MATH11400 Statistics 1 2010–11

Homepage http://www.stats.bris.ac.uk/%7Emapjg/Teach/Stats1/

Solution Sheet 6

1. (a) From your Parametric Families summary sheet (2.1 in handouts), $\mathrm{E}(X) = \theta/2$ and $\mathrm{Var}(X) = \theta^2/12$. Thus $\mathrm{E}(\bar{X}) = \mathrm{E}(\sum X_i/n) = \sum E(X_i)/n = n\theta/2n = \theta/2 = \tau$, so $\mathrm{bias}(\bar{X}) = \mathrm{E}(\bar{X} - \theta/2) = \mathrm{E}(\bar{X}) - \theta/2 = 0$, so $\bar{X}$ is unbiased as an estimator of τ.

   Also $\mathrm{Var}(\bar{X}) = \mathrm{Var}(\sum X_i/n) = \sum \mathrm{Var}(X_i)/n^2 = n\theta^2/12n^2 = \theta^2/12n$. Thus, as an estimator of the population median $\tau = \theta/2$, the method of moments estimator has mean square error $\mathrm{mse}\ (\bar{X}) = \mathrm{Var}(\bar{X}) + \mathrm{bias}(\bar{X})^2 = \theta^2/12n + 0 = \theta^2/12n$.

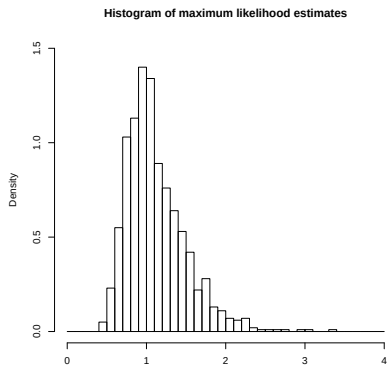
   (b) Here $\mathrm{E}(Y) = \int_0^\theta y f_Y(y;\theta)dy = \int_0^\theta ny^n/\theta^n dy = [ny^{n+1}/(n+1)\theta^n]_0^\theta = n\theta/(n+1)$. Thus $\mathrm{E}(\hat{\tau}_{\mathrm{mle}}) = \mathrm{E}(Y/2) = n\theta/2(n+1)$ and $\mathrm{bias}(\mathrm{E}(\hat{\tau}_{\mathrm{mle}}) = \mathrm{E}(\hat{\tau}_{\mathrm{mle}} - \theta/2) = n\theta/2(n+1) - \theta/2 = -\theta/2(n+1)$.

2. (a,b) I calculated the sample mean $\bar{x}$ and the maximum likelihood estimate $\hat{\theta} = 1/\bar{x}$ for each sample with the commands:

```
> sample.mean <- apply(xsamples,1,mean)
> theta.mle <- 1/sample.mean
```

   I used the command `summary(theta.mle)` to check the range of values in the array `theta.mle` and decided to start the cell break points at 0, finish at 4, and have cell widths of 0.1. The following commands then produced the histogram shown below, where `probability = T` gives a histogram of the normalised relative frequencies (i.e. approximating the probability density function) rather than the total counts. Since the process is random, your histogram may look slightly different, but the overall features should be the same.

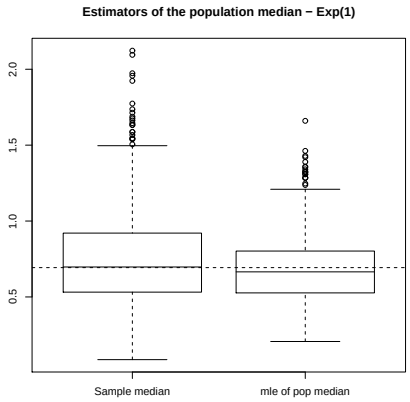
```
> hist(theta.mle, probability=T,
+ breaks=seq(0,4,0.1),ylim=c(0,1.5), xlab="",
+ main="Histogram of maximum likelihood estimates")
```



   (c) Recall from Probability 1 that the median of a continuous distribution is the value ξ such that $F_X(\xi) = 1/2$, i.e. $\xi = F_X^{-1}(1/2)$. For the Exp(θ) distribution, X takes values in $(0, \infty)$ and on this set $F_X(x) = 1 - e^{-\theta x}$ with inverse $F_X^{-1}(y) = -\log(1-y)/\theta$, so the population median is $\xi = F_X^{-1}(1/2) = -\log(1-1/2)/\theta = \log(2)/\theta$.

   To calculate the sample median and the maximum likelihood estimate of the population median $\tau(\theta) = \log(2)/\theta$ I used the commands:

```
> sample.median <- apply(xsamples,1,median)
> mle.median <- log(2)/theta.mle
```
To produce an annotated boxplot of the values of the sample median and the values of the mle of the population median I used the commands:
```
> boxplot(sample.median, mle.median,
+ names=c("Sample median","mle of pop median"),
+ main="Estimators of the population median - Exp(1)")
```
(d) To add a horizontal dashed line at height $\log(2)$ to the plot I used the command:
```
> abline(h=log(2), lty=2)
```
The boxplot shows that the distribution of the sample median is centered (i.e. has its median) on the true population median, whereas the median of the distribution of maximum likelihood estimates gives a slight systematic under-estimate of the true population median. However, the variability of the maximum likelihood estimate is clearly a lot smaller than that of the sample median, so overall the maximum likelihood estimate is likely to be much closer to the true value of the population median than the sample median.



(e) The true value of the population median is $\log(2) = 0.6931$. The sample mean and variance of the 1000 maximum likelihood estimates and the 1000 sample medians are given in the table below. We can approximate the bias of the estimators with the commands
```
> mean(mle.median-log(2))
> mean(sample.median-log(2))
```
and their mean square error (mse) with the commands
```
> var(mle.median) + (mean(mle.median-log(2)))^2
> var(median) + (mean(median-log(2)))^2.
```
Note that the outliers affect the mean of the sample medians so much that it is further away from the true population median than the mean of the mles - even though the median of the mles was further away from the population median than the median of the sample medians.

Overall, we see from the table that the mle has substantially smaller mean square error (mse) than the sample median, mainly because of its smaller variance, and for that reason, the mle would be preferred to the sample median as an estimator.

	approximate			
Estimators of the population median	mean	bias	variance	mse
mle	0.6845	-0.0086	0.0463	0.0463
sample median	0.7440	0.0509	0.0934	0.0960

3. For $i = 1, \ldots, 10$, let X_i denote the outcome of the ith toss, so $X_i = 1$ if a head is obtained and $Y_i = 0$ if a tail is obtained. Then the X_i are 10 independent Bernoulli(1/2) random variables, and the total number of heads is just $T = \sum_{i=1}^{10} X_i$.

(a) From your Probability notes, T has a Binomial(10,1/2) distribution. Thus the exact probabilities for A and B are given by

$$\begin{aligned} P(A) &= P(T \leq 1) = P(T = 0) + P(T = 1) = (\tfrac{1}{2})^{10} + 10(\tfrac{1}{2})^{10} &&= 0.0107. \\ P(B) &= P(T \geq 6) = \textstyle\sum_{k=6}^{10} P(T = k) = (210 + 120 + 45 = 10 + 1)(\tfrac{1}{2})^{10} &&= 0.3770. \end{aligned}$$

(b) If $X_1, \ldots, X_n$ are i.i.d. random variables with mean μ_X and variance σ_X^2, then the corollary to the central limit theorem says that the approximate distribution of $X_1 + \cdots + X_n$ is $N(n\mu_X, n\sigma_X^2)$. Here $\mathrm{E}(X_i) = 1/2$, $\mathrm{Var}(X_i) = 1/4$ and $n = 10$. Thus the distribution of $T = X_1 + \cdots + X_{10}$ is approximately $N(5, 5/2)$ and so $(T - 5)/\sqrt{5/2}$ has approximately the same distribution as $Z \sim N(0, 1)$.

Applying the central limit theorem without any continuity correction, we get

$$\begin{aligned} P(T \leq 1) &= P((T-5)/\sqrt{5/2} \leq (1-5)/\sqrt{5/2}) \simeq P(Z \leq -2.5298) \\ &= \Phi(-2.5298) = 1 - \Phi(2.5298) = 1 - 0.9943 = 0.0057. \\ P(T \geq 6) &= P((T-5)/\sqrt{5/2} \geq (6-5)/\sqrt{5/2}) \simeq P(Z \geq 0.6325) \\ &= 1 - \Phi(0.6325) = 0.2635. \end{aligned}$$

(c) If X is integer valued, then the continuity correction suggests that a better approximation to $P(X_1 + \cdots + X_{10} \leq k)$ is given by $P(S \leq k + 1/2)$, where $S \sim N(5, 5/2)$, so

$$\begin{aligned} P(T \leq 1) &= P(X_1 + \cdots + X_n \leq 1) \simeq P(S \leq 1.5) \\ &= P((S-5)/\sqrt{5/2} \leq (1.5-5)/\sqrt{5/2}) = P(Z \leq -2.2136) = 0.0134. \\ P(T \geq 6) &= P(X_1 + \cdots + X_n \geq 6) \simeq P(S \geq 5.5) \\ &= P((S-5)/\sqrt{5/2} \geq (5.5-5)/\sqrt{5/2}) = P(Z \geq 0.3162) = 0.3759. \end{aligned}$$

(d) The true probabilities are 0.0107 and 0.3770. The approximations obtained using the continuity correction (0.0134 and 0.3759) are appreciably better than the approximations (0.0057 and 0.2635) obtained without the correction. This is what we would expect, since the sample size n is small and the sum $X_1 + \cdots + X_{10}$ is integer valued.

4. (a) Let X be the number of parking places required by the residents of a randomly chosen apartment. Then X can take value $0, 1$ or 2. The information in the question tells us that $P(X = 2) = 0.2$, $P(X = 1) = 0.7$ and $P(X = 0) = 0.1$. Thus X has mean $\mu_X = \mathrm{E}(X) = (2 \times 0.2) + (1 \times 0.7) + (0 \times 0.1) = 1.1$.
In a very similar way, $\mathrm{E}(X^2) = 1.5$, so X has variance $\sigma_X^2 = \mathrm{E}(X^2) - [\mathrm{E}(X)]^2 = 0.29$.

(b) If we assume the demands of the residents of the 200 apartments are independent, then the total demand is T, where $T = X_1 + \cdots + X_n$, $n = 200$, and $X_1, X_2, \ldots, X_n$ are independent random variables with the same distribution as X. From (the corollary to) the central limit theorem, the distribution of T is approximately $N(n\mu_X, n\sigma_X^2)$. Here $n\mu_X = 220$ and $n\sigma_X^2 = 58$, so T is approximately $N(220, 58)$.

The event that there are not enough parking places to satisfy demand corresponds to $\{T > 230\}$. Since T is integer valued we use a continuity correction. Let $S \sim N(220, 58)$, then $P(T > 230) \simeq P(S > 230 + 1/2) = P((S - 220)/\sqrt{58} > (230.5 - 220)/\sqrt{58}) = P(Z > 1.3787) = 1 - \Phi(1.3787) = 0.0840$.

Without the continuity correction, the approximation indicated by the central limit theorem gives gives $P(T > 230) \simeq P((S - 220)/\sqrt{58} > (230.5 - 220)/\sqrt{58}) = P(Z > 1.3130) = 1 - \Phi(1.3130) = 0.0946$.

Since to have $P(Z > z) = 0.01$ implies that $z = 2.3263$, according to the first approximation we need k places where $(k + 0.5 - 220)/\sqrt{58} \geq 2.3263$, i.e. $k \geq 238$.

5. Let $X_1, X_2, X_3, \ldots$ be independent random variables with $P(X_i = 1) = 0.37 = 1 - P(X_i = 0)$. Thus each X_i has the same distribution as X where $X \sim$ Bernoulli(0.37), with $\mathrm{E}(X) = \mu_X = 0.37$ and $\mathrm{Var}(X) = \sigma_X^2 = 0.37 \times 0.63 = 0.2331$.

   Now let $T_n = X_1 + \cdots + X_n$ be the total number in the sample that say they support the government. Here $n = 1500$. From the central limit theorem, the integer valued r.v. T_n has approximately the same distribution as the continuous r.v. S_n, where $S_n \sim N(n\mu_X, n\sigma_X^2) = N(555, 349.65)$ and $(S_n - 555)/\sqrt{349.65}$ has the same distribution as $Z \sim N(0,1)$

   Then $P(|T_n/n - 0.37| \le 0.02) = P(|T_n - 555| \le 30) = P(525 \le T_n \le 585) \simeq P(525 - 0.5 \le S_n \le 585 + 0.5)$ [with the continuity correction] $= P(|(S_n - 555)/\sqrt{349.65}| \le 30.5/\sqrt{349.65}) = P(|Z| \le 1.6311) = 2\Phi(1.6311) - 1 = 2(0.94857) - 1 = 0.89713$. The exact Binomial probability is 0.89636. Without the continuity correction we would approximate $P(|T_n/n - 0.37| \le 0.02)$ by $P(525 \le S_n \le 585) = \cdots = 0.89137$.

6. (a) Let $Z = (U_1 + \cdots + U_{12}) - 6$, where $U_1, \ldots, U_{12}$ is a a random sample of size 12 from a $\mathrm{U}(0,1)$ distribution. A random variable U with $\mathrm{U}(0,1)$ distribution has mean $\mathrm{E}(U) = \mu_U = 1/2$ and variance $\mathrm{Var}(U) = \sigma_U^2 = 1/12$. Thus $\mathrm{E}(Z) = \sum_1^{12} \mathrm{E}(U_i) - 6 = (12 \times 1/2) - 6 = 0$ and $\mathrm{Var}(Z) = \sum_1^{12} \mathrm{Var}(U_i) = (12 \times 1/12) = 1$.

   (b) I used the following commands in *R*. Note that, to stop the graph of the Normal density going off the top of my histogram, I need to specify the height of the y-axis in my histogram with the sub-command `ylim=c(0,0.37)`.

```
> unif.dat <- runif(12000)
> unif.mat <- matrix(unif.dat,nrow=1000)
> unif.sum <- apply(unif.mat,1,sum)
> z.sample <- unif.sum - 6
> hist(z.sample, probability=T, ylim=c(0,0.4),
+ main = "Histogram of Z values")
> range <- seq(-3,3,0.01)
> lines(range, dnorm(range)) # pause to look at first plot
> qqnorm(z.sample)
> abline(0,1)
```

   The resulting plots are shown below. The fit to the Normal distribution is quite good. You could get a better idea of the detailed fit by specifying smaller cell widths in your histogram with a sub-command such as `breaks = seq(-4,4,0.1)` - where you would need to adjust these limits if any of your Z values were outside the range $(-4, 4)$.