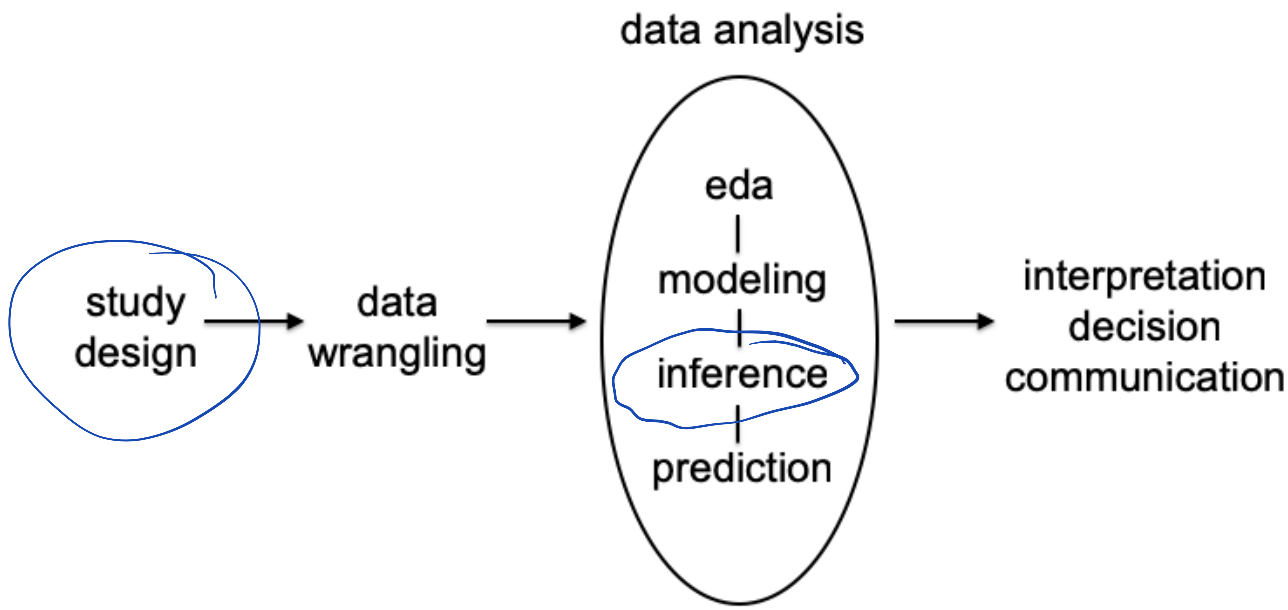
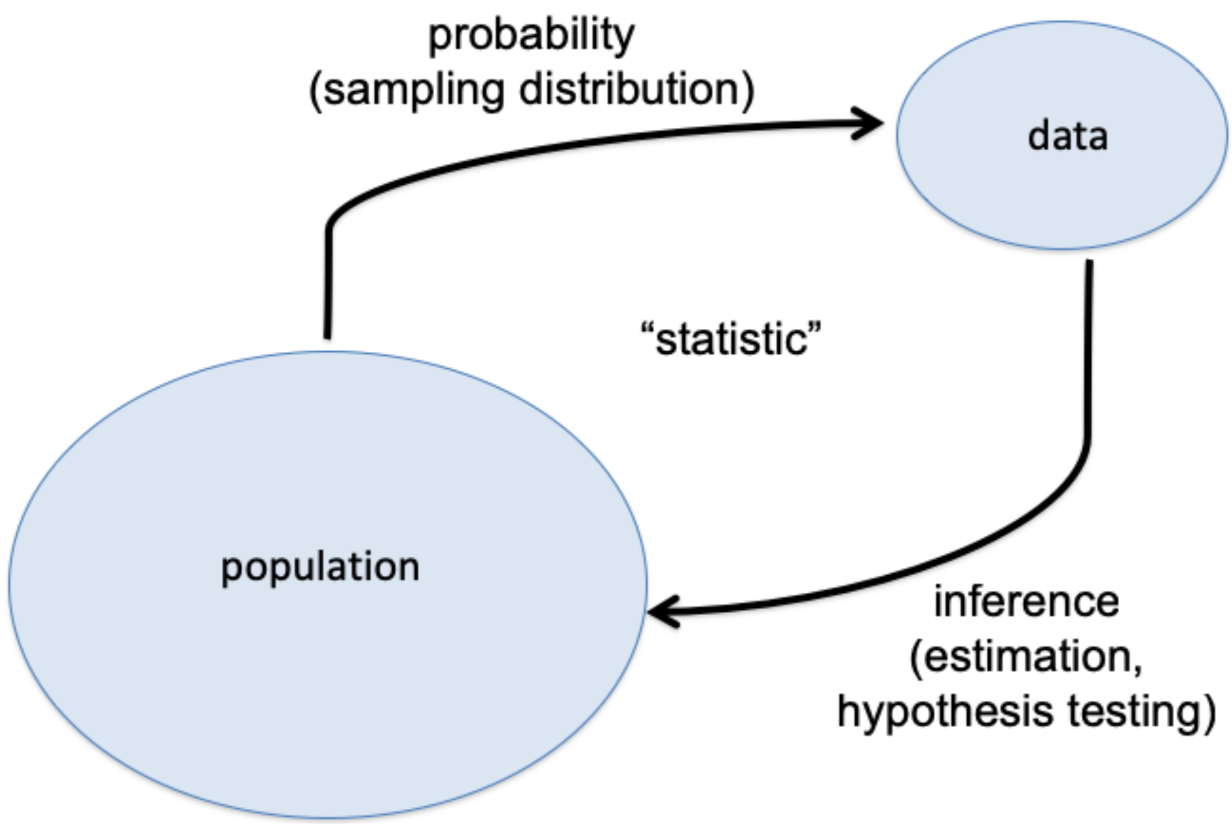
This Week's Topic: Frequentist Statistical Inference



Central Dogma of (Frequentist) Statistical Inference



Observing/ data according to a
Collecting/ probabilistic model
Generating

<u>Example</u>: Simple Random Sample

Units are uniformly and independently sampled from a population:

- Political survey (survey samples)
- Flipping a coin n times

<u>Example</u>: Randomized controlled study

Units are randomly sampled and then randomized to two or more treatment groups

- Clinical trial

Example: Fair Coin?

Suppose I claim that a specific coin is fair, i.e., that it lands on heads or tails with equal probability.

I flip it 20 times and it lands on heads 16 times.

$\hat{p}$ $\tilde{p}$

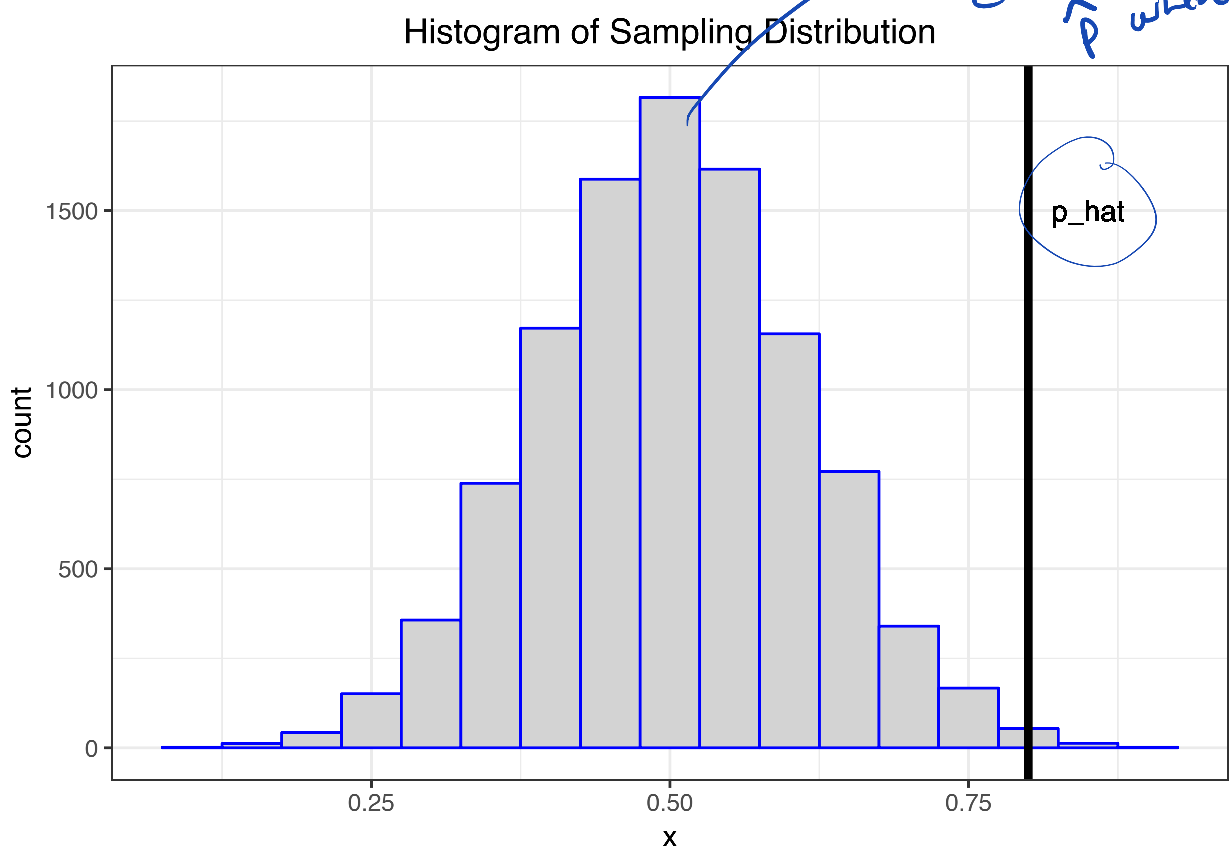
1. My data is $x = 16$ heads out of $n = 20$ flips.
2. My data generation model is $X \sim \text{Binomial}(20, p)$.
3. I form the statistic $\hat{p} = 16/20$ as an estimate of p.

Let's simulate 10,000 times what my estimate would look like if $p = 0.5$ and I repeated the 20 coin flips over and over.

```
> x <- replicate(n=1e4, expr=rbinom(1, size=20, prob=0.5))
> sim_p_hat <- x/20
> my_p_hat <- 16/20
```

sampling distribution of $\hat{p}$ when $p=0.5$

What can I do with this information?




```
> sum(abs(sim_p_hat-0.5) >= abs(my_p_hat-0.5))/1e4
[1] 0.0083
```

__Parameter__ a number that describes a population.

- Usually fixed
- We don't know its value
- appears in the probability model of how we collected data

__Statistic__ a number calculated from a sample of data

A statistic is used to estimate a parameter

__Sampling distribution__ of a statistic is the probability dist'n of the statistic under repeated realizations of the data from the assumed data generating probability dist'n

Goals of Inference

① Form estimates of the parameters

② Quantify uncertainty about the estimates

③ Test hypotheses on the parameters

Let's go through these goals in simple scenario:

Data generating process is

$X_1, X_2, \ldots, X_n \overset{iid}{\sim} \text{Normal}(\mu, \sigma^2)$

μ is unknown

σ^2 is known

① $\sum_{i=1}^{n} X_i \sim \text{Normal}(n\mu, n\sigma^2)$

(2) If $Z \sim$ Normal then

$a + bZ \sim$ Normal $\Longrightarrow$

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \sim \text{Normal}(\mu, \sigma^2/n)$$

Exercise: verify this

(3) $$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim \text{Normal}(0, 1)$$

Point estimate of μ:

$$\hat{\mu} = \bar{X}$$

$$\underbrace{\hat{\mu} \sim \text{Normal}(\mu, \sigma^2/n)}_{\text{sampling distribution}}$$

$$\frac{\hat{\mu} - \mu}{\sqrt{\sigma^2/n}} \sim \text{Normal}(0, 1)$$

Pivotal Statistic: a statistic whose sampling distribution does not depend on any unknown parameters

$$\frac{\hat{\mu} - \mu}{\sqrt{\sigma^2/n}} \text{ is pivotal}$$

Confidence Interval

Interval of the form

$$(\hat{\mu} - c_\ell, \hat{\mu} + c_u)$$

$$c_\ell, c_u \geq 0$$

where

$$\Pr(\mu - c_\ell \leq \hat{\mu} \leq \mu + c_u)$$

forms the "level" or coverage probability of the interval

If $Z \sim \text{Normal}(0,1)$ then

$$\Pr(-1.96 \le Z \le 1.96) = 0.95$$

Let $$Z = \frac{\hat{\mu} - \mu}{\sqrt{\sigma^2/n}}$$

$$\Pr\left(-1.96 \le \frac{\hat{\mu} - \mu}{\sqrt{\sigma^2/n}} \le 1.96\right) = 0.95$$

$$= \Pr\left(-1.96\sqrt{\sigma^2/n} \le \hat{\mu} - \mu \le 1.96\sqrt{\sigma^2/n}\right)$$

$$= \Pr\left(-1.96\sqrt{\sigma^2/n} \le \mu - \hat{\mu} \le 1.96\sqrt{\sigma^2/n}\right)$$

$$= \Pr\left(\hat{\mu} - 1.96\sqrt{\sigma^2/n} \le \mu \le \hat{\mu} + 1.96\sqrt{\sigma^2/n}\right)$$

$\Rightarrow$ $C_l = C_u = 1.96\sqrt{\sigma^2/n}$ gives me a 95% confidence interval

Confidence Interval Simulation

```
> mu <- 5
> n <- 20
> x <- replicate(10000, rnorm(n=n, mean=mu)) # 10000 studies
> m <- apply(x, 2, mean) # the estimate for each study
> ci <- cbind(m - 1.96/sqrt(n), m + 1.96/sqrt(n))
> head(ci)
         [,1]     [,2]
[1,] 4.797848 5.674386
[2,] 4.599996 5.476534
[3,] 4.472930 5.349468
[4,] 4.778946 5.655485
[5,] 4.778710 5.655248
[6,] 4.425023 5.301561
```

```
> cover <- (mu > ci[,1]) & (mu < ci[,2])
> mean(cover)
[1] 0.9512
```

Let z_α be the α-percentile of Normal$(0,1)$.

If $Z \sim$ Normal$(0,1)$ then

$$\Pr(Z \le z_\alpha) = \alpha$$

$(1-\alpha)$ level CI:

$$\left(\hat{\mu} - |z_{\alpha/2}|\frac{\sigma}{\sqrt{n}},\ \hat{\mu} + |z_{\alpha/2}|\frac{\sigma}{\sqrt{n}}\right)$$

$$z_{\alpha/2} = -z_{1-\alpha/2}$$

When $\alpha = 0.05$ then

$$z_{\alpha/2} = -1.96, \quad z_{1-\alpha/2} = 1.96$$

$(1-\alpha)$-level upper CI:

$$\Big(-\infty, \hat{\mu} + \underbrace{|z_\alpha| \frac{\sigma}{\sqrt{n}}}\Big) \quad \text{\small 1-α percentile}$$

$(1-\alpha)$-level lower CI:

$$\Big(\hat{\mu} - |z_\alpha| \frac{\sigma}{\sqrt{n}}, \infty\Big)$$

α percentile

Hypothesis Testing

Coin example I did a hypothesis test of $p = 0.5$ vs. $p \neq 0.5$

hypothesis test / significance test is a formal procedure for comparing observed data with a hypothesis whose truth we want to assess

the results of a test are expressed in terms of how well the data and one of the hypotheses agree

Null hypothesis (H_0) is the statement being tested

Alternative hypothesis (H_1) is the of the null, and it's the "interesting" state

H_0 and H_1 are defined in terms of parameter values (or probabilistic property)

Examples:

two-sided: $H_0: \mu = 5 \quad H_1: \mu \neq 5$

one-sided: $\begin{cases} H_0: \mu \leq 5 \quad H_1: \mu > 5 \\ H_0: \mu \geq 5 \quad H_1: \mu < 5 \end{cases}$

test statistic a statistic designed to quantify evidence against the H_0 in favor of H_1

$H_0: \mu = 5 \quad H_1: \mu \neq 5$

$$|z| = \left| \frac{\bar{X} - 5}{\sqrt{\sigma^2/n}} \right|$$

$$\hat{\mu} = \bar{X} \qquad \frac{\hat{\mu} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1)$$

If $H_0: \mu = 5$ is true then

$$Z = \frac{\bar{X} - 5}{\sqrt{\sigma^2/n}} \sim \text{Normal}(0,1)$$

is pivotal

Collect my n data points and calculate my <u>observed statistic</u>:

a number

$$z = \frac{\bar{x} - 5}{\sqrt{\sigma^2/n}}$$

Recall larger $|z|$ is, the more evidence against H_0 in favor of H_1.

<u>p-value</u> is the probability of observing a test statistic "as or more extreme" than the observed statistic under the sampling dist'n of the statistic when H_0 is true

Let $Z^* \sim \text{Normal}(0,1)$

$$\text{p-value} = \Pr(\underbrace{|Z^*|}_{\text{theoretical}} \geq \underbrace{|z|}_{\text{observed}})$$

Make a decision based on the p-value. Smaller it is, the more evidence against H_0 in favor of H_1 there is.

Let's say we call a test "significant" if p-value $\leq \alpha$.

<u>Type I error or false positive:</u>

Call test significant, i.e., reject H_0 in favor of H_1 when H_0 is actually true

<u>Type II error or false negative</u>

Fail to call test significant when H_1 is actually true

"Rate" is the probability of these errors under a decision rule.

If p-value $\leq \alpha$ is the rule then false positive rate is α

under two-sided test, is $\leq \alpha$

under a one-sided test

Exercise: convince yourself

p^* be the p-value under repeated studies. Sampling dist'n of p^* when H_0 is true is Uniform$(0,1)$ for a two-sided test.

Show $\Pr(p^* \leq t \,;\, H_0 \text{ true}) = t$

for $t \in [0,1]$.

$\Pr(p^* \leq t \,;\, H_0 \text{ true}) \leq t$

for one-sided

Power : Probability of a significant test when H_1 is true

Power = 1 − false negative rate

Power is calculated under a range of alternative parameter values

---

$X : \Omega \longrightarrow \mathbb{R}$

$(\Omega, \mathcal{F}, P)$

$\mathcal{R} = \{X(\omega) : \omega \in \Omega\}$

$\mathcal{R}$ discrete or continuous

$f(x)$ p.m.f. or p.d.f.

$\sum_{x \in \mathcal{R}} f(x) = 1$ or $\int_{x \in \mathcal{R}} f(x)\,dx = 1$

Joint Distributions

Distribution of two or more random variables

Bivariate joint dist'n :

rv's X, Y

have pmf or pdf $f(x,y)$

discrete $f(x,y) = Pr(X=x, Y=y)$

$= Pr(\{\omega : X(\omega)=x\} \cap \{\omega : Y(\omega)=y\})$

analogous for pdf's

$A_x \subseteq \mathbb{R}$, $A_y \subseteq \mathbb{R}$ then

$Pr(X \in A_x, Y \in A_y)$ is :

$\sum_{x \in A_x} \sum_{y \in A_y} f(x,y)$ discrete

$\int_{x \in A_x} \int_{y \in A_y} f(x,y)\, dx\, dy$ continuous

Bivariate cdf:

$$F(a, b) = Pr(X \le a, Y \le b)$$

Marginal Dist'n

$$f(x) = \sum_{y \in R_y} f(x, y)$$

$$f(x) = \int_{y \in R_y} f(x, y)\, dy$$

Independence of X and Y

$$f(x, y) = f(x) f(y)$$

Conditional dist'ns

$X \mid Y = y$

$$f(x \mid y) = \frac{f(x, y)}{f(y)}$$

$$\sum_{x \in R_x} f(x|y) = 1$$

$$\int_{x \in R_x} f(x|y)\, dx = 1$$

Law of Total Variance

$$E[X^k | Y=y] = \sum_{x \in R_x} x^k f(x|y)$$

or

$$= \int_{x \in R_x} x^k f(x|y)\, dx$$

General Joint Dist'ns

$X_1, X_2, \ldots, X_n$

$f(\boldsymbol{x}) = f(x_1, x_2, \ldots, x_n)$

$\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$

If the r.v.'s are independent

then $f(\boldsymbol{x}) = \prod_{i=1}^{n} f(x_i)$

Likelihood

$X_1, X_2, \ldots, X_n \sim \text{Bernoulli}(p)$

write their joint pmf as

$$f(\boldsymbol{x}; p) = \prod_{i=1}^{n} f(x_i; p)$$

Generically write θ as the parameter.

Joint pmf or pdf

$$f(\boldsymbol{x}; \theta)$$

$$\prod_{i=1}^{n} f(x_i; \theta) \quad \text{(independence)}$$

For observed data $x_1, x_2, \ldots, x_n$

I can calculate $f(\boldsymbol{x}; \theta)$

$\Longrightarrow$ this a function of θ

Likelihood

$$L(\theta;\mathbf{x}) = f(\mathbf{x};\theta)$$

viewed as a function of θ

for observed $\mathbf{x} = (x_1, x_2, \ldots, x_n)$

$$L(\theta;\mathbf{x}) = \prod_{i=1}^{n} L(\theta; x_i) \quad \text{independent}$$

log Likelihood

$$\log(L(\theta;\mathbf{x})) = \ell(\theta;\mathbf{x})$$

independence $\Longrightarrow$

$$\ell(\theta;\mathbf{x}) = \sum_{i=1}^{n} \ell(\theta, x_i)$$

Sufficient statistic

A sufficient statistic $T(\mathbf{x})$ is such that $\mathbf{X} \mid T(\mathbf{X})$ does depend on θ

If $f(\boldsymbol{x};\theta) = g(T(\boldsymbol{x});\theta)\, h(\boldsymbol{x})$

then $T(\boldsymbol{x})$ is sufficient

$$L(\theta;\boldsymbol{x}) = g(T(\boldsymbol{x}),\theta)\, h(\boldsymbol{x})$$
$$\propto L(\theta; T(\boldsymbol{x}))$$

Example $X_1, X_2, X_3, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$

$T(\boldsymbol{X}) = \bar{X}$ is sufficient for μ

$\bar{X} \sim \text{Normal}(\mu, \sigma^2/n)$

Likelihood Principle

Suppose $\boldsymbol{x}$ and $\boldsymbol{y}$ are two data sets such that

$$L(\theta;\boldsymbol{x}) \propto L(\theta;\boldsymbol{y})$$

i.e. $L(\theta;\boldsymbol{x}) = L(\theta;\boldsymbol{y})\, c(\boldsymbol{x},\boldsymbol{y})$

Then inference on θ should be the same for x and y.

Maximum Likelihood Estimation

Estimate θ as the value that maximizes $L(\theta; x)$

$$\hat{\theta}_{MLE} = \arg\max_\theta L(\theta; x)$$

$$= \arg\max_\theta \ell(\theta; x)$$

$$= \arg\max_\theta L(\theta; T(x))$$

Example:

$$X \sim \text{Binomial}(n, p)$$

$$L(p; x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$\propto p^x (1-p)^{n-x}$$

$$\ell(p;x) \propto x\log(p) + (n-x)\log(1-p)$$

- $\frac{d}{dp}\ell(p;x)$ set it to 0
- solve for p.

$$\Longrightarrow \hat{p} = \frac{x}{n}$$

It happens to be the case that

true parameter

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \,\dot{\sim}\, \text{Normal}(0,1)$$

MLE

standard error with $\hat{p}$ plugged

Approx 95% CI :

$$\left(\hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},\ \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right)$$

$$Var(X) = np(1-p)$$

$$Var\left(\frac{X}{n}\right) = \frac{1}{n^2} Var(X)$$

$$= \frac{np(1-p)}{n^2}$$

$$= \frac{p(1-p)}{n}$$

Properties of MLEs

Under certain "regularity conditions", we have the following:

Consistent: $\hat{\theta}_n$ MLE n observations

$$\hat{\theta}_n \xrightarrow{P} \theta$$

For all $\varepsilon > 0$,

$$Pr(|\hat{\theta}_n - \theta| > \varepsilon) \longrightarrow 0$$

as $n \longrightarrow \infty$

- Equivariant :

  If $\hat{\theta}$ is MLE for θ

  then $g(\hat{\theta})$ is MLE $g(\theta)$

  $$\left(\begin{array}{l} \hat{p} \text{ MLE } p \\ \dfrac{\hat{p}(1-\hat{p})}{n} \text{ MLE } \dfrac{p(1-p)}{n} \end{array} \right)$$

- Asymptotically Normal distributed

- Asymptotically Efficient

  Take any estimate $\tilde{\theta}_n$

  $$\frac{Var(\hat{\theta}_n)}{Var(\tilde{\theta}_n)} \longrightarrow \text{ to a quantity } \leq 1.$$

Fisher Information

Let $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F_\theta$

$$I_n(\theta) = Var\left(\frac{d}{d\theta} \log f(X;\theta)\right)$$

$$= \sum_{i=1}^{n} Var\left(\frac{d}{d\theta} \log f(X_i;\theta)\right)$$

$$= -E\left[\frac{d^2}{d\theta^2} \log f(X;\theta)\right]$$

$$= -\sum_{i=1}^{n} E\left[\frac{d^2}{d\theta^2} \log f(X_i;\theta)\right]$$

$$Var(\hat{\theta}_n) \approx \frac{1}{I_n(\theta)}$$

Standard Error

$$se(\hat{\theta}_n) = \sqrt{Var(\hat{\theta}_n)}$$

$$\text{se}(\hat{\theta}_n) \approx \frac{1}{\sqrt{I_n(\theta)}}$$

$$\hat{\text{se}}(\hat{\theta}_n) = \frac{1}{\sqrt{I_n(\hat{\theta}_n)}}$$

<u>MLE CLT</u>:

$$\frac{\hat{\theta}_n - \theta}{\text{se}(\hat{\theta}_n)} \xrightarrow{D} \text{Normal}(0,1)$$

$$\frac{\hat{\theta}_n - \theta}{\hat{\text{se}}(\hat{\theta}_n)} \xrightarrow{D} \text{Normal}(0,1)$$

$$Z = \frac{\hat{\theta}_n - \theta}{\hat{\text{se}}(\hat{\theta}_n)}$$ is approx pivotal

Normal(0,1)

Repeat our special case

where $\hat{\mu} \longleftarrow \hat{\theta}$

$\sqrt{\frac{\sigma^2}{n}} \longleftarrow \hat{se}(\hat{\theta})$