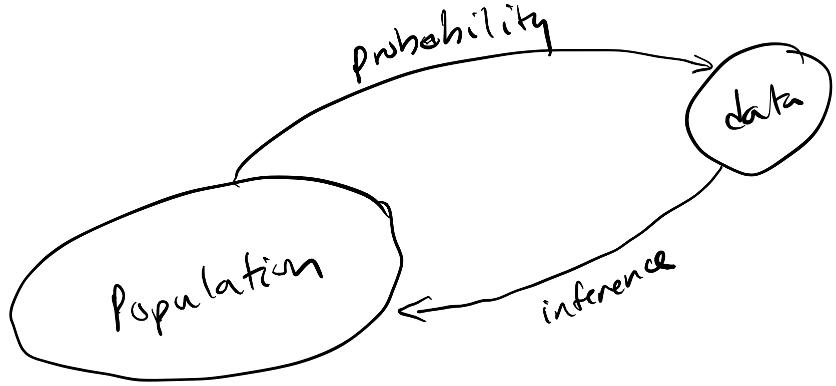
Week 4 QCB 408/508 Spring 2020

- Likelihood & maximum likelihood estimation (MLE)
- EFDs (exponential family distributions)
- Frequentist inference from pivotal statistics



rv's $X_1, X_2, \ldots, X_n \overset{iid}{\sim} F_\theta$

θ parameter should be informative about what we want to know about the population

$(X_1, X_2, \ldots, X_n) \sim F_\theta$ joint distribution

Two levels in a study:

1. Observed data $x_1, x_2, \ldots, x_n$
2. Random variables $X_1, X_2, \ldots, X_n$ that model obtained data

Let's suppose we have a joint pdf or a pmf for the rv model:

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

If we're going to evaluate this on observed data, it becomes a function of θ:

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

observed data

log-likelihood $\ell(\theta;\underline{x}) = \log L(\theta;\underline{x})$

Suppose $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f_\theta$

Then
$$\begin{aligned}
\ell(\theta;\underline{x}) &= \log(L(\theta;\underline{x})) \\
&= \log(f(\underline{x};\theta)) \\
&= \log\left(\prod_{i=1}^{n} f(x_i;\theta)\right) \\
&= \sum \log f(x_i;\theta) \\
&= \sum \ell(\theta;x_i)
\end{aligned}$$

Sufficient statistic

A statistic $T(\underline{x})$ is any function of the data

$T(\underline{x})$ is sufficient if $\underline{X} \mid T(\underline{X})$

does not depend on θ.

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

$T(\underline{x})$ is sufficient.

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

proportional in θ

Exercise

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

$\bar{X}_n$ is sufficient for μ.

Hint: $\sum_{i=1}^{n} (x_i - \mu)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 + n(\bar{x} - \mu)^2$.

$$f(\underline{x};\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ - \frac{\sum (x_i - \mu)^2}{2\sigma^2} \right\}$$

Other topics:

- Minimal sufficient statistics
- Complete sufficient statistics
- Ancillary statistics
- Basu's theorem

Maximum Likelihood Estimation

The MLE is the value of θ that maximizes the likelihood.

$$\hat{\theta}_{MLE} = \underset{\theta}{\operatorname{argmax}}\ L(\theta; \underset{\sim}{x})$$

$$= \underset{\theta}{\operatorname{argmax}}\ \ell(\theta; \underset{\sim}{x})$$

$$= \underset{\theta}{\operatorname{argmax}}\ L(\theta; T(\underset{\sim}{x}))$$

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

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

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

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

$$\frac{d}{dp} \ell(p; x) = 0 \quad , \quad \text{solve for } p$$

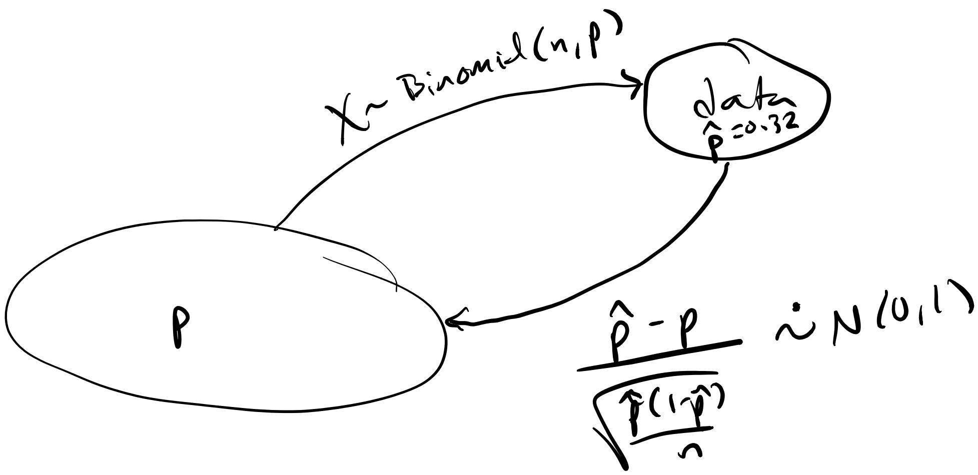
$$\Longrightarrow \hat{p}_{MLE} = \frac{X}{n}$$

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \overset{\cdot}{\sim} \text{Normal}(0,1) \quad \text{for large } n$$

Suppose my real life data says $\hat{p} = 0.32$

$$E[\hat{p}] = p \quad , \quad Var(\hat{p}) = \frac{p(1-p)}{n}$$

$$\hat{p} = \frac{X}{n}$$



I've observed $\hat{p} = 0.32$ $(x = 32,\ n = 100)$

But I want the "sampling distribution" of $\hat{p}$: the dist'n of $\hat{p} = \frac{X}{n}$ when I repeat the study (probability arm) over and over.

p is unknown, so the dist'n of $\hat{p}$ involves p. But a pivotal statistic does <u>not</u> involve p, such as:

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

Assumptions:

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

$\hat{\theta}_n$ is the MLE from the data.

"Regularity conditions" met.

(1) MLE is "consistent":

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

(2) Equivariance: If $\hat{\theta}_n$ is the MLE of θ, then $g(\hat{\theta}_n)$ is the MLE of $g(\theta)$.

Example: $X \sim \text{Binomial}(n, p)$.
$\hat{p} = X/n$. $n\hat{p}(1-\hat{p})$ is MLE of $Var(X) = np(1-p)$.

(3) Fisher Information is

$$
\begin{aligned}
I_n(\theta) &= Var\left(\frac{d}{d\theta} \ell(\theta; \underline{x})\right) \\
&= Var\left(\frac{d}{d\theta} \sum \ell(\theta; x_i)\right) \\
&= \sum Var\left(\frac{d}{d\theta} \ell(\theta; x_i)\right) \\
&\;\;\vdots \\
&= -E\left[\frac{d^2}{d\theta^2} \ell(\theta; \underline{x})\right] \\
&\;\;\vdots \\
&= -\sum_{i=1}^{n} E\left[\frac{d^2}{d\theta^2} \ell(\theta; x_i)\right]
\end{aligned}
$$

In general, the "standard error" of an estimator is the standard deviation of the sampling dist'n of the estimator.

For MLEs, the standard error of $\hat{\theta}_n$ is $\sqrt{Var(\hat{\theta}_n)} \equiv se(\hat{\theta}_n)$.

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

true θ

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

is defined as the usual standard error estimator of an MLE.

CLT for MLE :

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

as $n \to \infty$

$$\frac{\hat{\theta}_n - \theta}{\hat{se}(\hat{\theta}_n)} \xrightarrow{D} \text{Normal}(0,1) \quad \text{as } n \to \infty$$

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

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

Asymptotic Pivotal Statistic

Under above assumptions,

$$Z = \frac{\hat{\theta}_n - \theta}{\hat{se}(\hat{\theta}_n)} \xrightarrow{D} \text{Normal}(0,1) \quad \text{as } n \to \infty$$

is approximate pivotal statistic.

Optimality

The MLE is such that

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{D} \text{Normal}(0, \tau^2)$$

Suppose $\tilde{\theta}_n$ is any other estimator where $\sqrt{n}(\tilde{\theta}_n - \theta) \xrightarrow{D} \text{Normal}(0, \sigma^2)$

It follows that $\frac{\tau^2}{\sigma^2} \leq 1$.

Delta Method

If $g(\cdot)$ is a differentiable function and $g'(\theta) \neq 0$. We have

$$g(t) \approx g(\theta) + g'(\theta)(t-\theta).$$

Suppose I have $\hat{se}(\hat{\theta}_n)$, then

$$\hat{se}(g(\hat{\theta}_n)) = |g'(\hat{\theta}_n)| \, \hat{se}(\hat{\theta}_n).$$

Example: $X \sim \text{Binomial}(n, p)$. $\hat{p} = \frac{X}{n}$.

$\hat{se}(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. $g(p) = p(1-p)$.

$\widehat{g(p)} = \hat{p}(1-\hat{p})$. $g'(p) = 1 - 2p$.

$$\hat{se}(\hat{p}(1-\hat{p})) = |1-2\hat{p}| \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Distribution	MLE	Std Err	Z Statistic
Binomial(n,p)	$\hat{p} = X/n$	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$	$\frac{\hat{p}-p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}}$
Normal(μ,σ^2)	$\hat{\mu} = \overline{X}$	$\frac{\hat{\sigma}}{\sqrt{n}}$	$\frac{\hat{\mu}-\mu}{\hat{\sigma}/\sqrt{n}}$
Poisson(λ)	$\hat{\lambda} = \overline{X}$	$\sqrt{\frac{\hat{\lambda}}{n}}$	$\frac{\hat{\lambda}-\lambda}{\sqrt{\hat{\lambda}/n}}$

Exponential Family Distributions (EFD)

Simple case: Natural single Parameter EFD

$$f(x;\eta) = h(x)\exp\{\eta x - A(\eta)\}$$

Example: Bernoulli(p)

$$f(x;p) = p^x (1-p)^{1-x}$$

$$= \exp\{x\log(p) + (1-x)\log(1-p)\}$$

$$= \exp\{\log\left(\frac{p}{1-p}\right)x + \log(1-p)\}$$

$$\eta(p) = \log\left(\frac{p}{1-p}\right)$$

$$A(\eta) = \log(1+e^\eta)$$

$$= \exp\{\eta x - A(\eta)\}$$

A General Definition:

If X follows an EFD parameterized on the observed scale by $\underline{\theta}$ then it has pdf or pmf of the form:

d-dimensional

$$f(x;\underline{\theta}) = h(x)\exp\left\{\sum_{k=1}^{d} \eta_k(\underline{\theta})\,T_k(x) - A(\underline{\eta})\right\}$$

$$\underline{\eta} = \begin{pmatrix} \eta_1(\underline{\theta}) \\ \eta_2(\underline{\theta}) \\ \vdots \\ \eta_d(\underline{\theta}) \end{pmatrix}$$

$T_1(x), \dots, T_d(x)$ are sufficient statistics $\eta_1, \eta_2, \dots, \eta_d$ respectively

The functions $\eta_k(\underline{\theta})$ $k=1,\dots,d$ map the usual (observed) to the "natural parameters"

$A(\underline{\eta})$ is sometimes called "log normalizer"

$$A(\underline{\eta}) = \log \int h(x)\exp\left\{\sum_{k=1}^{d} \eta_k(\underline{\theta})\,T_k(x)\right\}dx$$

Example: Normal (μ, σ^2)

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{(x-\mu)^2}{2\sigma^2} \right\}$$

$$= \frac{1}{\sqrt{2\pi}} \exp\left\{ \frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} x^2 - \log(\sigma) - \frac{\mu^2}{2\sigma^2} \right\}$$

$$\eta(\mu, \sigma^2) = \begin{array}{cc} \begin{pmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{pmatrix} & \begin{array}{l} \eta_1 \\ \\ \eta_2 \end{array} \end{array} \qquad \begin{array}{l} T_1(x) = x \\ T_2(x) = x^2 \end{array}$$

$$A(\underline{\eta}) = \log(\sigma) + \frac{\mu^2}{2\sigma^2} = -\frac{1}{2}\log(-2\eta_2) - \frac{\eta_1^2}{4\eta_2}$$

<u>Calculating Moments</u>:

$$\frac{\partial}{\partial \eta_k} A(\underline{\eta}) = E[T_k(X)]$$

$$\frac{\partial^2}{\partial^2 \eta_k} A(\underline{\eta}) = \mathrm{Var}(T_k(X))$$

Maximum Likelihood :

$X_1, X_2, \ldots, X_n$ are iid from an EFD

$$\ell(\underset{\sim}{\eta}; \underset{\sim}{x}) = \sum_{i=1}^{n} \left\{ \left[\log\{h(x_i)\} \right] + \sum_{k=1}^{d} \eta_k(\underset{\sim}{\theta}) T_k(x_i) - A(\underset{\sim}{\eta}) \right\}$$

$$\frac{\partial}{\partial \eta_k} \ell(\underset{\sim}{\eta}; \underset{\sim}{x}) = \sum_{i=1}^{n} T_k(x_i) - n \frac{\partial}{\partial \eta_k} A(\underset{\sim}{\eta})$$

So MLE of η_k is the solution to:

$$\frac{1}{n} \sum_{i=1}^{n} T_k(x_i) = \frac{\partial}{\partial \eta_k} A(\underset{\sim}{\eta}) = E[T_k(X)]$$

Statistical Inference

We have observed data that is modeled by a probability generation process. The probability distribution has parameters informative about the population.

Statistical Inference

reverse engineers this forward to estimate parameters and provide measures of uncertainty about the estimates

- Parameter
- Statistic
- sampling distribution

sampling dist'n connects my calculated statistic to the population (probability model).

Goals and Strategies - Frequentist

data $x_1, x_2, \ldots, x_n$

Model $X_1, X_2, \ldots, X_n \sim F_\theta$

1. Point estimate of θ
2. Confidence interval of θ
   - uncertainty of point estimate
3. Hypothesis test - assesses specific value(s) of θ

Point estimation: Example MLE $\hat{\theta}_n$

Confidence Intervals (of MLEs):

CI has the form:

$$(\hat{\theta} - c_l, \hat{\theta} + c_u) \qquad c_l, c_u > 0$$

where

$$\Pr(\hat{\theta} - c_l \le \theta \le \hat{\theta} + c_u ; \theta)$$

forms the "level" or coverage probability.

Note: $\hat{\theta}$ is the rv and c_l, c_u are too usually

Approximate 95% CI for MLEs

$$0.95 \approx \Pr\left(-1.96 \le \frac{\hat{\theta} - \theta}{\hat{se}(\hat{\theta})} \le 1.96\right)$$

$$= \Pr\left(-1.96 \cdot \hat{se}(\hat{\theta}) \le \hat{\theta} - \theta \le 1.96 \cdot \hat{se}(\hat{\theta})\right.$$

$$= \Pr\left(\theta - 1.96\, \hat{se}(\hat{\theta}) \le \hat{\theta} \le \theta + 1.96\, \hat{se}(\hat{\theta})\right)$$

$$= \Pr\left(-1.96\,\hat{se}(\hat{\theta}) \le \theta - \hat{\theta} \le 1.96\,\hat{se}(\hat{\theta})\right)$$

$$= \Pr\left(\hat{\theta} - 1.96\,\hat{se}(\hat{\theta}) \le \theta \le \hat{\theta} + 1.96\,\hat{se}(\hat{\theta})\right)$$

$\Rightarrow$ 95% approx. CI is

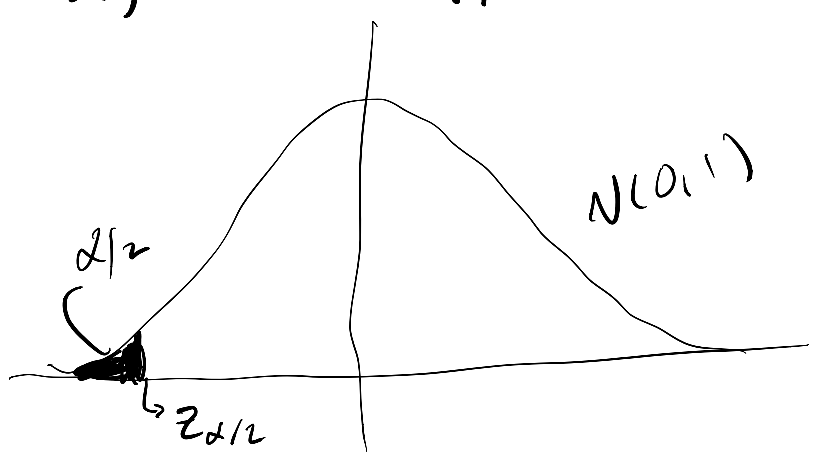
$$\left(\hat{\theta} - 1.96\,\hat{se}(\hat{\theta}),\ \hat{\theta} + 1.96\,\hat{se}(\hat{\theta})\right)$$

See reading for one-sided

$$(\hat{\theta} - c_\ell,\ \infty)$$

$$(-\infty,\ \hat{\theta} + c_u)$$

$(1-\alpha)\cdot 100\%$ approx CI:



z_α is the α-percentile of Normal$(0,1)$

$(1-\alpha)\cdot 100\%$ approx CI:

$$\left(\hat{\theta} - |z_{\alpha/2}|\,\widehat{se}(\hat{\theta}),\ \hat{\theta} + |z_{\alpha/2}|\,\widehat{se}(\hat{\theta})\right)$$

sample surveys:

$$|z_{\alpha/2}|\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le \overbrace{|z_{\alpha/2}|}^{1.96}\sqrt{\frac{0.5^2}{n}}$$

Next time: Hypothesis testing
+ Bayesian Inference