# Part A Statistics HT 2021

## Problem Sheet 1

**1.** (a) Suppose $X_1, \ldots, X_n$ are independent Bernoulli($p$) random variables. Use the delta method to find the asymptotic distribution of $\widehat{p}/(1-\widehat{p})$ where $\widehat{p}$ is the maximum likelihood estimator of $p$. (The quantity $p/(1-p)$ is the *odds* of a success.)

(b) Suppose $X_1, \ldots, X_n$ are independent Poisson($\lambda$) random variables. Find a function $g(\overline{X})$ such that the asymptotic variance of $g(\overline{X})$ does not depend on $\lambda$.

**2.** Let $X_1, \ldots, X_n$ be a random sample from a uniform distribution with probability density function

$$f(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Show that if $X_{(r)}$ is the $r^{th}$ order statistic, then

$$E(X_{(r)}) = \frac{r}{n+1}, \quad \text{var}(X_{(r)}) = \frac{r}{(n+1)(n+2)}\left(1 - \frac{r}{n+1}\right).$$

Define the median of the random sample, distinguishing between the two cases $n$ odd and $n$ even. Show that the median has expected value $\frac{1}{2}$ if the random sample is drawn from a uniform distribution on $(0,1)$. Find its variance in the case when $n$ is odd. What is the expected value of the median if the random sample is drawn from a uniform distribution on $(a,b)$?

[*Hint: remember that pdfs integrate to 1, there's no need to actually do any integration in this question.*]

**3.** Let $X$ be a continuous random variable with cumulative distribution function $F$ which is strictly increasing. If $Y = F(X)$, show that $Y$ is uniformly distributed on the interval $(0,1)$.

The *Weibull distribution* with parameters $\alpha > 0$ and $\lambda > 0$ has cumulative distribution function

$$F(x) = \begin{cases} 0 & \text{if } x < 0 \\ 1 - \exp(-(x/\lambda)^\alpha) & \text{if } x \geqslant 0. \end{cases}$$

It is typically used in industrial reliability studies in situations where failure of a system comprising many similar components occurs when the weakest component fails; it is also used in modelling survival times.

Explain why a probability plot for the Weibull distribution may be based on plotting the logarithm of the $r$th order statistic against $\log[-\log(1-\frac{r}{n+1})]$ and give the slope and intercept of such a plot.

**4.** Find the expected information for $\theta$, where $0 < \theta < 1$, based on a random sample $X_1, \ldots, X_n$ from:

(a) the geometric distribution $f(x;\theta) = \theta(1-\theta)^{x-1}$, $x = 1, 2, \ldots$

(b) the Bernoulli distribution $f(x;\theta) = \theta^x(1-\theta)^{1-x}$, $x = 0, 1$.

A statistician has a choice between observing random samples from the geometric or Bernoulli distributions with the same $\theta$. Which will give the more precise inference about $\theta$?

**5.** Suppose a random sample $Y_1, \ldots, Y_n$ from an exponential distribution with parameter $\lambda$ is rounded down to the nearest $\delta$, giving $Z_1, \ldots, Z_n$ where $Z_j = \delta \left\lfloor \frac{Y_j}{\delta} \right\rfloor$. Show that the likelihood contribution from the $j$th rounded observation can be written $(1 - e^{-\lambda\delta})e^{-\lambda z_j}$, and deduce that the expected information for $\lambda$ based on the entire sample is

$$\frac{n\delta^2 e^{-\lambda\delta}}{(1 - e^{-\lambda\delta})^2}.$$

Show that this has limit $n/\lambda^2$ as $\delta \to 0$, and that if $\lambda = 1$, the loss of information when data are rounded down to the nearest integer rather than recorded exactly, is less than 10%. Find the loss of information when $\delta = 0.1$, and comment briefly.

**6.** When $T_1$ and $T_2$ are estimators of a parameter $\theta$, the *asymptotic efficiency* of $T_1$ relative to $T_2$ is given by $\lim_{n\to\infty} \operatorname{avar}(T_2)/\operatorname{avar}(T_1)$, where $\operatorname{avar}(T_j)$ denotes the asymptotic variance of the approximating normal distribution of $T_j$, $j = 1, 2$.

Suppose $X_1, \ldots, X_n$ are independent and exponential with parameter $\theta$. Let $\#A$ denote the number of elements of a set $A$, and consider the two estimators

$$\widetilde{p} = \frac{\#\{i : X_i \geqslant 1\}}{n} \quad \text{and} \quad \widehat{p} = \overline{X}.$$

Find the asymptotic efficiency of $T_1 = -\log \widetilde{p}$ relative to $T_2 = 1/\widehat{p}$. Find the numerical value of the asymptotic efficiency when $\theta = 0.6, 1.6, 5.6$. Comment on the implications for using $T_1$ instead of $T_2$ to estimate $\theta$.

**7.** The figure below shows normal Q-Q plots for randomly generated samples of size 100 from four different densities: from a $N(0, 1)$ density, an exponential density, a uniform density, and a Cauchy density. (The Cauchy density is $f(x) = [\pi(1 + x^2)]^{-1}$ for $x \in \mathbb{R}$.)

Which Q-Q plot goes with which density?

Using R, you can try plots like these for yourself using commands like the following.

```
x1 <- rnorm(100)
qqnorm(x1)

x2 <- rexp(100)
qqnorm(x2)

x3 <- runif(100)
qqnorm(x3)

x4 <- rt(100, df = 1)
qqnorm(x4)
```

The symbol `<-` is the assignment operator in R, so `x1 <- rnorm(100)` sets `x1` equal to a randomly generated sample of size 100 from a $N(0, 1)$ density. (For `x4`, note that the Cauchy distribution is the same as the $t$-distribution with one degree of freedom.)

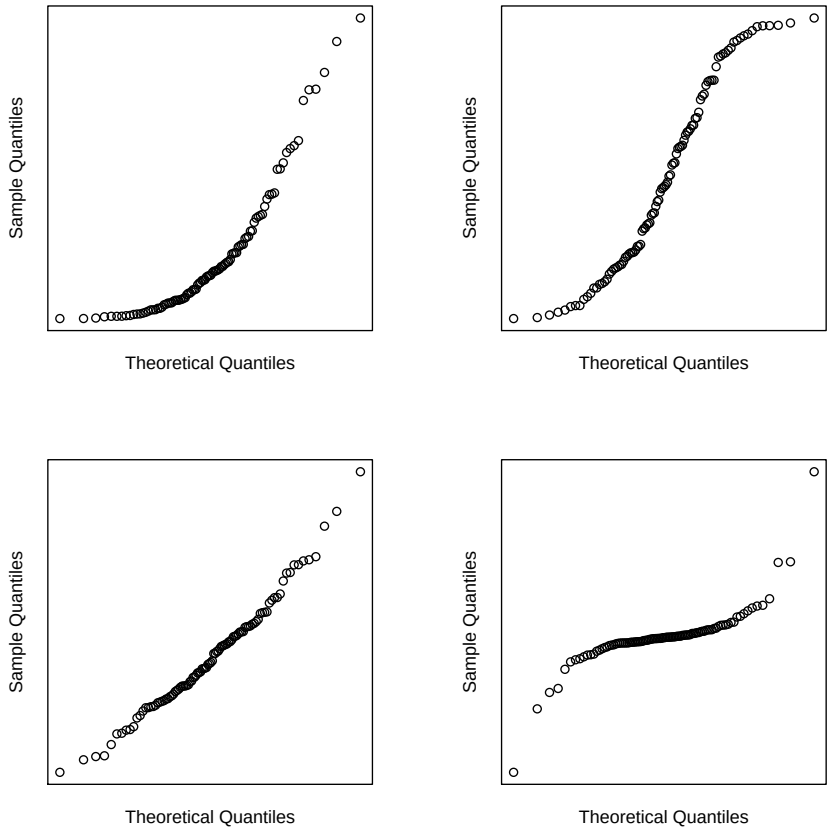


**Figure 1.** Normal Q-Q plots for four different samples, one from each of the following densities: $N(0,1)$, exponential, uniform, Cauchy. Which is which?

**8.** Read through the short document "Getting started with R" (available on the course webpage) and hopefully install R/RStudio and run the commands in that document yourself.

The R questions on these problem sheets need only a very small knowledge of R. R code will be supplied in questions. Each sheet will also be accompanied by a file containing further R code and more help, see the file `sheet1.R` for this sheet. There will be code you can cut and paste into R, and this code will also appear at the end of each sheet.

We can gain understanding from using R, for example:

(i) How typical is each of the Q-Q plots shown in Figure 1? Note that each time we generate a sample (e.g. using `rnorm`, `rexp`, ...) we get a different sample, so we can investigate how typical each one is by doing repeated Q-Q plots.

(ii) How much does Figure 1 change if the sample size is smaller (or larger) than 100?

To investigate (i) and (ii), run the R code in the previous question multiple times, and with different sample sizes. See `sheet1.R` for more.

*You can download* `sheet1.R` *(use "download" on RHS of course webpage), then view it in RStudio or R. If you try to view it directly in a browser you may get an error.*

*The contents of* `sheet1.R` *are also pasted in below.*

**9.** (See `sheet1.R` for more.) To generate a sample of size 100 from a $N(0,1)$ density and compare the sample with an exponential distribution, try the following:

```
n <- 100
x <- rnorm(n)
k <- 1:n
plot(-log(1 - k/(n+1)), sort(x), main = "Exponential Q-Q Plot",
     ylab = "Ordered data", xlab = "-log[1 - k/(n+1)]")
```

Can you explain the shape of this exponential Q-Q plot? What happens (and why) if you repeat but with the line `x <- rnorm(n)` replaced by `x <- rexp(n)`?

Try repeating using the data on insurance claim interarrival times:

```
x <- scan("http://www.stats.ox.ac.uk/~laws/partA-stats/data/interarrivals.txt")
n <- length(x)
k <- 1:n
```

followed by the plot command above. Try also using the data on insurance claim amounts:

```
x <- scan("http://www.stats.ox.ac.uk/~laws/partA-stats/data/amounts.txt")
n <- length(x)
k <- 1:n
```

Can you also do a Pareto Q-Q plot for each dataset? What do you conclude?

```
#############
## Sheet 1 ##
#############

#### question 7
x1 <- rnorm(100)
qqnorm(x1)

x2 <- rexp(100)
qqnorm(x2)

x3 <- runif(100)
qqnorm(x3)

x4 <- rt(100, df = 1)
qqnorm(x4)

#### question 8
# to see all four plots at once,
# i.e. to arrange the plots in a 2 x 2 array,
# use par(mfrow = c(2, 2)) and then the qqnorm commands
par(mfrow = c(2, 2))
# from now on plots will be in a 2 x 2 array

x1 <- rnorm(100)
qqnorm(x1, main = "Normal Q-Q plot: normal data")
x2 <- rexp(100)
qqnorm(x2, main = "Normal Q-Q plot: exponential data")
x3 <- runif(100)
qqnorm(x3, main = "Normal Q-Q plot: uniform data")
x4 <- rt(100, df = 1)
qqnorm(x4, main = "Normal Q-Q plot: Cauchy data")

# to get back to a 1 x 1 array of plots you would use
# par(mfrow = c(1, 1))

# try multiple plots to see how much variation there is
# from one sample to another
# normal data, n = 100, try running this a few times
for (i in 1:4) {
  x <- rnorm(100)
  qqnorm(x)
}

# and repeat but with x <- rexp(100)
# and with x <- runif(100)
# and with x <- rt(100, df = 1)

# next, vary the sample size
# normal data, n = 10
```

```
for (i in 1:4) {
  x <- rnorm(10)
  qqnorm(x)
}

# useful to also try n = 20, 50
# useful to also try exponential data (using rexp),
# and uniform data (using runif),
# and Cauchy, or t, data (using rt)

# e.g. uniform distribution, n = 20
for (i in 1:4) {
  x <- runif(20)
  qqnorm(x)
}

# can also look at t-distributions with different numbers
# of degrees of freedom
# e.g. t-distribution with 5 dgrees of freedom, n = 10
for (i in 1:4) {
  x <- rt(10, df = 5)
  qqnorm(x)
}

#### question 9
n <- 100
x <- rnorm(n)
k <- 1:n
plot(-log(1 - k/(n+1)), sort(x), main = "Exponential Q-Q Plot",
     ylab = "Ordered data", xlab = "-log[1 - k/(n+1)]")

# now try replacing x <- rnorm(n) by x <- rexp(n)
x <- rexp(n)
plot(-log(1 - k/(n+1)), sort(x), main = "Exponential Q-Q Plot",
     ylab = "Ordered data", xlab = "-log[1 - k/(n+1)]")

# are interarrival times exponential?
# exponential Q-Q plot with data on insurance claim interarrival times
x <- scan("http://www.stats.ox.ac.uk/~laws/partA-stats/data/interarrivals.txt")
n <- length(x)
k <- 1:n
plot(-log(1 - k/(n+1)), sort(x), main = "Exponential Q-Q Plot",
     ylab = "Ordered data", xlab = "-log[1 - k/(n+1)]")

# are claim amounts exponential?
# exponential Q-Q plot with data on insurance claim amounts
x <- scan("http://www.stats.ox.ac.uk/~laws/partA-stats/data/amounts.txt")
n <- length(x)
k <- 1:n
plot(-log(1 - k/(n+1)), sort(x), main = "Exponential Q-Q Plot",
```

```
     ylab = "Ordered data", xlab = "-log[1 - k/(n+1)]")

# are interarrival times Pareto?
# Pareto Q-Q plot for interarrival times
x <- scan("http://www.stats.ox.ac.uk/~laws/partA-stats/data/interarrivals.txt")
n <- length(x)
k <- 1:n
plot(-log(1 - k/(n+1)), sort(log(x)),
     main = "Pareto Q-Q Plot: interarrivals",
     ylab = "log(Ordered data)", xlab = "-log[1 - k/(n+1)]")

# are claim amounts Pareto?
# Pareto Q-Q plot for claim amounts
x <- scan("http://www.stats.ox.ac.uk/~laws/partA-stats/data/amounts.txt")
n <- length(x)
k <- 1:n
plot(-log(1 - k/(n+1)), sort(log(x)),
     main = "Pareto Q-Q Plot: amounts",
     ylab = "log(Ordered data)", xlab = "-log[1 - k/(n+1)]")
```

# Part A Statistics HT 2021

## Problem Sheet 2

**1.** What is the connection between Fisher's information and the asymptotic distribution of the maximum likelihood estimator?

Assume the individuals in a sample of size $n = 1029$ are independent and that each individual has blood type $M$ with probability $(1-\theta)^2$, type $MN$ with probability $2\theta(1-\theta)$, and type $N$ with probability $\theta^2$. For the following data (Rice, 2007) find the maximum likelihood estimate $\widehat{\theta}$ and use the asymptotic distribution of the MLE to find an approximate 95% confidence interval for $\theta$.

| Blood Type | $M$ | $MN$ | $N$ |
|---|---|---|---|
| Frequency | 342 | 500 | 187 |

**2.** Let $X_1, \ldots, X_n$ be independent $N(\mu, \sigma^2)$ random variables. Suppose that $\mu$ is known, $\sigma$ is unknown and that we want to estimate $\psi = \log \sigma$.

(a) Find the maximum likelihood estimator $\widehat{\sigma}$ and the asymptotic normal approximation to the distribution of $\widehat{\sigma}$.

(b) Use the delta method to find the asymptotic distribution of $\widehat{\psi}$ and hence find an approximate 95% confidence interval for $\psi$.

(c) Explain how the interval in (b) can be used to find an approximate confidence interval for $\sigma$.

**3.** The following data are time intervals in days between earthquakes which either registered magnitudes greater than 7.5 on the Richter scale or produced over 1,000 fatalities. Recording starts on 16 December, 1902 and ends on 4 March, 1977, a total period of 27,107 days. There were 63 earthquakes in all, and therefore 62 recorded time intervals.

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 840 | 1901 | 40 | 139 | 246 | 157 | 695 | 1336 | 780 | 1617 |
| 145 | 294 | 335 | 203 | 638 | 44 | 562 | 1354 | 436 | 937 |
| 33 | 721 | 454 | 30 | 735 | 121 | 76 | 36 | 384 | 38 |
| 150 | 710 | 667 | 129 | 365 | 280 | 46 | 40 | 9 | 92 |
| 434 | 402 | 209 | 82 | 736 | 194 | 99 | 599 | 220 | 584 |
| 759 | 556 | 304 | 83 | 887 | 319 | 375 | 832 | 263 | 460 |
| 567 | 328 | | | | | | | | |

Assuming the data to be a random sample $X_1, \ldots, X_n$ from an exponential distribution with parameter $\lambda$, obtain the maximum likelihood estimator $\widehat{\lambda}$ of $\lambda$ and calculate the maximum likelihood estimate.

Given that the moment generating function of a gamma distribution with parameters $(n, \lambda)$ is

$$M_n(t) = \left(\frac{\lambda}{\lambda - t}\right)^n$$

show that $Y = \sum_{i=1}^{n} X_i$ has a gamma distribution. Show that

$$\left(\frac{a}{n\overline{x}}, \frac{b}{n\overline{x}}\right)$$

is an exact 95% central confidence interval for $\lambda$ if

$$\int_0^a \frac{y^{n-1}e^{-y}}{\Gamma(n)}\,dy = \int_b^\infty \frac{y^{n-1}e^{-y}}{\Gamma(n)}\,dy = 0.025.$$

Obtain Fisher's information for $\lambda$ and use it to find an approximate 95% confidence interval for $\lambda$. The interval given by the exact method above is $(0.0018, 0.0029)$. Verify numerically that your approximate interval is close to this.

**4.** Let $X_1, \ldots, X_n$ be a random sample from a normal distribution with known mean $\mu$ and unknown variance $\sigma^2$. Three possible confidence intervals for $\sigma^2$ are

(a) $\left(\sum_{i=1}^n \frac{(X_i - \overline{X})^2}{a_1}, \sum_{i=1}^n \frac{(X_i - \overline{X})^2}{a_2}\right)$

(b) $\left(\sum_{i=1}^n \frac{(X_i - \mu)^2}{b_1}, \sum_{i=1}^n \frac{(X_i - \mu)^2}{b_2}\right)$

(c) $\left(\frac{n(\overline{X} - \mu)^2}{c_1}, \frac{n(\overline{X} - \mu)^2}{c_2}\right)$

where $a_1$, $a_2$, $b_1$, $b_2$, $c_1$, $c_2$ are constants.

Find values of these six constants which give confidence level 0.90 for each of the three intervals when $n = 10$ and compare the expected widths of the three intervals in this case.

With $\sigma^2 = 1$, what value of $n$ is required to achieve a 90% confidence interval of expected width less than 2 in cases (b) and (c) above?

[For a $\chi^2$ with e.g. 6 degrees of freedom, you can use `qchisq(0.05, 6)` to find the 0.05 quantile.]

**5.** Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be independent random samples from normal distributions $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, respectively, where the parameters $\mu_1, \mu_2, \sigma^2$ are unknown. Let

$$S^2 = (m + n - 2)^{-1}\left(\sum_{i=1}^m (X_i - \overline{X})^2 + \sum_{j=1}^n (Y_j - \overline{Y})^2\right).$$

Determine the distributions of both

$$(m + n - 2)S^2/\sigma^2 \quad \text{and} \quad \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{\sqrt{S^2(\frac{1}{m} + \frac{1}{n})}}.$$

Show how to construct a confidence interval for $\mu_1 - \mu_2$.

**6.** Ten students were asked to guess the width of a lecture room. Their guesses (in metres) were: 10, 11, 12, 13, 15, 16, 17, 18, 19, 25. The actual width of the room was 13.1 m.

(i) Assuming the data arise from a normal distribution, how would you test whether this distribution has the correct mean? State the appropriate null and alternative hypotheses, and any assumptions you need to make for the hypothesis test to be appropriate.

(ii) Carry out the test you suggested in (i) and state your conclusions.

(iii) Modify your test to test whether the data are from a distribution with a mean value *higher* than the true value and re-state your conclusions.

**7.** (Code in `sheet2.R`.) Read in the earthquake data from question 3 and try an exponential Q-Q plot:

```
x <- scan("http://www.stats.ox.ac.uk/~laws/partA-stats/data/quakes.txt")
n <- length(x)
k <- 1:n
plot(-log(1 - k/(n+1)), sort(x), main = "Exponential Q-Q Plot",
     ylab = "Ordered data", xlab = "-log[1 - k/(n+1)]")
```

Is an exponential model a reasonable assumption for this dataset?

The 2.5% and 97.5% quantiles for a gamma distribution with parameters $(n, 1)$ can be calculated as follows.

```
a <- qgamma(0.025, n)
b <- qgamma(0.975, n)
```

That is, the function `qgamma(p, n)` calculates the $p$th quantile of a gamma distribution with shape parameter $n$ and rate parameter 1.

Now calculate the exact confidence interval of question 3:

```
c(a, b) / sum(x)
```

Also use R to check that the approximate 95% confidence interval for $\lambda$ obtained using Fisher's information is as given in question 3 – you might have obtained one of two possible approx intervals, see `sheet2.R` for both.

**8.** (Code in `sheet2.R`.) For data that are exponential with parameter $\lambda$, there are three possible confidence intervals for $\lambda$ in question 3 – one based on the gamma distribution plus two approximation possibilities. These three are all different, but numerically they are almost the same in question 3, where $n = 62$.

Use the code in `sheet2.R` to investigate how the three differ when $n$ is small, e.g. $n = 10$.

What do you conclude?

**9.** (See `sheet2.R` for more.) To do question 6 you will need the sample mean $\overline{x}$ and sample standard deviation $s$:

```
x <- c(10, 11, 12, 13, 15, 16, 17, 18, 19, 25)
mean(x)
sd(x)
```

The functions `qt` and `pt` allow you to determine the significance (or otherwise) of your test statistic(s) in question 6. Use `qt` and/or `pt` to find the quantiles and/or probabilities that you need in question 6.

The $p$th quantile of a $t_r$-distribution can be calculated using `qt(p, r)`, so e.g. the 97.5% quantile of a $t_4$-distribution can be found using

```
qt(0.975, 4)
```

Alternatively, the cdf of a $t_r$-distribution at $y$ can be calculated using `pt(y, r)`, so e.g. the probability that a $t_4$ random variable is less than 1.96 is given by

```
pt(1.96, 4)
```

```
#############
## Sheet 2 ##
#############

#### question 7
x <- scan("http://www.stats.ox.ac.uk/~laws/partA-stats/data/quakes.txt")
n <- length(x)
k <- 1:n
plot(-log(1 - k/(n+1)), sort(x), main = "Exponential Q-Q Plot",
     ylab = "Ordered data", xlab = "-log[1 - k/(n+1)]")
abline(0, mean(x))
# abline above plots a line with intercept = 0 and gradient = mean(x)
# - from lecture notes the exponential Q-Q plot should have intercept 0
# and gradient mu if the data are exponential with mean mu
# use ?abline to see the help page for abline

a <- qgamma(0.025, n)
b <- qgamma(0.975, n)

# interval using the gamma distribution
c(a, b) / sum(x)

xbar <- mean(x)

# approx interval using lambda.hat +/- 1.96*I(lambda.hat)^{-1/2}
c(1 - 1.96/sqrt(n), 1 + 1.96/sqrt(n)) / xbar

# second approx interval from substituting I(lambda) = n/lambda^2
# and then solving the inequalities
# i.e. not replacing lambda by lambda.hat in order to estimate a variance
c(1/(1 + 1.96/sqrt(n)), 1/(1 - 1.96/sqrt(n))) / xbar

#### question 8
# above we have three slightly different intervals
# interval1 from gamma, interval2 from first approx method
# and interval3 from second approx

# n = 62 for the above data and the large sample properties are evident,
# the three intervals are almost the same

# now investigate how the three intervals perform in small samples,
# e.g. n = 10, using data generated from an exponential, parameter 1

# generate the sample, calculate and plot the three intervals
# repeat m times, e.g. m = 33 giving 99 intervals in total

# cut-and-paste the following chunk into R, you don't need to work out
# the details of what all the plotting commands are doing

# ---begin chunk---
```

```
n <- 10
a <- qgamma(0.025, n)
b <- qgamma(0.975, n)
m <- 33

plot(1, 1, type = "n", yaxt = "n", xlim = c(0, 5), ylim = c(0, 4*m),
     xlab = "lambda", ylab = "",
     main = paste("95% CIs: samples of size", n, "from exponential, parameter 1"))
abline(v = 1)
legend("topright", c("interval1", "interval2", "interval3"),
       lty = 1, lwd = 2, col = c(1, "orange2", "steelblue2"))

for (i in 1:m) {
  x <- rexp(n)
  ci1 <- c(a, b) / sum(x)
  ci2 <- c(1 - 1.96/sqrt(n), 1 + 1.96/sqrt(n)) / mean(x)
  ci3 <- c(1/(1 + 1.96/sqrt(n)), 1/(1 - 1.96/sqrt(n))) / mean(x)
  lines(ci1, rep(4*i-1, 2), lwd = 2)
  lines(ci2, rep(4*i-2, 2), lwd = 2, col = "orange2")
  lines(ci3, rep(4*i-3, 2), lwd = 2, col = "steelblue2")
}
# ---end chunk---

# x <- rexp(n) generates a sample of size n
# use ?rexp to see the help page for rexp - when no rate parameter is
# given, rate = 1 is the default, hence vertical line on the plot at
# the true value lambda = 1

# the three intervals behave differently in small samples
# try repeating with larger n, e.g. n = 20, 50
# - you only need to change the first line n <- 10 to a different value
# at n = 50 the three intervals are close, especially intervals 1 & 2
# (and n = 62 for the data in question 3)

#### question 9
x <- c(10, 11, 12, 13, 15, 16, 17, 18, 19, 25)

# test statistic
tobs <- sqrt(10)*(mean(x) - 13.1)/sd(x)

# two-sided p-value
2*(1 - pt(tobs, df = 9))

# one-sided p-value
1 - pt(tobs, df = 9)

# can check using t.test
# see ?t.test, by default it assumes two-sided, and also uses a method for
# unequal variances hence we want var.equal = TRUE
t.test(x, mu = 13.1, var.equal = TRUE)
```

```
# one-sided
t.test(x, mu = 13.1, alternative = "greater", var.equal = TRUE)

# or could compare tobs to the quantiles
qt(0.975, df = 9)
qt(0.95, df = 9)
```

# Part A Statistics HT 2021

## Problem Sheet 3

**1.** The heart rate (beats per minute) of 10 children was measured in two situations: (i) at rest, and (ii) in anticipation of them doing a minute's exercise. The data are given below.

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Rest, $x$ | 72 | 116 | 79 | 97 | 90 | 67 | 115 | 82 | 95 | 82 |
| Anticipation, $y$ | 76 | 120 | 84 | 99 | 93 | 75 | 116 | 83 | 98 | 87 |

The sample means and variances are $\overline{x} = 89.5$, $s_x^2 = 274.9$, $\overline{y} = 93.1$, $s_y^2 = 238.8$.

(a) Assuming the data are normally distributed, carry out a two-sample $t$-test of the null hypothesis that the mean heart rate for the two situations is the same, against the alternative that it is different. What further assumptions are required for the test to be valid?

How would you modify the test if the alternative is that the mean heart rate is *higher* in situation (ii)? Explain which alternative you think is more appropriate here.

(b) Suggest a more appropriate test than that in (a). Carry out this test and explain why you prefer it.

**2.** Let $X_1, \ldots, X_n$ be independent $N(\theta, \sigma_0^2)$ random variables, where $\sigma_0^2$ is known. Find the most powerful test of size $\alpha$ of $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$, where $\theta_1 > \theta_0$.

Show that the power function $w(\theta)$ of this test is given by

$$w(\theta) = 1 - \Phi\left(\frac{\sqrt{n}}{\sigma_0}(\theta_0 - \theta) + z_\alpha\right)$$

where $\Phi$ is the cumulative distribution function of the standard normal distribution and $\Phi(z_\alpha) = 1 - \alpha$.

If $\theta_0 = 0$, $\theta_1 = 0.5$ and $\sigma_0 = 1$, how large must $n$ be if $\alpha = 0.05$ and the power at $\theta_1$ is to be 0.975? [If $\Phi$ is the $N(0,1)$ cdf, then $\Phi(1.645) = 0.95$ and $\Phi(1.96) = 0.975$.]

**3.** A telephone receptionist for a large partnership of financial advisers is responsible for determining the precise nature of each incoming enquiry and connecting the client with an appropriate adviser. The number of inappropriate connections on any given day may be modelled by a random variable $X$ which has a Poisson distribution with mean $\mu$. If $Z$ is the number of inappropriate connections made over a period of $n$ days, determine the distribution of $Z$ and find its expected value.

Uhura, who has been such a receptionist for many years, has been found to have a mean rate of $\mu_U = 0.47$ inappropriate connections per day. For several months she has been training Spock, a new receptionist, with corresponding mean rate $\mu_S$. At a meeting of senior partners, it was conjectured that Spock was already as proficient as Uhura; accordingly they resolved to keep a daily record of the number of inappropriate connections made by him over his next 10 working days. Find a critical region of size 5% for a test of the hypothesis that Spock is as proficient as Uhura versus the alternative that he is less proficient.

For what values of $\mu_S$ does the probability of type II error fall below 10%?

[Note that if $\varphi_\mu(k) = \sum_{x=0}^{k} \mu^x e^{-\mu}/x!$, then $\varphi_{4.7}(8) = 0.95$, $\varphi_{13}(8) = 0.1$.]

**4.** When studying the sex ratio in a population using a sample of size $n$, it is usually assumed that, indpendently, each child is male with probability $p$. Renkonen (1956) observed 19,711 male births out of a total of 38,562 births in American families with two children each. Use the likelihood ratio statistic $\Lambda$ to test the hypothesis $H_0 : p = \frac{1}{2}$ against a suitable alternative which you should specify.

Renkonen also found 17,703 males out of 35,042 similar births in Finland. Use the generalised likelihood ratio test to test the hypothesis that $p$ has the same value in each country versus a suitable alternative.

**5.** (a) A random variable $X$ has a distribution given by

$$P(X = i) = \pi_i, \quad i = 1, \ldots, k$$

where $\sum_{i=1}^{k} \pi_i = 1$. In a sample of size $n$ from a population with distribution $X$, the frequency of outcome $i$ is $n_i$, where $n_i > 0$ and $\sum_{i=1}^{k} n_i = n$. Find the maximum likelihood estimates of $\pi_1, \ldots, \pi_k$.

(b) The leaves of the plant *Pharbitis nil* can be variegated or unvariegated and, at the same time, faded or unfaded. In an experiment reported by Bailey (1961), of 290 plants which were observed, 31 had variegated faded leaves, 37 had variegated unfaded leaves, 35 had unvariegated faded leaves and 187 had unvariegated unfaded leaves.

If the properties of variegated appearance and faded appearance are assumed independent, then a model for the above observations has respective probabilities $\frac{1}{16}, \frac{3}{16}, \frac{3}{16}, \frac{9}{16}$. The general alternative is that the probabilities $\pi_i$, $i = 1, \ldots, 4$, are restricted only by the constraint $\sum \pi_i = 1$. Use a $\chi^2$ goodness-of-fit test to show that the data offer strong evidence that the independence model is inappropriate.

(c) A genetic theory which allows for an effect called *genetic linkage* assumes a probability model for the above observations with respective probabilities

$$\frac{1}{16} + \theta, \ \frac{3}{16} - \theta, \ \frac{3}{16} - \theta, \ \frac{9}{16} + \theta.$$

Find the equation satisfied by the maximum likelihood estimate $\widehat{\theta}$ of $\theta$.

You may assume that $\widehat{\theta} = 0.058$.

Let $H_0$ be the null hypothesis that the genetic linkage model is appropriate, and let $H_1$ be the general alternative. If $L_0$ is the supremum of the likelihood under $H_0$ and if $L_1$ is the supremum of the likelihood under $H_1$, show that

$$\Lambda = 2 \sum_{i=1}^{4} n_i \log \left( \frac{n_i}{n \pi_i(\widehat{\theta})} \right)$$

where $\Lambda = -2(\log L_0 - \log L_1)$. Write down the approximate distribution of $\Lambda$.

What can you infer about the plausibility of the genetic linkage model?

**6.** The ordered pairs of random variables $(X_k, Y_k)$, $k = 1, \ldots, n$, are independent and

$$P\big((X_k, Y_k) = (i, j)\big) = \pi_{ij}, \quad i = 1, \ldots, r, \quad j = 1, \ldots, c$$

where $\sum_{i,j} \pi_{ij} = 1$. The frequency of the outcome $(i, j)$ is $n_{ij}$, where $n_{ij} > 0$.

Find the maximum likelihood estimates of the $\pi_{ij}$ assuming that

(i) $\pi_{ij} = \alpha_i \beta_j$ for $i = 1, \ldots, r$ and $j = 1, \ldots, c$, where $\sum_i \alpha_i = \sum_j \beta_j = 1$, and

(ii) without this assumption.

Hence find test statistics for testing the null hypothesis that the $X_k$ and the $Y_k$ are independent using

(a) the likelihood ratio method,

(b) Pearson's $\chi^2$ statistic.

What can you say about the distributions of these two statistics for large values of $n$?

The data below (Agresti, 2007) cross-classifies gender and political party identification in the USA: 2757 individuals indicated whether they identified more strongly with the Democratic or Republican party or as Independents. Is there an association between gender and political party identification?

| | Party Identification | | |
|---|---|---|---|
| | Democrat | Independent | Republican |
| Female | 762 | 327 | 468 |
| Male | 484 | 239 | 477 |

**7.** (See `sheet3.R` for more.) Can you carry out all of the numerical calculations required for this sheet using R? See `sheet3.R` for more help and for code that you can cut and paste into R – this will help with the numerical calculations on this sheet.

To find quantiles or values of the cdf of $t$-distributions, we can use the functions `qt` and `pt` as described at the end of Sheet 2. To do the same for the $N(0, 1)$ distribution use the similar functions `qnorm` and `pnorm`, and for chi-squared distributions use `qchisq` and `pchisq`.

For example, the 0.95 quantile of $N(0, 1)$, and $\Phi(1.96)$, can be found using

```
qnorm(0.95)
pnorm(1.96)
```

For the chi-squared case we need to supply the number of degrees of freedom: the 0.95 quantile of a $\chi^2_{10}$ distribution, and the probability that a $\chi^2_{10}$ is less than 13 can be found using

```
qchisq(0.95, df = 10)
pchisq(13, df = 10)
```

```
#############
## Sheet 3 ##
#############

#### question 1
x <- c(72, 116, 79, 97, 90, 67, 115, 82, 95, 82)
y <- c(76, 120, 84, 99, 93, 75, 116, 83, 98, 87)
m <- 10
n <- 10

xbar <- mean(x)
ssqx <- var(x)
ybar <- mean(y)
ssqy <- var(y)

ss <- ((m-1)*ssqx + (n-1)*ssqy) / (m+n-2)
s <- sqrt(ss)
tobs <- (xbar - ybar) / (s*sqrt(1/m + 1/n))

# since tobs is negative
2 * pt(tobs, df = 18)
pt(tobs, df = 18)

qt(0.1, df = 18)

# as a check
t.test(x, y, var.equal = TRUE)

# now paired
d <- y - x
t1 <- mean(d)/sqrt(var(d)/10)
1 - pt(t1, df = 9)
# as a check
t.test(d)

#### question 2
pnorm(1.96)
pnorm(1.645)

#### question 3
ppois(8, lambda = 4.7)
ppois(8, lambda = 13)

#### question 4
x1 <- 19711
n1 <- 38562
p1hat <- x1/n1
Lambda <- 2 * ( n1*log(2) + x1*log(p1hat) + (n1-x1)*log(1-p1hat) )
1 - pchisq(Lambda, df = 1)
```

```
x2 <- 17703
n2 <- 35042
p2hat <- x2/n2
phat <- (x1 + x2)/(n1 + n2)
term1 <- (phat/p1hat)^x1 * ((1-phat)/(1-p1hat))^(n1-x1)
term2 <- (phat/p2hat)^x2 * ((1-phat)/(1-p2hat))^(n2-x2)
ratio <- term1*term2
Lambda1 <- -2*log(ratio)
1 - pchisq(Lambda1, df = 1)

# same as Lambda1
Lambda2 <- -2 * ((x1+x2)*log(phat) + (n1+n2-x1-x2)*log(1-phat)
                 - x1*log(p1hat) - (n1-x1)*log(1-p1hat)
                 - x2*log(p2hat) - (n2-x2)*log(1-p2hat))

#### question 5
obs <- c(31, 37, 35, 187)
expect <- 290*c(1/16, 3/16, 3/16, 9/16)
L1 <- 2 * sum(obs * log(obs/expect))
P1 <- sum((obs - expect)^2/expect)
1 - pchisq(L1, df = 3)
1 - pchisq(P1, df = 3)

n1 <- 31
n2 <- 37
n3 <- 35
n4 <- 187
a <- - 16^2*n1 - 16^2*(n2+n3) - 16^2*n4
b <- - 96*n1 - 160*(n2+n3) + 32*n4
c <- 27*n1 - 9*(n2+n3) + 3*n4
theta1 <- (-b + sqrt(b^2-4*a*c))/(2*a)
theta2 <- (-b - sqrt(b^2-4*a*c))/(2*a)

# theta1 not a valid value of theta
c(1/16+theta1, 3/16-theta1, 3/16-theta1, 9/16+theta1)

# theta2 is a valid value
c(1/16+theta2, 3/16-theta2, 3/16-theta2, 9/16+theta2)

# the log-likelihood is maximised at theta2 - picture
theta <- seq(-0.05, 0.18, length.out=50)
plot(theta, n1*log(1+16*theta) + (n2+n3)*log(3-16*theta)
     + n4*log(9+16*theta), type = "l", ylab = "g(theta)")
abline(v = theta2, lty = 2)

expect2 <- 290*c(1/16+theta2, 3/16-theta2, 3/16-theta2, 9/16+theta2)
L2 <- 2 * sum(obs * log(obs/expect2))
P2 <- sum((obs - expect2)^2/expect2)
1 - pchisq(L2, df = 2)
1 - pchisq(P2, df = 2)
```

```
#### question 6
x <- matrix(c(762, 484, 327, 239, 468, 477), ncol = 3)
n <- sum(x)
alpha <- rowSums(x)/n
beta <- colSums(x)/n

# under the null, the expected number in cell (i,j) is n*alpha[i]*beta[j]
# an outer product, denoted by %o%, does exactly what we need
# e.g try
num <- 1:12
num %o% num

# so evaluate the expected numbers under the null by
expect <- n * alpha %o% beta
obs <- x

Lambda <- 2 * sum(obs * log(obs/expect))
Pearson <- sum((obs-expect)^2 / expect)
1 - pchisq(Lambda, df = 2)
1 - pchisq(Pearson, df = 2)

## as a check
chisq.test(x)
```

# Part A Statistics HT 2021

## Problem Sheet 4

**1.** Suppose that $X_1, \ldots, X_n$ each have a geometric distribution with probability mass function $f(x \mid \theta) = (1-\theta)^x \theta$ for $x = 0, 1, \ldots$. Suppose that the prior for $\theta$ is a Beta$(a, b)$ density. Find the posterior distribution of $\theta$.

**2.** Let $\theta > 0$ be an unknown parameter and let $c > 0$ be a known constant. Conditional on $\theta$, suppose $X_1, \ldots, X_n$ are independent each with probability density function

$$f(x|\theta) = \theta c^\theta x^{-(\theta+1)}, \quad x \geqslant c$$

and suppose the prior for $\theta$ is a Gamma$(\alpha, \beta)$ density. Find the posterior distribution of $\theta$.

**3.** Let $r \geqslant 1$ be a known integer and let $\theta \in [0, 1]$ be an unknown parameter. The negative binomial distribution with index $r$ and parameter $\theta$ has probability mass function

$$f(x \mid \theta) = \binom{x+r-1}{x} (1-\theta)^x \theta^r \quad \text{for } x = 0, 1, \ldots.$$

Let $\theta$ have a Beta$(a, b)$ prior density and suppose, given $\theta$, that $X_1, \ldots, X_n$ are independent each with the above negative binomial distribution.

(a) Show that the posterior density is also a Beta density.

(b) Explain how to construct a $100(1-\alpha)\%$ equal-tailed credible interval for $\theta$. Will this interval be a highest posterior density interval?

**4.** Suppose that $X$ has a $N(\theta, \phi)$ distribution, where $\phi$ is known, Suppose also that the prior distribution for $\theta$ is $N(\theta_0, \phi_0)$, where $\theta_0$ and $\phi_0$ are known.

(a) Find the posterior distribution of $\theta$ given $X = x$.

(b) Show that the posterior mean of $\theta$ always lies between the prior mean and the observed value $x$.

(c) Construct a $100(1-\alpha)\%$ highest posterior density interval for $\theta$.

(d) Let $\phi = 2$, $\theta_0 = 0$ and $\phi_0 = 2$.

(i) Suppose the observed value is $x = 4$. What are the mean and variance of the resulting posterior distribution? Sketch the prior, likelihood, and posterior on a single set of coordinate axes.

(ii) Repeat (i) assuming $\phi_0 = 18$. Explain any resulting differences. Which of these two priors would likely have more appeal for a frequentist statistician?

**5.** Let $X$ be the number of heads when a coin with probability $\theta$ of heads is flipped $n$ times.

(a) When the prior is $\pi(\theta)$, the prior predictive distribution for $X$ (the predictive distribution before observing any data) is given by

$$P(X = k) = \int_0^1 P(X = k|\theta)\pi(\theta)\, d\theta, \quad k = 0, 1, \ldots, n.$$

Find the prior predictive distribution when $\pi(\theta)$ is uniform on $(0, 1)$.

(b) Suppose you assign a Beta$(a, b)$ prior for $\theta$, and then you observe $x$ heads out of $n$ flips. Show that the posterior mean of $\theta$ is always lies between your prior mean, $a/(a+b)$, and the observed relative frequency of heads, $x/n$.

(c) Show that, if the prior distribution on $\theta$ is uniform, then the posterior variance is always less than the prior variance.

(d) Give an example of a Beta$(a, b)$ prior distribution and values of $x, n$ for which the posterior variance is larger than the prior variance. (Try $x = n = 1$.)

**6.** A coin, with probability $\theta$ of heads, is flipped $n$ times and $r$ heads are observed.

(a) If the prior for $\theta$ is a uniform distribution on $(0, 1)$, what is the probability that the next flip is a head?

(b) Can you generalise to the case where $\theta$ has a Beta$(a, b)$ prior and where we wish to find the probability of getting $k$ heads from $m$ further flips?

**7.** (a) Let $X \sim N(\theta, \sigma_0^2)$, where $\sigma_0^2$ is known. Find the Jeffreys' prior for $\theta$.

(b) Let $X \sim N(\mu_0, \sigma^2)$, where $\mu_0$ is known. Find the Jeffreys' prior for $\sigma$.

(c) Let $X$ be Poisson with parameter $\lambda$. Find the Jeffreys' prior for $\lambda$. Check that the posterior distribution of $\theta$ given $X = x$ is proper, but that the Jeffreys' prior is not.

**8.** Suppose $X$ is the number of successes in a binomial experiment with $n$ trials and probability of success $\theta$. Either $H_0 : \theta = \frac{1}{2}$ or $H_1 : \theta = \frac{3}{4}$ is true. Show that the posterior probability that $H_0$ is true is greater than the prior probability for $H_0$ if and only if

$$x \log 3 < n \log 2.$$

**9.** Let $X \sim$ Binomial$(n, \theta)$, where the prior for $\theta$ is uniform on $(0, 1)$. Suppose that we wish to compare the hypotheses $H_0 : \theta \leqslant \frac{1}{2}$ and $H_1 : \theta > \frac{1}{2}$.

What are the prior odds of $H_0$ relative to $H_1$?

Find an expression for the posterior odds of $H_0$ relative to $H_1$.

If we observe $X = n$, find the Bayes factor $B$ of $H_0$ relative to $H_1$.

Check that $B \to 0$ as $n \to \infty$. Why is this expected?

**10.** Suppose we have a random sample $X_1, \ldots, X_n$ from a Poisson distribution with mean $\theta$. Suppose we wish to test the hypothesis $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ and that, under $H_1$, the prior distribution $\pi(\theta|H_1)$ for $\theta$ is given by

$$\pi(\theta|H_1) = \frac{\beta^\alpha}{\Gamma(\alpha)} \theta^{\alpha-1} e^{-\beta\theta}, \quad \theta > 0.$$

Calculate the Bayes factor of $H_0$ relative to $H_1$.

When $n = 6$, $\sum x_i = 19$, $\theta_0 = 2$, find the numerical value of the Bayes factor (i) when $\alpha = 4$ and $\beta = \frac{2}{3}$, and (ii) when $\alpha = 36$ and $\beta = 6$. Compare and interpret the values of the Bayes factor in cases (i) and (ii).