Example sheet 2

Working with distributions

Network Performance—DJW—2010/2011

Question 1. I have taken a sample of n values, $X_1, \ldots, X_n$. For each value in the sample I know the value of an associated predictor variable $w_1, \ldots, w_n$. I believe that $X_i \sim \text{Exp}(\lambda w_i)$, where λ is unknown. Calculate the maximum likelihood estimator for λ.

Question 2. I have taken a series of measurements of flow sizes, and plotted their empirical distribution function. Based on my plot, I propose to fit the distribution

$$\log \mathbb{P}(X \geq x) = \begin{cases} -\lambda x & \text{if } x \leq 1024 \\ -\lambda x - \mu(x - 1024) & \text{if } x > 1024. \end{cases}$$

Show that this distribution has density function

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \leq 1024 \\ (\lambda + \mu) e^{-(\lambda+\mu)x + 1024\mu} & \text{if } x > 1024. \end{cases}$$

(i) Find formulae for the maximum likelihood estimators of λ and μ.
(ii) Give pseudocode for a random number generator that generates random samples from this distribution.

Question 3. (i) Let X be a continuous random variable, uniformly distributed between a and b. In other words, the range is $[a,b]$ and density function is $f(x) = 1/(b-a)$. What is the distribution function? Calculate the mean, median, mode and variance.
(ii) Let Y be the sum of two throws of a dice. The possible outcomes are $\Omega = \{2,3,\ldots,12\}$. Compute the mean, median, mode and variance of Y.
(iii) Let $Z \sim \text{Geom}(p)$. Calculate the distribution function of Z.

Question 4. (i) Let X be an Exponential random variable with parameter λ. Let $Y = aX$, for some constant $a > 0$. Calculate the distribution function, i.e. find $\mathbb{P}(Y \geq y)$ as a function of y. What is the common name for the distribution of Y?
(ii) Let $X_1, X_2, \ldots, X_n$ be independent Exponential random variables with parameters $\lambda_1, \lambda_2, \ldots, \lambda_n$ respectively. Let $Z = \min(X_1, X_2, \ldots, X_n)$. Calculate the distribution function for Z. Show that $Z \sim \text{Exp}(\lambda_1 + \cdots + \lambda_n)$.
(iii) Let $X_1, \ldots, X_n$ be as above, and let $Z = \max(X_1, X_2, \ldots, X_n)$. Calculate the distribution function for Z. *[Hint. First find $\mathbb{P}(Z < z)$.]*

Question 5. Let X, Y and Z be generated from the following three random number generators respectively:

```
def rexp(λ): return -1.0/λ * math.log(random.random())
def rpareto(α): return math.pow(random.random(), -1.0/α)
def rpareto2(α,m): return m*(1-1.0/α)*rpareto(α)
```

Then $X \sim \text{Exp}(\lambda)$, $Y \sim \text{Pareto}(\alpha)$, and Z has a scaled Pareto distribution. Given λ, find α and m such that Z has the same mean and variance as X.

Question 6. The makers of the breakfast cereal Filboid Studge are promoting their product by means of a tie-in with Star Wars. In each pack of cereal there is a collectible figure, selected at random from Yoda, Luke, and six others. I buy boxes of the cereal repeatedly until I have the complete collection, when I revert to chocolate-frosted sugar bombs. Let the random variable N be the number of boxes of Filboid Studge that I bought. What is the mean and standard deviation of N? *[Hint. Let $T_1 = 1$ be the number of boxes I need to buy until I obtain my first figure, let T_2 be the number of additional boxes until I have obtained two distinct figures, and so on. Find the distribution of each T_i.]*

Question 7. Consider this very simplified model of what happens when a source sends a packet to destination using TCP:

(i) Source transmits SYN. Then it waits for a reply, and in the meantime retransmits the SYN periodically using a timeout of 1.2 seconds.

(ii) Once the destination receives the SYN, it replies with a SYN/ACK. Then it waits until it hears from the source again, and in the meantime retransmits the SYN/ACK periodically using a timeout of 1.2 seconds.

(iii) Once the source receives the SYN/ACK, it transmits its data packet. It waits for the reply, and in the meantime it retransmits the data packet periodically. This time it knows the round trip time, so it uses a timeout value of $\max(\texttt{RTT}, 200\text{ ms})$.

(iv) Whenever the destination receives a data packet, it replies with an ACK.

The network has the property that, with probability p, a packet may be dropped. This same drop probability applies to packets travelling in either direction. Packet drops are independent.

Let the random variable T be the time until the source hears an ACK confirming that the data packet has been received. Find the mean and standard deviation of T, as a function of p and RTT.

[Advanced] Find a power-series expansion of $\mathbb{E}T$ of the form $\mathbb{E}T = a_0 + a_1 p + a_2 p^2 + \ldots$. Give an intuitive explanation of the a_0 and a_1 terms.

Question 8. *[Difficult]* I have a biased coin, which has probability p of yielding heads. How many tosses on average until I have tossed two heads in a row?

Question 9. Consider the log file of `www.wischik.com`, available at under the "Handouts" link at `http://www.cs.ucl.ac.uk/staff/d.wischik/Teach/NP` (or, if you prefer, use your own web server logs).

(i) It has been suggested that `Size+1` has a Pareto distribution, where `Size` is the size in bytes of the body of an http response. Fit this distribution.

(ii) It has also been suggested that `Size+1` has a lognormal distribution, i.e. that $\log(\texttt{Size}+1)$ has a normal distribution. Fit this distribution.

(iii) Plot the empirical distribution function (EDF) of `Size+1`. Generate random samples from the two fitted distributions, and superimpose their EDFs. Which looks to be a better fit?

Question 10. This quesion concerns the arrival process of requests to `www.wischik.com`. We wish to know if the arrival process is Poisson, i.e. if interarrival times are independent and exponentially distributed. Since arrival rates vary according to time of day and day of the week, restrict attention to records which apply to weekday afternoons, 2pm–4pm.

(i) Plot the EDF of interarrival time. Superimpose the cdf for the exponential distribution with the same mean. Do they agree?

(ii) A better method is to transform the scales of your EDF plot, so that if the interarrival times truly are exponential then the EDF should follow a straight line. Does it?

(iii) Split interarrival times into pairs, and produce a scatter-plot of the first time against the second time. Does it seem that successive interarrival times are independent?

(iv) Another method to visualize independence is as follows. Split the data set of interarrival times into three classes, depending on whether the preceding interarrival time was short, medium or large. (Choose the cutoff points so that the three classes have roughly the same number of data points.) Plot the EDF for each of the three classes. Are they the same?