Lab Exercises - Statistical Inference + Simulations

Problem 1. (Computing the MME and MLE using Newton-Raphson Method)

(a) Let $\mathbf{X} = (X_1, \ldots, X_n)$ be i.i.d. random variables with Kumaraswamy distribution with parameters $a > 0$ and $b > 0$. The probability density function of the Kumaraswamy distribution is

$$f(x; a, b) = abx^{a-1}\left(1 - x^a\right)^{b-1}, \quad \text{where } x \in [0, 1].$$

Find the method of moments for the Kumaraswamy distribution using numerical methods.

(b) Implement a function that takes as input a vector of data values X, performs the Newton-Raphson iterations to compute the MLEs $\hat{\alpha}$ and $\hat{\beta}$ in the Gamma (α, β) model, and outputs $\hat{\alpha}$ and $\hat{\beta}$. (You may use the form of the Newton-Raphson update equation derived in class. You may terminate the Newton-Raphson iterations when $|\alpha^{(t+1)} - \alpha^{(t)}|$ is sufficiently small.)

(c) For $n = 500$, use your function from part (a) to simulate the sampling distributions of $\hat{\alpha}$ and $\hat{\beta}$ computed from $X_1, \ldots, X_n \overset{\text{IID}}{\sim}$ Gamma(1,2). Plot histograms of the values of $\hat{\alpha}$ and $\hat{\beta}$ across 5000 simulations, and report the simulated mean and variance of $\hat{\alpha}$ and $\hat{\beta}$ as well as the simulated covariance between $\hat{\alpha}$ and $\hat{\beta}$. Compute the inverse of the Fisher Information matrix $I(\alpha, \beta)$ at $\alpha = 1$ and $\beta = 2$. Do your simulations support that $(\hat{\alpha}, \hat{\beta})$ is approximately distributed as $\mathcal{N}\left((1, 2), \frac{1}{n} I(1, 2)^{-1}\right)$? (You may use the formula for the Fisher information matrix $I(\alpha, \beta)$ and/or its inverse derived in class.)

Problem 2. (MLE in a misspecified model)

Suppose you fit the model Exponential(λ) to data $X_1, \ldots, X_n$ by computing the MLE $\hat{\lambda} = 1/\bar{X}$, but the true distribution of the data is $X_1, \ldots, X_n \overset{IID}{\sim}$ Gamma$(2, 1)$.

(a) Let $f(x \mid \lambda)$ be the PDF of the Exponential(λ) distribution, and let $g(x)$ be the PDF of the Gamma $(2, 1)$ distribution. Compute an explicit formula for the KL-divergence $D_{\mathrm{KL}}(g(x) \| f(x \mid \lambda))$ in terms of λ, and find the value λ^* that minimizes this KL-divergence.

(You may use the fact that if $X \sim$ Gamma(α, β), then $\mathbb{E}[X] = \alpha/\beta$ and $\mathbb{E}[\log X] = \psi(\alpha) - \log\beta$ where ψ is the digamma function.)

(b) Show directly, using the Law of Large Numbers and the Continuous Mapping Theorem, that the MLE $\hat{\lambda}$ converges in probability to λ^* as $n \to \infty$.

(c) Perform a simulation study for the standard error of $\hat{\lambda}$ with sample size $n = 500$, as follows: In each of $B = 10000$ simulations, sample $X_1, \ldots, X_n \overset{\text{IID}}{\sim}$ Gamma$(2, 1)$, compute the MLE $\hat{\lambda} = 1/\bar{X}$ for the exponential model, compute an estimate of the standard error of $\hat{\lambda}$ using the Fisher information $I(\hat{\lambda})$, and compute also the sandwich estimate of the standard error, $S_X / \left(\bar{X}^2 \sqrt{n}\right)$.

Report the true mean and standard deviation of $\hat{\lambda}$ that you observe across your 10000 simulations. Is the mean close to λ^* from part (a)? Plot a histogram of the 10000 estimated standard errors using the Fisher information, and also plot a histogram of the 10000 estimated standard errors using the sandwich estimate. Do either of these methods for estimating the standard error of $\hat{\lambda}$ seem accurate in this setting?

Problem 3. (Confidence intervals for a binomial proportion)

Let $X_1, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}(p)$ be n tosses of a biased coin, and let $\hat{p} = \bar{X}$. In this problem we will explore two different ways to construct a 95% confidence interval for p, both based on the Central Limit Theorem result

$$\sqrt{n}(\hat{p} - p) \to \mathcal{N}(0, p(1-p)). \tag{1}$$

(a) Use the plugin estimate $\hat{p}(1-\hat{p})$ for the variance $p(1-p)$ to obtain a 95% confidence interval for p. (This is the procedure discussed in class, yielding the Wald interval for p.)

(b) Instead of using the plugin estimate $\hat{p}(1-\hat{p})$, note that Eqn. (1) implies, for large n,

$$\mathbb{P}_p\left[-\sqrt{p(1-p)}Z_{\alpha/2} \leq \sqrt{n}(\hat{p} - p) \leq \sqrt{p(1-p)}Z_{\alpha/2}\right] \approx 1 - \alpha.$$

Solve the equation $\sqrt{n}(\hat{p}-p) = \sqrt{p(1-p)}Z_{\alpha/2}$ for p in terms of $\hat{p}$, and solve the equation $\sqrt{n}(\hat{p}-p) = -\sqrt{p(1-p)}Z_{\alpha/2}$ for p in terms of $\hat{p}$, to obtain a different 95% confidence interval for p.

(c) Perform a simulation study to determine the true coverage of the confidence intervals in parts (a) and (b), for the 9 combinations of sample size $n = 10, 40, 100$ and true parameter $p = 0.1, 0.3, 0.5$. (For each combination, perform at least $B = 100{,}000$ simulations. In each simulation, you may simulate $\hat{p}$ directly from $\frac{1}{n}\text{Binomial}(n, p)$ instead of simulating $X_1, \ldots, X_n$.) Report the simulated coverage levels in two tables. Which interval yields true coverage closer to 95% for small values of n?

Problem 4. (Power Comparisons)

Consider the problem of testing

$$H_0 : X_1, \ldots, X_n \overset{\text{IID}}{\sim} \mathcal{N}(0, 1)$$
$$H_1 : X_1, \ldots, X_n \overset{\text{IID}}{\sim} \mathcal{N}(\mu, 1)$$

at significance level $\alpha = 0.05$, where $\mu > 0$. We've seen four tests that may be applied to this problem, summarized below:

- Likelihood ratio test: Reject H_0 when $\bar{X} > \frac{1}{\sqrt{n}}Z_{0.05}$.
- t-test: Reject H_0 when $T := \sqrt{n}\bar{X}/S > t_{n-1;0.05}$, where $S^2 = \frac{1}{n-1}\sum_i \left(X_i - \bar{X}\right)^2$.
- Wilcoxon signed rank test: Reject H_0 when $W_+ > \frac{n(n+1)}{4} + \sqrt{\frac{n(n+1)(2n+1)}{24}}Z_{0.05}$, where W_+ is the Wilcoxon signed rank statistic.
- Sign test: Reject H_0 when $S > \frac{n}{2} + \sqrt{\frac{n}{4}}Z_{0.05}$, where S is the number of positive values in $X_1, \ldots, X_n$.

(For the Wilcoxon and sign test statistics, we are using the normal approximations for their null distributions.) These tests are successively more robust to violations of the $\mathcal{N}(0, 1)$ distributional assumption imposed by H_0.

(a) For $n = 100$, verify numerically that these tests have significance level close to α, in the following way: Perform 10,000 simulations. In each simulation, draw a sample of 100 observations from $\mathcal{N}(0, 1)$, compute the above four test statistics $\bar{X}, T$, W_+, and S on this sample, and record whether each test accepts or rejects H_0. Report the fraction of simulations for which each test rejected H_0, and confirm that these fractions are close to 0.05.

(b) For $n = 100$, numerically compute the powers of these tests against the alternative H_1, for the values $\mu = 0.1, 0.2, 0.3$, and 0.4. Do this by performing 10,000 simulations as in part (a), except now drawing each sample of 100 observations from $\mathcal{N}(\mu, 1)$ instead of $\mathcal{N}(0, 1)$. (You should be able to re-use most of your code from part (a).) Report your computed powers either in a table or visually using a graph.

(c) How do the powers of the four tests compare, when testing against a normal alternative? Your friend says, "We should always use the testing procedure that makes the fewest distributional assumptions, because we never know in practice, for example, whether the variance is truly 1 or whether data is truly normal." Comment on this statement. Rice says, "It has been shown that even when the assumption of normality holds, the [Wilcoxon] signed rank test is nearly as powerful as the t test. The [signed rank test] is thus generally preferable, especially for small sample sizes." Do your simulated results support this conclusion?

Problem 5. (Testing gender ratios)

In a classical genetics study, Geissler (1889) studied hospital records in Saxony and compiled data on the gender ratio. The following table shows the number of male children in 6115 families having 12 children:

Number of male children	Number of families
0	7
1	45
2	181
3	478
4	829
5	1112
6	1343
7	1033
8	670
9	286
10	104
11	24
12	3

Let $X_1, \ldots, X_{6115}$ denote the number of male children in these 6115 families.

(a) Suggest two reasonable test statistics T_1 and T_2 for testing the null hypothesis

$$H_0 : X_1, \ldots, X_{6115} \overset{IID}{\sim} \text{Binomial}(12, 0.5).$$

(This is intentionally open-ended; try to pick T_1 and T_2 to "target" different possible alternatives to the above null.) Compute the values of T_1 and T_2 for the above data.

(b) Perform a simulation to simulate the null distributions of T_1 and T_2. (For example: Simulate 6115 independent samples $X_1, \ldots, X_{6115}$ from Binomial$(12, 0.5)$, and compute T_1 on this sample. Do this 1000 times to obtain 1000 simulated values of T_1. Do the same for T_2.) Plot the histograms of the simulated null distributions of T_1 and T_2. Using your simulated values, compute approximate p-values of the hypothesis tests based on T_1 and T_2, for the above data. For either of your tests, can you reject H_0 at significance level $\alpha = 0.05$?

(c) In this example, why might the null hypothesis H_0 not hold? (Please answer this question regardless of your findings in part (b).)

Reference (Problem 5): Edwards, A. W. F. (1958). "An analysis of Geissler's data on the human sex ratio." Annals of human genetics, 23(1), 6-15.

Problem 6. (Monte Carlo Simulations)

(a) Consider the problem of approximating tail probabilities of a Standard Normal distribution. Find $\mathcal{P}(X > 2.04)$ for $X \sim N(0, 1)$ using Monte Carlo methods. While we can easily calculate this in `R` using the `pnorm` function, check how a Monte Carlo method would perform. How does the estimate change with the number of Monte Carlo samples?

(b) Consider the problem of calculating the expected present value of the payoff of a call option on a stock. Let $S(t)$ denote the price of the stock at time t. Consider a European call option, that is, the holder has the right to buy the stock at a fixed price K at a fixed time T in the future, $t = 0$ being the current time. If $S(T) > K$, the holder exercises the option for a profit of $S(T) - K$. If $S(T) \leq K$, the option expires worthless. The payoff at time T is thus

$$\max\{S(T) - K, 0\}.$$

The expected present value of this payoff is

$$E\left[e^{-rT} \max\{S(T) - K, 0\}\right],$$

where r is the compound interest rate. The evolution of the stock price $S(t)$ can be modeled via the Black-Scholes model expressed as

$$\frac{dS(t)}{S(t)} = r dt + \sigma dW(t),$$

where W is a standard Brownian motion. The solution to the above Stochastic Differential Equation is given by

$$S(T) = S(0) \exp\left\{\left(r - \frac{1}{2}\sigma^2\right)T + \sigma W(T)\right\},$$

where $S(0)$ is the current price of the stock and $W(T) \sim N(0, T)$, that is, $W(T) = \sqrt{T}Z, Z \sim N(0, 1)$. Thus, the logarithm of $S(T)$ is Normally distributed, or, in other words, the distribution of $S(T)$ is log-normal. The expected payoff $E\left[e^{-rT} \max\{S(T) - K, 0\}\right]$ is thus an integral w.r.t. to the lognormal density.

Can you perform Monte Carlo sampling to estimate this integral (that is, find the expected payoff)? This will require generating samples from the standard Normal distribution, followed by computing the value of the function inside the expectation and then taking the average. Can you also compute the estimated standard error of the estimator?

Some useful functions:

Here are R commands to find probabilities and quantiles for the "commonly used" distributions we have talked about.

Distribution	$p_Y(y) = P(Y = y)$	$F_Y(y) = P(Y \leq y)$	ϕ_c
$Y \sim b(n, p)$	dbinom $(y, \mathrm{n}, \mathrm{p})$	$\mathrm{pbinom}(\mathrm{y}, \mathrm{n}, \mathrm{p})$	qbinom $(\mathrm{c}, \mathrm{n}, \mathrm{p})$
$Y \sim$ geom(p)	$\mathrm{dgeom}(\mathrm{y}-1, \mathrm{p})$	$\mathrm{pgeon}(\mathrm{y}-1, \mathrm{p})$	$1+\mathrm{qgeom}(\mathrm{c}, \mathrm{p})$
$Y \sim \mathrm{mib}(r, p)$	dnbinom $(\mathrm{y}-\mathrm{r}, \mathrm{r}, \mathrm{p})$	$\mathrm{pnbinon}(\mathrm{y}-\mathrm{r}, \mathrm{r}, \mathrm{p})$	$\mathrm{r}+\mathrm{qnbinom}(\mathrm{c}, \mathrm{r}, \mathrm{p})$
$Y \sim \mathrm{hyper}(N, n, r)$	$\mathrm{dhyper}(\mathrm{y}, \mathrm{r}, \mathrm{N}-\mathrm{r}, \mathrm{n})$	$\mathrm{phyper}(\mathrm{y}, \mathrm{r}, \mathrm{N}-\mathrm{r}, \mathrm{n})$	qhyper $(\mathrm{c}, \mathrm{r}, \mathrm{N}-\mathrm{r}, \mathrm{n})$
$Y \sim \mathrm{Poisson}(\lambda)$	$\mathrm{dpois}(\mathrm{y}, \lambda)$	$\mathrm{ppois}(\mathrm{y}, \lambda)$	qpois (c, λ)

Table 1: Discrete distributions: Binomial, geometric, negative binomial, hypergeometric, Poisson.

Note that, in discrete distributions, the c^{th} quantile ϕ_c is defined as the smallest value satisfying $F_Y(\phi_c) = P(Y \leq \phi_c) \geq c$ $(0 < c < 1)$.

Distribution	$F_Y(y) = P(Y \leq y)$	ϕ_{p}
$Y \sim \mathcal{U}(\theta_1, \theta_2)$	punif $(\mathrm{y}, \theta_1, \theta_2)$	qunif $(\mathrm{p}, \theta_1, \theta_2)$
$Y \sim \mathcal{N}(\mu, \sigma^2)$	$\mathrm{pnorm}(\mathrm{y}, \mu, \sigma)$	$\mathrm{qnorm}(\mathrm{p}, \mu, \sigma)$
$Y \sim \mathrm{exponential}(\beta)$	$\mathrm{pexp}(\mathrm{y}, 1/\beta)$	$\mathrm{qexp}(\mathrm{p}, 1/\beta)$
$Y \sim \mathrm{gamma}(\alpha, \beta)$	$\mathrm{pgamna}(\mathrm{y}, \alpha, 1/\beta)$	$\mathrm{qgamma}(\mathrm{p}, \alpha, 1/\beta)$
$Y \sim \chi^2(\nu)$	$\mathrm{pchisq}(\mathrm{y}, \nu)$	$\mathrm{qchisq}(\mathrm{p}, \nu)$
$Y \sim \mathrm{beta}(\alpha, \beta)$	$\mathrm{pbeta}(\mathrm{y}, \alpha, \beta)$	$\mathrm{qbeta}(\mathrm{p}, \alpha, \beta)$
$Y \sim t(\nu)$	$\mathrm{pt}(\mathrm{y}, \nu)$	qt (p, ν)
$Y \sim F(\nu_1, \nu_2)$	pf $(\mathrm{y}, \nu_1, \nu_2)$	qf $(\mathrm{p}, \nu_1, \nu_2)$

Table 2: Continuous distributions: Uniform, normal, exponential, gamma, χ^2, beta, t, and F.

Note that, in continuous distributions, the p^{th} quantile ϕ_p satisfies $F_Y(\phi_p) = P(Y \leq \phi_p) = p$. Note that $0 < p < 1$. I used "c" above in the discrete distributions so as not to interfere with "p" in the binomial, geometric, and negative binomial distributions.