percentile = 1st quantile
$p = \frac{3}{4} = 0.75$ 75% percentile = 2nd quantile.

$X_1, X_2, \dots, X_5$ $\quad Y_1 < Y_2 < \dots < Y_5$ $\quad$ estimate $m(X)$ $(p = \frac{1}{2})$
estimate $\pi_p(X)$

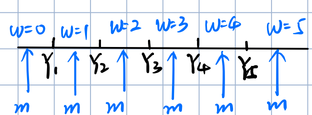
$(n+1)\cdot\frac{1}{2} = (5+1)\cdot\frac{1}{2} = 3$
Y_3 = sample median.

sample percentile $\tilde{\pi}_p = y_k + \delta(y_{k+1} - y_k)$ if $(n+1)p = k + \delta$ integer $[0,1)$.

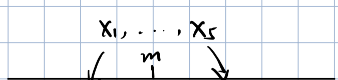
Given CI for $m(X)$, $\pi_p(X)$

Simple Idea: just use (Y_1, Y_5)
$\mathbb{P}(m(X) \in (Y_1, Y_5)) = \ ? \ = 1 - \alpha$.

Given: $\mathbb{P}(Y_1 < m \text{ and } Y_5 > m) = 1 - \mathbb{P}(Y_1 > m) - \mathbb{P}(Y_5 > m)$ $\quad X_1, X_2, X_3, X_4, X_5$



- $Y_1 < m$ means at least one of X_i's $< m$ $\quad Y_1 = \min(X_1, \dots, X_5)$
$Y_5 > m$ means at least one of X_i's $> m$ $\quad Y_5 = \max(X_1, \dots, X_5)$
- W = number of X_i's that are $< m$
$= \sum_{i=1}^{n} 1\{X_i < m\} = \text{Binom}(n, \frac{1}{2})$
i.i.d. sum of Bernoulli $(\frac{1}{2})$



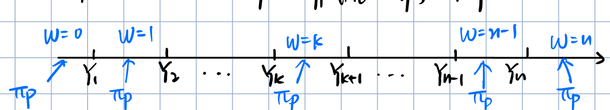
"success": $X_i < m$ $\quad \mathbb{P}(X_i < m) = \frac{1}{2}$
"failure": $X_i > m$ $\quad \mathbb{P}(X_i > m) = 1 - \frac{1}{2} = \frac{1}{2}$

$\Rightarrow W = k \iff Y_k < m < Y_{k+1}$

$\mathbb{P}(m \in (Y_1, Y_5)) = \sum_{k=1}^{4} \mathbb{P}(m \in (Y_k, Y_{k+1}))$
$= \sum_{k=1}^{4} \mathbb{P}(W = k)$
$= \mathbb{P}(1 \le W \le 4)$
$= 1 - \mathbb{P}(W=1) - \mathbb{P}(W=4) = 1 - \frac{1}{2^5} - \frac{1}{2^5}$

Generalize: $X_1, X_2, \dots, X_n$ $\quad Y_1 < Y_2 < \dots < Y_n$
want to find CI for π_p
If use (Y_i, Y_j) as my CI, $\mathbb{P}(\pi_p \in (Y_i, Y_j)) = 1 - \alpha = ?$
$W = \#\, X_i$'s that are $< \pi_p$ $= \text{Binom}(n, p) \approx \mathcal{N}(np, np(1-p))$

"success" = $X_i < \pi_p$ $\quad \mathbb{P}(X_i < \pi_p) = p$
"failure" = $X_i > \pi_p$ $\quad \mathbb{P}(X_i > \pi_p) = 1 - p$



$\Rightarrow W = k \Longleftrightarrow Y_k < \pi_p < Y_{k+1}$

$\mathbb{P}(\pi_p \in (Y_i, Y_j)) = 1 - \alpha$

$= \sum_{k=i}^{j-1} \mathbb{P}(\pi_p \in (Y_k, Y_{k+1}))$

$= \sum_{k=i}^{j-1} \mathbb{P}(W = k) = \mathbb{P}(i \le W \le j-1)$

$= \mathbb{P}(i - 0.5 \le W \le j - 0.5) \approx \mathbb{P}(i - 0.5 \le N(np, np(1-p)) \le j - 0.5)$

$= \mathbb{P}\left(\frac{i - 0.5 - np}{\sqrt{np(1-p)}} \le N(0,1) \le \frac{j - 0.5 - np}{\sqrt{np(1-p)}}\right)$

Example: $n = 27$ samples want CI for $\pi_{0.25}$

Let's compute $\hat{\pi}_{0.25} = y_7$

$(n+1)p = 28 \times 0.25 = 28 \times \frac{1}{4} = 7$

One reasonable choice for CI is (y_4, y_{10})

$np = 27 \times \frac{1}{4} = 6.75$

$\sqrt{np(1-p)} = \sqrt{27 \times \frac{1}{4} \times \frac{3}{4}} = \sqrt{\frac{81}{16}} = \frac{9}{4} = 2.25$.

$\mathbb{P}(\pi_{0.25} \in (Y_4, Y_{10})) = \mathbb{P}(4 \le W \le 9) = \mathbb{P}(4 - 0.5 \le W \le 10 - 0.5))$

$\approx \mathbb{P}\left(\frac{4 - 0.5 - 6.75}{2.25} \le Z \sim N(0,1) \le \frac{10 - 0.5 - 6.75}{2.25}\right)$

Hypthesis Testing

Suppose we're interested in an RV $X \sim N(\mu, 36)$. Based on external information, We've chosen to two competing hypothesis.

- Null hypothesis $H_1 : \mu = 50$
- Alternative hypothesis $H_1 : \mu = 55$

How can we test which one is more likely to be correct?

$H_0 : \mu = 50$ or $H_1 : \mu = 55$ $X_1, \dots, X_n$ i.i.d $N(\mu, 36) \xrightarrow{\text{experiment}}$ sample $x_1, x_2, \dots, x_n$ $\bar{x} = \frac{1}{n}(x_1 + x_2 + \dots + x_n)$

Intuitively larger $\bar{X}$ favor H_1 over H_0

Set up a "rejection threshold" $\mu_* = 53$.

Test: Reject H_0 (in favor of H_1) if $\bar{x} \geq \mu_* = 53$.
otherwise we "accept" (do not reject) H_0

This an example of a test for the sample null hypothesis $H_0 : \mu = 50$ against the simple alternative hypothesis $H_1 : \mu = 55$

The set of outcomes

$C := \{(x_1, x_2, \dots, x_n) : \bar{x} \geq \mu_* = 53\}$ is the critical region for this test.

It's specified by the test statistic $\bar{x}$.

$\alpha := \mathbb{P}(\text{type I error}) = \mathbb{P}((x_1, \dots, x_n) \in C ; H_0) = \mathbb{P}(\bar{X} \geq 53, \mu = 50)$

$\beta := \mathbb{P}(\text{type II error}) = \mathbb{P}((x_1, \dots, x_n) \notin C ; H_1) = \mathbb{P}(\bar{X} < 53, \mu = 55)$

	H_0 is true	H_1 is true
$\bar{x} \geq 53$	incorrectly reject H_0 "type I error"	correctly reject H_0
$\bar{x} < 53$	correctly accept H_0	incorrectly accept H_0 "type II error"

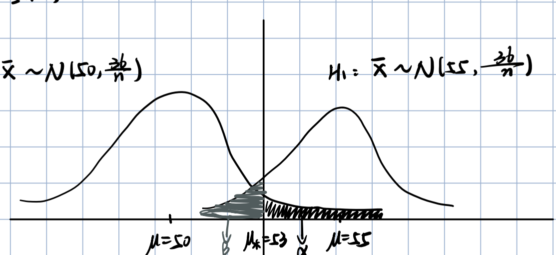
α is the "significance level" of this test.

- To compute these probabilities α, β, we need to know the distribution of the test static $\bar{X}$ under the hypothesis H_0, H_1 respectively.
- $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n) \sim N(\mu, \frac{36}{n})$

$$\frac{\bar{X} - \mu}{\sqrt{36/n}} \sim N(0,1) = Z$$

$$\begin{aligned}
\alpha &= \mathbb{P}(\bar{X} \geq 53 ; \mu = 50) \\
&= \mathbb{P}\left(\frac{\bar{X} - \mu}{\sqrt{36/n}} \geq \frac{53 - \mu}{\sqrt{36/n}} ; \mu = 50\right) \\
&= \mathbb{P}\left(Z \geq \frac{53 - 50}{\sqrt{36/n}}\right) \\
&= \mathbb{P}\left(Z \geq \frac{3}{6/\sqrt{n}}\right) \\
&= \mathbb{P}\left(Z \geq \tfrac{1}{2}\sqrt{n}\right)
\end{aligned}$$

$$\begin{aligned}
\beta &= \mathbb{P}(\bar{X} < 53 ; \mu = 55) \\
&= \mathbb{P}\left(\frac{\bar{X} - \mu}{\sqrt{36/n}} < \frac{53 - \mu}{\sqrt{36/n}} ; \mu = 55\right) \\
&= \mathbb{P}\left(\frac{\bar{X} - \mu}{\sqrt{36/n}} < \frac{53 - 55}{\sqrt{36/n}}\right) \\
&= \mathbb{P}\left(\frac{\bar{X} - \mu}{\sqrt{36/n}} < \frac{-2}{6/\sqrt{n}}\right) \\
&= \mathbb{P}\left(\frac{\bar{X} - \mu}{\sqrt{36/n}} < -\tfrac{1}{3}\sqrt{n}\right)
\end{aligned}$$

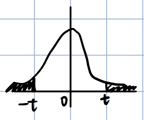


- Fixed sample size n, if we move up μ_*, then decrease α at the cost of increasing β.
- Fixed μ_*, if we increase sample size n, then α, β both decrease.
  In fact, $\alpha, \beta \longrightarrow 0$ exponentially fast.

Tail bound for standard normal $Z \sim N(0,1)$

Then $\mathbb{P}(Z > t) \le \frac{1}{\sqrt{2\pi}} \frac{e^{-\frac{t^2}{2}}}{t} \le \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$ for $t \ge 1$

$\int_t^\infty \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$



$\sim Z \sim N(0,1)$

$\alpha = \mathbb{P}(Z \ge \frac{1}{2}\sqrt{n})$

$\le \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{1}{2}\sqrt{n})^2}$

$= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{8}n} \xrightarrow[n\to\infty]{} 0$

exponentially decaying

$\beta = \mathbb{P}(Z < -\frac{1}{2}\sqrt{n})$

$= \mathbb{P}(-Z > \frac{1}{2}\sqrt{n})$

$\le \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{1}{2}\sqrt{n})^2}$

$= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{8}n} \xrightarrow[n\to\infty]{} 0$

exponentially decay.

General procedure for testing against simple alternative. $X \sim N(\mu, \sigma^2)$ σ known

$H_0 : \mu = \mu_0$

$H_1 : \mu = \mu_1$ $(\mu_1 \ne \mu_0)$

$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n) \sim N(\mu, \frac{\sigma^2}{n})$

Say $\mu_1 > \mu_0$

Choose some rejection threshold $\mu_* \in (\mu_0, \mu_1)$

Test: reject H_0 (in favor of H_1) if $\bar{X} \ge \mu_*$ otherwise don't reject / accept H_0.

Critical region. $C = \{(x_1, \dots, x_n) : \bar{x} \ge \mu_*\}$

significance level

$\alpha = \mathbb{P}(\text{type I region})$

$= \mathbb{P}(\bar{X} \ge \mu_*, \mu = \mu_0)$

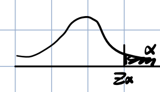
$= \mathbb{P}(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \ge \frac{\mu_*-\mu}{\sigma/\sqrt{n}} ; \mu = \mu_0)$

$= \mathbb{P}(Z \ge \frac{\mu_*-\mu_0}{\sigma/\sqrt{n}})$

Equivalently, $\frac{\mu_*-\mu_0}{\sigma/\sqrt{n}} = z_\alpha$

$\mu_* - \mu_0 = z_\alpha \cdot \frac{\sigma}{\sqrt{n}}$

$\mu_* = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$



This is rejection threshold to achieve significance level α.

For given significance level α, rejection threshold $\mu_* = \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$

$\beta = \mathbb{P}(\text{type II error})$

$= \mathbb{P}(\bar{X} < \mu_*, \mu = \mu_1)$

$= \mathbb{P}(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < \frac{\mu_*-\mu_1}{\sigma/\sqrt{n}} ; \mu = \mu_1)$

$= \mathbb{P}(Z < \frac{\mu_*-\mu_1}{\sigma/\sqrt{n}})$

$= \mathbb{P}(Z < \frac{\mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}} - \mu_1}{\sigma/\sqrt{n}})$

$= \mathbb{P}(Z < z_\alpha + \frac{\mu_0-\mu_1}{\sigma/\sqrt{n}})$

$= \mathbb{P}(Z < z_\alpha - \frac{\mu_1-\mu_0}{\sigma/\sqrt{n}})$

The quantity $\mu_1 - \mu_0$ is "effect size" $\theta := \frac{\mu_1-\mu_0}{\sigma}$ is "standardized effect size"
> 0

$= \mathbb{P}(Z < z_\alpha - \frac{\theta}{1/\sqrt{n}})$

$= \mathbb{P}(Z < z_\alpha - \sqrt{n}\theta)$

For n large enough. $z_\alpha \le \sqrt{n} \cdot \theta$

$= \mathbb{P}(-Z > \sqrt{n} \cdot \theta - z_\alpha)$

$\le e^{-\frac{1}{2}(\sqrt{n} \cdot \theta - z_\alpha)^2}$

$= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}[\sqrt{n}(\theta - \frac{z_\alpha}{\sqrt{n}})]^2}$

$= \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}n(\theta - \frac{z_\alpha}{\sqrt{n}})^2}$

$\approx \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}n \cdot \theta^2}$

Recall Fact: $\mathbb{P}(Z > t) \le \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$

$\sqrt{n} \cdot \theta - z_\alpha = \sqrt{n}(\theta - \frac{z_\alpha}{\sqrt{n}}) \approx \sqrt{n} \cdot \theta$