Statistics 2 Exercises

1. Using characteristic functions, show that as $n \to \infty$ and $p \to 0$ such that $np \to \lambda$, the binomial distribution with parameters n and p tends to the Poisson distribution.

2. Using characteristic functions, show that as the shape parameter $\alpha \to \infty$, the gamma distribution with shape parameter α and rate parameter λ, properly standardized, tends to the normal distribution.

3. The central limit theorem can be used to analyze round-off error. Suppose that the round-off error can be represented as a uniform random variable on $[-1/2, 1/2]$. If 100 numbers are added, approximate the probability that the round-off error exceeds (a) 1, (b) 2, and (c) 5.

4. Suppose $X_1, \ldots, X_{20}$ are independent random variables with density functions $f(x) = 2x$, $0 \le x \le 1$. Let $S = X_1 + \cdots + X_{20}$. Use the central limit theorem to approximate $\mathbb{P}(S \le 10)$.

5. (a) Use the Monte Carlo method with $n = 100$ and $n = 1000$ to estimate $\theta = \int_0^1 \cos(2\pi x)dx$. Compare the estimates with the exact answer.

   (b) Use Monte Carlo to evaluate $\int_0^1 \cos(2\pi x^2)dx$. Can you find the exact answer?

6. What is the variance of the estimate of an integral by the Monte Carlo method? Estimate $\theta = \int_0^1 \cos(2\pi x)dx$ by Monte Carlo, and compare the standard deviations of the estimates to the actual errors.

7. Suppose we wish to evaluate $\theta = \int_a^b g(x)dx$. Let f be a density function on $[a, b]$. Generate $X_1, \ldots, X_n$ from f and estimate θ by

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} \frac{g(X_i)}{f(X_i)}.$$

   (a) Show that $\mathbb{E}(\hat{\theta}) = \theta$.

   (b) Find an expression for $\mathrm{var}(\hat{\theta})$. Give an example for which it is finite and an example for which it is infinite. Note that if it is finite, the law of large numbers implies that $\hat{\theta} \to \theta$ as $n \to \infty$.

   (c) Consider estimating

$$\theta = \frac{1}{\sqrt{2\pi}} \int_0^1 e^{-x^2/2} dx$$

   by Monte Carlo. Taking $a = 0$, $b = 1$ and f as uniform on the unit interval one obtains the "straightforward" MC estimate

$$\frac{1}{n} \frac{1}{\sqrt{2\pi}} \sum_{i=1}^{n} e^{-U_i^2/2}.$$

   Can this estimate be improved by choosing f other than uniform?

8. Use the central limit theorem to find Δ such that $\mathbb{P}(|\hat{\theta} - \theta| \le \Delta) = 0.05$, where $\hat{\theta}$ is the MC estimate of $\theta = \int_0^1 \cos(2\pi x)dx$ based on $n = 1000$ points.

9. In addition to limit theorems that deal with sums, there are limit theorems that deal with extreme values such as maxima or minima. Here is an example. Let $U_1, \ldots, U_n$ be independent random variables distributed uniformly on $[0, 1]$, and let $U_{(n)}$ be their maximum. Find the cdf of $U_{(n)}$ and a standardized $U_{(n)}$, and show that the standardized variable tends to a limiting value.

10. Generate a sequence $U_1, \ldots, U_{1000}$ of independent uniform random variables on $[0, 1]$. Let $S_n = \sum_{i=1}^{n} U_i$ for $n = 1, 2, \ldots, 1000$. Plot each of the following versus n:

    (a) S_n

    (b) S_n/n

    (c) $S_n - n/2$

    (d) $(S_n - n/2)/n$

    (e) $(S_n - n/2)/\sqrt{n}$

    Explain the shapes of the resulting graphs.

11. Let X be a random variable with moment generating function $m(t) = \mathbb{E}(e^{tX})$ defined for all t in a neighborhood of $t = 0$. The cumulant generating function is $k(t) = \log(m(t))$, and defines the *cumulants* of X in the same fashion as the moment generating function defines the moments, i.e., $\kappa_j = k^{(j)}(0)$, where $k^{(j)}$ denotes the j-th derivative of k.

    (a) Derive the first four cumulants in terms of the first four (non-central) moments.

    (b) Express κ_2, κ_3 and κ_4 in terms of the central moments $\mu_i = \mathbb{E}((X - \mathbb{E}(X))^i)$.

    (c) Express the skewness coefficient $\text{Skew} = \mu_3/\mu_2^{3/2}$ and kurtosis coefficient $\text{Kurt} = \mu_4/\mu_2^2$ in terms of the cumulants.

12. Find the first four cumulants of the Poisson and normal distributions.

13. Suppose a sample is taken from a symmetric distribution whose tails decrease more slowly than those of the normal distribution. What would be the qualitative shape of a normal probability plot of this sample? What about the normal quantile plot?

14. Let $X_1, \ldots, X_n$ be an i.i.d. sample from a continuous distribution with distribution function $F(x|\theta)$. Show that $-2\sum_{i=1}^{n} \log(F(X_i|\theta))$ has a chi-squared distribution with $2n$ degrees of freedom.

15. The standard Gumbel distribution (type I extreme value distribution) has distributution function $F(x) = e^{-e^{-x}}$. Show that this has mode 0, median $-\log(\log(2))$ and moment generating function $m(t) = \Gamma(1 - t)$ for $t < 1$, where $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\, dt$ is the well-known Gamma function. Use the facts that $\Gamma(z + 1) = z\Gamma(z)$ and that the log-derivative $\psi(z) = d\log(\Gamma(z))/dz = \Gamma'(z)/\Gamma(z)$ satisfies $\psi(1) = -\gamma = 0.577215\ldots$ (the Euler-Mascheroni constant) and $\psi'(1) = \pi^2/6$ to show that if X has a standard Gumbel distribution, then

    $$\mathbb{E}(X) = \gamma, \qquad \text{var}(X) = \pi^2/6$$

    and

    $$\mathbb{E}(e^{-X}) = 1, \qquad \mathbb{E}(Xe^{-X}) = \gamma - 1, \qquad \mathbb{E}(X^2e^{-X}) = \pi^2/6 - 2\gamma + \gamma^2.$$

16. The location-scale extreme value distribution has distribution function

    $$F(t) = F_0((t - \mu)/\sigma), \qquad F_0(t) = \exp(-\exp(-x)).$$

    Give an expression for the p-th quantile $F^{-1}(p)$.

17. A popular model for claim size distributions in actuarial science is the Lomax ("Pareto") distribution with density

    $$f(x|\alpha, \lambda) = \frac{\alpha\lambda^\alpha}{(\lambda + x)^{\alpha+1}}, \qquad x > 0$$

(this is the standard Pareto distribution shifted so that its support starts at 0, and a special case of the generalized Pareto distribution with $\mu = 0$, $\sigma = \lambda/\alpha$ and $\xi = 1/\alpha$). Verify that the corresponding distribution function is

$$F(x|\alpha, \lambda) = 1 - \left(\frac{\lambda}{\lambda + x}\right)^{\alpha} = 1 - (1 + x/\lambda)^{-\alpha}, \qquad x > 0.$$

Show that if X has a Lomax distribution with parameters α and λ, then

$$\mathbb{E}(X) = \frac{\lambda}{\alpha - 1}, \qquad \mathrm{var}(X) = \frac{\lambda^2 \alpha}{(\alpha - 1)^2(\alpha - 2)}$$

provided that $\alpha > 1$ and $\alpha > 2$, respectively, and that conditional on $X > M$, the excess $X - M$ has a Lomax distribution with parameters α and $\lambda + M$.

18. Let $X_1, \ldots, X_n$ be i.i.d. positive random variables so that the Box-Cox transformed $X_i^{(\lambda)}$ are $N(\mu, \sigma^2)$, where

$$X^{(\lambda)} = \begin{cases} \dfrac{X^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log(X) & \lambda = 0. \end{cases}$$

(The normality assumption is actually only possible for $\lambda = 0$, but ignore this detail.) Derive the log-likelihood function $\ell(\mu, \sigma, \lambda | x_1, \ldots, x_n)$ of the observed data.

19. Let $X_1, \ldots, X_n$ be an i.i.d. sample from the family of distribution functions

$$F(x|p_0, p_1, \gamma) = p_0 I(0 \leq x) + (1 - p_0 - p_1) F_0(x|\gamma) + p_1 I(x \geq 1),$$

where the F_0 "live" on $(0, 1)$, i.e., $F_0(0|\gamma) = 0$ and $F_0(1|\gamma) = 1$ (e.g., the family of Beta distributions). Such data could for example be test scores standardized to $[0, 1]$, where n_0 of the sample values are exactly 0 (turned in a blank test), n_1 values are 1 (a perfect score), and the rest are between 0 and 1. Suppose F_0 has a density f_0. Determine the likelihood function for $\theta = (p_0, p_1, \gamma)$ via the "$2h$ method" as the limit for $h \to 0+$ of

$$\prod_{i=1}^{n} (F(x_i + h|\theta) - F(x_i - h|\theta))$$

suitably normalized (dividing each term where F is differentiable at x_i by $2h$) to obtain that

$$L(\theta | x_1, \ldots, x_n) = p_0^{n_0} p_1^{n_1} (1 - p_0 - p_1)^{n - n_0 - n_1} \prod_{i: 0 < x_i < 1} f_0(x_i|\gamma).$$

20. A company has manufactured certain objects and printed a serial number on each. The serial numbers start at 1 and end at N, where N is the number of objects that have been manufactured. One of these objects is selected at random, and its serial number is 888. What is the method of moments estimate of N? What is the MLE of N?

21. George spins a coin three times and observes no heads. He then gives the coin to Hilary. She spins it until the first head occurs, and ends up spinning it four times total. Let θ be the probability that the coin comes up heads. What are the likelihood and the MLE of θ?

22. For an i.i.d. sample $X_1, \ldots, X_n$, type II censoring occurs when we observe only the smallest r values. For example, in a study of light bulb lifetimes, we might stop the study after $r = 10$ lightbulbs failed. Assuming a continuous distribution with density $f(x|\theta)$, the likelihood is just the joint density of the smallest r order statistics evaluated at those statistics:

$$L(\theta | x_{(1)}, \ldots, x_{(r)}) = \frac{n!}{(n - r)!} \left(\prod_{i=1}^{r} f(x_{(i)}|\theta) \right) (1 - F(x_{(r)}|\theta))^{n-r}.$$

Suppose that $f(x|\sigma) = e^{-x/\sigma}/\sigma$. Determine the MLE of σ.

23. Let $X_1, \ldots, X_n$ be an i.i.d. sample from the shifted exponential distribution

$$f(x|\mu, \sigma) = \frac{1}{\sigma} e^{-(x-\mu)/\sigma}, \qquad x \geq \mu.$$

Estimate the parameters by the method of moments and maximum likelihood when (a) μ is known and (b) μ is unknown.

24. Let $X_1, \ldots, X_n$ be an i.i.d. sample from the Lomax distribution with parameters α and λ, where λ is known. Show that the MLE for α is given by

$$\hat{\alpha} = \frac{n}{\sum_{i=1}^n \log(1 + X_i/\lambda)}$$

and that the sampling distribution of $\hat{\alpha}/\alpha$ is that of the inverse of a Gamma distribution with shape and rate parameter n (i.e., an inverse Gamma distribution with shape parameter n and scale parameter n). What is the standard error of $\hat{\alpha}$? (Hint: verify first that if U is uniform on $[0, 1]$, then $\lambda(U^{-1/\alpha} - 1)$ has a Lomax distribution with parameters α and λ.)

25. Suppose the random variable X represents the monetary amount of damage caused by storms. The insurance company handling such claims will pay only $W = \max(X - 40000, 0)$, the excess of the damage ofter 40000 monetary units. In 2005, the payments made were 14000, 21000, 6000, 32000 and 2000. Assume that the density of X has the form

$$f(x|\alpha) = \frac{\alpha 2^\alpha 10^{4\alpha}}{(20000 + x)^{\alpha+1}}, \qquad x \geq 0,$$

where α is an unknown parameter.

   (a) Determine the density of W, its mean and its variance.
   (b) Determine the MLE $\hat{\alpha}$ of α based on the 2005 data, and give an estimate for the standard error of $\hat{\alpha}$.

26. Eire General Insurance has an arrangement with the reinsurance company SingapoRe, whereby the excess of any claim above M is handled by the reinsurer. Claim size X is traditionally modeled by a Lomax distribution with parameters α and $\lambda = 8400$, i.e.,

$$\mathbb{P}(X \geq x|\alpha, \lambda) = \left(\frac{\lambda}{\lambda + x}\right)^\alpha.$$

Show that the MLE of α based on a sample of $n + m$ claim payments (for Eire General) of the form $(x_1, \ldots, x_n, M, \ldots, M)$ is

$$\hat{\alpha} = n \Bigg/ \left(\sum_{i=1}^n \log(1 + x_i/\lambda) + m \log(1 + M/\lambda)\right).$$

If the amounts paid based on a sample with $n = 7$ and $m = 3$ were

$$(14.9, 775.7, 805.2, 993.9, 1127.5, 1602.5, 1998.3, 2000, 2000, 2000),$$

what would the MLE of α be?

27. Let $X_1, \ldots, X_n$ be an i.i.d. sample from the Lomax distribution. Determine the method of moments estimates and the (equations satisfied by the) MLEs for the paramaters α and λ of this distribution.

28. A random variable has a Weibull distribution if it has density

$$f(x|c,\gamma) = c\gamma x^{\gamma-1}e^{-cx^\gamma}, \qquad x > 0$$

or equivalently, if its distribution function for $x > 0$ is $F(x|c,\gamma) = 1 - e^{-cx^\gamma}$. Show that if X has this distribution, then X^γ has an exponential distribution with rate parameter c. How could this be used to estimate the parameters by the method of moments?

In practice, using the (seldom employed) method of percentiles may be more convenient. Here, one estimates the parameters by equating the first and third emporical and theoretical quartiles. Determine the corresponding estimates for c and γ.

29. One of the data sets obtained from a 1984 consulting session on max flow of rivers was $n = 35$ yearly maxima from one station with the following values:

```
R> Rivers

 [1]  5500  4380  2370  3220  8050  4560  2100  6840  5640  3500  1940
[12]  7060  7500  5370 13100  4920  6500  4790  6050  4560  3210  6450
[23]  5870  2900  5490  3490  9030  3100  4600  3410  3690  6420 10300
[34]  7240  9130
```

Find the MLEs of the parameters of the location-scale extreme value distribution model

$$f(x|\mu,\sigma) = \frac{1}{\sigma} f_0\left(\frac{x-\mu}{\sigma}\right), \qquad f_0(x) = e^{-x}e^{-e^{-x}}.$$

for this data, and use a Q-Q plot to graphically assess the goodness of fit.

$\boxed{\pi}$ 30. For 36 hurricanes that had moved far inland on the East Coast of the U.S.A. in 1900–1969, Larsen and Marx (2001) give maximum 24-hours precipitation levels during the time they were over mountains:

```
R> Hurricanes

 [1] 31.00  2.82  3.98  4.02  9.50  4.50 11.40 10.71  6.31  4.95  5.64
[12]  5.51 13.40  9.72  6.47 10.16  4.21 11.60  4.75  6.85  6.25  3.42
[23] 11.80  0.80  3.69  3.10 22.22  7.43  5.00  4.58  4.46  8.00  3.73
[34]  3.50  6.20  0.67
```

A histogram suggests that the Gamma distribution is a reasonable candidate model for these data. Use R to determine the MLEs for the parameters of the Gamma distribution (you could use the method of moments estimates of the parameters as starting points for the numerical optimization routine). Check the fit using a Q-Q plot: does the visual evidence suggest that the Gamma distribution is an appropriate model for the data?

31. On a particular class of policy, claim amounts coming into Surco Ltd. follow an exponential distribution with rate parameter λ. A reinsurance agreement has been made by Surco to handle the excess of any claim above 10000 monetary units. Over the past year, 80 claims were made, with 68 of these below 10000; these 68 in aggregate value amounted to 220000. The other 12 claims exceeded 10000.

Let x_i represent the amount of the i-th claim from the 68 claims beneath 10000. Show that the log-likelihood is

$$\ell(\lambda) = 68\log(\lambda) - \lambda\sum_{i=1}^{68} x_i - 120000\lambda.$$

Find the MLE $\hat{\lambda}$ and calculate an approximate 95 percent confidence interval for λ.

32. Suppose that $X_1, \ldots, X_n$ are i.i.d. $N(\mu_0, \sigma_0^2)$ and μ and σ^2 are estimated by the method of maximum likelihood, with resulting estimates $\hat{\mu}$ and $\hat{\sigma}^2$. Suppose the bootstrap is used to estimate the sampling distribution of $\hat{\mu}$.

    (a) Explain why the bootstrap estimate of the distribution of $\hat{\mu}$ is $N(\hat{\mu}, \hat{\sigma}^2/n)$.

    (b) Explain why the bootstrap estimate of the distribution of $\hat{\mu} - \mu_0$ is $N(0, \hat{\sigma}^2/n)$.

    (c) According to the previous part, what is the form of the bootstrap confidence interval for μ, and how does it compare to the exact confidence interval based on the t distribution?

33. Let $\hat{\theta}$ be a parameter estimate and θ_L and θ_U be the quantiles of the corresponding θ^* bootstrap distribution. Show that the bootstrap confidence interval for θ can be written as $(2\hat{\theta} - \theta_U, 2\hat{\theta} - \theta_L)$. Show that if the sampling distribution of θ^* is symmetric about $\hat{\theta}$, then the bootstrap confidence interval is (θ_L, θ_U).

34. The `gamma-arrivals` data file contains a set of gamma-ray data, consisting of the times between arrivals (interarrival times) of 3,935 photons measured in seconds.

    (a) Make a histogram of the interarrival times. Does it appear that a gamma distribution would be a plausible model?

    (b) Fit the parameters by the method of moments and by maximum likelihood. How do the estimates compare?

    (c) Plot the two fitted gamma densities on top of the histogram. Do the fits look reasonable?

    (d) For both maximum likelihood estimate and the method of moments, use the bootstrap to form approximate confidence intervals for the parameters. How do the confidence intervals for the two methods compare?

    (e) Is the interarrival time distribution consistent with a Poisson process model for the arrival times?

$\boxed{\pi}$ 35. The `Theft` data set gives the amounts of 120 theft claims made in a household insurance portfolio. Determine the method of moment estimates and MLEs for a Lomax ("Pareto") model for these data. Use the bootstrap to estimate the standard errors of these estimates, and obtain approximate 95% confidence intervals for the parameters, and compare the performances.

$\boxed{\pi}$ 36. The `Theft` data set gives the amounts of 120 theft claims made in a household insurance portfolio. Determine the method of percentiles estimates and MLEs for a Weibull model for these data. Use the bootstrap to estimate the standard errors of these estimates, and obtain approximate 95% confidence intervals for the parameters, and compare the performances.

37. The `bodytemp` data set contains normal body temperature readings (degrees Fahrenheit) and heart rates (beats per minute) of 65 males (coded by one) and 65 females (coded by two) from a study by Shoemaker. Assuming that the population distributions are normal, estimate the means and standard deviations of the males and females. Form 95% confidence intervals for the means. Standard folklore is that the average body temperature of 98.6 degrees Fahrenheit. Does this appear to be the case?

38. Let $X_1, \ldots, X_n$ be an i.i.d. sample from the uniform distribution on $[0, \theta]$, and let $X_{(n)} = \max(X_1, \ldots, X_n)$. Show that

    (a) $X_{(n)}/\theta$ is a pivot (i.e., its distribution does not depend on θ).

    (b) $(X_{(n)}, \alpha^{-1/n} X_{(n)})$ is a $100(1-\alpha)$ percent confidence interval for θ.

39. Let $X_1, \ldots, X_n$ be an i.i.d. sample from $N(\mu, a^2\mu^2)$, where a is known. Find a pivot and construct a $100(1-\alpha)$ percent confidence interval for μ.

40. Let $X_1, \ldots, X_n$ be an i.i.d. sample from $N(\mu, \sigma^2)$ with known σ, and let $0 < \beta < \alpha < 1$.

(a) Show that

$$\left(\bar{X} + z_{\alpha-\beta}\frac{\sigma}{\sqrt{n}}, \bar{X} - z_{\beta}\frac{\sigma}{\sqrt{n}}\right)$$

is a $100(1-\alpha)$ percent confidence interval for μ.

(b) Show that this confidence interval has minimum expected length when it is symmetric, i.e., if $\beta = \alpha/2$.

41. Suppose that $X_1, \ldots, X_n$ are a random sample from a lognormal distribution with unknown parameters. Construct a 95% confidence interval for the parameter μ. Use a Monte Carlo method to obtain an empirical estimate of the confidence level.

42. Suppose a 95% symmetric t-interval is applied to estimate a mean, but the data are non-normal. The coverage probability is not necessarily equal to 95%. Use a Monte Carlo experiment to estimate the coverage probability for samples of size 20 from the chi-squared distribution with 2 degrees of freedom.

43. Consider the univariate normal distribution as a location-scale family, i.e., parametrized by its mean μ and standard deviation σ. Determine the Fisher information matrix for $\theta = (\mu, \sigma)$. What does this imply for the asymptotic distribution of the MLEs of μ and σ?

44. The multinomial distribution with parameters n and $p_1, \ldots, p_k$, has pmf

$$\frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k},$$

where the p_i are non-negative and sum to one. Take $\theta = (p_1, \ldots, p_{k-1})$ and $p_k = p_k(\theta) = 1 - (p_1 + \cdots + p_{k-1})$. Show that the Fisher information matrix for θ is given by

$$I(\theta) = n(\operatorname{diag}(1/p_1, \ldots, 1/p_{k-1}) + 11'/p_k)$$

with inverse

$$I(\theta)^{-1} = (\operatorname{diag}(\theta) - \theta\theta')/n$$

(where 1 denotes a vector of all ones).

45. The location-scale family of extreme value distributions has densities

$$f(x|\mu, \sigma) = \frac{1}{\sigma} f_0\left(\frac{x-\mu}{\sigma}\right), \qquad f_0(x) = e^{-x}e^{-e^{-x}}.$$

Determine the Fisher information matrix for μ and σ. What does this imply for the asymptotic distribution of the MLEs of μ and σ?

46. Consider a location-scale extreme value model for the `Rivers` data.

(a) Find an estimate of the asymptotic covariance matrix of $\hat{\mu}$ and $\hat{\sigma}$.

(b) Estimate the median of the distribution of the largest flow rate in $N = 100$ years via

$$\hat{Q}_N = \hat{\mu} - \hat{\sigma} \log(-\log(\sqrt[N]{1/2})).$$

(c) Find an estimate for the variance of $\hat{Q}_N$.

47. The Gamma distribution is usually parametrized by its shape parameter α, and either its scale parameter s or its rate parameter $\lambda = 1/s$. Yet another parametrization is obtain by using α and the mean $\mu = \alpha/\lambda$. Show that this results in

$$f(x|\alpha, \mu = \alpha/\lambda) = \frac{(\alpha/\mu)^\alpha x^{\alpha-1} e^{-\alpha x/\mu}}{\Gamma(\alpha)}, \qquad x > 0.$$

Show that using this parametrization, the off-diagonal term in the Fisher information matrix vanishes, and use this to conclude that the MLEs $\hat{\alpha}$ and $\hat{\mu}$ are asymptotically independent.

48. Consider a location-scale model with densities

$$f(x|\mu, \sigma) = \frac{1}{\sigma} f_0\left(\frac{x-\mu}{\sigma}\right).$$

Show that the off-diagonal element of the Fisher information matrix is

$$\frac{1}{\sigma^2} \int_{-\infty}^{\infty} x \left(\frac{f_0'(x)}{f_0(x)}\right)^2 f_0(x)\, dx.$$

Conclude that if f_0 is symmetric about zero, the MLEs of μ and σ are asymptotically independent, so that there is no asymptotic loss in efficiency (no variance inflation) when estimating σ in addition to μ.

$\boxed{\pi}$ 49. Let $X_1, \ldots, X_n$ by i.i.d. positive random variables so that the Box-Cox transformed $X_i^{(\lambda)}$ are $N(\mu, \sigma^2)$, where

$$X^{(\lambda)} = \begin{cases} \dfrac{X^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log(X) & \lambda = 0. \end{cases}$$

(The normality assumption is actually only possible for $\lambda = 0$, but ignore this detail.) When all three parameters are estimated, the Fisher information matrix for $\lambda = 0$ (corresponding to the log-normal distribution) and a single observation is

$$I(\mu, \sigma, \lambda = 0) = \frac{1}{\sigma^2} \begin{bmatrix} 1 & 0 & -\tau_1 \\ 0 & 2 & -2\sigma\mu \\ -\tau_1 & -2\sigma\mu & \tau_2 \end{bmatrix}$$

where $\tau_1 = (\sigma^2 + \mu^2)/2$ and $\tau_2 = (7\sigma^4 + 10\sigma^2\mu^2 + \mu^4)/4$. Use simulation to verify that this is correct when $\mu = 1$ and $\sigma = 1$.

Hint: use for example $B = 1000$ replications of samples of size $n = 100$ from the log-normal distribution with parameters 1 and 1. In each sample, estimate the theoretical Fisher information matrix using the average Hessians of the log-densities.

50. Let $X_1, \ldots, X_n$ be an i.i.d. sample from $N(\mu, a^2\mu^2)$ with known a.

    (a) Construct an asymptotic $100(1-\alpha)$ percent confidence interval for μ based on the distribution of $\bar{X}_n$.

    (b) Repeat using a variance stabilization transformation $g(\mu)$.

Note: if $\sqrt{n}(\hat{\theta}_n - \theta)$ is asymptically distributed as $N(0, \sigma(\theta)^2)$ and g is differentiable with $g'(\theta) \neq 0$, then in general $\sqrt{n}(g(\hat{\theta}_n) - g(\theta))$ is asymptotically distributed as $N(0, g'(\theta)^2\sigma(\theta)^2)$. A variance stabilization transformation renders the asymptotic variance independent of θ, and hence must satisfy $g'(\theta) = c/\sigma(\theta)$ for some constant c.

51. Let $X_1, \ldots, X_n$ be an i.i.d. sample from the Bernoulli distribution with success probability p.

(a) Show that the variance stabilization transformation is $g(p) = \arcsin(\sqrt{p})$. What is the resulting asymptotic variance?

(b) Use these results to construct an asymptotic $100(1-\alpha)$ percent confidence interval for p.

52. Suppose one wants to estimate the variance of a normal distribution with unknown mean from a sample $X_1, \ldots, X_n$ of i.i.d. normal random variables. Recall that $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$ and the mean and variance of a chi-squared random variable with r degrees of freedom are r and $2r$, respectively.

(a) Which of the following estimates is unbiased?

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

(b) Which of these estimates has the smaller MSE?

(c) For what value of ρ does $\rho\sum_{i=1}^{n}(X_i - \bar{X})^2$ have the minimal MSE?

53. Let $X_1, \ldots, X_n$ be an i.i.d. sample from the Bernoulli distribution with success probability p. Consider the following three estimators:

$$\hat{p}_1 = \frac{1}{n}\sum_{i=1}^{n}X_i, \qquad \hat{p}_2 = \frac{1}{n+2}\left(\sum_{i=1}^{n}X_i + 1\right), \qquad \hat{p}_3 = X_1.$$

Find the MSEs of these three estimators. Is any estimate uniformly better (with respect to MSE) than the others?

54. Let $X_1, \ldots, X_n$ be an i.i.d. sample from the Bernoulli distribution with success probability p. Show that there is no unbiased estimator for the odds ratio $p/(1-p)$.

$\boxed{\pi}$ 55. Let $X_1, \ldots, X_n$ be i.i.d. uniform on $[0, \theta]$.

(a) Find the method of moments estimate of θ and its mean and its variance.

(b) Find the MLE of θ.

(c) Find the sampling distribution of the MLE, and calculate its mean and variance. Compare the variance, the bias, and the mean squared error to those of the method of moments estimate.

(d) Find a modification of the MLE that renders it unbiased.

56. Let $X_1, \ldots, X_n$ be an i.i.d. sample from the shifted (standard) exponential distribution

$$f(x|\theta) = e^{-(x-\theta)}, \qquad x \geq \theta.$$

(a) Find the MLE $\hat{\theta}_n$ for θ.

(b) Is $\hat{\theta}_n$ consistent in MSE?

(c) Show that $n(\hat{\theta}_n - \theta)$ has a standard exponential distribution.

(d) Find the asymptotic distribution of $\sqrt{n}(\hat{\theta}_n - \theta)$.

(e) Why does the asymptotic normality of the MLE not hold in this case?

57. Laplace's rule of succession. Laplace claimed that when an event happens n times in a row and never fails to happen, the probability that the event will occur the next time is $(n+1)/(n+2)$. Can you suggest a rationale for this claim?

58. Show that the gamma distribution is a conjugate prior for the exponential distribution. Suppose that the waiting time in a queue is modeled as an exponential random variable with unknown (rate) parameter λ and that the average time to serve a random sample of 20 customers is 5.1 minutes. A gamma distribution is used as a prior. Consider two cases: (1) the mean of the gamma is 0.5 and the standard deviation is 1, and (2) the mean is 10 and the standard deviation is 20. Plot the two posterior distributions and compare them. Find the two posterior means and compare them. Explain the differences.

59. Let the unknown probability that a basketball player makes a shot successfully be θ. Suppose your prior is uniform on $[0, 1]$ and that the player then makes two successful shots in a row. Assume that the outcomes of the shots are independent.

    (a) What is the posterior density of θ?

    (b) What would you estimate the probability that the player makes a third successful shot to be?

60. Three friends, Optimist, Realist and Pessimist, go to a casino. They decide to play a gambling game for which they do not know the probability p of winning. Motivated by an exciting lecture on Bayesian statistics, they decide to apply Bayesian analysis to the problem. Optimist chooses a prior on p from the conjugate family of distributions. In addition, he believes that the chances are 50-50 (i.e., the prior expectation is $1/2$) with a prior variance of $1/36$. Realist chooses a uniform prior on $[0, 1]$.

    (a) Show that both priors belong to the family of Beta distributions and find the parameters for these.

    (b) Pessimist does not have any prior beliefs and decides to use the non-informative Jeffreys prior $\pi(p) \propto \sqrt{I(p)}$. What is his prior? Does it also belong to the Beta family?

    (c) Being poor students of Quantitative Finance, the friends do not have enough money to gamble individually, so they decide to play together. They play the game 25 times and win 12 times. What is the posterior distribution for each of them?

    (d) Find the corresponding posterior means and calculate 95 percent Bayesian credible intervals for p for each student. (You could use the fact that if p has a Beta distribution with parameters α and β and $\rho = p/(1-p)$ is the odds ratio, then $\rho\beta/\alpha$ has an F distribution with parameters 2α and 2β.)

    (e) The three friends tell their classmate Skeptic about the exciting Bayesian analysis each of them has performed. However, Skeptic is naturally skeptical about the Bayesian approach. He does not believe in any priors and decides to perform a "classical" (non-Bayesian) analysis of the same data. What is his estimate and 95 percent confidence interval for p? Compare the results and comment on them.

61. The waiting time (at a certain time of day) for a bus at a given bus stop is known to have a uniform distribution on $[0, \theta]$. From similar routes it is known that θ has a Pareto distribution with parameters 7 and 4 (remember that the density of the Pareto satisfies $f(\theta|\alpha, \beta) \propto \theta^{-(\alpha+1)}$, $\theta \geq \beta$). During the last 5 days, waiting times of $10, 3, 2, 5$ and 14 minutes were observed.

    (a) Show that the Pareto distribution provides a conjugate prior for uniform data, and find the posterior distribution of θ.

    (b) Estimate θ with respect to squared error (i.e., find the posterior mean of θ).

    (c) Find a 95 percent HPD for θ.

    (d) Test the hypothesis $H_0 : 0 \leq \theta \leq 15$ versus $H_A : \theta > 15$ by choosing the (a posteriori) more likely hypothesis.

62. Let $X_1, \ldots, X_n$ be an i.i.d. sample from the geometric distribution

$$f(x|p) = p(1-p)^x, \qquad x = 0, 1, \ldots$$

(a) Show that the family of Beta distributions with parameters α and β gives a conjugate prior for p, derive the corresponding posterior distribution, and find the Bayes estimator of p with respect to squared error.

(b) A popular non-informative prior is the Jeffreys prior $\pi(p) \propto \sqrt{I(p)}$, where I is the Fisher information. Find this prior, and repeat the previous question for it.

63. Suppose that X is a discrete random variable with

$$\mathbb{P}(X=0) = 2\theta/3, \quad \mathbb{P}(X=1) = \theta/3, \quad \mathbb{P}(X=2) = 2(1-\theta)/3, \quad \mathbb{P}(X=3) = (1-\theta)/3$$

where $0 \le \theta \le 1$ is a parameter. The following 10 independent observations were taken from such a distribution: (3, 0, 2, 1, 3, 2, 1, 0, 2, 1).

(a) Find the method of moments estimate of θ.

(b) Find an approximate standard error of your estimate.

(c) What is the maximum likelihood estimate of θ?

(d) What is an approximate standard error of the MLE?

(e) If the prior distribution of Θ is uniform on $[0, 1]$, what is the posterior density? Plot it. What is the mode of the posterior?

$\boxed{\pi}$ 64. Suppose that X follows a geometric distribution

$$\mathbb{P}(X = x) = p(1-p)^x, \qquad x = 0, 1, \ldots$$

and assume an i.i.d. sample of size n.

(a) Find the method of moments estimate of p.

(b) Find the MLE of p.

(c) Find the asymptotic variance of the MLE.

(d) Let p have a uniform prior distribution on $[0, 1]$. What is the posterior distribution of p? What is the posterior mean?

65. Suppose that $X_1, \ldots, X_n$ are i.i.d. $N(\mu, \sigma^2)$.

(a) If μ is known, what is the MLE of σ?

(b) If σ is known, what is the MLE of μ?

(c) In the case above (σ known), does any other unbiased estimate of μ have smaller variance?

$\boxed{\pi}$ 66. Let $X_1, \ldots, X_n$ be an i.i.d. sample from an exponential distribution with density function

$$f(x|\tau) = \frac{1}{\tau} e^{-x/\tau}, \qquad x \ge 0.$$

(a) Find the MLE of τ.

(b) What is the exact sampling distribution of the MLE? (Hint: the sum of the X_i follows a gamma distribution.)

(c) Use the central limit theorem to find a normal approximation to the sampling distribution.

(d) Show that the MLE is unbiased, and find its exact variance.

(e) Is there any other unbiased estimate with smaller variance?

(f) Find the form of an approximate confidence interval for τ.

(g) Find the form of an exact confidence interval for τ.

67. Consider a simple linear regression model without an intercept:

$$Y_i = \beta x_i + \epsilon_i, \qquad i = 1, \ldots, n,$$

where the ϵ_i are i.i.d. $N(0, \sigma^2)$.

(a) Find the MLE for β. Is it unbiased? Find its MSE.

(b) Find the MLE for σ.

(c) Find the MLE for β/σ.

(d) Derive the Fisher information matrix $I(\beta, \sigma)$.

(e) Find the Cramer-Rao lower bound for an unbiased estimator of β/σ.

Hint: for (e), use the multivariable version of the Rao-Cramer inequality: if T is unbiased for the scalar valued function $g(\theta)$ (i.e., $E_\theta(T) = g(\theta)$) and the usual regularity conditions hold, then $\mathrm{var}_\theta(T) \geq (\nabla g(\theta))' I(\theta)^{-1} \nabla g(\theta)$.

68. Let $X_1, \ldots, X_n$ be an i.i.d. sample from a Poisson distribution with mean λ, and let $T = \sum_{i=1}^n X_i$.

(a) Show that the distribution of $X_1, \ldots, X_n$ given T is independent of λ, and conclude that T is sufficient for λ.

(b) Show that X_1 is not sufficient.

(c) Use the factorization theorem to show that T is sufficient. Identify the functions g and h of that theorem.

69. Show that $\prod_{i=1}^n X_i$ and $\sum_{i=1}^n X_i$ are sufficient statistics for the gamma distribution.

70. Find sufficient statistics based on samples of size n for the parameters of the following distributions: (a) the uniform distribution on $[0, \theta]$; (b) the uniform distribution on $[-\theta, \theta]$; (c) the uniform distribution on $[-\theta_L, \theta_U]$.

71. Find sufficient statistics based on samples of size n for the parameters of the shifted exponential distribution

$$f(x|\tau, \mu) = \frac{1}{\tau} e^{-(x-\mu)/\tau}, \qquad x \geq \mu,$$

where (a) μ is known and (b) μ is unknown.

$\boxed{\pi}$ 72. Suppose that $X_1, \ldots, X_n$ are i.i.d. random variables on the interval $[0, 1]$ which have a beta distribution with parameters α and 2α, i.e., density

$$f(x|\alpha) = \frac{\Gamma(3\alpha)}{\Gamma(\alpha)\Gamma(2\alpha)} x^{\alpha-1}(1-x)^{2\alpha-1}$$

where $\alpha > 0$ is a parameter to be estimated from the sample. It can be shown that

$$\mathbb{E}(X) = \frac{1}{3}, \qquad \mathrm{var}(X) = \frac{2}{9(3\alpha+1)}.$$

(a) How could the method of moments be used to estimate α?

(b) What equation does the MLE of α satisfy? (Hint: use the digamma and trigamma functions.)

(c) What is the asymptotic variance of the MLE?

(d) Find a sufficient statistic for α.

73. Suppose that $X_1, \ldots, X_n$ are i.i.d. with density function

$$f(x|\theta) = e^{-(x-\theta)}, \qquad x \geq \theta.$$

(a) Find the method of moments estimate of θ.

(b) Find the MLE of θ. (Hint: be careful, and do not differentiate before thinking.)

(c) Find a sufficient statistic for θ.

$\boxed{\pi}$ 74. The Pareto distribution has been used in finance and economics as a model for a density function with a slowly decaying tail:

$$f(x|x_0, \theta) = \theta x_0^\theta x^{-\theta-1}, \qquad x \geq x_0$$

where $\theta > 0$. Assume that $x_0 > 0$ is given and that $X_1, \ldots, X_n$ is an i.i.d. sample from the Pareto distribution.

(a) Find the method of moments estimate of θ.

(b) Find the MLE of θ.

(c) Find the asymptotic variance of the MLE.

(d) Find a sufficient statistic for θ.

$\boxed{\pi}$ 75. Let $X_1, \ldots, X_n$ be i.i.d. random variables with density function

$$f(x|\theta) = (\theta + 1)x^\theta, \qquad 0 \leq x \leq 1.$$

(a) Find the method of moments estimate of θ.

(b) Find the MLE of θ.

(c) Find the asymptotic variance of the MLE.

(d) Find a sufficient statistic for θ.

76. Let X have one of the following distributions:

X	H_0	H_A
x_1	.2	.1
x_2	.3	.4
x_3	.3	.1
x_4	.2	.4

(a) Compute the likelihood ratio Λ for each possible of X value and order the x_i according to Λ.

(b) What is the likelihood ratio test of H_0 versus H_A at level $\alpha = .2$? What is the test at level $\alpha = .5$?

(c) If the prior probabilities are $\mathbb{P}(H_0) = \mathbb{P}(H_1)$, which outcomes favor H_0?

(d) What prior probabilities correspond to the decision rules with $\alpha = .2$ and $\alpha = .5$?

77. True or false, and state why.

(a) The significance level of a statistical test is equal to the probability that the null hypothesis is true.

(b) If the significance level of a test is decreased, the power would be expected to increase.

(c) If a test is rejected at the significance level α, the probability that the null hypothesis is true equals α.

(d) The probability that the null hypothesis is falsely rejected is equal to the power of the test.

(e) A type I error occurs when the test statistic falls in the rejection region of the test.

(f) A type II error is more serious than a type I error.

(g) The power of a test is determined by the null distribution of the test statistic.

(h) The likelihood ratio is a random variable.

78. Consider the introductory coin tossing example from the lecture. Suppose that instead of tossing the coin 10 times, the coin was tossed until a head came up and the total number of tosses X was recorded.

    (a) If the prior probabilities are equal, which outcomes favor H_0 and which outcomes favor H_1?

    (b) Suppose $\mathbb{P}(H_0)/\mathbb{P}(H_1) = 10$. What outcomes favor H_0?

    (c) What is the significance level of the test that rejects H_0 if $X \geq 8$?

    (d) What is the power of this test?

79. Let $X_1, \ldots, X_n$ be am i.i.d. sample from a Poisson distribution. Find the likelihood ratio for testing $H_0 : \lambda = \lambda_0$ versus $H_A : \lambda = \lambda_A$ where $\lambda_A > \lambda_0$. Use the fact that the sum of independent Poisson random variables follows a Poisson distribution to explain how to determine a rejection region for a test at level α. Show that the obtained test is UMP for $H_0 : \lambda = \lambda_0$ versus $H_A : \lambda > \lambda_0$.

80. Let $X_1, \ldots, X_n$ be an i.i.d. sample from the Pareto distribution

$$f(x|\theta, \gamma) = \frac{\theta\gamma^\theta}{x^{\theta+1}}, \qquad x \geq \gamma$$

    where γ is known.

    (a) Find the most powerful level α test for $H_0 : \theta = \theta_0$ versus $H_A : \theta = \theta_A$, where $\theta_A > \theta_0$.

    (b) Is there a UMP test at level α for testing the composite hypotheses $H_0 : \theta \leq \theta_0$ versus $H_A : \theta > \theta_0$? If so, what is its power function?

    (Hint: show that $\log(X_i/\gamma)$ has an exponential distribution.)

81. Suppose that $X_1, \ldots, X_n$ form a random sample from a density function $f(x|\theta)$ for which T is a sufficient statistic for θ. Show that the likelihood ratio test of $H_0 : \theta = \theta_0$ versus $H_A : \theta = \theta_A$ is a function of T. Explain how, if the distribution of T is known under H_0, the rejection region of the test may be chosen so that the test has level α.

82. Let $X \sim N(0, \sigma^2)$ and consider testing $H_0 : \sigma = \sigma_0$ versus $H_A : \sigma = \sigma_A$, where $\sigma_A > \sigma_0$.

    (a) What is the likelihood ratio as a function of x? What values favor H_0? What is the rejection region of a level α test?

    (b) For a sample $X_1, \ldots, X_n$ distributed as above, repeat the previous question.

    (c) Is the test in the previous question uniformly most powerful for testing $H_0 : \sigma = \sigma_0$ versus $H_A : \sigma > \sigma_0$?

83. Suppose that a single observation X is taken from a uniform density on $[0, \theta]$ and consider testing $H_0 : \theta = 1$ versus $H_A : \theta = 2$.

    (a) Find a test that has significance level $\alpha = 0$. What is its power?

(b) For $0 < c < 1$, consider the test that rejects when $X \leq c$. What is its significance level and power?

(c) What is the significance level and power of the test that rejects when $1 - c \leq X \leq 1$?

(d) Does the likelihood ratio test determine a unique rejection region?

(e) What happens if the null and alternative hypothesis are interchanged?

84. Let $X_1, \ldots, X_n$ be an i.i.d. sample from

$$f(x|\theta) = \theta x^{\theta - 1}, \qquad 0 \leq x \leq 1$$

(i.e., a Beta distribution with parameters θ and 1).

(a) Show that the uniform distribution on $[0, 1]$ is a particular case of $f(x|\theta)$.

(b) Find the generalized likelihood ratio test for testing the null hypothesis that the data comes from the uniform distribution on $[0, 1]$.

Hint: show that $-\log(X)$ has an exponential distribution.

85. Let X be a binomial random variable with n trials and probability p of success.

(a) What is the generalized likelihood ratio for testing $H_0 : p = .5$ versus $H_A : p \neq .5$?

(b) Show that the test rejects for large values of $|X - n/2|$.

(c) Using the null distribution of X, show how the significance level corresponding to a rejection region $|X - n/2| > c$ can be determined.

(d) If $n = 10$ and $c = 2$, what is the significance level of the test?

86. True or false:

(a) The generalized likelihood ratio statistic Λ never exceeds 1.

(b) If the p-value is .03, the corresponding test will reject at the significance level .02.

(c) If a test rejects at significance level .06, then the p-value does not exceed .06.

(d) The p-value of a test is the probability that the null hypothesis is correct.

(e) In testing a simple versus simple hypothesis using the likelihood ratio, the p-value equals the likelihood ratio.

(f) If a chi-squared test statistic with 4 degrees of freedom has a value of 8.5, the p-value is less than 0.05.

87. Suppose that a test statistic T has a standard normal null distribution.

(a) If the test rejects for large values of $|T|$, what is the p-value corresponding to $T = 1.50$?

(b) Answer the same question if the test rejects for large T.

88. Suppose that a level α test based on a test statistic T rejects if $T > t_0$. Suppose that g is a monotone-increasing function and let $S = g(T)$. Is the test that rejects if $S > g(t_0)$ a level α test?

89. Suppose that the null hypothesis is true, that the distribution of the test statistic T is continuous with cdf F and that the test rejects for large values of T. Let V denote the p-value of the test.

(a) Show that $V = 1 - F(T)$.

(b) Conclude that the null distribution of V is uniform.

(c) If the null hypothesis is true, what is the probability that the p-value is greater than .1?

(d) Show that the test that rejects if $V < \alpha$ has significance level α.

90. A popular alternative to the generalized likelihood ratio test is the *score test* (also known as Lagrange multiplier test), which tests H_0 versus H_A : not H_0 using the score statistic

$$T_S = S(\tilde{\theta})' I(\tilde{\theta})^{-1} S(\tilde{\theta})$$

where $\tilde{\theta}$ is the (restricted) MLE under the null, $S(\theta) = \nabla_\theta \log(L(\theta|x_1, \ldots, x_n))$ is the score function, and I the Fisher information matrix.

Suppose that $X_1, \ldots, X_n$ are independent Poisson random variables with means $\lambda_1, \ldots, \lambda_n$, respectively. Show that the score statistic for testing $H_0 : \lambda_1 = \cdots = \lambda_n$ (i.e., that all X_i have the same distribution) is given by

$$T_S = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\bar{X}}.$$

91. There is a great deal of folklore about the effects of the full moon on humans and animals. Do animals bite humans more during a full moon? In an attempt to study this question, Bhattacharjee et al. (2000) collected data on admissions to a medical facility for treatment of bites by animals: cats, rats, horses, and dogs. 95% of the bites were by man's best friend, the dog. The lunar cycle was divided into 10 periods, and the number of bites in each period is shown in the following table. Day 29 is the full moon. Is there a temporal trend in the incidence of bites?

Lunar Day	# of Bites
16, 17, 18	137
19, 20, 21	150
22, 23, 24	163
25, 26, 27	201
28, 29, 1	269
2, 3, 4	155
5, 6, 7	142
8, 9, 10	146
11, 12, 13	148
14, 15	110

92. Consider testing goodness of fit for a multinomial distribution with two cells. Denote the number of observations in each cell by X_1 and X_2 and let the hypothesized probabilities be p_1 and p_2. Pearson's chi-squared statistic is equal to

$$X^2 = \sum_{i=1}^{2} \frac{(X_i - np_i)^2}{np_i}.$$

Show this may be expressed as

$$\frac{(X_1 - np_1)^2}{np_1(1 - p_1)}.$$

Show that under the null hypothesis,

$$\frac{X_1 - np_1}{\sqrt{np_1(1 - p_1)}}$$

approximately has a standard normal distribution and hence X^2 as its square has a chi-squared distribution with 1 degree of freedom.

93. Let $X_i \sim \text{binomial}(n_i, p_i)$, $i = 1, \ldots, m$ be independent. Derive a likelihood ratio test for the null hypothesis

$$H_0 : p_1 = p_2 = \cdots = p_m$$

against the alternative hypothesis that the p_i are not all equal. What is the large-sample distribution of this test statistic?

$\boxed{\pi}$ 94. (a) In 1965, a newspaper carried a story about a high school student who reported getting 9207 heads and 8743 tails in 17,950 coin tosses. Is this a significant discrepancy from the null hypothesis $H_0 : p = 1/2$?

(b) Jack Youden, a statistician at the National Bureau of Standards, contacted the student and asked him exactly how he had performed the experiment (Youden, 1974). To save time, the student had tossed groups of five coins at a time, and a younger brother had recorded the results, shown in the following table:

# of Heads	Frequency
0	100
1	524
2	1080
3	1126
4	655
5	105

Are the data consistent with the hypothesis that all coins were fair ($p = 1/2$)?

(c) Are the data consistent with the hypothesis that all five coins had the same probability of heads but that this probability was not necessarily 1/2? (Hint: use the binomial distribution.)

$\boxed{\pi}$ 95. The following 30 claims are for vandal damage to cars in a certain community over a period of six months:

38	56	77	110	112	138	152	168	188	210
228	241	252	273	283	288	291	299	305	317
321	356	374	422	485	527	529	559	567	656

Use the method of percentiles (based on quartiles) to fit a Weibull distribution of the form $F(x|c,\gamma) = 1 - \exp(-cx^\gamma)$ to the data. Cut the data at breakpoints 145, 225, 310 and 420, compute the observed and expected counts for the 5 intervals thus obtained, and perform a chi-squared goodness-of-fit test for the Weibull model.

$\boxed{\pi}$ 96. Use Kolmogorov-Smirnoff tests to test the fitness of the Weibull (ML and/or method of percentiles) and log-normal distributions to the Theft claim data.

$\boxed{\pi}$ 97. In a general insurance portfolio, the frequencies of claim events according to policyholder are as follows:

Number of Claims	0	1	2	3	4	5	6
Frequency	65623	12571	1644	148	13	1	0

Fit both a Poisson and a negative binomial model to this data, and comment on which model provides a better fit.

98. The `bodytemp` data set contains normal body temperature readings (degrees Fahrenheit) and heart rates (beats per minute) of 65 males (coded by one) and 65 females (coded by two) from a study by Shoemaker (1996).

(a) Assess the normality of the male and female body temperatures by using quantile plots. In other to judge the inherent variability of these plots, simulate several samples from normal distributions with matching means and standard deviations, and make quantile plots. What do you conclude?

(b) Repeat the preceding problem for heart rates.

(c) For the males, test the null hypothesis that the mean body temperature is $98.6°$ versus the alternative that the mean is not equal to $98.6°$. Do the same for the females. What do you conclude?

99. Estimate the 0.025, 0.05, 0.95 and 0.975 quantiles of the sample coefficient of skewness under normality by a Monte Carlo experiment. Compare the estimated quantiles with that of the large sample normal approximation $N(0, 6/n)$.

100. The "skewness test of normality" tests normality via H_0: skewness is zero (as is the case for the normal) against H_A: skewness is non-zero, with critical values obtained using the large sample normal approximation $N(0, 6/n)$ of the sample skewness under normality. Estimate the power of this test against symmetric Beta(α, α) distributions and comment on the results. Are the results different for heavy-tailed symmetric alternatives such as the Student t distribution (with few degrees of freedom)?

101. The 2000 U.S. Presidential election was very close and hotly contested. George W. Bush was ultimately appointed to the Presidency by the U.S. Supreme Court. Among the issues was a confusing ballot in Palm Beach County, Florida, the so-called butterfly ballot, for which, although the Democrats were listed in the second row on the left, a voter wishing to specify them would have to punch the third hole—punching the second would result in a vote for the Reform Party (Pat Buchanan). After the election, many distraught Democratic voters claimed that they had inadvertently voted for Buchanan, a right-wing candidate.

    The data set `palmbeach` contains relevant data: vote counts by county in Florida and for four presidential candidates in 2000, the total vote counts in 2000, the presidential vote counts for three presidential candidates in 1996, the vote count for Buchanan in the 1996 Republican primary, the registration in Buchanan's reform party, and the total registrations in the county. Does this data support voters' claims that they were misled by the form of the ballot? Start by making two scatterplots: a plot of Buchanan's votes versus Bush's votes in 2000, and a plot of Buchanan's votes in 2000 versus his votes in the 1996 primary.

102. The `bodytemp` data set contains normal body temperature readings (degrees Fahrenheit) and heart rates (beats per minute) of 65 males (coded by one) and 65 females (coded by two) from a study by Shoemaker (1996).

    (a) For both males and females, make scatterplots of heart rate versus body temperature. Comment on the relationship or the lack thereof.

    (b) Quantify the strengths of the relationships by computing Pearson and rank correlation coefficients.

    (c) Does the relationship for males appear to be the same as for females? Examine this question graphically, by making a scatterplot showing both females and males and identifying females and males by different plotting symbols.

103. In 1970, the U.S. Congress instituted a lottery for the military draft to support the unpopular war in Vietnam. All 366 possible birth dates were placed in plastic capsules in a rotating drum and were selected one by one. Eligible males born on the first day drawn were first in line to be drafted followed by those on the second day drawn, etc. The results were criticized by some who claimed that government incompetence at running a fair lottery resulted in a tendency to men born later in the year being more likely to be drafted. Indeed, later investigations revealed that the birthdates were placed in the drum month by month and

were not thoroughly mixed. The variables in the data set `1970lottery` are month, month number, day of the year, and draft number.

(a) Plot draft number versus day number. Do you see any trend?

(b) Calculate the Pearson and rank correlation coefficients. What do they suggest?

(c) Is the correlation statistically significant? One way to assess this is via a permutation test. Randomly permute the draft numbers and find the correlation of this random permutation with the day numbers. Do this 100 times and see how many resulting correlation coefficients exceed the one observed in the data. If you are not satisfied with 100 times, do it 1,000 times.

(d) Make parallel boxplots of the draft numbers by month. Do you see any pattern?

104. Two independent samples are to be compared to see if there is a difference in the population means. If a total of s subjects are available in the experiment, how should this total be allocated between the two samples in order to (a) provide the shortest confidence interval for $\mu_X - \mu_Y$ and (b) make the test of $H_0 : \mu_X = \mu_Y$ as powerful as possible? Assume that the observations in the two samples are normally distributed with the same variance.

105. Let $X_1, \ldots, X_n$ be i.i.d. with cdf F, and $Y_1, \ldots, Y_m$ be i.i.d. with cdf G. The hypothesis to be tested is that $F = G$. Suppose for simplicity that $m + n$ is even so that in the combined sample of X's and Y's, $(m + n)/2$ observations are less than the median and $(m + n)/2$ are greater.

(a) As a test statistic, consider T, the number of X's less than the median of the combined sample. Show that under the null hypothesis T follows a hypergeometric distribution:

$$\mathbb{P}(T = t) = \frac{\binom{(m+n)/2}{t}\binom{(m+n)/2}{n-t}}{\binom{m+n}{n}}.$$

Explain how to form a rejection region for this test.

(b) Show to to find a confidence interval for the difference between the median of F and the median of G under the shift model $G(x) = F(x - \Delta)$. (Hint: use the order statistics.)

(c) Apply the results to the `bearings` data set recording the performances of engine bearings made of different compounds (McCool, 1979).

106. (Permutation Test for Means.) Consider the fusion of ice data (data files `icea` and `iceb` for methods A and B, respectively). We ask whether the measurements provided by methods A and B are identical, or exchangeable in the following sense. There are $13 + 8 = 21$ measurements in all and there are $\binom{21}{8}$ ways that 8 of these could be assigned to method N. Is the particular assignment we have observed unusual among these in the sense that the means of the two samples are unusually different?

(a) It is not inconceivable but may be asking too much to generate all $\binom{21}{8}$ partitions. So just choose a random sample of these partitions, say of size 1000, and make a histogram of the resulting values of $\bar{X}_A - \bar{X}_B$. Where on this distribution does the value of $\bar{X}_A - \bar{X}_B$ that was actually observed fall?

(b) In what way is this procedure similar to the Mann-Whitney test?

107. For the fusion of ice data, use a bootstrap to estimate the standard error of and a confidence interval for $\bar{X}_A - \bar{X}_B$.

108. If $X \sim N(\mu_X, \sigma_X^2)$ and Y is independent $N(\mu_Y, \sigma_Y^2)$, what is $\pi = \mathbb{P}(X < Y)$ in terms of the parameters μ_X, μ_Y, σ_X and σ_Y?

109. This problem contrasts the power functions of paired and unpaired designs. Graph and compare the power curves for testing $H_0 : \mu_X = \mu_Y$ for the following two designs.

(a) Paired: $\text{cov}(X_i, Y_i) = 50$, $\sigma_X = \sigma_Y = 10$, $i = 1, \ldots, 25$.

(b) Unpaired: $X_1, \ldots, X_{25}$ and $Y_1, \ldots, Y_{25}$ are independent with variances as above.

110. Date file `phonelines` contains the results of an experiment to test a method for reducing faults on telephone lines (Welch, 1987). Fourteen matched pairs of areas were used, and fault rates for the test and control areas were recorded.

   (a) Plot the differences versus the control rates and summarize what you see.

   (b) Calculate the mean difference, its standard deviation, and a a confidence interval.

   (c) Calculate the median difference and a confidence interval and compare to the previous result.

   (d) Do you think it is more appropriate to use a t test or a nonparametric method to test whether the apparent difference between test and control could be due to chance? Why? Carry out both tests and compare.

111. Biological effects of magnetic fields are a matter of current research and concern. Data file `magfield` is from an early study of the effects of a strong magnetic field on the development of mice (Barnothy, 1964). 10 cages, each containing 30-day-old albino female mice, were subjected for a period of 12 days to a field with an average strength of 80 Oe/cm. Thirty other mice housed in 10 similar cages were not placed in a magnetic field and served as controls. The weights gains in grams for each of the cages were recorded.

   (a) Display the data using parallel dotplots.

   (b) Find a 95% confidence interval for the difference of the mean weight gains.

   (c) Use a t test to assess the statistical significance of the observed difference. What is the p-value of the test?

   (d) Repeat using a nonparametric test.

   (e) What is the difference between the median weight gains?

   (f) Use the bootstrap to estimate the standard error of the difference of median weight gains.

   (g) Form a confidence interval for the difference of median weight gains based on the bootstrap approximation to the sampling distribution.

112. The *Hodges-Lehmann shift estimate* is defined to be $\hat{\Delta} = \text{median}(X_i - Y_j)$, where $X_1, \ldots, X_n$ are independent observations from a distribution F and $Y_1, \ldots, Y_m$ are independent observations from a distribution G, independent of the X_i.

   (a) Show that if F and G are normal distributions, then $\mathbb{E}(\hat{\Delta}) = \mu_X - \mu_Y$.

   (b) Why is $\hat{\Delta}$ robust to outliers?

   (c) What the $\hat{\Delta}$ for the magnetic field data, and how does it compare to the differences of the means and the medians?

   (d) Use the bootstrap to approximate the sampling distribution and the standard error of $\hat{\Delta}$.

   (e) From the bootstrap approximation to the sampling distribution, form an approximate 90% confidence interval for $\hat{\Delta}$.

113. The `bodytemp` data set contains normal body temperature readings (degrees Fahrenheit) and heart rates (beats per minute) of 65 males (coded by one) and 65 females (coded by two) from a study by Shoemaker (1996).

   (a) Using normal theory, form a 95% confidence interval for the difference of mean body temperatures between males and females. Is the use of the normal approximation reasonable?

(b) Using normal theory, form a 95% confidence interval for the difference of mean heart rates between males and females. Is the use of the normal approximation reasonable?

(c) Use both parametric and nonparametric tests to compare the body temperatures and heart rates. What to you conclude?

114. It is conventional wisdom in military squadrons that pilots tend to father more girls than boys. Snyder (1961) gathered data for military fighter pilots. The sex of the pilots' offspring was tabulated for three kinds of flight duty during the month of conception, as shown in the following table.

Father's Activity	Female	Male
Flying Fighters	51	38
Flying Transports	14	16
Not Flying	38	46

Is there any significant difference between the three groups? In the United States in 1950, 105.37 males were born for every 100 females. Are the data consistent with this sex ratio?

115. Grades in an elementary statistics class were classified by the students' majors.

	A	B	C	D-F
Psychology	8	14	15	3
Biology	15	19	4	1
Other	13	15	7	4

Is there any relationship between grade and major?

116. This problem considers some more data on Jane Austen and her imitator (Morton, 1978). The following table gives the relative frequency of the word *a* preceeded by (PB) and not preceeded by (NPB) the word *such*, the word *and* followed by (FB) or not followed by (NFB) *I*, and the word *the* preceded by and not preceded by *on*.

	Sense and Sensibility	*Emma*	*Sandition I*	*Sandition II*
a PB *such*	14	16	8	2
a NPB *such*	133	180	93	81
and FB *I*	12	14	12	1
and NFB *I*	241	285	139	153
the PB *on*	11	6	8	17
the NPB *on*	259	265	221	204

Was Austen consistent in these habits of style from one work to another? Did her imitator successfully copy this aspect of her style?

117. A market research team conducted a survey to investigate the relationship of personality to attitude towards small cars. A sample of 250 adults in a metropolitan area were asked to fill out a 16-item self-perception questionnaire, on the basis of which they were classified into three types: cautious conservative, middle-of-the-roader, and confident explorer. They were then asked to given their overall opinion of small cars: favorable, neutral, or unfavorable.

	Cautious	Midroad	Explorer
Favorable	79	58	49
Neutral	10	8	9
Unfavorable	10	34	42

Is there a relationship between personality type and attitude towards small cars? If so, what is the nature of the relationship?

$\boxed{\pi}$ 118. Is it advantageous to wear the color red in a sporting contest? Hill and Barton (2005) explain that it is specifically the color red that correlates with male dominance and testosterone levels. In the 2004 Olympic Games, contestants in four combat sports were randomly assigned red or blue outfits (or body protectors), with winners tabulated according to color as follows.

	Red	Blue
Boxing	148	120
Freestyle Wrestling	27	24
Greco-Roman Wrestling	25	23
Tae Kwan Do	45	35

(There are several `red-blue` data files with more details.)

(a) Let π_R denote the probability that the contestant wearing red wins. Test the null hypothesis that $\pi_R = 1/2$ versus the alternative hypothesis that π_R is the same in each sport, but $\pi_R \neq 1/2$.

(b) Test the null hypothesis $\pi_R = 1/2$ against the alternative hypothesis that allows π_R to be different in different sports, but not equal to $1/2$.

(c) Are either of these hypothesis tests equivalent to that which would test the null hypothesis $\pi_R = 1/2$ versus the alternative hypothesis $\pi_R \neq 1/2$, using as data the total numbers of wins summed over all the sports?

(d) Is there any evidence that wearing red is more favorable in some of the sports than others?

119. If X is not full rank, then minimizing $\|y - X\beta\|^2$ does not have a unique solution. Show that $\hat{\beta} = (X'X)^{+}X'y$ gives the minimizer with minimum length, where $(X'X)^{+}$ is the Moore-Penrose inverse of $X'X$.

Hint: if X has rank r and $X = U_r D_r V_r'$ is the compact SVD of X, then $(X'X)^{+} = V_r D_r^{-2} V_r'$.

120. Let X have full column rank p with QR decomposition $X = QR$. Show that the least squares estimator can be obtained by solving $R\beta = Q'y$. How can this be solved efficiciently?

121. Consider the linear model $\mathbb{E}(y) = X\beta$, $\text{cov}(y) = \sigma^2 I_n$ with $\text{rank}(X) = p$. Show that if $\tilde{\beta} = Ay$ is any unbiased linear estimator of β, then $\text{cov}(\tilde{\beta}) \geq \text{cov}(\hat{\beta})$ in the sense that the difference of the covariance matrices is non-negative definite, with equality iff $\tilde{\beta} = \hat{\beta}$.

122. A k-parameter exponential family has densities (with respect to a suitable reference measure) of the form

$$f(y|\theta) = \exp\left(\sum_{j=1}^{k} c_j(\theta)T_j(y) + d(\theta) + S(y)\right),$$

where $c_1(\theta), \ldots, c_k(\theta)$ are called the natural parameters of the family.

Consider the normal linear regression model $y \sim N(X\beta, \sigma^2 I_n)$. Show that the distribution of y belongs to a $(p+1)$-parameter exponential family with natural parameters $\beta_1/\sigma^2, \ldots, \beta_p/\sigma^2, 1/\sigma^2$.

Show also that $(X'y, y'y)$ is sufficient for (β, σ^2).

123. Consider the normal linear regression model $y \sim N(X\beta, \sigma^2 I_n)$. Show that

$$\hat{y} \sim N(X\beta, \sigma^2 P_X), \qquad \hat{e} \sim N(0, \sigma^2 Q_X)$$

and that $\hat{y}$ and $\hat{e}$ are independent.

124. Consider the normal linear regression model $y \sim N(X\beta, \sigma^2 I_n)$. Partition $X = [X_1, X_2]$ and similarly $\beta = [\beta_1', \beta_2']'$. Derive the generalized LRT for the significance of any of the predictors in X_2.

In R, we can use `anova()` for this. Use this to test whether for the German data, the linear model of Amount on Duration can significantly be improved by adding Age as a predictor, and verify that the R output agrees with the theoretical result.

125. Consider the normal linear regression model $y \sim N(X\beta, \sigma^2 I_n)$. Partition $X = [X_1, X_2]$ and similarly $\beta = [\beta_1', \beta_2']'$. What is the restricted MLE of β under $H_0 : \beta_2 = 0$?

(If necessary, use the block matrix inversion formula from `https://en.wikipedia.org/wiki/Block_matrix`.)

126. Suppose that $X \in \mathbb{R}^{n \times p}$ contains the constant predictor. Show that $\hat{y} = X\hat{\beta}$ maximizes the sample correlation between y and any linear combination $y = X\beta$ of the columns of X.

127. Consider the linear regression model $\mathbb{E}(y) = X\beta$, $\operatorname{cov}(y) = \sigma^2 I_n$. Let L be an $r \times p$ matrix of rank $r \leq p$. Show that $L\hat{\beta}$ is a BLUE for $L\beta$.

128. Consider the normal linear regression model $y \sim N(X\beta, \sigma^2 I_n)$. Suppose that $y_0 \sim N(x_0'\beta, \sigma^2)$ is independent from y. Construct a $100(1-\alpha)\%$ confidence interval for $x_0'\beta$ and a $100(1-\alpha)\%$ prediction interval for y_0.

129. Let $y_i \sim N(\mu, \sigma^2)$ be i.i.d. Determine the generalized LRT for $H_0 : \mu = 0$ versus $H_A : \mu \neq 0$.

130. Suppose we have independent random variables

$$y_{j,k} \sim N(\mu_j, \sigma^2), \qquad k = 1, \dots, n_j, \quad j = 1, \dots, p.$$

Write $n = \sum_{j=1}^p n_j$ and

$$\bar{y}_{j,\cdot} = \frac{1}{n_j} \sum_{k=1}^{n_j} y_{j,k}, \qquad \bar{y}_{\cdot,\cdot} = \frac{1}{n} \sum_{j=1}^{p} \sum_{k=1}^{n_j} y_{j,k}$$

for the sample means of group j and the total sample mean, respectively. Show that the generalized LRT for

$$H_0 : \mu_1 = \cdots = \mu_K$$

against the H_A that not all μ_j are the same rejects H_0 if

$$F = \frac{\sum_{j=1}^p n_j (\bar{y}_{j,\cdot} - \bar{y}_{\cdot,\cdot})^2 \Big/ (p-1)}{\sum_{j=1}^p \sum_{k=1}^{n_j} (y_{j,k} - \bar{y}_{j,\cdot})^2 \Big/ (n-p)}$$

is large, and that under H_0, $F \sim F_{p-1,n-p}$. This is called *one-way analysis of variance* (ANOVA).

Hint: remember that for testing a linear hypothesis in a normal linear regression model, the GLRT has

$$F = \frac{\|X(\hat{\beta} - \hat{\beta}_0)\|^2 / r}{\|Q_X y\|^2 / (n-p)}.$$