UPPSALA UNIVERSITY
Dept of Mathematics
Rolf Larsson

Matematical statistics
Inference II 1MS037
2018–11–29

Exam 2017, solutions

1. (a) For a given day, drawing from m units (where $m = 2n$ on working days and $m = n$ during weekends) *with replacement*, the number of defective pens X sampled is Binomial with parameters m and p.

   Let $\mathbf{X} = (X_1, ..., X_7)$ where $X_i \sim \text{Bin}(2n, p)$ if $i = 1, 2, 3, 4, 5$ and $X_i \sim \text{Bin}(n, p)$ if $i = 6, 7$. Assuming that the X_i are simultaneously dependent, for a sample $\mathbf{x} = (x_1, ..., x_7)$ we get the probability function

$$
\begin{aligned}
p(\mathbf{x}) &= \prod_{i=1}^{5} \binom{2n}{x_i} p^{x_i}(1-p)^{2n-x_i} \prod_{i=6}^{7} \binom{n}{x_i} p^{x_i}(1-p)^{n-x_i} \\
&= \left\{ \prod_{i=1}^{5} \binom{2n}{x_i} \right\} p^{\sum_{i=1}^{5} x_i}(1-p)^{10n-\sum_{i=1}^{5} x_i} \\
&\quad \left\{ \prod_{i=6}^{7} \binom{n}{x_i} \right\} p^{\sum_{i=6}^{7} x_i}(1-p)^{2n-\sum_{i=6}^{7} x_i} \\
&= \left\{ \prod_{i=1}^{5} \binom{2n}{x_i} \right\} \left\{ \prod_{i=6}^{7} \binom{n}{x_i} \right\} p^{\sum_{i=1}^{7} x_i}(1-p)^{12n-\sum_{i=1}^{7} x_i}.
\end{aligned}
$$

   (b) We may rewrite the equation above as

$$
p(\mathbf{x}) = (1-p)^{12n} \exp\left\{ \sum_{i=1}^{7} x_i \log\left(\frac{p}{1-p}\right) \right\} \left\{ \prod_{i=1}^{5} \binom{2n}{x_i} \right\} \left\{ \prod_{i=6}^{7} \binom{n}{x_i} \right\}.
$$

   This is on one-parameter exponential family form,

$$
p(\mathbf{x}) = A(\theta) \exp\{\zeta(\theta)T(\mathbf{x})\}h(\mathbf{x}),
$$

   with $\theta = p$, natural parameter $\zeta(p) = \log\left(\frac{p}{1-p}\right)$ and sufficient statistic $T(\mathbf{x}) = \sum_{i=1}^{7} x_i$.

   (c) The sum of independent binomial random variables with the same p parameter is binomial, summing the values of the n parameters. Hence, the sufficent statistic (seen as a random variable) $T(\mathbf{X}) = \sum_{i=1}^{7} X_i$ is $\text{Bin}(12n, p)$.

(d) The log likelihood is

$$
\begin{aligned}
l(p) &= C + 12n\log(1-p) + \sum_{i=1}^{7} x_i \log\left(\frac{p}{1-p}\right) \\
&= C + (12n - s)\log(1-p) + s\log(p),
\end{aligned}
$$

where C is a constant and $s = \sum_{i=1}^{7} x_i$. The first two derivatives are

$$
\begin{aligned}
l'(p) &= -\frac{12n-s}{1-p} + \frac{s}{p}, \\
l''(p) &= -\frac{12n-s}{(1-p)^2} - \frac{s}{p^2}.
\end{aligned}
$$

Denoting the random counterpart of s by $S \sim \mathrm{Bin}(12n, p)$, the Fisher information may be obtained as

$$
\begin{aligned}
I(p) = -E\{l''(p;\mathbf{X})\} &= \frac{12n - E(S)}{(1-p)^2} + \frac{E(S)}{p^2} = \frac{12n - 12np}{(1-p)^2} + \frac{12np}{p^2} \\
&= 12n\left(\frac{1}{1-p} + \frac{1}{p}\right) = \frac{12n}{p(1-p)}.
\end{aligned}
$$

(e) For an estimator T that is unbiased for p, the Cramér-Rao lower bound for its variance is

$$
\frac{1}{I(p)} = \frac{p(1-p)}{12n}.
$$

2. (a) Let the observed sample be $\mathbf{x} = (x_1, ..., x_n)$. The likelihood function is

$$
L(p;\mathbf{x}) = \prod_{i=1}^{n} pq^{x_i} = p^n q^{\sum_{i=1}^{n} x_i}.
$$

(b) The log likelihood is

$$
l(p;\mathbf{x}) = n\log p + s\log(1-p),
$$

where $s = \sum_{i=1}^{n} x_i$. This gives the score function

$$
V(p;\mathbf{x}) = l'(p;\mathbf{x}) = \frac{n}{p} - \frac{s}{1-p}.
$$

(c) Solving $l'(p; \mathbf{x}) = 0$ w.r.t p, we find $p = n/(s+n)$. Moreover,

$$l''(p; \mathbf{x}) = -\frac{n}{p^2} - \frac{s}{(1-p)^2} < 0,$$

and so, the solution above is the MLE, i.e. $\hat{p}_{MLE} = n/(s+n)$.

(d) The expectation of $\hat{p}_{MLE}$, seen as a random variable, is

$$E(\hat{p}_{MLE}) = E\left(\frac{n}{S+n}\right),$$

where $S = \sum_{i=1}^n X_i$. This is *not* equal to

$$\frac{n}{E(S)+n} = \frac{n}{n(1-p)/p+n} = p.$$

Hence, $\hat{p}_{MLE}$ is not unbiased.

(e) Efficiency is only defined for unbiased estimators, so the answer is no. (However, being an MLE in a regular exponential family, $\hat{p}_{MLE}$ is *asymptotically* efficient.)

3. (a) For a day i, where $i = 1, ..., 5$ for working days and $i = 6, 7$ for weekends, drawing from n units with replacement, the number of defective pens X_i sampled is Binomial with parameters n and p_i. Here, $p_i = p$ for $i = 1, ..., 5$ whereas $p_i = 2p$ for $i = 6, 7$.

Let $\mathbf{X} = (X_1, ..., X_7)$ where $X_i \sim \text{Bin}(n, p)$ if $i = 1, 2, 3, 4, 5$ and $X_i \sim \text{Bin}(n, 2p)$ if $i = 6, 7$. Assuming that the X_i are simultaneously independent, for a sample $\mathbf{x} = (x_1, ..., x_7)$ we get the probability function

$$\begin{aligned}
p(\mathbf{x}) &= \prod_{i=1}^{5} \binom{n}{x_i} p^{x_i}(1-p)^{n-x_i} \prod_{i=6}^{7} \binom{n}{x_i} (2p)^{x_i}(1-2p)^{n-x_i} \\
&= \left\{\prod_{i=1}^{5} \binom{n}{x_i}\right\} p^{\sum_{i=1}^5 x_i}(1-p)^{5n-\sum_{i=1}^5 x_i} \\
&\quad \left\{\prod_{i=6}^{7} \binom{n}{x_i}\right\} (2p)^{\sum_{i=6}^7 x_i}(1-2p)^{2n-\sum_{i=6}^7 x_i} \\
&= \left\{\prod_{i=1}^{7} \binom{n}{x_i}\right\} 2^{\sum_{i=6}^7 x_i} p^{\sum_{i=1}^7 x_i}(1-p)^{5n-\sum_{i=1}^7 x_i}(1-2p)^{2n-\sum_{i=6}^7 x_i}.
\end{aligned}$$

(b) We may rewrite the equation above as

$$
\begin{aligned}
p(\mathbf{x}) =&(1-p)^{5n}(1-2p)^{2n} \exp\left\{\sum_{i=1}^{5} x_i \log\left(\frac{p}{1-p}\right) + \sum_{i=6}^{7} x_i \log\left(\frac{p}{1-2p}\right)\right\} \\
&\left\{\prod_{i=1}^{7}\binom{n}{x_i}\right\} 2^{\sum_{i=6}^{7} x_i}.
\end{aligned}
$$

This is on two-parameter exponential family form,

$$p(\mathbf{x}) = A(\theta)\exp\left\{\sum_{j=1}^{2}\zeta_j(\theta)T_j(\mathbf{x})\right\} h(\mathbf{x}),$$

with $\theta = p$, natural parameters $\zeta_1(p) = \log\left(\frac{p}{1-p}\right)$, $\zeta_2(p) = \log\left(\frac{p}{1-2p}\right)$ and sufficient statistics $T_1(\mathbf{x}) = \sum_{i=1}^{5} x_i$ and $T_2(\mathbf{x}) = \sum_{i=6}^{7} x_i$.

(c) Consider the sufficient statistics as random variables, $T_1(\mathbf{X}) = \sum_{i=1}^{5} X_i$ and $T_2(\mathbf{X}) = \sum_{i=6}^{7} X_i$.
Because the X_i are independent and $\mathrm{Var}(X_i) = np(1-p)$ for $i = 1, ..., 5$, it follows that $\mathrm{Var}\{T_1(\mathbf{X})\} = 5np(1-p)$.
For $i = 6, 7$, $\mathrm{Var}(X_i) = 2np(1-2p)$, and so, $\mathrm{Var}\{T_2(\mathbf{X})\} = 4np(1-2p)$.
Moreover, beacause of independence, $\mathrm{Cov}\{T_1(\mathbf{X}), T_2(\mathbf{X})\} = 0$.

(d) As was seen above, the sufficient statistics $T_1(\mathbf{X})$ and $T_2(\mathbf{X})$ are independent. To prove that we have a strictly 2-dimensional exponential family, it remains to verify that $1, \zeta_1(p), \zeta_2(p)$ are linearly independent, i.e. that there is no linear combination of them which is zero. If such a one should exist, then for all p,

$$a + b\log\left(\frac{p}{1-p}\right) + c\log\left(\frac{p}{1-2p}\right) = 0$$

for some constants a, b, c, not all zero. But this is impossible. (It is easily seen that for this to hold, different p require different sets of constants a, b, c.)
Hence, we have a strictly 2-dimensional exponential family.

4. (a) No, since the paramaters (a, h) enter the (log) probability function together with x in a non linear way.

   (b) Because $E(X) = a$, $\hat{a}_{MME} = \bar{x}$, the mean of the sample $\mathbf{x} = (x_1, ..., x_n)$. Moreover,
   $$E(X^2) = \text{Var}(X) + \{E(X)\}^2 = \frac{1}{5}h^2 + a^2,$$
   i.e.
   $$\frac{1}{n}\sum_{i=1}^{n} x_i^2 = \frac{1}{5}\hat{h}^2_{MME} + \hat{a}^2_{MME} = \frac{1}{5}\hat{h}^2_{MME} + \bar{x}^2,$$
   which yields
   $$\hat{h}^2_{MME} = 5\left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2\right) = \frac{5}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2,$$
   i.e.
   $$\hat{h}_{MME} = \sqrt{\frac{5}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

   (c) Yes, beacause $E(\bar{X}) = E(X) = a$.

   (d) We have $\text{Var}(\bar{X}) = \frac{1}{n}\text{Var}(X) = \frac{h^2}{5n}$.

   (e) No. It is not a function of a sufficient statistic. It is, however, unbiased, so by the Rao-Blackwell theorem, an estimator with smaller variance can be constructed from it as the conditional expectation given a sufficient statistic. The result is then a function of this sufficient statistic, so it can not equal $\hat{a}_{MME}$ with probability one. This means that the so constructed statistic has strictly smaller variance than $\hat{a}_{MME}$.

5. (a) Yes. The null model is $N(1, 1)$.

   (b) The statistic
   $$T(\mathbf{x}) = \frac{\bar{X} - \mu}{\mu/\sqrt{n}}$$
   seems appropriate. Other possible choices are $T_1(\mathbf{x}) = \bar{X} - \mu$ or $T_2(\mathbf{x}) = S^2 - \mu$, where S is the sample standard deviation.

   (c) The null distributions for the statistics in (b) are $N(0, 1)$ and $N(0, 1/n)$ for the first two. For the third, $Y = (n-1)S^2/\sigma^2$ is χ^2 with $n-1$ degrees of freedom, where in our case, $\sigma^2 = \mu$. Hence, under the null, $T_2(\bar{X})$ is distributed as $Y/(n-1) - 1$.

   (d) It is ambiguous to define the p value for a test using $T(\mathbf{x})$, since for $\mu < 1$, small values of T are extreme in the sense of the mean, but less extreme in the sense of the variance.
   For T_1, the p value is obtained from the fact that $\sqrt{n}T_1(\mathbf{X})$ is $N(0, 1)$ under H_0. Let $t_{obs} = \bar{x} - 1$. By definition, the p value is
   $$2\min\{P(T_1 < t_{obs}), P(T_1 > t_{obs})\} = 2\min\{\Phi(\sqrt{n}t_{obs}), \Phi(-\sqrt{n}t_{obs})\},$$
   where Φ is the standard normal distribution function.
   The p value for T_2 may be obtained analogously from the χ^2 distribution.

(e) Uniform on $(0,1)$, since it is always so for test statistics with continuous distributions.

(f) Strongly reject H_0. There is strong evidence in favour of H_1, i.e. that $\mu \neq 1$.

6. We complement the table with $p_0(x)/p_1(x)$:

x	2	3	4	5	6	7	8
$p_0(x)$	0.1	0.02	0.33	0.05	0.2	0.1	0.2
$p_1(x)$	0	0.03	0.17	0.3	0.2	0.2	0.1
$\frac{p_0(x)}{p_1(x)}$	∞	0.67	1.94	0.17	1	0.5	2

(a) The smallest possible value of $\frac{p_0(x)}{p_1(x)}$ is $\frac{p_0(5)}{p_1(5)} = 0.17$. Under H_0, this has probability $p_0(5) = 0.05$. The second smallest is $\frac{p_0(7)}{p_1(7)} = 0.5$, which under H_0 has probability $p_0(7) = 0.1$. To achieve the significance level $\alpha = 0.1$, we then reject with probability $1/2$ in the latter case. Hence, we get the test function

$$\varphi(x) = \begin{cases} 1 & \text{if } x = 5, \\ 1/2 & \text{if } x = 7, \\ 0 & \text{otherwise.} \end{cases}$$

We may check that the probability of error of the first type is

$$E_0\{\varphi(X)\} = p_0(5) + \frac{1}{2}p_0(7) = 0.05 + \frac{1}{2} * 0.1 = 0.05.$$

(b) The error of second type is not to reject when the alternative is true. The probability of *not* committing this error (the test power) is calculated under the p_1 distribution to be

$$E_1\{\varphi(X)\} = p_1(5) + \frac{1}{2}p_1(7) = 0.3 + \frac{1}{2} * 0.2 = 0.4.$$

hence, the probability of the error is $\beta = 1 - 0.4 = 0.6$.

(c) An immedeate alternative is the test that rejects if.f. $x = 7$. The probability of this to happen under the null is $p_0(7) = 0.1$.

Another (maybe less sensible) alternative is to reject if.f. $x = 2$.

(d) As we saw under (b), the power of the NP test is 0.4. The test with critical region $\{x = 7\}$ has power $p_1(7) = 0.2$, which is lower (as it should be). The (not so sensible) test with critical region $\{x = 2\}$ has power $p_1(2) = 0$.

7. (a) The error of the first type is to reject a true null hypothesis ($\mu = 1$ here).

   (b) The error of the second type is not to reject when the alternative hypothesis is true ($\mu = -1$ here).

   (c) Let the observations be $\mathbf{x} = (x_1, ..., x_n)$. The likelihood ratio is

$$\frac{p_0(\mathbf{x})}{p_1(\mathbf{x})} = \frac{(2\pi)^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i - 1)^2\right\}}{(2\pi)^{-n/2} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(x_i + 1)^2\right\}}$$

$$= \exp\left[-\frac{1}{2}\left\{\sum_{i=1}^{n}(x_i - 1)^2 - (x_i + 1)^2\right\}\right] = \exp\left(2\sum_{i=1}^{n} x_i\right).$$

   Hence, rejecting if the likelihood ratio is smaller than a constant is equivalent to rejecting if $\bar{x} < C$ for some constant C. This is the NP test.

   (d) Since $\sqrt{n}(\bar{X} - \mu) \sim N(0, 1)$, we get the probability of error of type I as

$$\alpha = P_{\mu=1}(\bar{X} < C) = P_{\mu=1}\{\sqrt{n}(\bar{X} - 1) < \sqrt{n}(C - 1)\} = \Phi\{\sqrt{n}(C - 1)\},$$

   where Φ is the standard normal distribution function.
   Similarly, the power function is

$$1-\beta = P_{\mu=-1}(\bar{X} < C) = P_{\mu=-1}\{\sqrt{n}(\bar{X}+1) < \sqrt{n}(C+1)\} = \Phi\{\sqrt{n}(C+1)\}.$$

   We may derive C explicitly from the first equation via $-\lambda_\alpha = \sqrt{n}(C-1)$, which yields $C = 1 - \lambda_\alpha/\sqrt{n}$. The second equation then yields

$$1 - \beta = \Phi(2\sqrt{n} - \lambda_\alpha).$$

   This gives the probability of error of the second type as

$$\beta = 1 - \Phi(2\sqrt{n} - \lambda_\alpha).$$

   We find that this is an increasing function of λ_α, hence it is decreasing in α. Greater α correspond to smaller β and vice versa.

   (e) A plot of $(\alpha, \beta) = [\Phi\{\sqrt{n}(C - 1)\}, 1 - \Phi\{\sqrt{n}(C + 1)\}]$ where $n = 10$ and C runs through suitable real numbers is shown in figure 1.

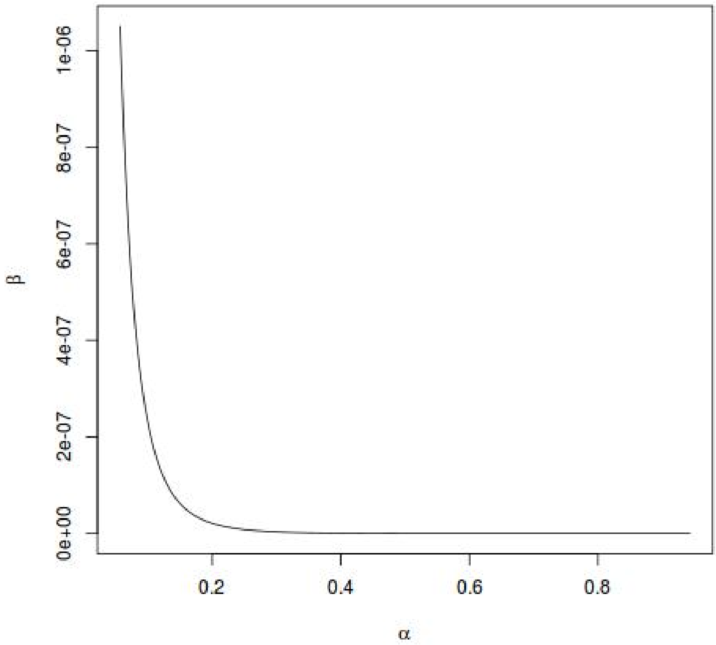

Figure 1: Plot of (α, β) for $n = 10$, problem 7.

8. Observe: the density should read

$$f(x;\alpha,\beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} I_{[0,1]}(x).$$

(a) Yes, as is seen by writing ($\mathbf{x} = (x_1, ..., x_n)$)

$$\begin{aligned} f(\mathbf{x};\alpha,\beta) &= \prod_{i=1}^{n} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} e^{\alpha\log(x_i)+\beta\log(1-x_i)} \frac{I_{[0,1]}(x_i)}{x_i(1-x_i)} \\ &= \left\{\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\right\}^n e^{\alpha\sum_{i=1}^n \log(x_i)+\beta\sum_{i=1}^n \log(1-x_i)} \prod_{i=1}^{n} \frac{I_{[0,1]}(x_i)}{x_i(1-x_i)}. \end{aligned}$$

By inspection, the natural parameters are α and β, and the corresponding sufficient statistics are $\sum_{i=1}^n \log(x_i)$ and $\sum_{i=1}^n \log(1-x_i)$.

(b) For a symmetric distribution, the mean equals the mode. Hence, in this case,

$$\frac{\alpha}{\alpha+\beta} = \frac{\alpha-1}{\alpha+\beta-2}.$$

If $\alpha+\beta \neq 2$, this is equivalent to $\alpha(\alpha+\beta) - 2\alpha = \alpha(\alpha+\beta) - (\alpha+\beta)$, i.e. $\alpha = \beta$. We then have

$$x^{\alpha-1}(1-x)^{\beta-1} = \{x(1-x)\}^{-1/2},$$

which is symmetric about $x = 1/2$, since the function $x(1-x)$ is. Since $I_{[0,1]}(x)$ is also symmetric about $x = 1/2$, the result follows.

(c) We may consider β as fixed. From the likelihood in the solution of (a), it is seen that we have an exponential family with monotone likelihood ratio. The sufficient statistic for α is $T(\mathbf{x}) = \sum_{i=1}^n \log(x_i)$. It follows from Blackwell's theorem that the UMP test is given by

$$\varphi(\mathbf{x}) = \begin{cases} 1 & \text{if } \sum_{i=1}^n \log(x_i) > C, \\ 0 & \text{if } \sum_{i=1}^n \log(x_i) \leq C, \end{cases}$$

i.e. we reject if $\sum_{i=1}^n \log(x_i) > C$ for some constant C. The value of this constant is given by $\alpha = P\{\sum_{i=1}^n \log(X_i) > C\}$.

(d) A UMP α-test is a size α test φ^* which has uniformly highest power among all size α tests, i.e. $E_\theta\{\varphi^*(\mathbf{X})\} \geq E_\theta\{\varphi(\mathbf{X})\}$ for all θ belonging to the alternative and all size α tests φ.