Assignment #1 STA355H1S

due Friday, January 31, 2020

Instructions: Solutions to problems 1 and 2 are to be submitted on Quercus (PDF files only). You are strongly encouraged to do problems 3 through 7 but these are **not** to be submitted for grading.

1. Suppose that $\boldsymbol{Y} = (Y_1, \cdots, Y_n)^T$ where $Y_1, \cdots, Y_n$ are independent Normal random variables where $Y_i \sim \mathcal{N}(\mu_i, \sigma^2)$. If Γ is an $n \times n$ orthogonal matrix (that is, $\Gamma^{-1} = \Gamma^T$) then $\boldsymbol{Z} = \Gamma \boldsymbol{Y}$ is a random vector whose elements $Z_1, \cdots, Z_n$ are independent Normal random variables each with variance σ^2 whose means $\boldsymbol{\nu} = (\nu_1, \cdots, \nu_n)^T$ are defined by $\boldsymbol{\nu} = \Gamma \boldsymbol{\mu}$. It is often convenient to assume that the mean vector $\boldsymbol{\nu}$ is "sparse" in the sense that all but a small fraction of its components are exactly 0. (In practice, the matrix Γ is chosen so that the sparsity of $\boldsymbol{\nu} = \Gamma \boldsymbol{\mu}$ is a reasonable assumption.)

Half-normal plots (which are often called Daniel plots) are used in some statistical models to distinguish values of $Z_1, \cdots, Z_n$ coming from a $\mathcal{N}(0, \sigma^2)$ distribution from those coming from Normal distributions with non-zero means. Suppose for example that $\nu_{i_1}, \cdots, \nu_{i_k}$ are non-zero with the remaining components equal to 0; then we would expect the values of $|Z_{i_1}|, \cdots, |Z_{i_k}|$ to be larger than other values of $\{|Z_i|\}$. Defining $W_i = |Z_i|$, we plot the ordered values $W_{(1)} \le \cdots \le W_{(n)}$ versus the corresponding quantiles of a standard "half-normal" distribution (the distribution of the absolute value of a $\mathcal{N}(0, 1)$ random variable); if $Z_1, \cdots, Z_n$ come from a $\mathcal{N}(0, \sigma^2)$ distribution then the points should lie close to a straight line whose slope is σ; on the other hand, if $\nu_{i_1}, \cdots, \nu_{i_k}$ are non-zero then we might expect the largest values $W_{(n-k+1)}, \cdots, W_{(n)}$ to lie noticeably above the line whose slope is σ. However, since σ is unknown, we need to estimate it and we do not want this estimate influenced (that is, biased upwards) by larger values of W_i; in part (b) below, we define possible "robust" estimators of σ.

(a) If $Z \sim \mathcal{N}(0, \sigma^2)$, show that

(i) the cdf of $|Z|$ is $G(x) = 2\Phi(x/\sigma) - 1$ where $\Phi(t)$ is the cdf of a $\mathcal{N}(0, 1)$ random variable;

(ii) the τ quantile of the distribution of $|Z|$ is $G^{-1}(\tau) = \sigma \Phi^{-1}((\tau + 1)/2)$.

(b) Suppose that $Z_1, \cdots, Z_n$ are independent $\mathcal{N}(0, \sigma^2)$ random variables and define $W_i = |Z_i|$ for $i = 1, \cdots, n$ and the order statistics $W_{(1)} \le W_{(2)} \le \cdots \le W_{(n)}$. The result of part (a) suggests that we could estimate σ using an order statistic $W_{(k)}$ as follows:

$$\widehat{\sigma}_k = \frac{W_{(k)}}{\Phi^{-1}((\tau_k + 1)/2)}$$

where (for example) $\tau_k = k/(n+1)$. If $\tau_k \to \tau \in (0,1)$ as $k, n \to \infty$ then

$$\sqrt{n}(\widehat{\sigma}_k - \sigma) \xrightarrow{d} \mathcal{N}(0, \gamma^2(\tau)).$$

Give an expression for $\gamma^2(\tau)$. For what value of τ is $\gamma^2(\tau)$ minimized? (You can determine the minimizing value of τ graphically.)

(c) A random variable U is said to be stochastically greater than a random variable V if $P(U \le x) \le P(V \le x)$ for all x with $P(U \le x) < P(V \le x)$ for some x. (This definition seems strange but note that it implies that values of V will tend to be less than values of U.) Suppose that $U \sim \mathcal{N}(\mu_1, \sigma^2)$ and $V \sim \mathcal{N}(\mu_2, \sigma^2)$ where $|\mu_1| > |\mu_2|$. Show that $|U|$ is stochastically greater than $|V|$. (Hint: First of all, show that the distribution of $|U|$ depends on $|\mu_1|$ so that we can assume that $\mu_1 > \mu_2 \ge 0$. Then show that if $X \sim \mathcal{N}(\mu, \sigma^2)$ for $\mu \ge 0$ then $P(|X| \le x)$ decreases as μ increases. Calculus is your friend here!)

(d) The function `halfnormal.txt` on Quercus contains a function to do half-normal plots. This function `halfnormal` has three arguments: the data `x`, the value of τ, `tau` (which defaults to $\tau = 0.5$) used to estimate σ, and an optional parameter `ylim`, which allows you to define the minimum and maximum y-axis values. The file `data.txt` contains 1000 observations from Normal distributions whose means are almost all 0. Using half-normal plots, try to estimate how many of the 1000 means are non-zero. There is no right or wrong approach here so feel free to be creative.

2. The hazard or failure rate function of a non-negative continuous random variable X is defined to be

$$h(x) = \frac{f(x)}{1 - F(x)} \quad \text{for } x \ge 0$$

where $f(x)$ is the pdf of X and $F(x)$ is its cdf. We can also define $h(x)$ by

$$h(x) = \lim_{\delta \downarrow 0} \frac{1}{\delta} P(x \le X \le x + \delta | X \ge x).$$

(a) A useful formula for the expected value of any non-negative random variable is

$$E(X) = \int_0^\infty (1 - F(x))\, dx.$$

If X is also continuous with pdf $f(x)$ then this formula can be derived as follows:

$$\begin{aligned}
E(X) &= \int_0^\infty x f(x)\, dx \\
&= \int_0^\infty \int_0^x f(x)\, dt\, dx \\
&= \int_0^\infty \int_t^\infty f(x)\, dx\, dt \\
&= \int_0^\infty (1 - F(t))\, dt.
\end{aligned}$$

If $h(x)$ is the hazard function of X, show that

$$E(X) = \int_0^1 \frac{1}{h(F^{-1}(\tau))}\, d\tau.$$

(Hint: Make the change of variables $u = F^{-1}(\tau)$.)

(b) Suppose that $X_{(k)}$ is the k-th order statistic where $k \approx \tau n$ (for some $\tau \in (0,1)$) and define $D_k = X_{(k)} - X_{(k-1)}$. From lecture, we know that the distribution of $n\,D_k$ is approximately Exponential with mean $1/f(F^{-1}(\tau))$. Use this fact to show that the distribution of $(n-k+1)D_k$ is approximately Exponential with mean $1/h(F^{-1}(\tau))$. (Hint: Note that $h(F^{-1}(\tau)) = f(F^{-1}(\tau))/(1-\tau)$.)

(c) The shape of $h(x)$ provides useful information about the distribution not readily obvious from the pdf and cdf; for example, if X represents the lifetime of some (say) electronic component then a decreasing hazard function would indicate that the component improves with age.

The **total time on test (TTT) plot** provides one to assess the rough shape of $h(x)$ based on a sample $x_1, \cdots, x_n$. To construct this plot, we define

$$\begin{aligned} d_1 &= n x_{(1)} \\ d_k &= (n-k+1)(x_{(k)} - x_{(k-1)}) \quad \text{for } k = 2, \cdots, n \end{aligned}$$

and plot $(d_1 + \cdots + d_k)/(x_1 + \cdots + x_n)$ versus k/n for $k = 1, \cdots, n$. Using the result from part (b), we might argue that $(d_1 + \cdots + d_k)/(x_1 + \cdots + x_n)$ is an estimate of

$$\frac{1}{E(X)} \int_0^\tau \frac{1}{h(F^{-1}(\tau))}\, d\tau$$

for $\tau = k/n$. If the underlying hazard function $h(x)$ is decreasing then the shape of these points will be roughly convex (and lie below the 45^o line) while if $h(x)$ is increasing then the shape of the points will be roughly concave (and lie above the 45^o line).

Given data in a vector `x`, the TTT plot can be constructed as follows:

```
> x <- sort(x) # order elements from smallest to largest
> n <- length(x) # find length of x
> d <- c(n:1)*c(x[1],diff(x))
> plot(c(1:n)/n, cumsum(d)/sum(x), xlab="t", ylab="TTT")
> abline(0,1) # add 45 degree line to plot
```

Data on the lifetimes (in hours) of Kevlar 373/epoxy strands (subjected to constant pressure at 90% stress level) are contained in the file `kevlar.txt`. Construct a TTT plot for these data. Does the hazard function appear to be increasing or decreasing with time?

Supplemental problems (not to be handed in):

3. (a) Suppose that X has a Gamma distribution with shape parameter α and scale parameter λ; the density of X is

$$f(x) = \frac{\lambda^\alpha x^{\alpha-1} \exp(-\lambda x)}{\Gamma(\alpha)} \quad \text{for } x > 0$$

Find expressions for the skewness and kurtosis of X in terms of α and λ. (Do these depend on λ?) What happens to the skewness and kurtosis as $\alpha \to \infty$?
(b) Suppose that $X_1, \cdots, X_n$ are independent and define $S_n = X_1 + \cdots + X_n$. Assuming that $E(X_i^3)$ is well-defined for all i, show that the skewness of S_n is given by

$$\text{skew}(S_n) = \left(\sum_{i=1}^n \sigma_i^2\right)^{-3/2} \sum_{i=1}^n \sigma_i^3 \, \text{skew}(X_i)$$

where $\sigma_i^2 = \text{Var}(X_i)$. (Hint: Follow the proof given for the kurtosis identity assuming for simplicity that $E(X_i) = 0$; this is more simple since $E(S_n)$ involves a triple summation, most of whose terms are 0.)

4. Suppose that $X_1, \cdots, X_n$ are independent random variables with distribution function F where $\mu = E(X_i)$ and $\sigma^2 = \text{Var}(X_i)$. For some families of distributions, the variance is a function of the mean so that $\sigma^2 = \sigma^2(\mu)$. A function g is said to be a variance stabilizing transformation for the family of distributions if

$$\sqrt{n}(g(\bar{X}_n) - g(\mu)) \xrightarrow{d} \mathcal{N}(0, 1)$$

(a) Show that g defined above must satisfy the differential equation

$$g'(\mu) = \pm \frac{1}{\sigma(\mu)}.$$

(Note that g is not unique.)
(b) Find variance stabilizing transformations for

(i) Poisson distributions;

(ii) Exponential distributions;

(iii) Bernoulli distributions.

5. Suppose that $X_1, \cdots, X_n$ are independent random variables with some continuous distribution function F. Given data $x_1, \cdots, x_n$ (outcomes of $X_1, \cdots, X_n$), we can make a boxplot to graphically represent the data — observations beyond the "whiskers" (which extend to at most $1.5 \times$ interquartile range from the upper and lower quartiles) are flagged as possible outliers. When n is large enough, we can obtain a crude estimate for the expected number of outliers as follows:

(i) Compute the lower and upper quartiles of F, $F^{-1}(1/4)$ and $F^{-1}(3/4)$ and define $\mathrm{IQR} = F^{-1}(3/4) - F^{-1}(1/4)$.

(ii) Compute the probability of an outlier by

$$F(F^{-1}(1/4) - 1.5 \times \mathrm{IQR}) + 1 - F(F^{-1}(3/4) + 1.5 \times \mathrm{IQR})$$

(iii) The expected number of outliers is simply n times the probability in part (ii).

Compute the expected number of outliers for the following distributions.
(a) Normal distribution – note that the probability in (ii) will not depend on the mean and variance so you can assume a standard normal distribution. (The R functions `pnorm` and `qnorm` can be used to compute the distribution function and quantiles, respectively, for the normal distribution.)
(b) Laplace distribution with density

$$f(x) = \frac{1}{2}\exp(-|x|).$$

(No R functions for the distribution functions and quantiles seem to exist for the Laplace distribution. However, both are easy to evaluate analytically.)
(c) Cauchy distribution with density

$$f(x) = \frac{1}{\pi(1+x^2)}$$

(The R functions `pcauchy` and `qcauchy` can be used to compute the distribution function and quantiles, respectively, for the Cauchy distribution.)
(d) Comment on the differences between the 3 distributions considered in parts (a)–(c). In particular, how does the proportion of outliers change as the "tails" (i.e. the rate at which $f(x)$ goes to 0 as $|x| \to \infty$) of the distributions change?

6. Suppose that $X_1, X_2, \cdots$ is a sequence of independent random variables with mean μ and variance $\sigma^2 < \infty$; define $\bar{X}_n = n^{-1}(X_1 + \cdots + X_n)$. Describe the limiting behaviour (that is, either convergence in probability or convergence in distribution as well as the limit as $n \to \infty$) of the following random variables.
(a) $S_n^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$.
(b) $\sqrt{n}(\bar{X}_n - \mu)/S_n$.
(c) $\sqrt{n}(\exp(\bar{X}_n) - \exp(\mu))/S_n$.
(d) $\dfrac{1}{n}\sum_{i=1}^{n}|X_i - \bar{X}_n|$. (The limit here should be intuitively clear; however, proving it is not easy!)

7. Suppose that $a_n(X_n - \theta) \xrightarrow{d} Z$ (where $a_n \uparrow \infty$) and that $g(x)$ is an infinitely differentiable function (that is, it has derivatives of all orders). The Delta Method says that

$$a_n(g(X_n) - g(\theta)) \xrightarrow{d} g'(\theta)Z;$$

if $g'(\theta) = 0$ then the right hand side above is 0 and so $a_n(g(X_n) - g(\theta)) \xrightarrow{p} 0$.
(a) Suppose that $g'(\theta) = 0$ and $g''(\theta) \neq 0$. Use the Taylor series expansion

$$g(x) = g(\theta) + g'(\theta)(x - \theta) + \frac{1}{2}g''(\theta)(x - \theta)^2 + r_n$$

(where $r_n/(x - \theta)^2 \to 0$ as $x \to \theta$) to find the limiting distribution of $a_n^2(g(X_n) - g(\theta))$.
(b) Extend the result of part (a) to the case where $g'(\theta) = g''(\theta) = \cdots = g^{(k-1)}(\theta) = 0$ but $g^{(k)}(\theta) \neq 0$ ($g^{(k)}$ denotes the k-th derivative of g).