Stat 135 Lab 7
Midterm Review

GSI: Yutong Wang
Mar 6, 2020

Berkeley
UNIVERSITY OF CALIFORNIA

About COVID-19

George Takei
@GeorgeTakei

Live longer and prosper.

HOW TO PROPERLY GREET SOMEONE DURING THE CORONAVIRUS OUTBREAK

9:32 AM · Mar 5, 2020 · Twitter for iPhone

For students who are in my sections:

I will post lab slides including more detailed explanation starting from this week. So those of you who feel unwell will have access to the lab material.

Also, if you intended but felt unwell to go to my office hours, you are always welcome to post your questions in piazza. Please specifically write [Yutong's OH] in your title or post private questions if you'd prefer, and I'll try to accommodate and answer your questions in Piazza.

Please take care of yourselves, and let us know if you are affected in any way.

Berkeley
UNIVERSITY OF CALIFORNIA

Proper handwashing for 20 sec

Take 90 seconds and watch this short video:

WHO: How to handwash? With soap and water

This is something we can all do! 🧼

🐻 Review Part 1: Estimation

- Estimation
  - Method of moments (MoM) 8.4
  - Maximum likelihood estimator (MLE) 8.5
    - Large sample theory 8.5.2
    - Confidence interval 8.5.3
  - Difference between the estimators
    - Efficiency, CRLB 8.7
    - Sufficiency, Rao-Blackwell theorem 8.8
  - Approximation using Delta method 4.6
- Hypothesis testing

MoM

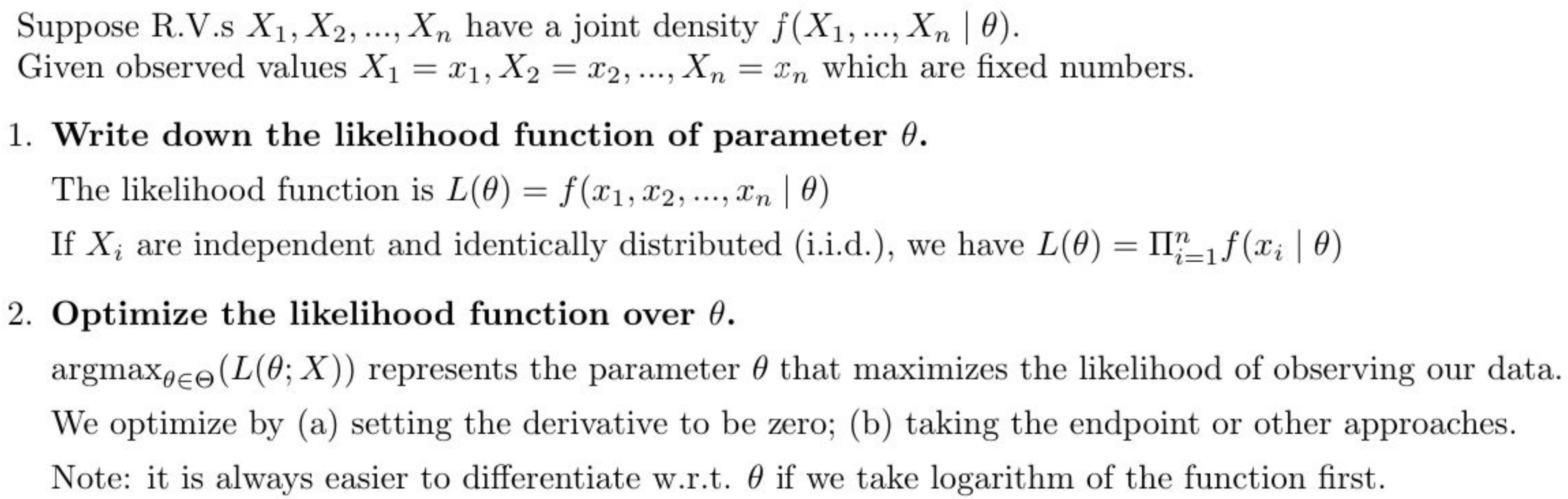
- The k-th **population moment** of a probability law is a function of parameters, defined as $\mu_k = E(X^k)$
- The k-th **sample moment** is the average of k-th powers of the sample, defined as $\hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$
- **To calculate MoM**:

1. Calculate the first k population moments in terms of the parameters
2. Invert the expressions in step 1 and express the parameters as a function of the population moments
3. Plug in the sample moments in place of the population moments to obtain MoM estimates of the parameters

🐻 Review Part 1: Estimation

- Estimation
  - Method of moments (MoM) 8.4
  - Maximum likelihood estimator (MLE) 8.5
    - Large sample theory 8.5.2
    - Confidence interval 8.5.3
  - Difference between the estimators
    - Efficiency, CRLB 8.7
    - Sufficiency, Rao-Blackwell theorem 8.8
  - Approximation using Delta method 4.6
- Hypothesis testing

MLE



Suppose R.V.s $X_1, X_2, ..., X_n$ have a joint density $f(X_1, ..., X_n \mid \theta)$.
Given observed values $X_1 = x_1, X_2 = x_2, ..., X_n = x_n$ which are fixed numbers.

1. **Write down the likelihood function of parameter θ.**

   The likelihood function is $L(\theta) = f(x_1, x_2, ..., x_n \mid \theta)$

   If X_i are independent and identically distributed (i.i.d.), we have $L(\theta) = \Pi_{i=1}^{n} f(x_i \mid \theta)$

2. **Optimize the likelihood function over θ.**

   $\text{argmax}_{\theta \in \Theta}(L(\theta; X))$ represents the parameter θ that maximizes the likelihood of observing our data.

   We optimize by (a) setting the derivative to be zero; (b) taking the endpoint or other approaches.

   Note: it is always easier to differentiate w.r.t. θ if we take logarithm of the function first.

Adam's note: Don't compute the log right away - first look at the likelihood itself and see if it's easy to maximize directly.

Properties of MLE

- Under appropriate smoothness conditions on f , the MLE from an i.i.d. sample is consistent.
- As the sample size n goes to infinity, The large sample distribution of MLE is approximately normal with mean θ_0 and variance $\frac{1}{nI(\theta_0)}$.
  Thus, MLE is asymptotically unbiased, and asymptotically normal, with asymptotic variance equal to $\frac{1}{nI(\theta_0)}$, where $I(\theta)$ is the Fisher information.
- Fisher information:

$$I(\theta) = \mathbb{E}\left[\frac{\partial}{\partial\theta}\log f(X \mid \theta)\right]^2 = -E\left[\frac{\partial^2}{\partial\theta^2}\log f(X \mid \theta)\right]$$

🐻 Review Part 1: Estimation

- Estimation
  - Method of moments (MoM) 8.4
  - Maximum likelihood estimator (MLE) 8.5
    - Large sample theory 8.5.2
    - Confidence interval 8.5.3
  - Difference between the estimators
    - Efficiency, CRLB 8.7
    - Sufficiency, Rao-Blackwell theorem 8.8
  - Approximation using Delta method 4.6
- Hypothesis testing

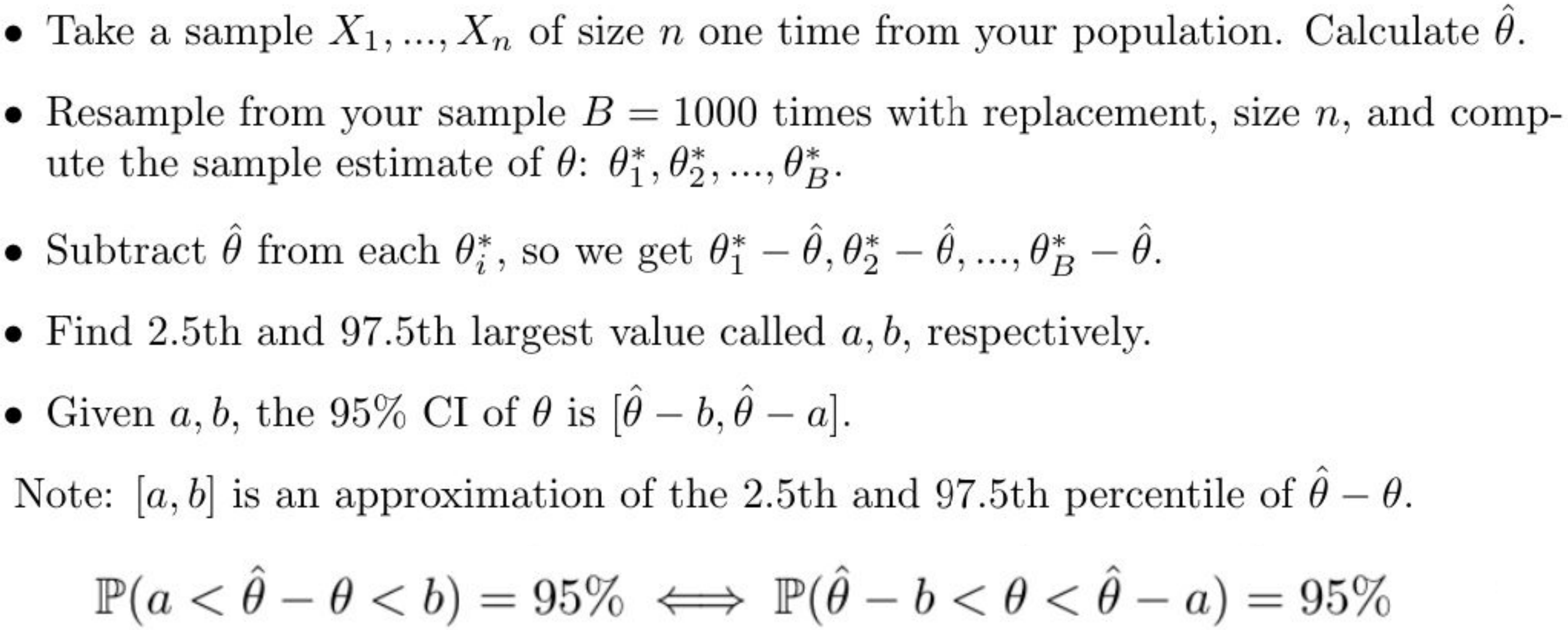
Confidence intervals

1. Exact methods
2. Approximation based on large sample theory of MLE
3. Bootstrap confidence intervals (Adam’s lecture notes in lecture 5, 10)

👀 What is the difference between parametric and nonparametric bootstrap?

Non-parametric bootstrap

Adam’s lecture notes #10: resample with replacement



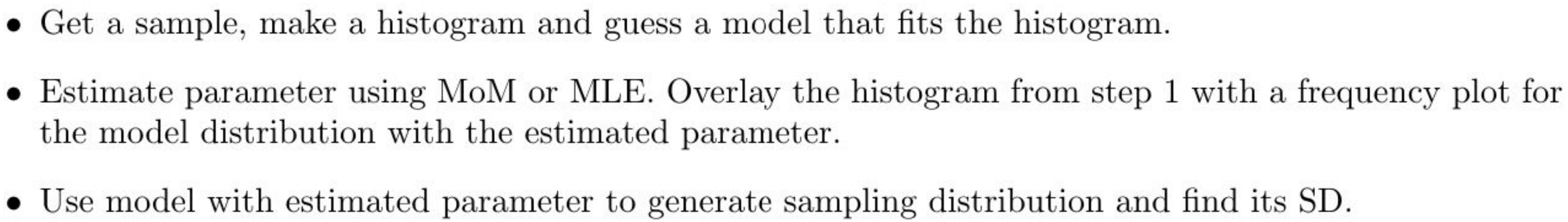
- Take a sample $X_1, ..., X_n$ of size n one time from your population. Calculate $\hat{\theta}$.
- Resample from your sample $B = 1000$ times with replacement, size n, and compute the sample estimate of θ: $\theta_1^*, \theta_2^*, ..., \theta_B^*$.
- Subtract $\hat{\theta}$ from each θ_i^*, so we get $\theta_1^* - \hat{\theta}, \theta_2^* - \hat{\theta}, ..., \theta_B^* - \hat{\theta}$.
- Find 2.5th and 97.5th largest value called a, b, respectively.
- Given a, b, the 95% CI of θ is $[\hat{\theta} - b, \hat{\theta} - a]$.

Note: $[a, b]$ is an approximation of the 2.5th and 97.5th percentile of $\hat{\theta} - \theta$.

$$\mathbb{P}(a < \hat{\theta} - \theta < b) = 95\% \iff \mathbb{P}(\hat{\theta} - b < \theta < \hat{\theta} - a) = 95\%$$

Parametric bootstrap

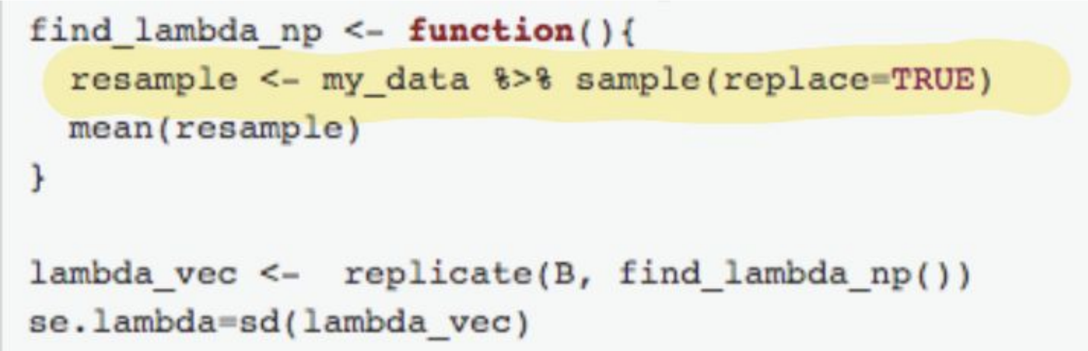
Adam’s lecture notes #10: resample from the model parametrized by the MoM or MLE.



- Get a sample, make a histogram and guess a model that fits the histogram.
- Estimate parameter using MoM or MLE. Overlay the histogram from step 1 with a frequency plot for the model distribution with the estimated parameter.
- Use model with estimated parameter to generate sampling distribution and find its SD.

bootstrap in R (Adam’s lecture note 10)

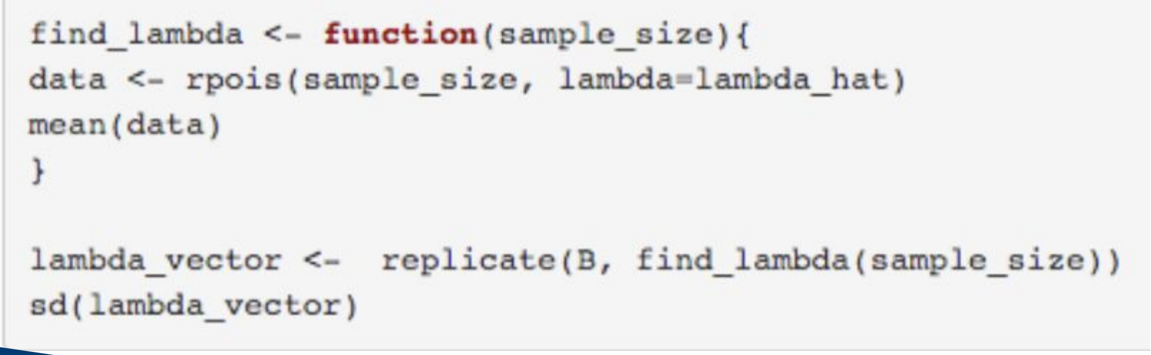
Non-parametric bootstrap:



```r
find_lambda_np <- function(){
  resample <- my_data %>% sample(replace=TRUE)
  mean(resample)
}

lambda_vec <-  replicate(B, find_lambda_np())
se.lambda=sd(lambda_vec)
```

Parametric bootstrap:

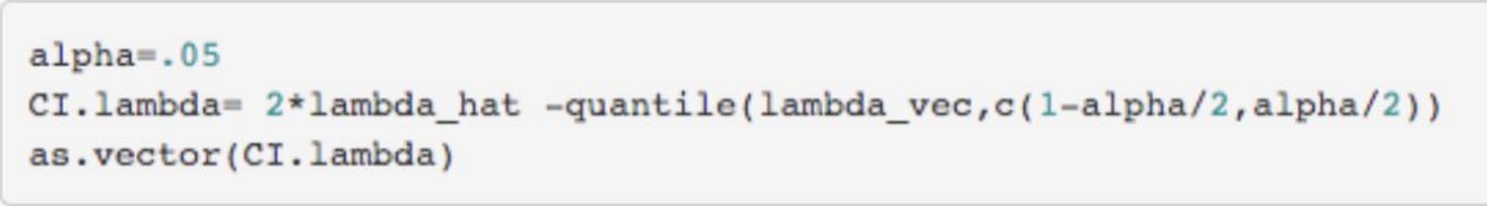


```r
find_lambda <- function(sample_size){
data <- rpois(sample_size, lambda=lambda_hat)
mean(data)
}

lambda_vector <-  replicate(B, find_lambda(sample_size))
sd(lambda_vector)
```

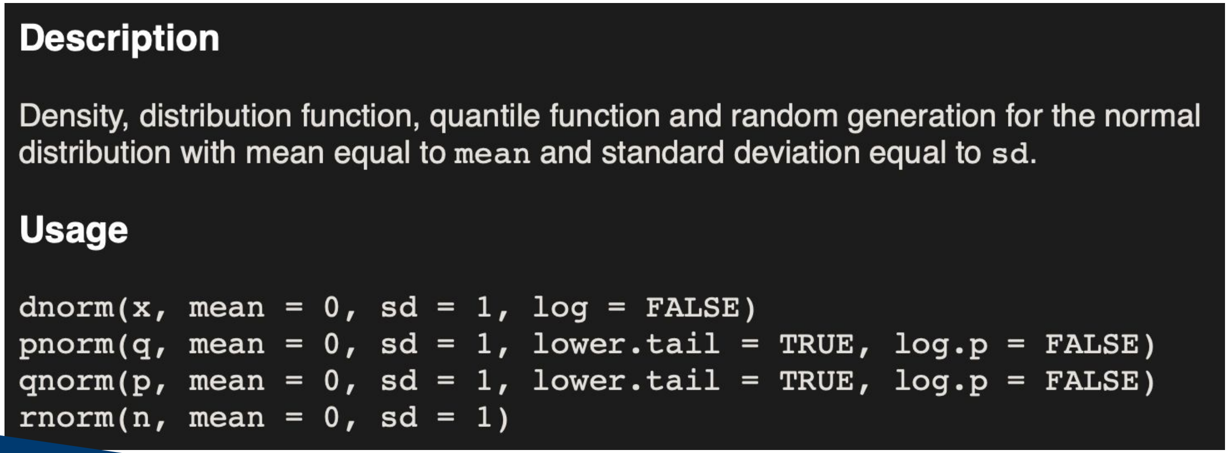
bootstrap in R (contd)

Bootstrap confidence interval:



```
alpha=.05
CI.lambda= 2*lambda_hat -quantile(lambda_vec,c(1-alpha/2,alpha/2))
as.vector(CI.lambda)
```

Distributions in R:



Description

Density, distribution function, quantile function and random generation for the normal distribution with mean equal to `mean` and standard deviation equal to `sd`.

Usage

```
dnorm(x, mean = 0, sd = 1, log = FALSE)
pnorm(q, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
qnorm(p, mean = 0, sd = 1, lower.tail = TRUE, log.p = FALSE)
rnorm(n, mean = 0, sd = 1)
```

🐻 Review Part 1: Estimation

- Estimation
  - Method of moments (MoM) 8.4
  - Maximum likelihood estimator (MLE) 8.5
    - Large sample theory 8.5.2
    - Confidence interval 8.5.3
  - Difference between the estimators
    - Efficiency, CRLB 8.7
    - Sufficiency, Rao-Blackwell theorem 8.8
  - Approximation using Delta method 4.6
- Hypothesis testing

Efficiency/CRLB

- Given two estimates, $\hat{\theta}$ and $\tilde{\theta}$, of a parameter θ, the efficiency of $\hat{\theta}$ relative to $\tilde{\theta}$ is defined to be $\frac{\mathrm{Var}(\tilde{\theta})}{\mathrm{Var}(\hat{\theta})}$.
- Let $X_1, ..., X_n$ be i.i.d. with density function $f(x \mid \theta)$. Let $T = t(X_1, ..., X_n)$ be an unbiased estimate of θ. Then, under smoothness assumptions on $f(x \mid \theta)$,

$$\mathrm{Var}(T) \geq \frac{1}{nI(\theta)}$$

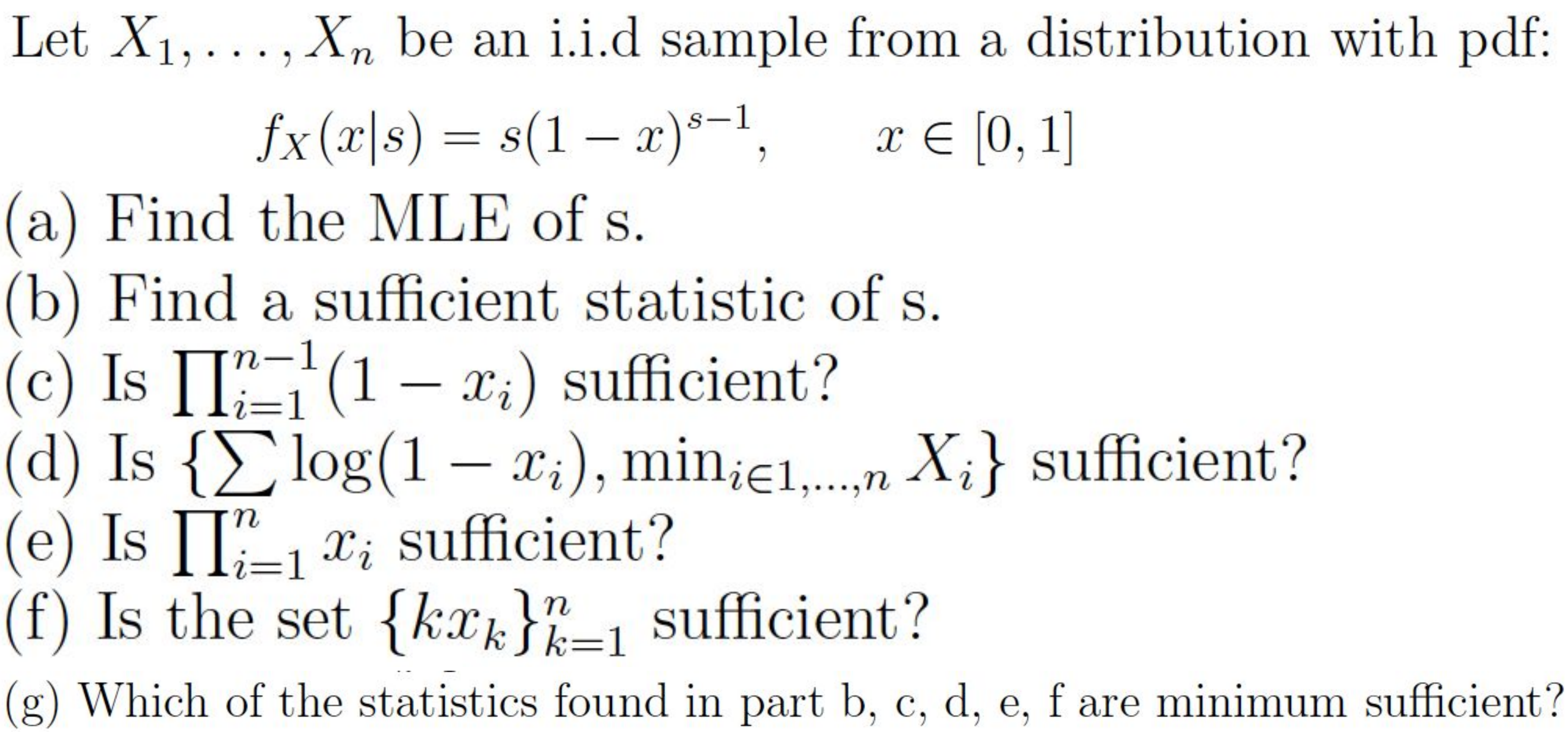
🐻 Review Part 1: Estimation

- Estimation
  - Method of moments (MoM) 8.4
  - Maximum likelihood estimator (MLE) 8.5
    - Large sample theory 8.5.2
    - Confidence interval 8.5.3
  - Difference between the estimators
    - Efficiency, CRLB 8.7
    - Sufficiency, Rao-Blackwell theorem 8.8
  - Approximation using Delta method 4.6
- Hypothesis testing

Sufficiency & Rao-Blackwell

- A **statistic** $T = T(X_1, ..., X_n)$ is a function of the data only. (no parameter θ involved!)
- A statistic $T = T(X_1, ..., X_n)$ is **sufficient** for $\mathcal{P} = \{P_\theta, \forall\theta \in \Theta\}$ if $P_\theta(X_1, ..., X_n \mid T = t)$ does not depend on θ.
- **(Factorization)** T is sufficient for $\mathcal{P}$ if and only if (iff) there exists functions g_θ, h such that $P_\theta(X^n) = g_\theta(T(X^n))h(X^n)$,where X^n denotes $(X_1, X_2, ..., X_n)$.
- If T is sufficient for θ, the maximum likelihood estimate is a function of T.
- *Definition:* T is **minimal sufficient** for $\mathcal{P}$ if (1) T is sufficient, (2) for any sufficient $S = S(X^n)$, there exists f with $T = f(S)$
- *Criterion:* T is **minimal sufficient** for $\mathcal{P}$ iff $\frac{P_\theta(X^n)}{P_\theta(Y^n)}$ does not depend on $\theta \iff T(X^n) = T(Y^n)$
- Rao-Blackwell Theorem: Let $\hat{\theta}$ be an estimator of θ, with $E(\hat{\theta}^2) < \infty$. Suppose that T is sufficient for θ, and let $\tilde{\theta} = \mathbb{E}(\hat{\theta} \mid T)$. Then, for all θ, $\text{MSE}[\tilde{\theta}] \leq \text{MSE}[\hat{\theta}]$.

Problem 1: MLE & Sufficient Statistic



Let $X_1, \ldots, X_n$ be an i.i.d sample from a distribution with pdf:

$$f_X(x|s) = s(1-x)^{s-1}, \qquad x \in [0,1]$$

(a) Find the MLE of s.

(b) Find a sufficient statistic of s.

(c) Is $\prod_{i=1}^{n-1}(1-x_i)$ sufficient?

(d) Is $\{\sum \log(1-x_i), \min_{i\in 1,\ldots,n} X_i\}$ sufficient?

(e) Is $\prod_{i=1}^{n} x_i$ sufficient?

(f) Is the set $\{kx_k\}_{k=1}^{n}$ sufficient?

(g) Which of the statistics found in part b, c, d, e, f are minimum sufficient?

Problem 1 (hints)

- (a) Don’t compute the log right away - first look at the likelihood itself and see if it’s easy to maximize directly.

  If not, we will compute Lik -> logLik -> set derivative to be 0
- (b) Factorization theorem
- (c-f) To check if some statistic T is sufficient, we can
  - Check if T is a function of another sufficient statistic.
  - Apply the corollary about MLE: if T is sufficient for s, MLE is a function of T. i.e., If MLE is not a function of T, T is not sufficient.
  - Remember the whole sample is always sufficient.
  - If none of the above works, can you construct any counterexample to show it is not sufficient?
- (g) Use the criterion of MSS (compute likelihood ratio) to find MSS.

🐻 Review Part 1: Estimation

- Estimation
  - Method of moments (MoM) 8.4
  - Maximum likelihood estimator (MLE) 8.5
    - Large sample theory 8.5.2
    - Confidence interval 8.5.3
  - Difference between the estimators
    - Efficiency, CRLB 8.7
    - Sufficiency, Rao-Blackwell theorem 8.8
  - Approximation using Delta method 4.6
- Hypothesis testing

Berkeley
UNIVERSITY OF CALIFORNIA

Delta method (Propagation of error)

The Central Question is: What is the mean and variance of Y = g(X) for some fixed function g?

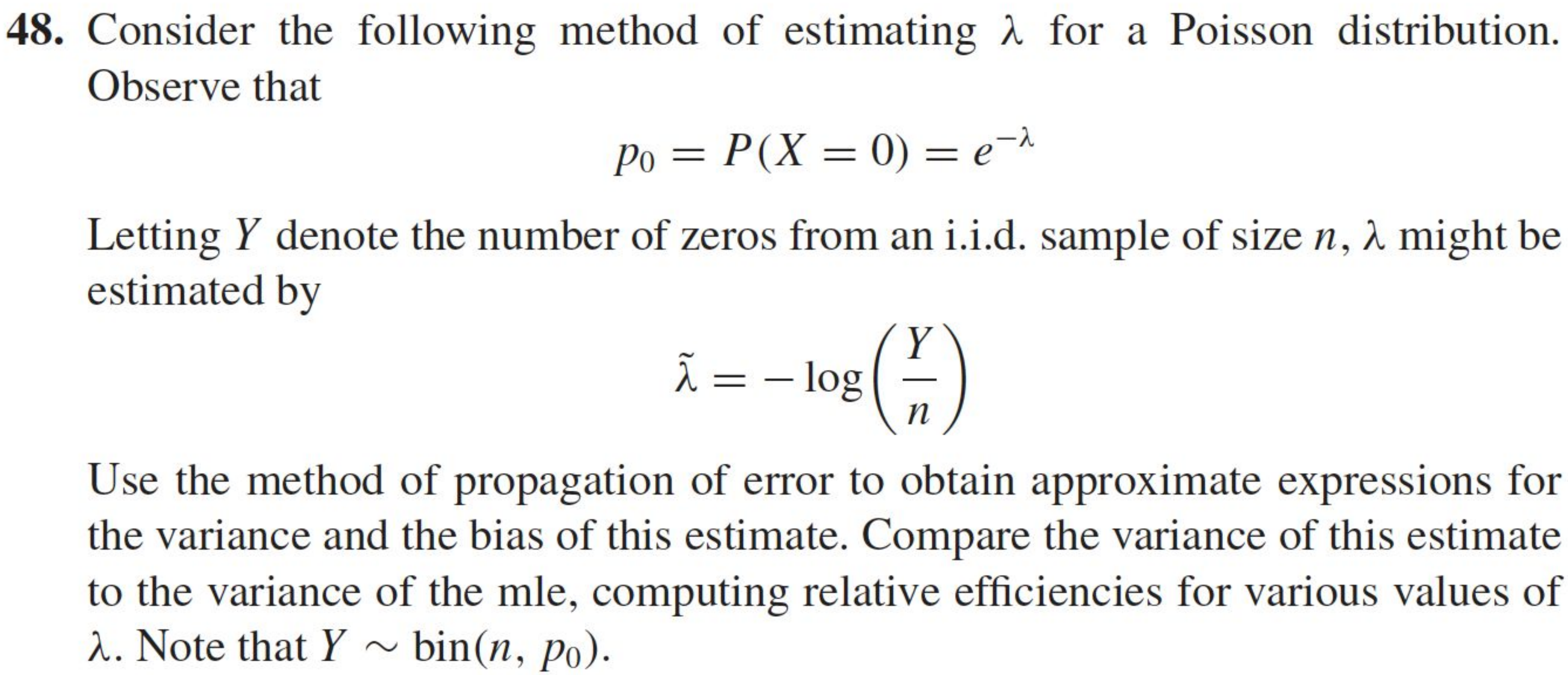
🚫 Note that E(g(X)) is NOT equal to g(E(X))!

✓

$$E(Y) \approx g(\mu_X) + \frac{1}{2}\sigma_X^2 g''(\mu_X)$$

$$\sigma_Y^2 \approx \sigma_X^2 [g'(\mu_X)]^2$$

Example: Delta method (HW 5E)



48. Consider the following method of estimating λ for a Poisson distribution. Observe that

$$p_0 = P(X = 0) = e^{-\lambda}$$

Letting Y denote the number of zeros from an i.i.d. sample of size n, λ might be estimated by

$$\tilde{\lambda} = -\log\left(\frac{Y}{n}\right)$$

Use the method of propagation of error to obtain approximate expressions for the variance and the bias of this estimate. Compare the variance of this estimate to the variance of the mle, computing relative efficiencies for various values of λ. Note that $Y \sim \text{bin}(n, p_0)$.

Example (hints)

- What to estimate: λ for a Poisson distribution
- Estimation methods:
  - MLE
  - $\tilde{\lambda} = -\log\left(\frac{Y}{n}\right)$, which is some nonlinear function of Y.

We have two methods in parallel for estimation. The first one (MLE) is w.r.t. X, and the second one is w.r.t. Y.

We know that $Y \sim \mathrm{Binom}(n, e^{-\lambda})$, then in order to calculate the bias and variance of $\tilde{\lambda}$, are you going to calculate the integration or apply Delta method?

🐻 Review Part 2: Hypothesis testing

- Estimation
- Hypothesis testing
  - Terminology
  - LRT & Neyman Pearson lemma
  - Uniformly most powerful test
  - Generalized LRT
  - Duality of CI and hypothesis testing

Terminology, LRT, Neyman Pearson Lemma

- A **hypothesis** is a statement about the parameter. One hypothesis $H_0 : \theta \in \Theta_0$ is singled out as the **null hypothesis**, and the other complementary one is $H_1 : \theta \in \Theta_1$ as the **alternative hypothesis**.
- Rejecting H_0 when it is true is called a **type I error**.
- The probability of a type I error is called the **significance level** of the test and is usually denoted by $\alpha = \mathbb{P}_0(d(X) = 1)$
- Accepting the null hypothesis when it is false is called a **type II error** and its probability is usually denoted by $\beta = \mathbb{P}_1(d(X) = 0)$
- The probability that the null hypothesis is rejected when it is false is called the **power** of the test, and equals $1 - \beta = 1 - \mathbb{P}_1(d(X) = 0) = \mathbb{P}_1(d(X) = 1)$
- A **test statistic** is a function of your data that leads you to a decision whether to reject or not reject the null hypothesis.
- For some fixed significance level α, the **likelihood ratio test** says: reject H_0 if $\Lambda < c$, where $\Lambda = \frac{P_0(X)}{P_1(X)}$ is the likelihood ratio, and c is some function of α, with $\alpha = P_0(d(X) = 1) = P_0(\Lambda < c)$.
- The **rejection region** is the set of values of the test statistic that leads to rejection of H_0.
- **Neyman-Pearson Lemma**: Suppose that H_0 and H_1 are simple hypotheses and that LRT rejects H_0 with significance level α, then any other level-α test has smaller power.

UMP test & GLRT

- A **simple hypothesis** is one that fully specifies the sampling distribution. (Θ_0 or Θ_1 is a singleton.) If a hypothesis does not completely specify the probability distribution, the hypothesis is called a **composite hypothesis**.

- If the null H_0 is simple and the alternative H_1 is composite, a test that is most powerful for every simple alternative in H_1 is said to be **uniformly most powerful**.

- The likelihood ratio test is optimal for testing a simple hypothesis versus a simple hypothesis. And **generalized LRT** is used when the hypotheses are not simple.

- Suppose $H_0 : \theta \in \Theta_0, H_1 : \theta \in \Theta_1$, where $\Theta_0 \cap \Theta_1 = \emptyset, \Omega = \Theta_0 \cup \Theta_1$. Define

$$\Lambda = \frac{\max_{\theta \in \Theta_0}[\text{Lik}(\theta)]}{\max_{\theta \in \Omega}[\text{Lik}(\theta)]}$$

  , where $\max_{\theta \in \Omega}[\text{Lik}(\theta)] = \text{Lik}(\hat{\theta}_{\text{ML}})$.

- GLRT: reject H_0 if $\Lambda < c$, where c is some function of α.

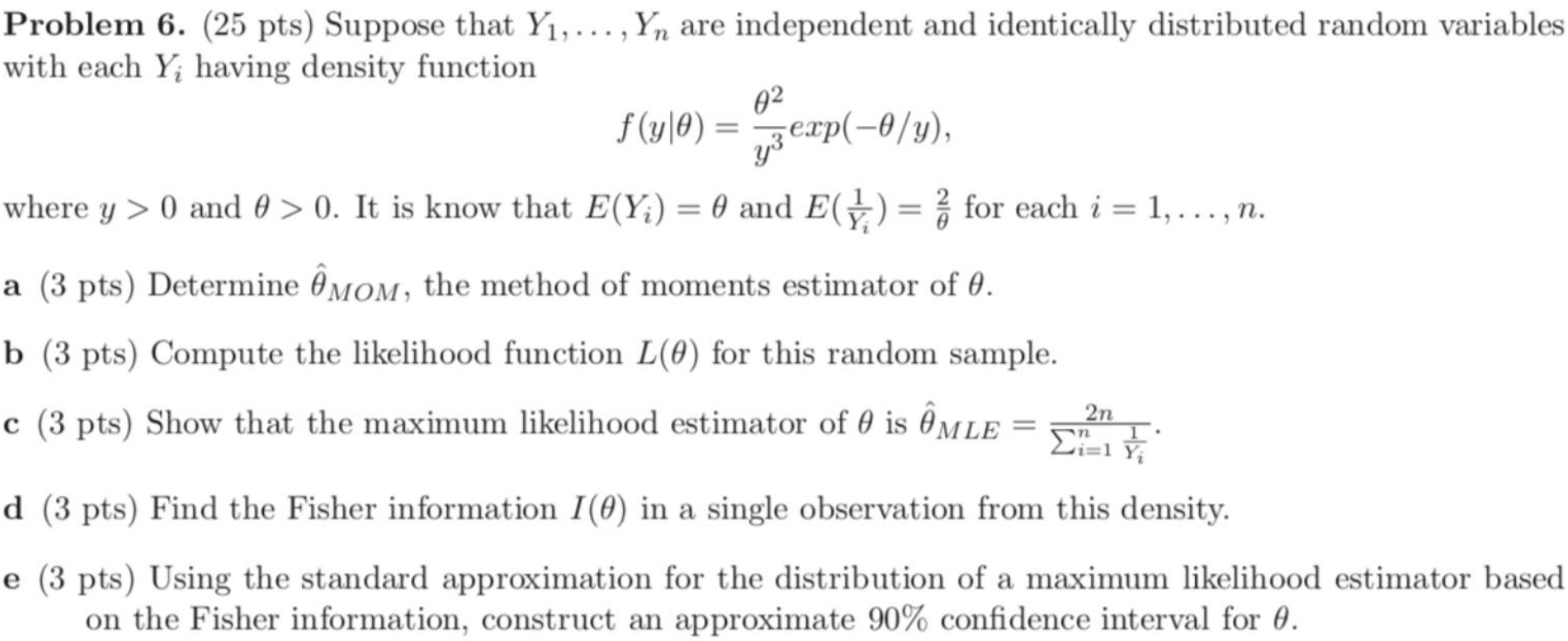
- **p-value** is the probability of getting a test statistic as or more extreme as what you observed, given the null hypothsis being true.

Applications of LRT
(from Adam’s Midterm Review Notes)

- Tests for the population mean
  - Large sample size: normal approximation, no matter what the population distribution is
  - Small sample size: What is the population distribution?
    - If it is normal distribution, is the variance known?
    - What if it is dichotomous?
    - If neither, what facts do we know about the population distribution?
    - Otherwise, we at least still have Bootstrap!!

      🥾👢🥾👢🥾👢

Problem 2



Problem 6. (25 pts) Suppose that $Y_1, \ldots, Y_n$ are independent and identically distributed random variables with each Y_i having density function

$$f(y|\theta) = \frac{\theta^2}{y^3} exp(-\theta/y),$$

where $y > 0$ and $\theta > 0$. It is know that $E(Y_i) = \theta$ and $E(\frac{1}{Y_i}) = \frac{2}{\theta}$ for each $i = 1, \ldots, n$.

a (3 pts) Determine $\hat{\theta}_{MOM}$, the method of moments estimator of θ.

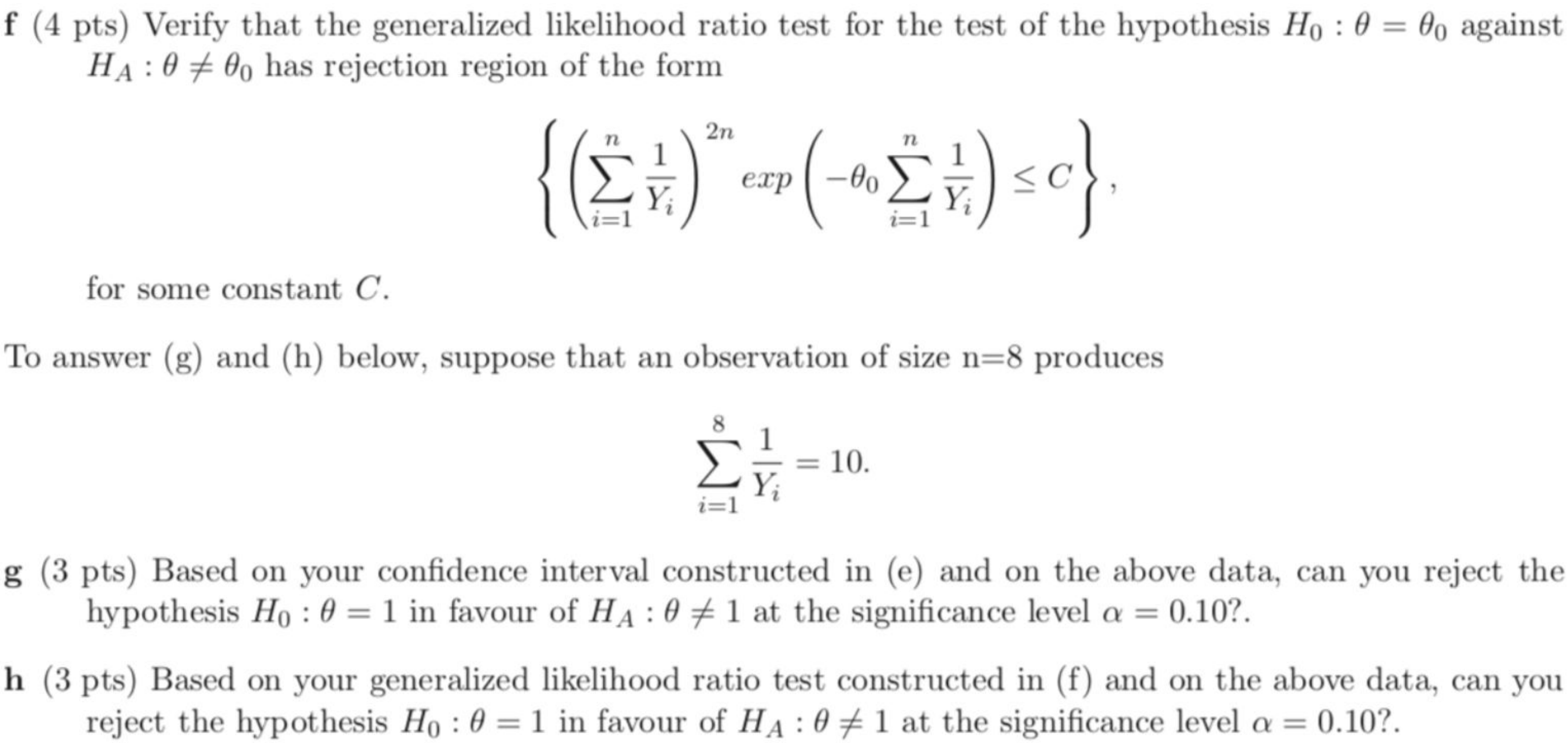
b (3 pts) Compute the likelihood function $L(\theta)$ for this random sample.

c (3 pts) Show that the maximum likelihood estimator of θ is $\hat{\theta}_{MLE} = \frac{2n}{\sum_{i=1}^{n} \frac{1}{Y_i}}$.

d (3 pts) Find the Fisher information $I(\theta)$ in a single observation from this density.

e (3 pts) Using the standard approximation for the distribution of a maximum likelihood estimator based on the Fisher information, construct an approximate 90% confidence interval for θ.

Problem 2 (contd)



f (4 pts) Verify that the generalized likelihood ratio test for the test of the hypothesis $H_0 : \theta = \theta_0$ against $H_A : \theta \neq \theta_0$ has rejection region of the form

$$\left\{ \left(\sum_{i=1}^{n} \frac{1}{Y_i} \right)^{2n} exp\left(-\theta_0 \sum_{i=1}^{n} \frac{1}{Y_i} \right) \leq C \right\},$$

for some constant C.

To answer (g) and (h) below, suppose that an observation of size n=8 produces

$$\sum_{i=1}^{8} \frac{1}{Y_i} = 10.$$

g (3 pts) Based on your confidence interval constructed in (e) and on the above data, can you reject the hypothesis $H_0 : \theta = 1$ in favour of $H_A : \theta \neq 1$ at the significance level $\alpha = 0.10$?.

h (3 pts) Based on your generalized likelihood ratio test constructed in (f) and on the above data, can you reject the hypothesis $H_0 : \theta = 1$ in favour of $H_A : \theta \neq 1$ at the significance level $\alpha = 0.10$?.

Problem 2 (hints)

- (e) Approximated CI is

$$[\hat{\theta}_{\text{MLE}} - z_{\alpha/2} \cdot \frac{1}{\sqrt{nI(\hat{\theta}_{\text{MLE}})}}, \hat{\theta}_{\text{MLE}} + z_{\alpha/2} \cdot \frac{1}{\sqrt{nI(\hat{\theta}_{\text{MLE}})}}]$$

- (g) Remember the duality between CI and hypothesis tests. The null hypothesis is accepted if 1 lies in the confidence region.

Problem 3

Let X be a *single* observation from the probability density function $f(x) = \theta x^{\theta-1}, 0 < x < 1$.

(a) Find the most powerful test using significance level $\alpha = 0.05$ for testing the hypothesis $H_0 : \theta = 1$ and $H_1 : \theta = 2$ (sketch the densities $f(x \mid H_0)$ and $f(x \mid H_1)$ for the two hypothesis).

(b) What is the power of the test?

(c) What is the p-value of $X = 0.8$?

(d) For fixed $\alpha = 0.05$, is the test uniformly most powerful against the alternative hypothesis $H_1 : \theta > 1$?

Problem 3 (hint)

- (d) To find the UMP test, we should consider a simple alternative hypothesis $\theta = \theta_1, \theta_1 > 1$