# MAST20005/MAST90058: Week 2 Problems

1. Let $X_1, X_2, \ldots, X_9$ be a random sample. We are told that $\mathbb{E}(X_3) = 7$ and $\text{var}(X_4) = 4$.
   (a) What is $\text{sd}(X_2)$?
   (b) What is $\text{var}(X_7 + X_8)$?
   (c) What is $\text{cov}(X_3, X_4)$?
   (d) What is an approximate distribution of the sample mean, $\bar{X}$?

2. Let $Y = X_1 + \cdots + X_{15}$ be the sum of iid rvs, each with pdf $f(x) = (3/2)x^2$ where $-1 < x < 1$.
   (a) What is $\mathbb{E}(X_1)$?
   (b) Calculate $\mathbb{E}(Y)$ and $\text{var}(Y)$.
   (c) We would like to calculate $\Pr(-0.3 < Y < 1.5)$. Use the Central Limit Theorem to approximate this probability. *Hint*: $\Phi(-0.1) = 0.4602$ and $\Phi(0.5) = 0.6915$, where $\Phi(\cdot)$ is the standard normal cdf.

3. In each of the following scenarios, is the sample that is described a random sample? What assumptions are being made? Are they realistic? What is the 'population' in each case?
   (a) Tingjin runs a plant experiment. He creates ten pots, plants an identical seed in each one, and leaves them in the same spot in the sun. After 6 weeks he measures the height of each plant, giving measurements $x_1, x_2, \ldots, x_{10}$.
   (b) Damjan measures the height of all of his immediate family members, giving measurements $y_1, y_2, y_3, y_4$.
   (c) Every day, Robert counts the number of people sitting down on South Lawn. He does this for 100 days in a row, giving counts $z_1, z_2, \ldots, z_{100}$.

4. Consider the following realisations from $X$:

   43.1 48.9 42.6 43.7 41.0

   Note that: $\sum_{i=1}^5 x_i = 219.3$, $\sum_{i=1}^5 x_i^2 = 9654.27$.
   (a) Reminder from the lectures: the sample mean, $\bar{x} = n^{-1}\sum_{i=1}^n x_i$, is a measure of location; the sample variance, $s^2 = (n-1)^{-1}\sum_{i=1}^n (x_i - \bar{x})^2$, is a measure of spread; the sample standard deviation, $s = \sqrt{s^2}$, is another measure of spread.
      i. Show that $\sum_{i=1}^n (x_i - \bar{x}) = 0$ and $s^2 = (n-1)^{-1}(\sum_{i=1}^n x_i^2 - n\bar{x}^2)$.
      ii. Compute the mean and standard deviation for the dataset above.
   (b) How would your answers to part (a) change if we multiplied each data point by 2?
   (c) Reminder: the default ('Type 7') sample quantiles in R are defined as $\hat{\pi}_p = x_{(k)}$, where $k = 1 + (n-1)p$. The set of statistics {minimum, $\hat{q}_1$, $\hat{q}_2$, $\hat{q}_3$, maximum} is often called the 'five-number summary', where the $\hat{q}_i$ are the sample quartiles.
      i. Calculate the five-number summary of the above dataset.
      ii. The interquartile range (IQR) is $\hat{q}_3 - \hat{q}_1$. Calculate the IQR of the above dataset.
      iii. Draw a box plot for this dataset using these quantiles.

(d) An alternate definition of sample quantiles is given by: $\widetilde{\pi}_p = x_{(i)} + r \cdot (x_{(i+1)} - x_{(i)})$, where $(n+1)p = i + r$ such that $i$ is an integer and $0 \leqslant r < 1$.

i. Re-compute the five-number summary using this definition.

ii. Show that $\widetilde{\pi}_p$ are the 'Type 6' quantiles (as defined in the lectures).

(e) Outliers are observations that don't seem to belong with the rest of the data. They can occur through data entry errors or problems with an experiment. Extreme observations are not necessarily errors but there is a crude convention for when to identify and label them as 'outliers': an observation $x$ is an outlier if $x < \hat{\pi}_{0.25} - k \times \text{IQR}$ or $x > \hat{\pi}_{0.75} + k \times \text{IQR}$, where typically $k = 1.5$. These are often depicted graphically on a box plot, by only extending the whiskers up to $k \times \text{IQR}$ from each quartile and plotting each outlier as an individual point. (This is the default way that R will draw box plots.) According to this convention, are there any outliers in the sample above?

5. Create a sample of 4 numbers from $\{1, 2, 3, 4\}$, with repeats allowed, that maximises the sample variance.

6. The following are Prostaglandin-endoperoxide synthase 2 (COX2) measurements on tissue samples from 10 mice (COX2 is a protein involved in inflamatory processes related to cancer):

```
10.39  10.43  9.99  11.17  8.91  11.20  11.38  7.74  10.61  11.11
```

(a) Calculate the five-number summary (you may use either Type 6 or Type 7 quantiles).

(b) Are there any outliers in this sample (using R's default convention)?

(c) Draw a box plot for this dataset.

7. The following are observations on maximum rainfall (cm/day) in a year measured by a weather station in Tasmania (King Island Airport) in 10 consecutive years:

```
9.9  4.7  20.5  1.8  4.7  9.8  20.5  20.2  6.5  3.0
```

(a) Draw a histogram of these data.

(b) Suppose that the random variable $Y$ follows the extreme value (EV) distribution which depends on a location parameter $\theta$, and a scale parameter $\xi$. This distribution is obtained as maximum of a set of values (maximum wind speed, precipitation, peak flow, etc.). This distribution has the property that we can write $Y$ as

$$Y = \theta + \xi Z,$$

where $Z$ has the standard EV distribution with cdf $F(z) = e^{-e^{-z}}$. Thus, the inverse cdf function is $F^{-1}(p) = -\ln(-\ln p)$, so we expect that,

$$x_{(k)} \approx \theta + \xi F^{-1}\left(\frac{k}{n+1}\right).$$

Use a QQ plot to assess whether the EV model looks correct for these data. How could you estimate the parameters $\theta$ and $\xi$ based on your plot?

**(Drawing the plot will be easier with a computer in the lab class, but you can discuss this problem in the tutorial beforehand.)**

# MAST20005/MAST90058: Week 3 Problems

1. (a) Let $X_1, \ldots, X_n$ be a random sample from $\mathrm{N}(\mu, \sigma^2)$ where $-\infty < \mu < \infty$ and $\sigma^2 > 0$. Assume that $\sigma^2$ is known (i.e. it is a fixed, known value). Show the maximum likelihood estimator of $\mu$ is $\hat{\mu} = \bar{X}$.

   (b) A random sample $X_1, \ldots, X_n$ of size $n$ is taken from a Poisson distribution with mean $\lambda > 0$.

   i. Show the maximum likelihood estimator of $\lambda$ is $\hat{\lambda} = \bar{X}$.

   ii. Suppose with $n = 40$ we observe 5 zeros, 7 ones, 12 twos, 9 threes, 5 fours, 1 five, and 1 six. What is the maximum likelihood estimate of $\lambda$?

   (c) Let $X_1, \ldots, X_n$ be random samples from the following probability density functions. In each case find the maximum likelihood estimator $\hat{\theta}$.

   i. $f(x \mid \theta) = \frac{1}{\theta^2} x \exp(-x/\theta), \quad 0 < x < \infty, \quad 0 < \theta < \infty$

   ii. $f(x \mid \theta) = \frac{1}{2\theta^3} x^2 \exp(-x/\theta), \quad 0 < x < \infty, \quad 0 < \theta < \infty$

   iii. $f(x \mid \theta) = \frac{1}{2} \exp(-|x - \theta|), \quad -\infty < x < \infty, \quad -\infty < \theta < \infty$

   *Hint*: The last part involves minimizing $\sum_{i=1}^n |x_i - \theta|$, which is tricky. Try $n = 5$ and the sample {6.1, -1.1, 3.2, 0.7, 1.7}. Then deduce the MLE in general.

2. Consider a random sample of $n$ observations on $X$ having the following pmf:

| $x$ | 0 | 1 | 2 |
|---|---|---|---|
| $p(x)$ | $1-\theta$ | $3\theta/4$ | $\theta/4$ |

   (a) Find two unbiased estimators for $\theta$: one based on $X$, and one based on $Z = \text{freq}(0) = \sum_{i=1}^n I(X_i = 0)$ (i.e. the frequency of $x = 0$).

   (b) Compare the two estimators above in terms of their variance.

3. Let $f(x \mid \theta) = \theta x^{\theta - 1}$, $0 < x < 1$, $0 < \theta < \infty$ and let $X_1, \ldots, X_n$ denote a random sample from this distribution. Note that,

$$\int_0^1 x\theta\, x^{\theta-1} dx = \frac{\theta}{\theta + 1}.$$

   (a) Sketch the pdf of $X$ for $\theta = 1/2$ and $\theta = 2$.

   (b) Show that $\hat{\theta} = -n / \left(\sum_{i=1}^n \ln X_i\right)$ is the maximum likelihood estimator of $\theta$.

   (c) For each of the following three sets of observations from this distribution, compute the maximum likelihood estimates and the method of moments estimates.

| $X$ | $Y$ | $Z$ |
|---|---|---|
| 0.0256 | 0.9960 | 0.4698 |
| 0.3051 | 0.3125 | 0.3675 |
| 0.0278 | 0.4374 | 0.5991 |
| 0.8971 | 0.7464 | 0.9513 |
| 0.0739 | 0.8278 | 0.6049 |
| 0.3191 | 0.9518 | 0.9917 |
| 0.7379 | 0.9924 | 0.1551 |
| 0.3671 | 0.7112 | 0.0710 |
| 0.9763 | 0.2228 | 0.2110 |
| 0.0102 | 0.8609 | 0.2154 |

   ($\sum_{i=1}^n \ln(x_i) = -18.2063$, $\sum_{i=1}^n \ln(y_i) = -4.5246$, $\sum_{i=1}^n \ln(z_i) = -10.42968$, $\sum_{i=1}^n x_i = 3.7401$, $\sum_{i=1}^n y_i = 7.0592$, $\sum_{i=1}^n z_i = 4.6368$)

4. Let $X_1, \ldots, X_n$ be a random sample from the exponential distribution whose pdf is $f(x \mid \theta) = (1/\theta)\exp(-x/\theta)$, $0 < x < \infty$, $0 < \theta < \infty$.

   (a) Show that $\bar{X}$ is an unbiased estimator of $\theta$.

   (b) Show that the variance of $\bar{X}$ is $\theta^2/n$.

   (c) Calculate an estimate of $\theta$ if a random sample gave the following values:

   3.5 8.1 0.9 4.4 0.5

5. Let $X_1, \ldots, X_n$ be a random sample from a distribution having finite variance $\sigma^2$. Show that

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

   is an unbiased estimator of $\sigma^2$.

   *Hint*: Use a result from question 4(a)(i) from week 2 to derive an alternative expression for $S^2$ and then compute $E(S^2)$.

6. Let $X_1, \ldots, X_n$ be iid observations from $\mathrm{N}(0, \theta^2)$. Consider the estimators $S^2$ and $\hat{\theta}^2 = n^{-1}\sum_{i=1}^{n} X_i^2$. Show that $\hat{\theta}^2$ is unbiased and $\mathrm{var}(\hat{\theta}^2) < \mathrm{var}(S^2)$ for any $n > 1$.

7. Let $X_1, \ldots, X_n$ be iid observations from $X \sim \mathrm{N}(\mu, \sigma^2)$. Since $X$ has a symmetric pdf, we might expect that both the sample mean $\bar{X}$ and the sample median $\hat{\pi}_{0.5}$ will be good estimators of the population mean $\mu$.

   (a) Find the variance of $\bar{X}$.

   (b) In general, the sample median will approximately follow a normal distribution, $\hat{\pi}_{0.5} \sim \mathrm{N}(\pi_{0.5}, \pi/2 \times \sigma^2/n)$, where $\pi_{0.5}$ is the true median (we will learn more about this later in the semester). How does the variance of the sample median compare with that of the sample mean?

   (c) Are the estimators biased?

   (d) Which estimator do you expect to be more accurate?

# MAST20005/MAST90058: Week 4 Problems

*Some useful information for many of the problems is shown at end of this problem sheet.*

1. Consider the COX2 mouse data from problem 6 in week 2. Here are some descriptive statistics from an analysis of them in R:

```
> x <- c(10.39, 10.43, 9.99, 11.17,  8.91,
+        11.20, 11.38, 7.74, 10.61, 11.11)
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   7.74   10.09   10.52   10.29   11.15   11.38
> sd(x)
[1] 1.159109
```

Assume that these measurements are a random sample from a normal distribution.

   (a) Provide two estimates of the population mean, one based on the sample mean and one on the sample median.

   (b) What is the standard error of each of these estimates?
   *Hint*: One of the problems from last week is helpful here.

2. A random sample of size 16 from $N(\mu, 25)$ gave $\bar{x} = 73.8$. Find a 95% CI for $\mu$.

3. A nut shop sells peanuts in 2 kg bags that are weighed on an old 25 kg scale. Suppose it is known that the standard deviation of weights is 0.12 kg. If a sample of 16 bags of peanuts were carefully weighed in a laboratory and the average weight was 2.09 kg, find an approximate 95% confidence interval for the mean weight of peanuts in the '2 kg' bags sold by the nut shop.

4. To determine whether the bacteria count at St Kilda Beach is at normal levels, $n = 37$ samples of water were taken at the beach and the number of bacteria colonies in 100 milliliters of water from each sample were counted. The sample characteristics were $\bar{x} = 11.95$ and $s = 11.80$, measured in hundreds of colonies. Find the approximate 95% confidence interval for the mean number of colonies (measured in hundreds of colonies) in 100 milliliters of water at the beach.

5. Thirteen tons of cheese is stored in some old gypsum mines, including some wheels labelled '22 kg'. A random sample of 9 of these wheels yields $\bar{x} = 20.9$ and $s = 1.858$. Assuming that the weights of these wheels follows $N(\mu, \sigma^2)$, find a 95% confidence interval for $\mu$. Is the claim these are '22 kg' wheels reasonable?

6. The length of life of brand X light bulbs is assumed to be $N(\mu_X, 784)$. The length of life of brand Y light bulbs is assumed to be $N(\mu_Y, 627)$ and these lifetimes are independent of $X$. If a random sample of $n = 56$ brand X light bulbs yielded $\bar{x} = 937.4$ hours and a random sample of size $m = 57$ brand Y light bulbs yielded $\bar{y} = 988.9$, find a 95% confidence interval for the difference in mean lifetimes, $\mu_X - \mu_Y$. Is it reasonable to conclude that the two brands of light bulb have the same mean lifetimes?

Quantiles of a standard normal: $\Phi^{-1}(0.975) = 1.960$, $\Phi^{-1}(0.95) = 1.645$
Quantiles of $t_{37}$: $F^{-1}(0.975) = 2.026$, $F^{-1}(0.95) = 1.687$
Quantiles of $t_{36}$: $F^{-1}(0.975) = 2.028$, $F^{-1}(0.95) = 1.688$
Quantiles of $t_9$: $F^{-1}(0.975) = 2.262$, $F^{-1}(0.95) = 1.833$
Quantiles of $t_8$: $F^{-1}(0.975) = 2.306$, $F^{-1}(0.95) = 1.859$

# MAST20005/MAST90058: Week 5 Problems

*Some useful information for many of the problems is shown at end of this problem sheet.*

1. Let $X$ be the length in centimeters of a species of fish when caught in the spring. A random sample of 13 observations yielded the sample variance $s^2 = 37.751$. Find a 95% confidence interval for $\sigma$.

2. A test was conducted to determine if a wedge on the end of a plug designed to hold a seal onto that plug was operating correctly. The data were the force required to remove a seal from the plug with the wedge in place ($X$) and without the wedge ($Y$). Assume the distributions of $X$ and $Y$ are $N(\mu_X, \sigma^2)$ and $N(\mu_Y, \sigma^2)$ respectively. Samples of size 10 on each variable yielded:

| | $n$ | $\bar{x}$ | $s$ |
|---|---|---|---|
| $X$ | 10 | 2.548 | 0.323 |
| $Y$ | 10 | 1.564 | 0.210 |

   (a) Find a 95% confidence interval for $\mu_X - \mu_Y$.

   (b) Do you think the wedge is operating correctly?

   (c) How could you support the assumption $\text{var}(X) = \text{var}(Y)$ in part (a)?

3. A candy maker produces mints that have a mean weight of 20.4 grams. For quality assurance, $n = 16$ mints were selected at random from the Wednesday morning shift, yielding $\bar{x} = 21.95$ grams and $s_x = 0.197$. On Wednesday afternoon $m = 13$ mints were selected at random, giving $\bar{y} = 21.88$ grams and $s_y = 0.318$. Find a 90% confidence interval for $\sigma_x/\sigma_y$, the ratio for the standard deviations of the mints produced by the morning and afternoon shifts respectively. Is it reasonable to suppose the standard deviations are the same in the two shifts?

4. A machine shop manufactures toggle levers. A lever is flawed if a standard nut cannot be screwed onto the threads. Let $p$ be the proportion of flawed toggle nuts the shop manufactures. There were 24 flawed levers out of a sample of 642 that were selected randomly from the production line.

   (a) Give a point estimate of $p$.

   (b) Find an approximate 95% confidence interval.

   (c) Find a one-sided 95% confidence interval that gives an upper bound for $p$.

5. Let $X$ be the length of a male grackle (a type of bird). Suppose $X \sim N(\mu, 4.84)$. Find the sample size that is needed to produce an estimate of $\mu$ that is accurate to within $\pm 0.4$, as measured by a 95% confidence interval.

6. We wish to hold a public opinion poll for a close election. Let $p$ denote the proportion of votes who favour candidate A. How large a sample should be taken if we want the maximum error of the estimate of $p$ to be equal to:

   (a) 0.03, with 95% confidence interval?

   (b) 0.02, with 95% confidence interval?

   (c) 0.03, with 90% confidence interval?

7. We obtain the following random sample on $X \sim \mathrm{N}(\mu, \sigma^2)$.

33.8 32.2 30.7 35.4 31 30.3 26.8 33.2 27.8 27.2

(a) Give a point estimate and 90% confidence interval for $\mu$.

(b) Give a point estimate and 95% confidence interval for $\sigma$.

(c) Give a 90% prediction interval for a future observation on $X$.

8. We observe the outcome of $n$ Bernoulli trials with parameter $p$. In the lectures we obtained an approximate confidence interval for $p$ by using two approximations: (i) the Central Limit Theorem to get a sampling distribution for the sample proportion, $\hat{p}$; and (ii) the 'plug-in' approximation to get a standard error. If we avoid making the second approximation we can derive a more accurate confidence interval. The exercises below take you through the steps in the derivation. We will refer to this as the *quadratic approximation*.

(a) Write an appropriate probability interval for deriving this confidence interval.

(b) We need to rearrange this to be in terms of $p$. Show that this is equivalent to solving the following quadratic inequality:

$$(\hat{p} - p)^2 < c^2 p(1-p)/n,$$

where $c = \Phi^{-1}(1 - \alpha/2)$.

(c) Show that the solution to this inequality (in terms of $p$) is an interval with the following endpoints:

$$\frac{\hat{p} + \frac{c^2}{2n} \pm c\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{c^2}{4n^2}}}{1 + \frac{c^2}{n}}.$$

**Quantiles of various distributions**

Standard normal: $\Phi^{-1}(0.975) = 1.960$, $\Phi^{-1}(0.95) = 1.645$

$\chi^2_{20}$: $F^{-1}(0.025) = 9.591$, $F^{-1}(0.975) = 34.17$
$\chi^2_{19}$: $F^{-1}(0.025) = 8.907$, $F^{-1}(0.975) = 32.85$
$\chi^2_{18}$: $F^{-1}(0.025) = 8.231$, $F^{-1}(0.975) = 31.53$
$\chi^2_{13}$: $F^{-1}(0.025) = 5.009$, $F^{-1}(0.975) = 24.74$
$\chi^2_{12}$: $F^{-1}(0.025) = 4.404$, $F^{-1}(0.975) = 23.34$
$\chi^2_{10}$: $F^{-1}(0.025) = 3.247$, $F^{-1}(0.975) = 20.48$
$\chi^2_{9}$ : $F^{-1}(0.025) = 2.700$, $F^{-1}(0.975) = 19.02$

$t_{20}$: $F^{-1}(0.975) = 2.086$, $F^{-1}(0.95) = 1.725$
$t_{19}$: $F^{-1}(0.975) = 2.093$, $F^{-1}(0.95) = 1.729$
$t_{18}$: $F^{-1}(0.975) = 2.101$, $F^{-1}(0.95) = 1.734$
$t_{13}$: $F^{-1}(0.975) = 2.160$, $F^{-1}(0.95) = 1.771$
$t_{12}$: $F^{-1}(0.975) = 2.179$, $F^{-1}(0.95) = 1.782$
$t_{10}$: $F^{-1}(0.975) = 2.228$, $F^{-1}(0.95) = 1.812$
$t_{9}$ : $F^{-1}(0.975) = 2.262$, $F^{-1}(0.95) = 1.833$

$\mathrm{F}_{12,15}$: $F^{-1}(0.05) = 0.3821$, $F^{-1}(0.95) = 2.475$
$\mathrm{F}_{13,16}$: $F^{-1}(0.05) = 0.3976$, $F^{-1}(0.95) = 2.397$
$\mathrm{F}_{15,12}$: $F^{-1}(0.05) = 0.4040$, $F^{-1}(0.95) = 2.617$
$\mathrm{F}_{16,13}$: $F^{-1}(0.05) = 0.4171$, $F^{-1}(0.95) = 2.515$

# MAST20005/MAST90058: Week 6 Problems

1. To analyse the data from a brain study, we use the regression model: $Y_i = \alpha + \beta x_i + \varepsilon_i$, $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$, $i = 1, \ldots, n$. The response is the brain weight (on a log-scale) for $n = 62$ terrestrial mammals, while the predictor is the body weight (also on a log-scale). Consider the following (partial) R output:

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.13479
x             0.75169
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
Residual standard error: 0.6943 on 60 degrees of freedom
Multiple R-squared:  0.9208,    Adjusted R-squared:  0.9195
F-statistic: 697.4 on 1 and 60 DF,  p-value: < 2.2e-16
```

Recall that $\mathrm{var}(\hat{\beta}) = \sigma^2/K$, where $K = \sum_i (x_i - \bar{x})^2$. Find a 95% confidence interval for $\beta$. You may use the following information:

```
> sd(x)
[1] 3.123128
> qt(c(0.999, 0.99, 0.975, 0.95), df = 60)
[1] 3.231 2.39 2.00 1.67
```

2. Consider random variables $X_1, X_2, X_3$ having joint density $f(x_1, x_2, x_3)$. Suppose that

$$\mathbb{E}(X_3 \mid X_1 = x_1, X_2 = x_2) = \alpha + \beta_1(x_1 - \mu_1) + \beta_2(x_2 - \mu_2)$$

where $\mu_i = \mathbb{E}(X_i)$. Show that:

(a) $\alpha = \mu_3$,

(b) both of:

$$\beta_1 = \frac{\sigma_{13}\sigma_2^2 - \sigma_{12}\sigma_{23}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}, \quad \beta_2 = \frac{\sigma_{23}\sigma_1^2 - \sigma_{12}\sigma_{13}}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}$$

where $\sigma_i^2 = \mathrm{var}(X_i)$ and $\sigma_{ij} = \mathrm{cov}(X_i, X_j)$.

3. Consider the simple linear model $Y = \alpha_0 + \beta(x_i - \bar{x}) + \varepsilon$ where $\varepsilon \sim \mathrm{N}(0, \sigma^2)$.

(a) Show that

$$\sum_{i=1}^{n} [Y_i - \alpha_0 - \beta\,(x_i - \bar{x})]^2 = n(\hat{\alpha}_0 - \alpha_0)^2 + (\hat{\beta} - \beta)^2 \sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{i=1}^{n} [Y_i - \hat{\alpha}_0 - \hat{\beta}(x_i - \bar{x})]^2$$

(b) For an appropriate value of $c$ (which one?), show that the endpoints for a $100 \cdot (1-\gamma)\%$ confidence interval for $\alpha_0$ are:

$$\hat{\alpha}_0 \pm c\frac{\hat{\sigma}}{\sqrt{n}}.$$

(c) Letting $F^{-1}$ be the inverse cdf of $\chi^2_{n-2}$, show that a $100 \cdot (1-\gamma)\%$ confidence interval for $\sigma^2$ is:

$$\left( \frac{(n-2)\hat{\sigma}^2}{F^{-1}(1-\gamma/2)}, \frac{(n-2)\hat{\sigma}^2}{F^{-1}(\gamma/2)} \right).$$

4. Explain why the model $\mu(x) = \beta_1 e^{\beta_2 x}$ is not a linear model.

5. To fit the quadratic curve $y = \beta_1 + \beta_2 x + \beta_3 x^2$ to a set of points, we minimise

$$h(\beta_1, \beta_2, \beta_3) = \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i - \beta_3 x_i^2)^2.$$

By setting the three first partial derivatives of $h$ with respect to $\beta_1$, $\beta_2$ and $\beta_3$ to zero, show that $\beta_1$, $\beta_2$ and $\beta_3$ satisfy the normal equations:

$$\sum y_i = \beta_1 n + \beta_2 \sum x_i + \beta_3 \sum x_i^2$$
$$\sum x_i y_i = \beta_1 \sum x_i + \beta_2 \sum x_i^2 + \beta_3 \sum x_i^3$$
$$\sum x_i^2 y_i = \beta_1 \sum x_i^2 + \beta_2 \sum x_i^3 + \beta_3 \sum x_i^4$$

6. (a) Show that:

$$\sum_{i=1}^{n} (x_i - \bar{x}) y_i = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i (y_i - \bar{y})$$
$$= \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} = \sum_{i=1}^{n} x_i y_i - \frac{1}{n} \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i$$

(b) Prove the following identity for the sum of squared residuals:

$$d^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 - \frac{\left[\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})\right]^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

7. The following table gives the leaf area, $y$, of a particular type of tree at age $x$ years.

| $x$ | 8 | 11 | 17 | 20 | 23 | 26 |
|---|---|---|---|---|---|---|
| $y$ | 14.8 | 17.3 | 20.8 | 24.4 | 29.3 | 35.0 |
| | 9.0 | | 23.7 | 28.9 | | 33.4 |
| | 11.0 | | | 27.8 | | 37.8 |

Some useful statistics:

$$n = 13 \qquad \sum x_i y_i = 6282.3 \qquad \sum (x_i - \bar{x})(y_i - \bar{y}) = 741.1$$
$$\sum x_i = 230 \qquad \sum x_i^2 = 4648 \qquad \sum (x_i - \bar{x})^2 = 578.8$$
$$\sum y_i = 313.2 \qquad \sum y_i^2 = 8546.0 \qquad \sum (y_i - \bar{y})^2 = 1000.2$$

Using a simple linear regression model, calculate the following:

(a) Estimates of all of the parameters

(b) Standard errors for all of the regression coefficients

(c) A 95% confidence interval for the expectation of $Y$ when $x = 18$

(d) A 95% prediction interval for $Y$ when $x = 18$

# MAST20005/MAST90058: Week 7 Problems

*Some useful information for many of the problems is shown at end of this problem sheet.*

1. A ball is drawn from one of two bowls. Bowl A contains 100 red balls and 200 white balls; Bowl B contains 200 red balls and 100 white balls. Let $p$ denote the probability of drawing a red ball from the chosen bowl. Then $p$ is unknown as we don't know which bowl is being used. We shall test the simple null hypothesis, $H_0: p = 1/3$, against the simple alternative, $H_1: p = 2/3$. We draw three balls at random with replacement from the selected bowl. Let $X$ be the number of red balls drawn. Let the critical region be $x \in \{2, 3\}$. What are the probabilities $\alpha$ and $\beta$ respectively of Type I and Type II errors?

2. Let $Y \sim \text{Bi}(100, p)$. To test $H_0: p = 0.08$ against $H_1: p < 0.08$, we reject $H_0$ if $Y \leqslant 6$.
   (a) Determine the significance level $\alpha$ of the test.
   (b) Find the probability of a Type II error if, in fact, $p = 0.04$.

3. If a newborn baby has a birth weight that is less than 2500 grams we say the baby has a low birth weight. The proportion of babies with low birth weight is an indicator of nutrition for the mothers. In the USA, approximately 7% of babies have a low birth weight. Let $p$ be the proportion of babies born in the Sudan with low birth weight. Test the null hypothesis $H_0: p = 0.07$ against the alternative $H_1: p > 0.07$. If $y = 23$ babies out of a random sample of $n = 209$ babies had low birth weight, what is your conclusion at the following significance levels:
   (a) $\alpha = 0.05$?
   (b) $\alpha = 0.01$?
   (c) Find the p-value of this test. (Recall the p-value is the probability of the observed value or something more extreme when the null hypothesis is true).

4. Let $p_m$ and $p_f$ be the respective proportions of male and female white crowned sparrows that return to their hatching site. Give the endpoints for a 95% confidence interval for $p_m - p_f$, given that 124 out of 894 males and 70 out of 700 females returned. Does this agree with the conclusion of the test of $H_0: p_m = p_f$ against $H_1: p_m \neq p_f$ with $\alpha = 0.05$?

5. Vitamin $B_6$ is one of the vitamins in a multivitamin pill manufactured by a pharmaceutical company. The pills are produced with a mean of 50 milligrams of vitamin $B_6$ per pill. The company believes there is a deterioration of 1 milligram per month, so that after 3 months they expect that $\mu = 47$. A consumer group suspects that $\mu < 47$ after 3 months.
   (a) Define a critical region to test $H_0: \mu = 47$ against $H_1: \mu < 47$ with a significance level of $\alpha = 0.05$ based on a random sample of size $n = 20$.
   (b) If the 20 pills resulted in a mean of $\bar{x} = 46.94$ and a standard deviation of $s = 0.15$, what is your conclusion?
   (c) Give limits for the p-value of this test.

6. Let $X$ represent the height of professional female volleyball players. Assume that $X \sim \text{N}(\mu, \sigma^2)$ approximately. Suppose it is known that $\mu = 1.9$ metres worldwide. Do Australian female players differ from this? We explore this using a random sample of size $n = 9$.
   (a) Define the null hypothesis.

(b) Define the alternative hypothesis.

(c) Define a critical region for which $\alpha = 0.05$.

(d) Calculate the value of the test statistic if $\bar{x} = 2.05$ and $s = 0.17$.

(e) What is your conclusion?

(f) Give limits for the p-value of this test.

7. In May, the weights of 2-kilogram boxes of laundry detergent had a mean of 2.13 kilograms with a standard deviation of 0.095. The goal was to decrease the standard deviation. The company decided to adjust the filling machines and then test $H_0: \sigma = 0.095$ against $H_1: \sigma < 0.095$. In June, a random sample of size $n = 20$ gave $\bar{x} = 2.10$ and $s = 0.065$.

   (a) At an $\alpha = 0.05$ significance level, was the company successful?

   (b) What is an approximate p-value for this test?

8. The World Health Organisation collects air quality data around the world, which includes a measurement of suspended particles in $\mu$g/m$^3$. Let $X$ and $Y$ equal the concentration of suspended particles in the city centres of Melbourne and Sydney, respectively. Using $n = 13$ observations of $X$ and $m = 16$ observations of $Y$, we shall test $H_0: \mu_X = \mu_Y$ against $H_1: \mu_X \neq \mu_Y$ using a significance level of 5%.

   (a) Define the test statistic and the critical region assuming the variances are equal.

   (b) If $\bar{x} = 72.9$, $s_X = 25.6$, $\bar{y} = 81.7$ and $s_Y = 28.3$, calculate the value of the test statistic and state your conclusion.

   (c) Give limits for the p-value of this test.

   (d) Test whether the assumption of equal variances is valid. Let $\alpha = 0.05$.

Some potentially helpful R output:

```
> dbinom(0:3, 3, 1/3)
[1] 0.29629630 0.44444444 0.22222222 0.03703704
> dbinom(0:3, 3, 2/3)
[1] 0.03703704 0.22222222 0.44444444 0.29629630
> pnorm(c(-0.737, -0.553, 1.276, 1.531, 2.269))
[1] 0.2305612 0.2901317 0.8990222 0.9371153 0.9883658

> p1 <- c(0.75, 0.9, 0.95, 0.975, 0.99)
> qnorm(p1)
[1] 0.6744898 1.2815516 1.6448536 1.9599640 2.3263479
> qt(p1, 27)
[1] 0.683685 1.313703 1.703288 2.051831 2.472660
> qt(p1, 20)
[1] 0.6869545 1.3253407 1.7247182 2.0859634 2.5279770
> qt(p1, 19)
[1] 0.6876215 1.3277282 1.7291328 2.0930241 2.5394832
> qt(p1, 8)
[1] 0.7063866 1.3968153 1.8595480 2.3060041 2.8964594

> p2 <- c(0.025, 0.05, 0.95, 0.975)
> qchisq(p2, 19)
[1]  8.906516 10.117013 30.143527 32.852327
> qf(p2, 12, 15)
[1] 0.3147424 0.3821387 2.4753130 2.9632824
```

# MAST20005/MAST90058: Week 8 Problems

*Some useful information for many of the problems is shown at end of this problem sheet.*

1. **Random digits (group activity).** A volunteer in each group should write down a string of 51 random digits, such as:

   3 7 2 4 1 6 8 5...

   For each digit, record whether: (i) the next digit is the same as the preceding one (e.g. '2 2') and (ii) the next digit differs from the preceding one by 1 (e.g. '2 3'). For the purpose of this exercise, assume that the digits 0 and 9 only differ by 1. After the volunteer writes down 51 numbers, the entire group should carry out a hypothesis test at the 5% level of significance to determine whether the volunteer's sequence of numbers is truly random (you will need to think carefully about what this means and how to translated it into an appropriate null hypothesis).

2. Strawberries are being packed for the market. It is claimed that the median weight, $m$, of these boxes is 40 kilograms.

   (a) Use the following data and a Wilcoxon test statistic at an approximate significance level of $\alpha = 0.05$ to test the null hypothesis $H_0: m = 40$ against $H_1: m < 40$

   41.195, 39.485, 41.229, 36.840, 38.050, 40.890, 38.345, 34.930, 39.245, 31.031, 40.780, 38.050, 30.906

   It may help to complete the following table. Ties are assigned the average rank.

| $i$ | $x_i$ | $x_i - m$ | Rank | Sign |
|---|---|---|---|---|
| 1 | 41.195 | 1.195 | 5 | 1 |
| 2 | 39.485 | −0.515 | 1 | −1 |
| 3 | 41.229 | 1.229 | 6 | 1 |
| 4 | 36.840 | −3.160 | 10 | −1 |
| 5 | 38.050 | −1.950 | 8.5 | −1 |
| 6 | 40.890 | 0.89 | 4 | 1 |
| 7 | 38.345 | −1.655 | 7 | −1 |
| 8 | 34.930 | −5.070 | 11 | −1 |
| 9 | 39.245 | −0.755 | 2 | −1 |
| 10 | 31.031 | −8.969 | 12 | −1 |
| 11 | 40.780 | 0.780 | 3 | 1 |
| 12 | 38.050 | −1.950 | 8.5 | −1 |
| 13 | 30.906 | −9.094 | 13 | −1 |

   (b) Give limits for the p-value of this test.

   (c) Use the sign test to test the same hypothesis.

   (d) Compare the results of the two tests.

3. In a biology laboratory the mating of two red-eye fruit flies yielded $n = 432$ offspring among which 254 were red-eyed, 69 were brown-eyed, 87 were scarlet-eyed, and 22 were white-eyed. Use these data to test, with $\alpha = 0.05$, the hypothesis that the ratio among the offspring would be 9:3:3:1 respectively.

4. We wish to determine if two groups of nurses distribute their time in six different categories about the same way. That is, the hypothesis under consideration is $H_0: p_{i1} = p_{i2}$, $i = 1, \ldots, 6$. To test this, nurses are observed at random throughout several days, each observation resulting in a mark in one of the six categories. The summary data are given in the following frequency table:

| | Category | | | | | | |
|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | Total |
| Group I | 95 | 36 | 71 | 21 | 45 | 32 | 300 |
| Group II | 53 | 26 | 43 | 18 | 32 | 28 | 200 |

Do a chi-squared test with $\alpha = 0.05$.

Some potentially helpful R output:

```
> pbinom(1:6, 13, 0.5)
[1] 0.001708984 0.011230469 0.046142578 0.133422852 0.290527344 0.500000000

> p <- seq(0.9, 0.975, 0.025)
> qnorm(p)
[1] 1.281552 1.439531 1.644854 1.959964
> qchisq(p, 1)
[1] 2.705543 3.170053 3.841459 5.023886
> qchisq(p, 2)
[1] 4.605170 5.180534 5.991465 7.377759
> qchisq(p, 3)
[1] 6.251389 6.904644 7.814728 9.348404
> qchisq(p, 4)
[1]  7.779440  8.496282  9.487729 11.143287
> qchisq(p, 5)
[1]  9.236357 10.008315 11.070498 12.832502
```

# MAST20005/MAST90058: Week 9 Problems

*Some useful information for many of the problems is shown at end of this problem sheet.*

1. In a one-way ANOVA with $I$ treatments and $J$ observations per treatment, let $\mu = I^{-1}\sum \mu_i$.

   (a) Express $\mathbb{E}(\bar{X}_{\cdot\cdot})$ in terms of $\mu$. (Hint: $\bar{X}_{\cdot\cdot} = I^{-1}\sum \bar{X}_{i\cdot}$)

   (b) Compute $\mathbb{E}(\bar{X}_{i\cdot}^2)$

   (c) Compute $\mathbb{E}(\bar{X}_{\cdot\cdot}^2)$

   (d) Compute $\mathbb{E}(SS(T))$ and then show that

   $$\mathbb{E}(MS(T)) = \sigma^2 + \frac{J}{I-1}\sum(\mu_i - \mu)^2$$

   (e) Using the result of (d), what is $\mathbb{E}(MS(T))$ when $H_0$ is true? When $H_0$ is false, how does $\mathbb{E}(MS(T))$ compare with $\sigma^2$?

2. In an experiment to compare the tensile strengths of five different types of copper wire, four samples of each type were used. In an ANOVA, the between-groups and within-groups mean squares statistics were computed as $MS(T) = 2573.3$ and $MS(E) = 1394.2$ respectively. Use the F-test at a 5% significance level to test $H_0$: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$ against the alternative $H_1$: $H_0$, where $\mu_i$ is the mean tensile strength of copper wire of type $i$.

3. Consider the following partial output from a regression of the average brain and body weights for 62 species of mammals (variables are transformed on the log-scale).

```
lm(formula = brain ~ body)

             Estimate Std. Error
(Intercept)   2.13479    0.09604
Body          0.75169    0.02846
---
Residual standard error: 0.6943 on 60 degrees of freedom
Multiple R-squared:  0.9208,    Adjusted R-squared:  0.9195
F-statistic: 697.4 on 1 and 60 DF,  p-value: < 2.2e-16
```

   (a) Test the null hypothesis of no association between body and brain weights at the $\alpha = 0.01$ level of significance.

   (b) Use the following approximate distribution to obtain a test of size $\alpha$ for the null hypothesis $H_0$: $\rho = 0$ against $H_1$: $\rho \neq 0$ based on $R$, the sample correlation coefficient.

   $$\frac{1}{2}\ln\left(\frac{1+R}{1-R}\right) \approx \mathrm{N}\left(\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right), \frac{1}{n-3}\right)$$

   (c) What is the sample correlation coefficient for these data?

   (d) Apply the procedure in (b) to the mammals data using the significance level $\alpha = 0.01$.

   (e) Based on the above results, state your conclusion about the relationship between body and brain weight of mammals.

4. Let $X_1, \ldots, X_n$ be a random sample from $\mathrm{N}(\mu, \sigma^2)$, where $\mu$ and $\sigma^2$ are unknown. We wish to test $H_0\colon \sigma^2 = \sigma_0^2$ against $H_1\colon \sigma^2 \neq \sigma_0^2$.

   (a) Find $L_0$ and $L_1$, the maximised likelihoods under $H_0$ and $H_1$, that are required in order to write a likelihood ratio.

   (b) Show that the likelihood ratio test rejects $H_0$ if $w > c_1$ or $w < c_2$ (for some constants $c_1$ and $c_2$), where $w = \sum_i (x_i - \bar{x})^2/\sigma_0^2$.

Some potentially helpful R output:

```
> p <- c(0.95, 0.975, 0.99, 0.995)
> qnorm(p)
[1] 1.644854 1.959964 2.326348 2.575829
> qt(p, 60)
[1] 1.670649 2.000298 2.390119 2.660283
> qchisq(p, 60)
[1] 79.08194 83.29767 88.37942 91.95170
> qf(p, 4, 15)
[1] 3.055568 3.804271 4.893210 5.802907
> qf(p, 5, 20)
[1] 2.710890 3.289056 4.102685 4.761574
> qf(p, 15, 4)
[1]  5.857805  8.656541 14.198202 20.438268
> qf(p, 20, 5)
[1]  4.558131  6.328555  9.552646 12.903488
```

# MAST20005/MAST90058: Week 10 Problems

1. Let $X_{(1)} < \cdots < X_{(5)}$ be the order statistics of 5 independent observations from an exponential distribution that has a mean of $\theta = 3$.

   (a) Find the pdf of the sample median $X_{(3)}$.

   (b) Compute the probability that $X_{(4)} < 5$.

   (c) Determine $\Pr(1 < X_{(1)})$.

2. Let $X_1, \ldots, X_{10}$ be a random sample from a shifted exponential distribution with pdf $f(x \mid \theta) = e^{-(x-\theta)}$, $\theta \leqslant x < \infty$.

   (a) Show that $Y = \min(X_i) = X_{(1)}$ is the maximum likelihood estimator of $\theta$.

   (b) Find the pdf of $Y$.

   (c) Show that $\mathbb{E}(Y) = \theta + \frac{1}{10}$ and that $Y - \frac{1}{10}$ is an unbiased estimator of $\theta$.

   (d) Compute $\Pr(\theta < Y < \theta + c)$ and use it to construct a 95% confidence interval for $\theta$.

   (e) Where have you seen this example before?

3. Let $X_{(1)} < \cdots < X_{(n)}$ be the order statistics of $n$ independent observations from the uniform distribution Unif$(0, 1)$.

   (a) Find the pdf of $X_{(1)}$.

   (b) Verify that $\mathbb{E}(X_{(1)}) = \frac{1}{n+1}$.

4. Let $X$ have a Laplace distribution with pdf $f(x \mid \theta) = \frac{1}{2}e^{-|x-\theta|}$. (This is also known as a *double exponential distribution*, can you see why?) Suppose we have a random sample of $n$ observations on $X$.

   (a) Show that $\mathbb{E}(X) = \theta$ and $\text{var}(X) = 2$. (Hint: $\int_0^\infty z^2 e^{-z} dz = 2$)

   (b) Consider the estimator, $\hat{\theta}_1 = \bar{X}$. Find its mean and variance.

   (c) Consider the estimator, $\hat{\theta}_2 = \hat{M}$. Find its approximate mean and variance.

   (d) Which estimator is better?

   (e) What is the maximum likelihood estimator of $\theta$?

5. The following times (in minutes) between tram arrivals were observed at a particular tram stop:

   ```
   0.67, 2.46, 1.00, 8.89, 8.85, 28.45, 2.95,
   2.36, 0.37, 5.66, 6.26, 1.80,  1.88, 4.66
   ```

   Find an approximate 95% confidence interval for the median and state its exact confidence level. You may use the following information:

   ```
   > pbinom(0:6, size = 14, prob = 0.5)
   [1] 0.0001 0.0009 0.0065 0.0287 0.0898 0.2120 0.3953
   > pbinom(13:7, size = 14, prob = 0.5)
   [1] 0.9999 0.9991 0.9935 0.9713 0.9102 0.7880 0.6047
   ```

# MAST20005/MAST90058: Week 11 Problems

1. A food hygiene inspection of Melbourne restaurants is carried out routinely. The outcome of each inspection is either a pass or a fail. The inspectors are not always reliable at picking up all food hygiene issues. From past experience, it is known that the probability of a pass if food hygiene is satisfactory is 0.9, and if it is unsatisfactory is 0.3.

   (a) Fast food restaurants are known to have unsatisfactory standards about 20% of the time. If an inspection of such a restaurant results in a fail, what is the probability it has unsatisfactory food hygiene standards?

   (b) Fine dining restaurants are known to have unsatisfactory standards only about 2% of the time. If an inspection of such a restaurant results in a fail, what is the probability it has unsatisfactory food hygiene standards?

   (c) Draw tree diagrams to represent this problem.

2. You are running a clinical trial of a new treatment for depression. Of the first 40 patients in your trial, 4 have resulted in a successful outcome.

   (a) Using a uniform prior distribution for the probability of a successful outcome, what is the posterior mean?

   (b) Your colleagues overseas, who are running similar trials, have observed successful outcomes in about 4% of their patients. You know their results are relevant and their trials are larger than yours, but there are some differences that make them only partly comparable. You decide to use their results as a prior to get a sense of the likely outcome of your own trial. You deem their information to be equivalent to about 60 pseudo-observations.

      i. Construct a prior that encapsulates this information.

      ii. Using this prior, what is your new posterior distribution?

      iii. What is your new posterior mean?

3. Let $X \sim \mathrm{N}(\theta, \sigma^2)$ where $\sigma$ is a known value. Use a normal distribution as a prior for $\theta$. Given a single observation on $X$, derive the posterior distribution. Show full working.

4. A standard mathematics test is given to all students about to start high school in Australia. The test scores are known to be normally distributed with a variance of 25. You need to review a random selection of 16 of the test papers. The mean score for this subset of students was 70. Using an appropriate uninformative prior distribution, give a 95% credible interval for the overall mean score for all students.

5. Let $Y$ be the sum of $n$ observations from a Poisson distribution with mean $\theta$. Let the prior for $\theta$ be a gamma distribution with parameters $\alpha$ and $\beta$, parameterised as follows:

$$f(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\theta\beta}, \quad 0 \leqslant \theta < \infty \quad (\text{Note: } \mathbb{E}(\theta) = \alpha/\beta,\ \mathrm{var}(\theta) = \alpha/\beta^2)$$

   (a) Find the posterior pdf of $\theta$ given $Y = y$. (Hint: You only need to consider the terms that involve $\theta$. What distribution do those terms correspond to?)

   (b) What is the posterior mean?

   (c) Show that the posterior mean is a weighted average of the maximum likelihood estimate and the prior mean.

6. Consider a random sample $X_1, \ldots, X_n$ from a distribution with pdf

$$f(x \mid \theta) = 3\theta x^2 e^{-\theta x^3}, \quad 0 < x < \infty.$$

Let the prior for $\theta$ be a gamma distribution with $\alpha = 4$ and $\beta = 4$. Find the posterior mean of $\theta$.

7. Suppose that the time in minutes required to serve a customer at a certain shop has an exponential distribution with mean $1/\theta$. As a prior for $\theta$ use a gamma distribution with mean 0.2 and standard deviation 0.1. If the average time to serve 20 customers is observed to be 3.8 minutes, what is the posterior distribution of $\theta$?

8. Take a random sample $X_1, \ldots, X_n \sim \text{Unif}(0, \theta)$. Use an improper uniform prior for $\theta$.

   (a) Derive the posterior distribution of $\theta$.

   (b) Show that the posterior mode is $x_{(n)}$.

   (c) Derive a 95% credible interval of the form $(x_{(n)}, c \cdot x_{(n)})$.

   (d) Derive a 95% confidence interval of the same form.
   (Hint: you will get a different value of $c$.)

   (e) What prior distribution would result in these two interval estimates being identical (i.e. the same value of $c$)?

8. (a). improper uniform prior $f(\theta) = 1$.

$f(x) = \frac{1}{\theta}$

$f(x|\theta) = \prod_{i=1}^{n} \frac{1}{\theta} = \theta^{-n}$ (whether ...)

$f(\theta|x) = \frac{f(x|\theta) f(\theta)}{f(x)}$

$\propto \theta^{-n} \cdot 1, \quad x_{(n)} < \theta$

normalizing constant

$F(\theta|x) = \int_{x_{(n)}}^{\infty} \theta^{-n} d\theta = \left[\frac{1}{1-n} \theta^{-n+1}\right]_{x_{(n)}}^{\infty} = 0 - \frac{1}{1-n} x_{(n)}^{-n+1} = \frac{1}{(n-1) x_{(n)}^{n-1}}$

$\therefore f(\theta|x) = \frac{\theta^{-n}}{\frac{1}{(n-1) \cdot x_{(n)}^{n-1}}} = (n-1) \cdot x_{(n)}^{n-1} \cdot \theta^{-n}, \quad x_{(n)} < \theta$

(b) decreasing, $x_{(n)} < \theta \Rightarrow \theta = x_{(n)}$

(c) $F(x) = \int_{x_{(n)}}^{\infty} (n-1) \cdot x_{(n)}^{n-1} \cdot \theta^{-n} d\theta = \left[\frac{(n-1) x_{(n)}^{n-1}}{1-n} \cdot \theta^{1-n}\right]_{x_{(n)}}^{\infty} = -x_{(n)}^{n-1} \cdot \left(\frac{1}{\theta}\right)^{n-1}$

$0 - \left[-x_{(n)}^{n-1} \cdot x_{(n)}^{1-n}\right]$

$= 1.$

# MAST20005/MAST90058: Week 12 Problems

1. Let $X_1, \ldots, X_n$ be a random sample from $\text{Bi}(1, p)$.
   (a) Find the Cramér–Rao lower bound for unbiased estimators of $p$.
   (b) We know that $\bar{X}$ is an unbiased estimator of $p$.
   Show that $\bar{X}$ attains the Cramér–Rao lower bound.

2. Let $X_1, \ldots, X_n$ be a random sample from $\text{N}(\mu, \theta)$ where $\mu$ is known.
   (a) Show that the maximum likelihood estimator of $\theta$ is,
   $$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2.$$
   (b) Find the Cramér–Rao lower bound for unbiased estimators of $\theta$.
   (c) What is the approximate distribution of $\hat{\theta}$?
   (d) What is the exact distribution of $n\hat{\theta}/\theta$?

3. Let $X_1, \ldots, X_n$ be a random sample from the density:
   $$f(x \mid \theta) = \frac{x}{\theta^2}e^{-x/\theta}, \quad 0 < x < \infty, \quad 0 < \theta < \infty.$$
   (a) Find a sufficient statistic for $\theta$.
   (b) Write down the log-likelihood function and the score function.
   (c) Determine the maximum likelihood estimator of $\theta$.
   (d) Find the Cramér–Rao lower bound for unbiased estimators of $\theta$.
   (Hint: some information from previous week's tutorial will help you to find $\mathbb{E}(X)$.)
   (e) A random sample of size $n = 35$ gave $\bar{x} = 10.5$. Determine the maximum likelihood estimate of $\theta$ and an approximate 95% confidence interval for $\theta$.

4. Find a sufficient statistic for $p$ when you toss a coin 10 times and $p$ is the probability of a head. Also do this for the case where $p$ is the probability of a head for the first 5 tosses and changes to $(1 - p)$ for the last five tosses.

5. Find sufficient statistics for $\theta$ (where $\theta > 0$) when we observe data from:
   (a) $X \sim \text{Unif}(0, \theta)$
   (b) $X \sim \text{Unif}(-\frac{\theta}{2}, \frac{\theta}{2})$

6. Find sufficient statistics for $\theta$ (where $\theta > 0$) when we observe $X$ from the following pdfs:
   (a) $f(x \mid \theta) = \frac{1}{\theta}e^{-x/\theta}, \quad 0 < x < \infty$
   (b) $f(x \mid \theta) = e^{-(x-\theta)}, \quad \theta < x < \infty$
   (c) $f(x \mid \theta) = \frac{1}{\theta}e^{-(x-\theta)/\theta}, \quad \theta < x < \infty$

7. Refer back to problem 2 from week 3.
   (a) What is a sufficient statistic for $\theta$?
   (b) What does that suggest about the relative merits of the two estimators we derived?