10-602: Statistical Approaches to Learning and Discovery

Problem Set 1 Solutions

---

1. (a) The variables x and y are independent, because their joint density

$$p(x,y) = \begin{cases} 1 & \text{if } x,y \in [0,1] \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

can be factored into $p(x)p(y)$, where $p(x)$ is uniform in $[0,1]$ and so is $p(y)$.

(b) The variables are dependent. Consider the circle centered at $(1/2,1/2)$ with unit diameter (a subset of the square above). If $y=1$, then x must be $1/2$ in order for (x,y) to be in the circle. But if $y=1/2$, then x can be anywhere in $[0,1]$. Thus x depends on y. This can also be seen geometrically by taking horizontal slices of the circle and seeing that they are not equivalent.

(c) The variables are dependent. Again, consider extreme cases. If $x=2$, then we must have $a=b=1$, so that $y=0$. But if $x=1$ then y can have any value in $[-1,1]$. Thus y depends on x.

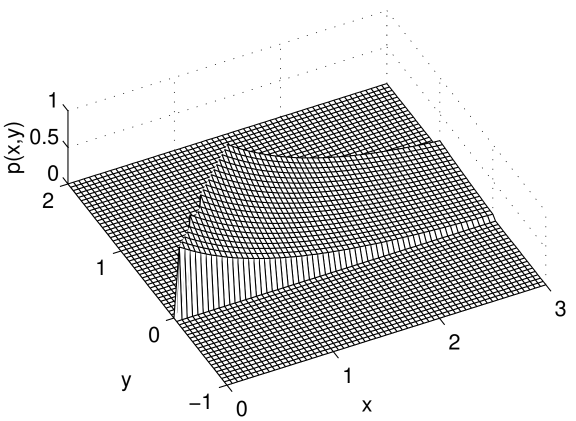
(d) The variables are independent. Since the transformation from (a,b) to (x,y) is linear, x and y are jointly Gaussian. Their covariance is

$$\begin{aligned} \text{cov}(x,y) &= E[xy] - E[x]E[y] && (2) \\ &= E[(a+b)(a-b)] - 0 && (3) \\ &= E[a^2] - E[b^2] = 0 && (4) \end{aligned}$$

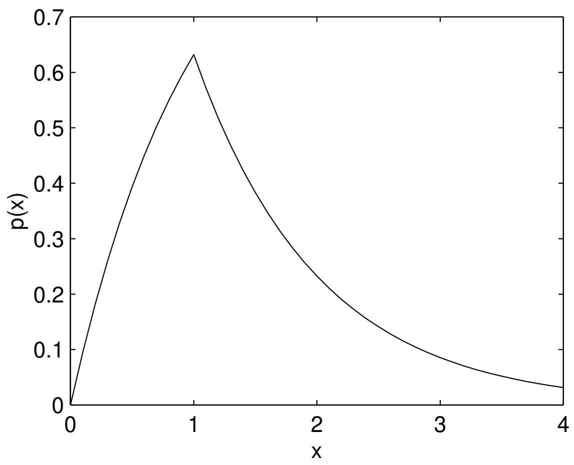
Jointly Gaussian variables with zero covariance must be independent.

2. (a)

$$p(x,y) = \begin{cases} e^{-(x-y)} & 0 \le y \le 1 \text{ and } y \le x \\ 0 & \text{otherwise} \end{cases} \tag{5}$$



(b)



(c)

$$\text{If } x < 1\text{: } p(x) = \int_0^x e^{-(x-y)} dy = 1 - e^{-x} \tag{6}$$

$$\text{If } x \geq 1\text{: } p(x) = \int_0^1 e^{-(x-y)} dy = e^{-x}(e-1) \tag{7}$$

3. (a) There are several ways to derive this. One approach is to consider a specific sequence of balls and then generalize. Consider the sequence 10011, where 1 means yellow. This has probability

$$\frac{Y}{N}\frac{(N-Y)}{(N-1)}\frac{(N-Y-1)}{(N-2)}\frac{(Y-1)}{(N-3)}\frac{(Y-2)}{(N-4)} = \frac{Y(Y-1)(Y-2)(N-Y)(N-Y-1)}{N(N-1)(N-2)(N-3)(N-4)} \tag{8}$$

From this it is immediate that any sequence with y yellow balls will have probability

$$\frac{Y(Y-1)\cdots(Y-y+1)\ \ (N-Y)(N-Y-1)\cdots(N-Y-(n-y)+1)}{N(N-1)\cdots(N-n+1)}$$

$$= \frac{Y!/(Y-y)!\ \ (N-Y)!/(N-Y-(n-y))!}{N!/(N-n)!} = \frac{\binom{N-n}{Y-y}}{\binom{N}{Y}} \tag{9}$$

Since there are $\binom{n}{y}$ such sequences, the probability of getting y yellow balls is thus

$$p(y|n,N,Y) = \frac{\binom{n}{y}\binom{N-n}{Y-y}}{\binom{N}{Y}} \text{ or } \frac{\binom{Y}{y}\binom{N-Y}{n-y}}{\binom{N}{n}} \tag{10}$$

When $Y = 1$, the probability that $y = 1$ should intuitively be n/N, and this is supported by the formula. Another way to prove this formula is to consider the urn as a sequence of N balls from which we take the first n. Given y, there are $\binom{n}{y}$ distinguishable sequences for the balls taken out of the urn and $\binom{N-n}{Y-y}$ distinguishable sequences for the balls left in the urn, out of $\binom{N}{Y}$ distinguishable sequences to begin with.

(b) As shown above, the probability of any sequence of n balls depends only on y. Thus y is sufficient for p.

(c) Let $y_i = 1$ if ball i is yellow, and $y_i = 0$ otherwise. Then $y = \sum_i y_i$. An important property of the y_i is that they are exchangeable. From this we find that $E[y_i] = p$ and $\text{var}(y_i) = p(1-p)$. (This is obvious for the first ball, and since the balls are exchangeable it must be true for all of them.) What is the covariance between y_i and y_j? Again, it suffices to consider y_1 and y_2:

$$E[y_1 y_2] = \Pr(y_1 = 1, y_2 = 1) = \frac{Y}{N}\frac{Y-1}{N-1} \tag{11}$$

$$\text{cov}(y_i, y_j) = E[y_i y_j] - E[y_i]E[y_j] = -\frac{p(1-p)}{N-1} \tag{12}$$

Now we can deduce the mean and variance of y:

$$E[y] = \sum_i E[y_i] = np \tag{13}$$

$$\text{var}(y) = E[\left(\sum_i (y_i - E[y_i])\right)^2] = \sum_i \text{var}(y_i) + \sum_{i \neq j} \text{cov}(y_i, y_j) \tag{14}$$

$$= np(1-p)\frac{N-n}{N-1} \tag{15}$$

Finally,

$$E[\bar{p}] = p \qquad (\bar{p} \text{ is unbiased}) \tag{16}$$

$$E[(\bar{p}-p)^2] = \mathrm{var}(\bar{p}) = \frac{p(1-p)}{n}\frac{N-n}{N-1} \tag{17}$$

(d) In this problem it is important to get the representation right so that the calculations are easy. Using the y_i's, we can write

$$
\begin{aligned}
p^* &= \frac{1+\sum_{i=2}^n y_i}{n} && (18)\\
E[p^*] &= \frac{1+(n-1)p}{n} = \frac{1}{n} + \frac{n-1}{n}p \neq p \qquad (p^* \text{ is biased}) && (19)\\
\mathrm{var}\Big(\sum_{i=2}^n y_i\Big) &= (n-1)p(1-p)\frac{N-n+1}{N-1} \qquad (\text{from last part}) && (20)\\
\mathrm{var}(p^*) &= \frac{(n-1)p(1-p)}{n^2}\frac{N-n+1}{N-1} && (21)\\
E[(p^*-p)^2] &= \mathrm{var}(p^*) + (E[p^*]-p)^2 && (22)\\
&= \frac{(n-1)p(1-p)}{n^2}\frac{N-n+1}{N-1} + \frac{(1-p)^2}{n^2} && (23)
\end{aligned}
$$

The mean square error is not always worse than that for $\bar{p}$—see part (e). Another way of representing p^* is with two separate cases:

$$p^* = \begin{cases} \bar{p} & \text{if the first ball was yellow} \\ (y+1)/n = \bar{p} + 1/n & \text{if the first ball was not yellow} \end{cases} \tag{24}$$

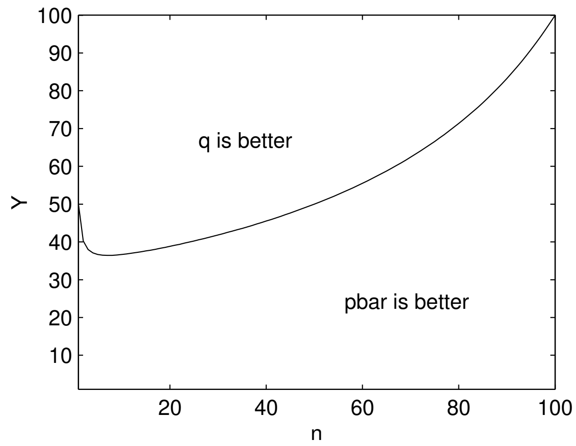
This easily leads to the same expression for $E[p^*]$, but calculating $\mathrm{var}(p^*)$ takes some care. It is not correct to simply add the variances of the two cases (or add weighted variances). The mere fact that p^* is composed of two cases adds between-case variance on top of the within-case variance.

(e) Given that y out of the n balls drawn are yellow, the marginal probability that any of the drawn balls is yellow is $y/n = \bar{p}$. Thus

$$
\begin{aligned}
E[y_i|y] &= \bar{p} && (25)\\
q &= E[p^*|y] = \frac{1}{n} + \frac{n-1}{n}\bar{p} && (26)\\
E[q] &= \frac{1}{n} + \frac{n-1}{n}p \qquad (\text{same as } E[p^*]) && (27)\\
\mathrm{var}(q) &= \left(\frac{n-1}{n}\right)^2 \mathrm{var}(\bar{p}) && (28)\\
E[(q-p)^2] &= \mathrm{var}(q) + (E[q]-p)^2 && (29)\\
&= \left(\frac{n-1}{n}\right)^2 \mathrm{var}(\bar{p}) + \frac{(1-p)^2}{n^2} && (30)
\end{aligned}
$$

The MSE of q is less than the MSE of p^*, which is what the Rao-Blackwell theorem tells us. Compared to $\bar{p}$, the MSE of q is lower whenever p is close enough to 1. This makes sense since q is biased toward 1. For example, when $n = 1$, $MSE(\bar{p}) = p(1-p)$ and

$MSE(q) = MSE(p^*) = (1-p)^2$, making $MSE(q) < MSE(\bar{p})$ when $p > 1/2$. The cutoff for other values of n is summarized in the following plot (for $N = 100$):



4. (a) Let $D = \{Y_1, ..., Y_n\}$. The likelihood function is

$$L(D|\theta_1, \theta_2) = \begin{cases} \frac{1}{(\theta_2-\theta_1)^n} & \text{if } \theta_1 < \min(D) \text{ and } \theta_2 > \max(D) \\ 0 & \text{otherwise} \end{cases} \tag{31}$$

or more formally

$$L(D|\theta_1, \theta_2) = \frac{\mathrm{lt}(\theta_1, \min(D))\mathrm{lt}(\max(D), \theta_2)}{(\theta_2 - \theta_1)^n} \tag{32}$$

where

$$\mathrm{lt}(y, z) = \begin{cases} 1 & \text{if } y < z \\ 0 & \text{otherwise} \end{cases} \tag{33}$$

The MLE for θ_1 is $\min(D)$ and the MLE for θ_2 is $\max(D)$. These are sufficient statistics since the likelihood function depends only on them. To prove they are minimal, notice that the ratio of the likelihood function for two sets of data is constant as a function of θ_1 and θ_2 only if the min and the max are the same for the two data sets. It then follows by the Lehmann-Scheffé Theorem that these are minimal sufficient statistics.

(b) Now the likelihood function is

$$L(D|\theta_1) = \begin{cases} \frac{1}{a^n} & \text{if } \theta_1 < \min(D) \text{ and } \theta_1 > \max(D) - a \\ 0 & \text{otherwise} \end{cases} \tag{34}$$

Even though there is one parameter, there is no single sufficient statistic. The minimal set of sufficient statistics is $(\min(D), \max(D))$. This is related to the fact that the MLE for θ_1 is not unique: all values between $\max(D) - a$ and $\min(D)$ have equal likelihood.

This problem illustrates that an MLE is not always a sufficient statistic. "Sufficient" does not mean "can be used to define an estimator." Sufficiency is a statement about all possible estimators—that there is nothing else the data can tell you about a. In this problem, the statistic $\min(D)$ alone does not tell you everything about a that you could learn from the data.

5. (a) In exponential family form, the distribution is

$$p(y|a, b) = \frac{b^a}{\Gamma(a)y} \exp(-by + a\log(y)) \tag{35}$$

so that the sufficient statistics are $d_1(y) = y$ and $d_2(y) = \log(y)$. The maximum-likelihood equations must be of the form $E[d_j(y)] = \frac{1}{n}\sum_i d_j(y_i)$, so by inspection of the likelihood:

$$E[y] = \frac{a}{b} = \frac{1}{n}\sum_i y_i \tag{36}$$

$$E[\log(y)] = \Psi(a) - \log(b) = \frac{1}{n}\sum_i \log(y_i) \tag{37}$$

You can also take the derivatives of the log-likelihood to get these equations, but that would not be a solution "by inspection."

(b) Following the hint, take derivatives to get

$$\frac{\int_0^\infty \log(y) y^{a-1} \exp(-by)dy}{\int_0^\infty y^{a-1} \exp(-by)dy} = \Psi(a) - \log(b) \tag{38}$$

$$E[\log(y)] = \Psi(a) - \log(b) \tag{39}$$

This is a handy trick in statistics.