STA355
UTSG

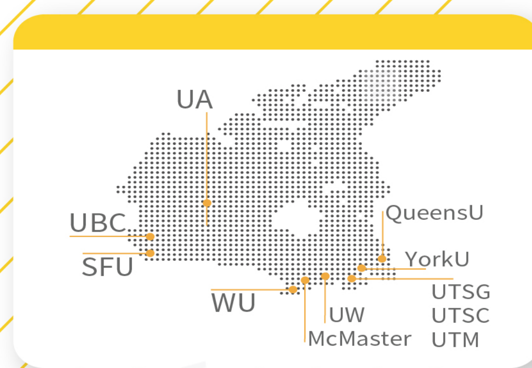
Week 6 周课

EASY GROUP 易维集团 | Powered by Easy Group EASY EDU 易维集团旗下公司

关于Easy Edu

Easy Education Inc（易途教育）为Easy Group旗下品牌，是加拿大最专业，最具规模及影响力的华人教育培训机构，在多伦多、滑铁卢、温哥华、阿尔伯塔等地均设有培训基地。

截至目前共开设**十一大校区**，提供近三百门大学科目培训辅导。自2014年成立至今，帮助过数万名海外留学生度过学术难关。



累计线下学员 **10万+**

线上学员人数 **16万+**

线上关注人数 **17万+**

想要获得更多加拿大留学生相关资讯
还请大家**关注**我们哦

这里不仅仅有**超多有趣**的小故事
还有 **许多干货**和**最新发布** 的一手消息
涵盖全加拿大**吃喝玩乐学**
让你再也不用费力查找资讯

不仅如此，我们也会不定期放出**超多福利**
甭管你是本科在读、找工作、申研
在**Easy Edu**都能获得最专业的帮助

Easy Edu
为大家的成长、学习保驾护航
在这里，**你永远不是一座孤岛**

4.0

商务合作洽谈 请联系

Disclaimer

This complementary study package is provided by Easy Education Inc. and its affiliated mentors. This study package seeks to support your study process and should be used as a complement, **NOT** substitute to course material, lecture notes, problem sets, past tests and other available resources.

We acknowledge that this package contains some materials provided by professors and staff of the University of Toronto, and the sources of these materials are cited in details wherever they appear.

This package is distributed for free to students participating in Easy Education's review seminars, and are not for sale or other commercial uses whatsoever. We kindly ask you to refrain from copying or selling in part or in whole any information provided in this package.

Thank you for choosing Easy Education. We sincerely wish you the best of luck in all of your exams.

Easy Education Inc.

今日内容:

1. Week 7 学校正课内容 (lectures 9-10)

From point estimation to interval estimation

- **Model:** We observe $X_1, \cdots, X_n$ assumed to have a distribution depending on some unknown parameter θ.
- **Point estimation:** Estimate the value of θ by a **point estimator** $\widehat{\theta} = \widehat{\theta}(X_1, \cdots, X_n)$.
- $\widehat{\theta}$ will have a **sampling distribution**, which will depend on θ as well as possibly other unknown parameters.
- The standard deviation of the sampling distribution is called the **standard error**.
    - We can often estimate the standard error using the Delta Method or jackknife method.
- **Interval estimation:** Define an interval

$$\mathcal{I} = [\ell(X_1, \cdots, X_n), u(X_1, \cdots, X_n)]$$

that we believe will contain θ with probability close to 1.

Interval estimation

- In this course, we'll talk about two approaches to interval estimation.
- **Confidence intervals:** These are typically defined in terms of the sampling distribution of a point estimator $\widehat{\theta}$ (or some related statistic).
  - We will often need to use approximations (e.g. normal approximations) to the sampling distributions.
  - The "confidence level" is defined in terms of repeated sampling.
- **Credible intervals:** These are based on the posterior distribution of θ given the observed data $x_1, \cdots, x_n$.
  - If $\pi(\theta | x_1, \cdots, x_n)$ is the posterior density of θ then $\mathcal{I}$ is a $100p\%$ credible interval if

$$\int_{\mathcal{I}} \pi(\theta | x_1, \cdots, x_n)\, d\theta = p.$$

  - Note that the "credible level" is defined in terms of the posterior distribution (which depends on the prior distribution and the data).

Confidence intervals

- **Definition:** An interval $\mathcal{I} = [\ell(X_1, \cdots, X_n), u(X_1, \cdots, X_n)]$ is a **confidence interval** (CI) with coverage $100p\%$ (or a $100p\%$ CI) if

$$\underbrace{P_\theta \left[\ell(X_1, \cdots, X_n) \le \theta \le u(X_1, \cdots, X_n)\right]}_{P_\theta(\theta \in \mathcal{I})} = p \ \text{ for all } \theta \in \Theta.$$

- If the probability statement above holds approximately (e.g. if n is large) then we often say that the interval is an approximate $100p\%$ CI for θ.
  - Typically, we have $\mathcal{I}_n = [\ell_n(X_1, \cdots, X_n), u_n(X_1, \cdots, X_n)]$ with

$$P_\theta(\theta \in \mathcal{I}_n) \to p \ \text{ as } n \to \infty$$

for all $\theta \in \Theta$.

Some comments on CIs

- In the definition of a CI, the interval $\mathcal{I}$ is actually a random interval (depending on the random variables $X_1, \cdots, X_n$) and a CI is defined in terms of the probability that the random interval contains θ.
- The data-based interval $[\ell(x_1, \cdots, x_n), u(x_1, \cdots, x_n)]$ **cannot** be interpreted in terms of the probability distribution of $(X_1, \cdots, X_n)$ – the interval either contains θ or it doesn't!
- However, the length of the CI gives us an idea about the uncertainty in the estimation of θ (much like an estimate of the standard error).
  - Many CIs are formed in terms of an estimator and its standard error.
  - For example, $\widehat{\theta} \pm z \times \widehat{\text{se}}(\widehat{\theta})$ for some z.

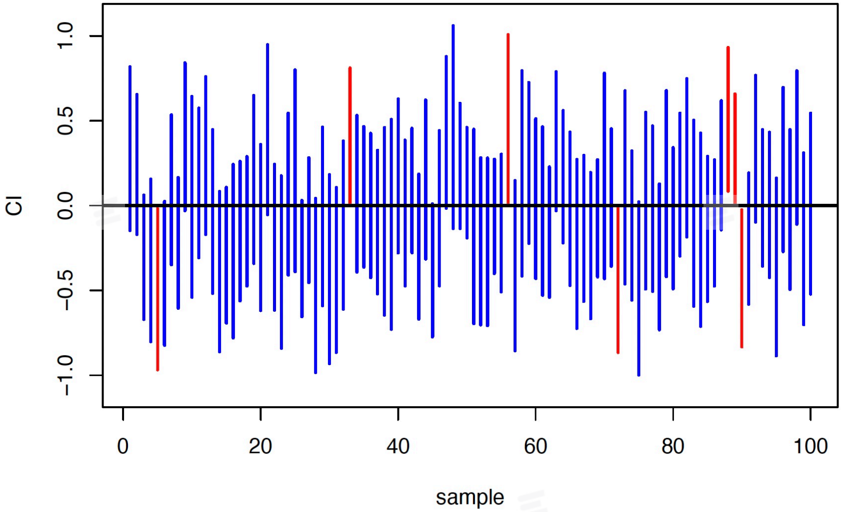
Example: CI demonstration

- **Model:** $X_1, \cdots, X_{20}$ independent $\mathcal{N}(\mu, \sigma^2)$ with both μ and σ^2 unknown.
- Classic 95% CI for μ:

$$\left[\bar{X} - 2.093\frac{S}{\sqrt{20}}, \bar{X} + 2.093\frac{S}{\sqrt{20}}\right]$$

  where $\bar{X}$ and S^2 are the sample mean and variance respectively.
  - 2.093 is the 0.975 quantile of Student's t distribution with 19 $(= 20 - 1)$ degrees of freedom.
  - Note that $\widehat{\text{se}}(\bar{X}) = S/\sqrt{n}$.
- **Simulation experiment:** Generate 100 samples of size 20 from a $\mathcal{N}(0, 1)$ distribution and compute 95% CIs for each sample.
  - Given the theory, we would expect that approximately 95 of the 100 constructed CIs will contain the true mean 0.

100 95% CIs for μ



CIs that do not include 0 are indicated by red bars.

Determining CIs: Heuristics

- **Model:** $X_1, \cdots, X_n$ independent with some (unknown) cdf F.
- Estimate $\theta = \theta(F)$ by $\widehat{\theta}$ where

$$\widehat{\theta} \approx \mathcal{N}(\theta, [\text{se}(\widehat{\theta})]^2).$$

- Thus if $\widehat{\text{se}}(\widehat{\theta})$ is a "good" estimator of $\text{se}(\widehat{\theta})$, we should have

$$\frac{\widehat{\theta} - \theta}{\widehat{\text{se}}(\widehat{\theta})} \approx \mathcal{N}(0, 1).$$

- Thus, for example,

$$P\left(-1.96 \leq \frac{\widehat{\theta} - \theta}{\widehat{\text{se}}(\widehat{\theta})} \leq 1.96\right) \approx 0.95$$

and so $\widehat{\theta} \pm 1.96\,\widehat{\text{se}}(\widehat{\theta})$ are the limits of an approximate 95% CI for θ.

The pivotal method

- It's not too difficult to formalize these heuristics.
- **Idea:** Find a random variable $g(X_1, \cdots, X_n, \theta)$ whose distribution is independent of θ and any other unknown parameters.
  - $P_\theta[g(X_1, \cdots, X_n, \theta) \leq x] = G(x)$ where $G(x)$ is completely known.
  - $g(X_1, \cdots, X_n, \theta)$ is called a **pivot**.
- Given the pivot, we choose a and b so that

$$p = P_\theta\left[a \leq g(X_1, \cdots, X_n, \theta) \leq b\right] = \underbrace{G(b) - G(a-)}_{independent\ \ of\ \ \theta}.$$

- From this, we get

$$\begin{aligned} p &= P_\theta\left[a \leq g(X_1, \cdots, X_n, \theta) \leq b\right] \\ &\quad \vdots \ (\textit{manipulation}!) \\ &= P_\theta\left[\ell(X_1, \cdots, X_n) \leq \theta \leq u(X_1, \cdots, X_n)\right]. \end{aligned}$$

Details of the pivotal method

- **Choice of pivot:**
  - If we have a point estimator $\widehat{\theta}$ then we can often define the pivot to be $g(\widehat{\theta}, \theta)$ where g is chosen to make its distribution independent of θ.
  - If we cannot find an exact pivot, we can sometimes find an approximate pivot $g(\widehat{\theta}, \theta)$ whose distribution is approximately independent of θ – for example,

$$\frac{\widehat{\theta} - \theta}{\widehat{\text{se}}(\widehat{\theta})} \approx \mathcal{N}(0, 1).$$

- **Choice of *a* and *b*:**
  - Ideally, we'd like to choose a and b to make the CI as short as possible.
  - However, if G is the cdf of the pivot then a good default is to define a so that $G(a) = (1 - p)/2$ and b so that $1 - G(b) = (1 - p)/2$.

Example: CIs for normal parameters

- **Model:** $X_1, \cdots, X_n$ independent $\mathcal{N}(\mu, \sigma^2)$.
- CI for μ: Pivot
$$\frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\widehat{\mu} - \mu}{\widehat{\text{se}}(\widehat{\mu})} \sim \mathcal{T}(n-1)$$
where $\mathcal{T}(n-1)$ is a Student's t distribution with $n-1$ degrees of freedom.
- Define t_p so that
$$P\left(-t_p \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq t_p\right) = p.$$
- Then the $100p\%$ CI for μ is
$$\left[\bar{X} - t_p \frac{S}{\sqrt{n}}, \bar{X} + t_p \frac{S}{\sqrt{n}}\right].$$

Example: CIs for normal parameters (cont'd)

- For the variance, we have

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \chi^2(n-1).$$

- Define a and b so that

$$P\left(a \le \frac{(n-1)S^2}{\sigma^2} \le b\right) = p$$

  from which a $100p\%$ CI for σ^2 is

$$\left[\frac{(n-1)S^2}{b}, \frac{(n-1)S^2}{a}\right].$$

- Choice of a and b:
  - Equal tail areas: $G(a) = 1 - G(b)$ with $G(b) - G(a) = p$.
  - Minimum length: minimize $a^{-1} - b^{-1}$ subject to $G(b) - G(a) = p$.

Example: Exponential distribution

- **Model:** $X_1, \cdots, X_n$ independent Exponential random variables with pdf
$$f(x;\lambda) = \lambda \exp(-\lambda x) \text{ for } x \geq 0$$
where $\lambda > 0$ is unknown.
- We can estimate λ by $\widehat{\lambda} = 1/\bar{X}$.
- We can approximate the distribution of $\widehat{\lambda}$ as follows:
$$\begin{aligned} \sqrt{n}(\widehat{\lambda} - \lambda) &\approx \mathcal{N}(0, \lambda^2) \\ \text{or } \widehat{\lambda} &\approx \mathcal{N}(\lambda, \lambda^2/n). \end{aligned}$$
- From this, it follows that $\mathrm{se}(\widehat{\lambda})$ is approximately $\lambda/\sqrt{n}$ and this can be estimated by $\widehat{\mathrm{se}}(\widehat{\lambda}) = \widehat{\lambda}/\sqrt{n}$.

Example: Exponential distribution (cont'd)

- Look at 4 pivots for λ:
  1. $\lambda \sum_{i=1}^{n} X_i \sim \text{Gamma}(n, 1)$ (exact pivot).
  2. $\dfrac{\widehat{\lambda} - \lambda}{\lambda / \sqrt{n}} \approx \mathcal{N}(0, 1)$ (approximate pivot).
  3. $\dfrac{\widehat{\lambda} - \lambda}{\widehat{\lambda} / \sqrt{n}} \approx \mathcal{N}(0, 1)$ (approximate pivot).
  4. $\sqrt{n}\{\ln(\widehat{\lambda}) - \ln(\lambda)\} \approx \mathcal{N}(0, 1)$ (approximate pivot).

- Pivot 1 uses the fact that a sum of independent Exponential random variables has a Gamma distribution.

- Pivot 4 uses the variance stabilizing transformation discussed earlier.

Example: Exponential distribution (cont'd)

- The CIs resulting from the 4 pivots are:
  1. $\left[\frac{a}{\sum_{i=1}^n X_i}, \frac{b}{\sum_{i=1}^n X_i}\right]$ where a and b are such that
  $$P[a \leq \text{Gamma}(n,1) \leq b] = p;$$
  2. $\left[\frac{\widehat{\lambda}}{1 + z_p/\sqrt{n}}, \frac{\widehat{\lambda}}{1 - z_p/\sqrt{n}}\right]$ where z_p satisfies
  $$P[-z_p \leq \mathcal{N}(0,1) \leq z_p] = p;$$
  3. $\left[\widehat{\lambda}(1 - z_p/\sqrt{n}), \widehat{\lambda}(1 + z_p/\sqrt{n})\right]$;
  4. $\left[\widehat{\lambda}\exp(-z_p/\sqrt{n}), \widehat{\lambda}\exp(z_p/\sqrt{n})\right]$.

Simulation experiment: Coverage of the 4 CIs

- Generate $n = 20$ and $n = 100$ observations from an Exponential distribution and compute the four 95% CIs for 10000 samples.
  - We can estimate the coverage of each interval to within 0.005 with probability close to 1. ($\sqrt{0.05 \times 0.95/10000} \approx 0.002$).
  - Note that the true coverage for pivot 1 is exactly 0.95.

	Est'd coverage	
Pivot	$n = 20$	$n = 100$
1	0.949	0.949
2	0.922	0.945
3	0.954	0.950
4	0.946	0.947

- **Conclusions:**
  - Pivot 2 (i.e. $\sqrt{n}(\widehat{\lambda} - \lambda)/\lambda$) has the poorest coverage of the four CIs.
  - For $n = 100$, the coverage of all four is very close to the nominal 0.95.

Example: Unif$(0, \theta)$

- $X_1, \cdots, X_n$ independent Unif$(0, \theta)$ random variables with pdf

$$f(x; \theta) = \frac{1}{\theta} \quad \text{for } 0 \leq x \leq \theta.$$

- θ represents the maximum possible value of $X_1, \cdots, X_n$.
- **Pivot:** $X_{(n)}/\theta$. Why?
  - Since θ is the maximum possible value, $X_{(n)}$ should contain the most information about θ.
  - Distribution of $X_{(n)}/\theta$ is independent of θ:

$$P_\theta\left(\frac{X_{(n)}}{\theta} \leq x\right) = x^n \ \text{ for } 0 \leq x \leq 1.$$

Example: Unif$(0, \theta)$ (cont'd)

- How to construct a $100p\%$ CI for θ:
  - Take $0 \le a < b \le 1$ such that

$$P\left(a \le \frac{X_{(n)}}{\theta} \le b\right) = b^n - a^n = p.$$

  - Inverting this, we get

$$\left[\frac{X_{(n)}}{b}, \frac{X_{(n)}}{a}\right].$$

- "Optimal" choice of a, b:

$$b = 1 \quad a = (1-p)^{1/n}.$$

Example: CIs for quantiles $F^{-1}(\tau)$

- **Model:** $X_1, \cdots, X_n$ independent random variables with continuous cdf F and pdf f.
- **Goal:** Find a CI (e.g. 95% CI) for the quantile $\theta = F^{-1}(\tau)$.
- **Approach:** Use the pivot

$$g(X_1, \cdots, X_n, \theta) = \sum_{i=1}^{n} I(X_i \leq \theta) \sim \text{Binomial}(n, \tau).$$

- To use this pivot, note that if $a < b$ are integers between 1 and n then

$$\left\{\theta : a \leq \sum_{i=1}^{n} I(X_i \leq \theta) \leq b\right\} = \left\{\theta : X_{(a)} \leq \theta \leq X_{(b)}\right\}.$$

- Thus $[X_{(a)}, X_{(b)}]$ is a $100p\%$ CI for $\theta = F^{-1}(\tau)$ where

$$p = \sum_{k=a}^{b} \binom{n}{k} \tau^k (1-\tau)^{n-k}.$$

- This is a **distribution-free** CI for θ.

Example: CIs for $F^{-1}(\tau)$ (cont'd)

- For a given p (e.g. 0.95), we can find a and b using a normal approximation to the Binomial distribution so that

$$p \approx \sum_{k=a}^{b} \binom{n}{k} \tau^k (1-\tau)^{n-k}.$$

- If n is large enough then $\text{Binomial}(n,\tau) \approx \mathcal{N}(n\tau, n\tau(1-\tau))$.
- Using the normal approximation (with a continuity correction) we get

$$\begin{aligned} a &= \left\lfloor n\tau + \frac{1}{2} - z_p\sqrt{n\tau(1-\tau)} \right\rfloor \\ b &= \left\lceil n\tau - \frac{1}{2} + z_p\sqrt{n\tau(1-\tau)} \right\rceil \end{aligned}$$

where $\lfloor x \rfloor$ and $\lceil x \rceil$ round x, respectively, down and up to the nearest integer, for example:

$$\lfloor 3.5 \rfloor = 3 \quad \lceil 3.5 \rceil = 4.$$

补充笔记：

Consider the function:

$$g(X_1, X_2, \dots, X_n, \theta) = \sum_{i=1}^{n} I(X_i \leq \theta)$$

where $I(\cdot)$ is the indicator function, which equals 1 if the condition inside is true and 0 otherwise.
Given $\theta = F^{-1}(\tau)$, we have:

$$P(X_i \leq \theta) = P(X_i \leq F^{-1}(\tau)) = F(F^{-1}(\tau)) = \tau$$

Since the X_i are independent:

$$g(X_1, \dots, X_n, \theta) = \sum_{i=1}^{n} I(X_i \leq \theta) \sim \text{Binomial } (n, \tau)$$

This means g follows a Binomial distribution with parameters n (number of trials) and τ (probability of success).

The order statistics $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ are the sorted values of the sample.
If a and b are integers satisfying $1 \leq a < b \leq n$, consider the event:

$$a \leq g(X_1, \dots, X_n, \theta) \leq b$$

This event can be translated in terms of order statistics:

$$a \leq \sum_{i=1}^{n} I(X_i \leq \theta) \leq b \iff X_{(a)} \leq \theta \leq X_{(b)}$$

Explanation:

- $\sum_{i=1}^{n} I(X_i \leq \theta) \geq a$ implies that at least a observations are less than or equal to θ, meaning θ is at least as large as the a-th smallest observation $X_{(a)}$.
- Similarly, $\sum_{i=1}^{n} I(X_i \leq \theta) \leq b$ implies that at most b observations are less than or equal to θ, meaning θ is no larger than the b-th smallest observation $X_{(b)}$.

Thus, the interval $[X_{(a)}, X_{(b)}]$ serves as a confidence interval for θ.

We seek:

$$P(X_{(a)} \leq \theta \leq X_{(b)}) = P(a \leq g(X_1, \dots, X_n, \theta) \leq b) = p$$

where p is the desired confidence level (e.g., 95%).

Given that $g \sim \text{Binomial } (n, \tau)$, the probability p is:

$$p = \sum_{k=a}^{b} \binom{n}{k} \tau^k (1-\tau)^{n-k}$$

This sum represents the probability that the Binomial random variable falls between a and b, inclusive.

Thus, the interval $[X_{(a)}, X_{(b)}]$ is a $100p\%$ confidence interval for θ, independent of the underlying distribution F (hence, distribution-free).

When n is sufficiently large, the Binomial distribution $\text{Binomial}\ (n, \tau)$ can be approximated by a Normal distribution:

$$\text{Binomial}\ (n, \tau) \approx \mathcal{N}(\mu, \sigma^2)$$

where:

$$\mu = n\tau \ \text{ and } \ \sigma^2 = n\tau(1 - \tau)$$

The Binomial distribution is discrete, while the Normal distribution is continuous. To improve the approximation, especially for smaller n, we apply a continuity correction of ± 0.5.

$$p \approx P(a \leq \text{Binomial}\ (n, \tau) \leq b)$$

using the Normal approximation, solve for a and b such that:

$$P\left(\mu - z_p\sigma - \frac{1}{2} \leq \text{Binomial}\ (n, \tau) \leq \mu + z_p\sigma + \frac{1}{2}\right) \approx p$$

Rearranging, we obtain:

$$\begin{aligned} a &\approx \mu - z_p\sigma - \frac{1}{2} = n\tau - z_p\sqrt{n\tau(1 - \tau)} - \frac{1}{2} \\ b &\approx \mu + z_p\sigma + \frac{1}{2} = n\tau + z_p\sqrt{n\tau(1 - \tau)} + \frac{1}{2} \end{aligned}$$

Since a and b must be integers, we round them appropriately:

$$a = \left\lfloor n\tau + \frac{1}{2} - z_p\sqrt{n\tau(1 - \tau)} \right\rfloor$$
$$b = \left\lceil n\tau - \frac{1}{2} + z_p\sqrt{n\tau(1 - \tau)} \right\rceil$$

where:

- $\lfloor x \rfloor$ denotes the floor of x (rounding down to the nearest integer).
- $\lceil x \rceil$ denotes the ceiling of x (rounding up to the nearest integer).

Example of Rounding:

$$\lfloor 3.5 \rfloor = 3 \ \text{ and } \ \lceil 3.5 \rceil = 4$$

Example: Comparison of CIs for the normal mean

- **Model:** $X_1, \cdots, X_n$ independent $\mathcal{N}(\mu, \sigma^2)$ random variables.
  - Note that $\mu = E(X_i) = F^{-1}(1/2)$.
- Compare the two CIs:

$$\bar{X} \pm t_p \frac{S}{\sqrt{n}} \quad \text{versus} \quad [X_{(a)}, X_{(b)}].$$

- Compare the lengths of the two intervals when n is large:
  - $t_p \to z_p$ and $S \xrightarrow{p} \sigma$ so that length of parametric CI is approximately $2 z_p \sigma / \sqrt{n}$.
  - $X_{(b)} - X_{(a)} \approx z_p \sigma \sqrt{2\pi} / \sqrt{n}$.
- Thus

$$\frac{\text{length of dist. free CI}}{\text{length of parametric CI}} \approx \frac{\sqrt{2\pi}}{2} = 1.253$$

- **Tradeoff:** Distribution free CI is always valid but parametric CI will be shorter if that model is correct.

Example: CIs for the normal mean: $n = 20$

- Compare $\bar{X} \pm 2.093 \times S/\sqrt{n}$ to $[X_{(6)}, X_{(14)}]$.
  - Coverage for the latter CI is $\sum_{k=6}^{14} \binom{20}{k} \frac{1}{2^{20}} = 0.9586$.
- R code:

```
> len.t <- NULL
> len.df <- NULL
> for (i in 1:1000) {
+     x <- sort(rnorm(20))
+     len.t <- c(len.t, 2*qt(0.975,19)*sqrt(var(x)/20))
+     len.df <- c(len.df,x[14]-x[6])
+     }
> sum(len.t<len.df)/1000 # proportion of samples where len.t < len.df
[1] 0.664
> mean(len.t)
[1] 0.9239648
> mean(len.df)
[1] 1.029677
> mean(len.df/len.t)
[1] 1.114631
```

Example: Gamma pivot

- Sometimes, we may not be able to find pivotal quantities based on a point estimator $\hat{\theta}$.
- For samples from a continuous distribution with only **one unknown parameter**, we can always find at least one pivotal value.
- If X is continuous with CDF $F(x)$ and $U = F(X)$, then $U \sim \text{Unif}(0, 1)$.
- Suppose that $U \sim \text{Unif}(0, 1)$ and let $Y = -\ln U$. Therefore $y = g(u) = -\ln u$, $u = g^{-1}(y) = e^{-y}$. Note the range of Y here is $(0, \infty)$. The density of Y becomes

$$f_Y(y) = f_U(e^{-y})|-e^{-y}| = e^{-y} \quad \text{for } y \geq 0.$$

  That is $Y \sim \text{Exp}(1)$.
- For continuous X, we have $Y = -\ln F(X) \sim \text{Exp}(1)$.

Maximum likelihood estimation

- **Model:** $(X_1, \cdots, X_n)$ random variables with joint pdf or pmf $f(x_1, \cdots, x_n; \theta_1, \cdots, \theta_k)$ where $\theta_1, \cdots, \theta_k$ are unknown parameters.
- Given the data $x_1, \cdots, x_n$, we can define the **likelihood function**

$$\mathcal{L}(\theta_1, \cdots, \theta_k) = f(\underbrace{x_1, \cdots, x_n}_{data}; \theta_1, \cdots, \theta_k)$$

  which is a function over the parameter space (for fixed $x_1, \cdots, x_n$).
- **Definition:** Suppose that for each $\mathbf{x} = (x_1, \cdots, x_n)$, $(T_1(\mathbf{x}), \cdots, T_k(\mathbf{x}))$ maximize $\mathcal{L}(\theta_1, \cdots, \theta_k)$. Then **maximum likelihood estimators** (MLEs) of $\theta_1, \cdots, \theta_k$ are

$$\widehat{\theta}_j = T_j(X_1, \cdots, X_n) \text{ for } j = 1, \cdots, k.$$

Example: A non-regular model

- **Model:** $X_1, \cdots, X_n$ independent random variables with pdf

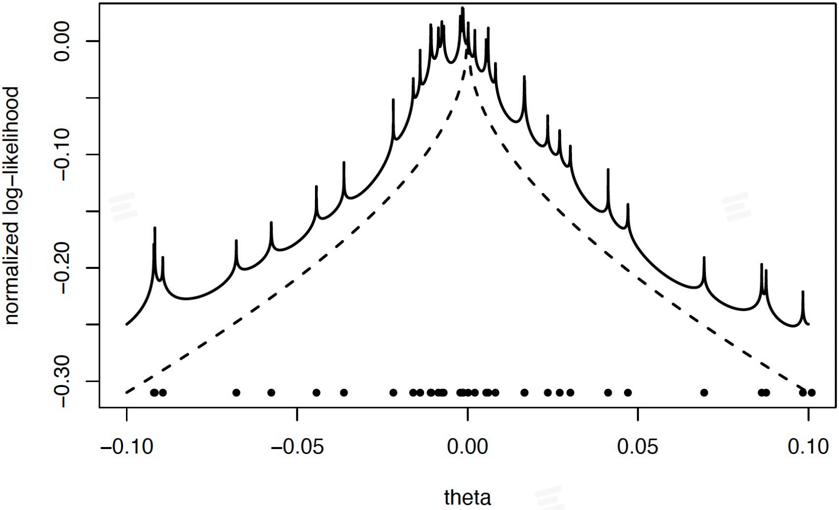
$$f(x;\theta) = \frac{|x-\theta|^{-1/2}}{2\sqrt{\pi}} \exp\left(-|x-\theta|\right).$$

- The likelihood function is

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \left\{ \frac{|x_i-\theta|^{-1/2}}{2\sqrt{\pi}} \exp\left(-|x_i-\theta|\right) \right\}.$$

- Note that as $\theta \to$ any x_i, $\mathcal{L}(\theta) \uparrow \infty$.
  - This suggests that either the MLE does not exist or that any X_i is an MLE of θ.
  - However, the likelihood function still provides a lot of information about θ.
- **Illustration:** Generate 100 observations from the model with $\theta = 0$ and plot $\ln \mathcal{L}(\theta)$.

Plot of log-likelihood function



- $\mathcal{L}(x_i) = \infty$ for $i = 1, \cdots, 100$.
- Note that $\mathcal{L}(\theta)$ is generally largest for θ close to 0.

Example: The Neyman-Scott problem

- **Model:** $(X_1, Y_1), \cdots, (X_n, Y_n)$ independent pairs of independent Normal random variables.
  - Within each pair X_i and Y_i are independent $\mathcal{N}(\mu_i, \sigma^2)$ random variables.
- **Context:** X_i and Y_i are independent measurements of the same quantity.
  - σ represents the standard deviation of the measurement error.
- Unknown parameters: $\mu_1, \cdots, \mu_n$ and σ^2.
  - The number of unknown parameters $(n + 1)$ tends to infinity with n.
  - For each i, we have only two observations to estimate μ_i.
  - More information for estimating σ^2.

ML estimation of σ^2 and $\{\mu_i\}$

- Given data $(x_1, y_1), \cdots, (x_n, y_n)$, the likelihood function is

$$\mathcal{L}(\mu_1, \cdots, \mu_n, \sigma) = \prod_{i=1}^{n} \left\{ \frac{1}{2\pi\sigma^2} \exp\left[-\frac{(x_i - \mu_i)^2 + (y_i - \mu_i)^2}{2\sigma^2} \right] \right\}.$$

- $\widehat{\sigma}$ and $\{\widehat{\mu}_i\}$ are solutions to the following equations (**likelihood equations**):

$$\begin{aligned} \frac{1}{\widehat{\sigma}^2}(x_i + y_i - 2\widehat{\mu}_i) &= 0 \quad (i = 1, \cdots, n) \\ -\frac{2n}{\widehat{\sigma}} + \frac{1}{\widehat{\sigma}^3} \sum_{i=1}^{n} \left[(x_i - \widehat{\mu}_i)^2 + (y_i - \widehat{\mu}_i)^2 \right] &= 0. \end{aligned}$$

- MLEs:

$$\begin{aligned} \widehat{\mu}_i &= \frac{X_i + Y_i}{2} \quad (i = 1, \cdots, n) \\ \widehat{\sigma}^2 &= \frac{1}{4n} \sum_{i=1}^{n} (X_i - Y_i)^2. \end{aligned}$$

Inconsistency of $\widehat{\sigma}^2$

- Note that $X_i - Y_i \sim \mathcal{N}(0, 2\sigma^2)$ since X_i and Y_i are independent.
- Thus by the WLLN, we have

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i) \xrightarrow{p} 2\sigma^2.$$

- Thus for the MLE of σ^2, we have

$$\widehat{\sigma}^2 = \frac{1}{4n}\sum_{i=1}^{n}(X_i - Y_i)^2 \xrightarrow{p} \frac{\sigma^2}{2}.$$

- Thus the MLE of σ^2 is not consistent.

Sufficiency

- Suppose we have $(X_1, \cdots, X_n)$ with joint pdf or pmf $f(x_1, \cdots, x_n; \theta)$.
  - θ may consist of k parameters so that $\theta = (\theta_1, \cdots, \theta_k)$.
- **Question:** Can we reduce $\mathbf{X} = (X_1, \cdots, X_n)$ to $\mathbf{T} = (T_1(\mathbf{X}), \cdots, T_m(\mathbf{X}))$ without losing any information about θ?
  - Ideally, we'd like m much smaller than n.
- For a given $\mathbf{T}$, look at the conditional distribution of $\mathbf{X}$ given $\mathbf{T} = \mathbf{t}$:
  - If this conditional distribution does not depend on θ then $\mathbf{T}$ contains the same information about θ as $\mathbf{X}$.
  - If $\mathbf{X}$ and $\mathbf{T}$ are both discrete, we can write

$$f(\mathbf{x}; \theta) = f_T(\mathbf{t}; \theta) f(\mathbf{x}|\mathbf{t}) \ \text{ where } \mathbf{t} = \mathbf{T}(\mathbf{x}).$$

- **Definition:** A statistic $\mathbf{T} = (T_1(\mathbf{X}), \cdots, T_m(\mathbf{X}))$ is a **sufficient statistic** for θ (or more generally a model) if the conditional distribution of $\mathbf{X}$ given $\mathbf{T} = \mathbf{t}$ depends only on $\mathbf{t}$ (and not θ).

Example: Order statistics

- $X_1, \cdots, X_n$ independent continuous random variables with unknown cdf F.
  - F is the parameter in this model.
- **Claim:** The order statistics $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ are a sufficient statistic for F.
- **Proof:** Look at the conditional distribution of $(X_1, \cdots, X_n)$ given $X_{(1)} = x_1, \cdots, X_{(n)} = x_n$ (where $x_1 \leq x_2 \leq \cdots \leq x_n$).
  - If $(y_1, \cdots, y_n)$ is a permutation of $(x_1, \cdots, x_n)$ then

$$P(X_1 = y_1, \cdots, X_n = y_n | X_{(1)} = x_1, \cdots, X_{(n)} = x_n) = \frac{1}{n!}$$

since all permutations of the ordered sample are equally likely given the order statistics.

$$\sum_{all\ \mathbf{y}} P(X_1 = y_1, \cdots, X_n = y_n | X_{(1)} = x_1, \cdots, X_{(n)} = x_n) = 1.$$

Sufficiency and the likelihood function

- **Neyman Factorization Theorem:** Suppose that the joint pdf or pmf of $\mathbf{X} = (X_1, \cdots, X_n)$ is $f(\mathbf{x}; \theta)$. Then the statistic $\mathbf{T} = (T_1(\mathbf{X}), \cdots, T_m(\mathbf{X}))$ is a sufficient statistic for θ if, and only if

$$f(\mathbf{x}; \theta) = g(\mathbf{T}(\mathbf{x}); \theta)h(\mathbf{x})$$

  where the function h does not depend on θ.
- Note that the likelihood function is

$$\begin{aligned} \mathcal{L}(\theta) &= f(\mathbf{x}; \theta) \\ &= g(\mathbf{T}(\mathbf{x}); \theta)h(\mathbf{x}). \end{aligned}$$

- Since $h(\mathbf{x})$ does not depend on θ, it is just a multiplicative constant in the likelihood function:
  - Maximizing $\mathcal{L}(\theta)$ is equivalent to maximizing $g(\mathbf{T}(\mathbf{x}); \theta)$.
  - Effectively, $\mathcal{L}(\theta)$ depends on the data $\mathbf{x}$ only through the value of $\mathbf{T}(\mathbf{x})$.
  - If the MLE is unique, it depends on $\mathbf{X}$ only through $\mathbf{T}(\mathbf{X})$.

Computing MLEs

- Likelihood function $\mathcal{L}(\theta)$.
  - Assume for simplicity that θ is real-valued.
- **Question:** How do we find $\widehat{\theta}$ maximizing $\mathcal{L}(\theta)$?
- Two general scenarios:
  1. $\mathcal{L}(\theta)$ is differentiable and the parameter space Θ is an open set (i.e. every point of Θ is an interior point). Then $\widehat{\theta}$ (if it exists) satisfies the **likelihood equation**
  $$\frac{d}{d\theta} \ln \mathcal{L}(\widehat{\theta}) = 0.$$
  In some cases, we can use the 2nd derivative to estimate the standard error.
  2. $\widehat{\theta}$ occurs at a “boundary”:
     - Boundary of Θ (if Θ is not an open set);
     - an extreme of the data (e.g. $\widehat{\theta} = X_{(n)}$).

  In these cases, we need to directly maximize $\mathcal{L}(\theta)$.

Example: Uniform distribution on $[0, \theta]$

- **Model:** $X_1, \cdots, X_n$ independent Unif$(0, \theta)$ random variables with $\theta > 0$ unknown:

$$
\begin{aligned}
f(x;\theta) &= \begin{cases} \theta^{-1} & \text{if } 0 \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases} \\
&= \frac{1}{\theta} I(0 \leq x \leq \theta).
\end{aligned}
$$

- The likelihood function is

$$
\begin{aligned}
\mathcal{L}(\theta) &= \prod_{i=1}^{n} \left\{ \frac{1}{\theta} I(0 \leq x_i \leq \theta) \right\} \\
&= \frac{1}{\theta^n} \prod_{i=1}^{n} I(0 \leq x_i \leq \theta) \\
&= \frac{1}{\theta^n} I(\theta \geq \max\{x_1, \cdots, x_n\}).
\end{aligned}
$$

Example: Uniform distribution on $[0, \theta]$ (cont'd)

- Thus $\mathcal{L}(\theta) = 0$ for $\theta < \max\{x_1, \cdots, x_n\}$ while for $\theta \geq \max\{x_1, \cdots, x_n\}$, $\mathcal{L}(\theta) = \theta^{-n}$, which is decreasing as θ increases.
- Thus $\mathcal{L}(\theta)$ is maximized at $\max\{x_1, \cdots, x_n\}$ and so the MLE of θ is

$$\widehat{\theta} = \max\{X_1, \cdots, X_n\} = X_{(n)}.$$

- **Question:** What happens if we define $f(x; \theta) = \theta^{-1}$ for $0 < x < \theta$?
  - Then $\mathcal{L}(\theta) = 0$ for $\theta \leq \max\{x_1, \cdots, x_n\}$ and $\mathcal{L}(\theta) = \theta^{-n}$ for $\theta > \max\{x_1, \cdots, x_n\}$.
  - Technically, the MLE doesn't exist!
  - However, the likelihood function is largest for θ close to (and greater than) $\max\{x_1, \cdots, x_n\}$.

Example: Geometric distribution

- **Model:** $X_1, \cdots, X_n$ are independent Geometric(θ) random variables:
$$f(x;\theta) = \theta(1-\theta)^x \text{ for } x = 0, 1, 2, \cdots$$
where $0 < \theta < 1$.
- The likelihood and log-likelihood functions are
$$\begin{aligned} \mathcal{L}(\theta) &= \prod_{i=1}^{n} \{\theta(1-\theta)^{x_i}\} \\ \ln \mathcal{L}(\theta) &= n \ln(\theta) + \ln(1-\theta) \sum_{i=1}^{n} x_i. \end{aligned}$$
- Differentiating (with respect to θ), we get the likelihood equation
$$\frac{n}{\widehat{\theta}} - \frac{1}{1-\widehat{\theta}} \sum_{i=1}^{n} x_i = 0.$$

Example: Geometric distribution (cont'd)

- Provided that $\sum_{i=1}^{n} x_i > 0$, the likelihood equation has a unique solution

$$\widehat{\theta} = \frac{1}{1 + \bar{x}}.$$

- Does this maximize the log-likelihood function? Check second derivative:

$$\frac{d^2}{d\theta^2} \ln \mathcal{L}(\theta) = -\frac{n}{\theta^2} - \frac{1}{(1-\theta)^2} \sum_{i=1}^{n} x_i < 0.$$

- Therefore, the MLE is $\widehat{\theta} = (1 + \bar{X})^{-1}$ (provided $\bar{X} > 0$).

Example: Geometric distribution (cont'd)

- **Question:** How can we estimate the standard error of $\widehat{\theta} = (1 + \bar{X})^{-1}$?
  - We could use the Delta Method – we need to find $\mathrm{Var}_\theta(X_i)$.
  - We can also base an estimate on the second derivative of the log-likelihood.
- Given the MLE $\widehat{\theta}$, we define the **observed Fisher information** as

$$-\frac{d^2}{d\theta^2} \ln \mathcal{L}(\widehat{\theta}).$$

- We can then estimate the standard error of $\widehat{\theta}$ by

$$\widehat{\mathrm{se}}(\widehat{\theta}) = \left\{-\frac{d^2}{d\theta^2} \ln \mathcal{L}(\widehat{\theta})\right\}^{-1/2} = \left(\frac{\bar{X}}{n(1 + \bar{X})^3}\right)^{1/2}.$$

- **Question:** Why does this work?

Example: Exponential distribution

- **Model:** $X_1, \cdots, X_n$ are independent Exponential random variables with pdf
$$f(x; \lambda) = \lambda \exp(-\lambda x) \text{ for } x \geq 0$$
where $\lambda > 0$ is unknown.
- The log-likelihood function is
$$\ln \mathcal{L}(\lambda) = n \ln(\lambda) - \lambda \sum_{i=1}^{n} x_i.$$
- Differentiating we get the likelihood equation
$$\frac{n}{\widehat{\lambda}} - \sum_{i=1}^{n} x_i = 0$$
and note that the 2nd derivative is $-n/\lambda^2 < 0$.
- Thus the MLE is $\widehat{\lambda} = 1/\bar{X}$.

Example: Exponential distribution (cont'd)

- We noted in the Geometric example that we could estimate the standard error of the MLE using the observed Fisher information obtained from the 2nd derivative of the log-likelihood function.
- In this case, the observed Fisher information is

$$-\frac{d^2}{d\lambda^2} \ln \mathcal{L}(\widehat{\lambda}) = \frac{n}{\widehat{\lambda}^2}.$$

- Therefore, we obtain the standard error estimator

$$\widehat{\text{se}}(\widehat{\lambda}) = \left(\frac{n}{\widehat{\lambda}^2}\right)^{-1/2} = \frac{\widehat{\lambda}}{\sqrt{n}}.$$

- Note that this is the same as the estimator obtained via the Delta Method.

Example: Gamma distribution

- **Model:** $X_1, \cdots, X_n$ are independent Gamma random variables with pdf
$$f(x;\alpha,\beta) = \frac{x^{\alpha-1}\exp(-x/\beta)}{\beta^\alpha \Gamma(\alpha)} \quad \text{for } x \geq 0$$
where $\alpha, \beta > 0$ are unknown.
- The log-likelihood function is
$$\ln \mathcal{L}(\alpha,\beta) = -n\alpha \ln(\beta) - n \ln \Gamma(\alpha) + (\alpha - 1) \ln \left(\prod_{i=1}^{n} x_i \right) - \frac{1}{\beta} \sum_{i=1}^{n} x_i.$$
- The partial derivatives are
$$\frac{d}{d\alpha} \ln L(\alpha,\beta) = -n \ln(\beta) - n\frac{\Gamma'(\alpha)}{\Gamma(\alpha)} + \ln \left(\prod_{i=1}^{n} x_i \right);$$
$$\frac{d}{d\beta} \ln L(\alpha,\beta) = -\frac{n\alpha}{\beta} + \frac{1}{\beta^2} \sum_{i=1}^{n} x_i.$$

Example: Gamma distribution (cont'd)

- Now, define

$$\psi(\alpha) := \frac{d}{d\alpha} \ln \Gamma(\alpha) = \frac{\Gamma'(\alpha)}{\Gamma(\alpha)} \quad \text{and } \tilde{x}_n = \left(\prod_{i=1}^{n} x_i \right)^{1/n}$$

  where $\psi(\cdot)$ is called the **digamma function** and $\tilde{x}_n$ is the **geometric mean** of $x_1, \ldots, x_n$.

- Setting the partial derivatives to 0, we obtain the maximum likelihood equations

$$\beta = \frac{\overline{x}_n}{\alpha}$$

$$\ln(\alpha) - \psi(\alpha) - \ln(\overline{x}_n / \tilde{x}_n) = 0.$$

- There is no closed form solution for these equations, but they can be solved numerically.

MLE: Invariance property

Invariance property of MLEs

If $\hat{\theta}$ is the MLE of θ and if $u(\theta)$ is a function of θ, then $u(\hat{\theta})$ is an MLE for $u(\theta)$.

- **Examples:**
  1. Let $X_1, \ldots, X_n$ be a random sample from an Exponential distribution with scale parameter β. What is the MLE for estimating $p(\beta) = P(X \geq 1) = e^{-1/\beta}$? Since $\overline{X}_n$ is the MLE for β, we have $\widehat{p(\beta)} = p(\hat{\beta}) = e^{-1/\overline{X}_n}$.
  2. Let $X_1, \ldots, X_n$ be a random sample from a Poisson distribution with parameter $\lambda > 0$. What is the MLE for estimating $p(\lambda) = P(X = 0) = e^{-\lambda}$? Since $\overline{X}_n$ is the MLE for λ, we have $\widehat{p(\lambda)} = p(\hat{\lambda}) = e^{-\overline{X}_n}$.