A9: Statistics

Sheet 1 — HT 2024

(Lectures 1–4, Notes section 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(λ) random variables. Find a function $g(\overline{X})$ such that the asymptotic variance of $g(\overline{X})$ does not depend on λ.

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

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

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

Neil Laws: neil.laws@stats.ox.ac.uk

4. Find the expected information for θ, 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 θ. Which will give the more precise inference about θ?

5. Suppose a random sample $Y_1, \ldots, Y_n$ from an exponential distribution with parameter λ is rounded down to the nearest δ, 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 jth rounded observation can be written $(1-e^{-\lambda\delta})e^{-\lambda z_j}$, and deduce that the expected information for λ based on the entire sample is

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

   Show that this has limit n/λ^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 θ, 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 θ. 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 θ.

Neil Laws: neil.laws@stats.ox.ac.uk

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?

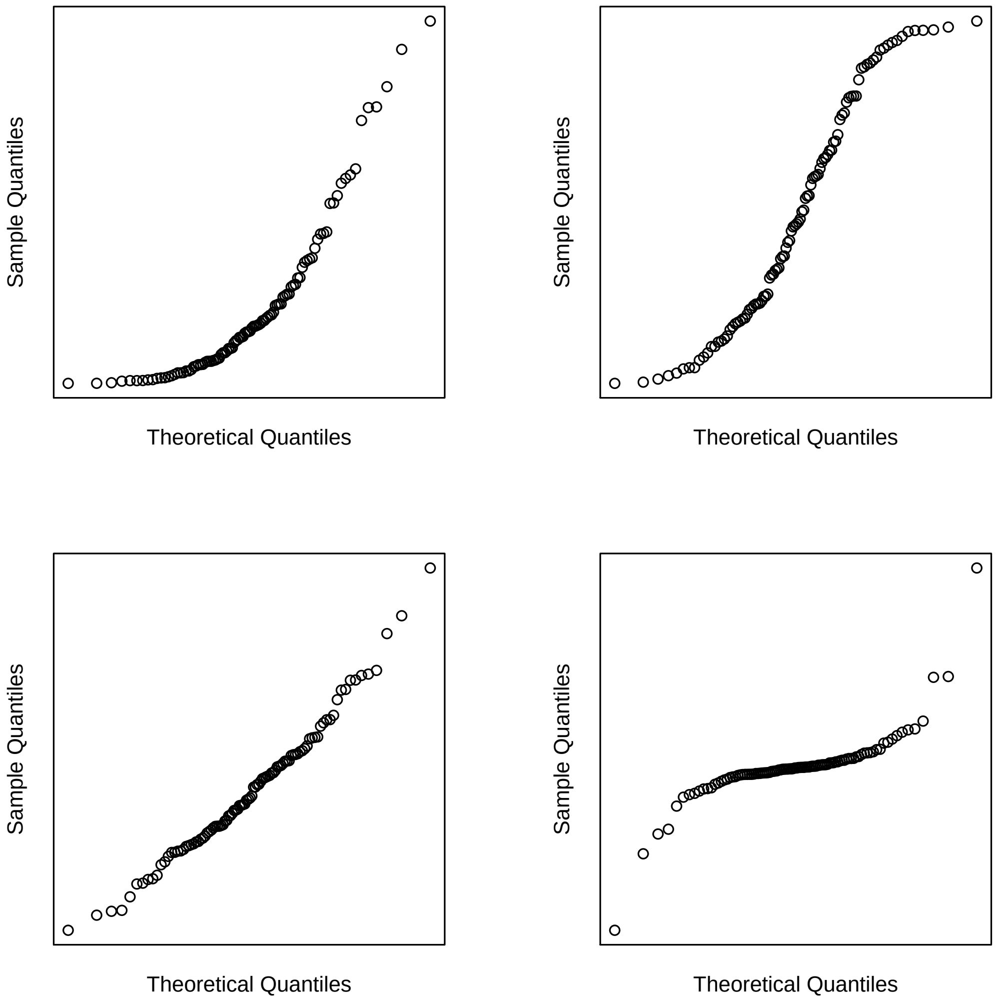

Figure 1. Normal Q-Q plots for four different samples, one from each of the following densities: $N(0,1)$, exponential, uniform, Cauchy. Which is which?

8. Read through the *Getting started with R* document (on Moodle) and install R/RStudio and run the commands in the document yourself.

   The R questions on these problem sheets need only a very small knowledge of R. R code will be supplied in questions and/or at the end of the sheet – this will be code you can copy and paste into R.

   (a) How typical is each of the Q-Q plots shown in Figure 1 above? 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.

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

   Use the R code at the end of the sheet to investigate.

9. (See the code at the end of this sheet 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?

Neil Laws: neil.laws@stats.ox.ac.uk

```
#### Question 8
# to do one example plot for each distribution in question 7:
x1 <- rnorm(100)
qqnorm(x1)

x2 <- rexp(100)
qqnorm(x2)

x3 <- runif(100)
qqnorm(x3)

# the Cauchy distribution is the t-distribution with 1 degree of freedom
x4 <- rt(100, df = 1)
qqnorm(x4)

# to see all four plots at once 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)
}
```


Neil Laws: neil.laws@stats.ox.ac.uk

```
# 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 degrees 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",
```


Neil Laws: neil.laws@stats.ox.ac.uk

```
     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)]")
```


Neil Laws: neil.laws@stats.ox.ac.uk