# HW1_Solution

*Sara Taheri*

*1/17/2017*

## Problem 1

Childhood lead poisoning is a public health concern in the United States. In a certain population, one child in 8 has a high blood lead level. In a randomly selected group of 16 children from the population, provide the probability for the following statements. Show your calculations.

This is a binomial distribution with the following parameters:

$p$ = The probability of having a high blood lead level = $1/8 = 0.125$,

$1 - p$ = The probability of not having a high blood lead level = $1 - 1/8 = 0.875$,

$n = 16$ = The number of children,

$k$ = The number of children with high blood lead level

a) None has high blood lead level?

$$p(k = 0) = \binom{16}{0}(0.125)^0(0.875)^{16} = 0.1180671$$

b) One child has high blood lead level?

$$p(k = 1) = \binom{16}{1}(0.125)^1(0.875)^{15} = 0.2698676$$

c) Two children have high blood lead level?

$$p(k = 2) = \binom{16}{2}(0.125)^2(0.875)^{14} = 0.2891438$$

d) Three or more have high blood lead level?

in this case, $k = 3, or 4, or 5, ..., or 16$, so we have to calculate $p(k = 3) + p(k = 4) + p(k = 5) + ... + p(k = 16)$ which is hard to compute. However this is equivalent to $1 - (p(k = 0) + p(k = 1) + p(k = 2))$. So we have,

$$1 - (p(k = 0) + p(k = 1) + p(k = 2)) = 1 - (0.1180671 + 0.2698676 + 0.2891438) = 0.3229215$$

---

## Problem 2

The hight of corn plants follows a Normal distribution with mean 145cm and standard deviation 22cm.

a) What percentage of plants are between 135cm and 155cm tall?

Assume X is a random variable with a Normal distribution with mean 145cm and standard deviation 22cm. $X \sim \mathcal{N}(145, 22^2)$. We know that $Z = \frac{X-145}{22} \sim \mathcal{N}(0, 1)$. So,

$$\begin{aligned} P(135 < X < 155) &= P(\frac{135-145}{22} < Z < \frac{155-145}{22}) = P(-0.45 < Z < 0.45) \\ &= P(Z < 0.45) - P(Z < -0.45) = 0.6736448 - 0.3263552 = 0.3472896 \end{aligned}$$

So, around 34.72% of the plants have a hight between 135cm and 155cm.

Note that I calculate the amount of $P(Z < 0.45)$ and $P(Z < -0.45)$ with R. Here is the code and the result:

```
#P(Z < 0.45)
pnorm(0.45)
```

```
## [1] 0.6736448
```

```
#P(Z < -0.45)
pnorm(-0.45)
```

```
## [1] 0.3263552
```

b) If we choose a random sample of 16 plants, what is the probability that the mean hight would be between 135cm and 155cm tall?

Random samples of size $n$ are repeatedly drawn from a fixed population with mean $\mu$ and standard deviation $\sigma$. Each sample has it's own mean $\bar{y}$. The variation of the $\bar{y}$ is specified by the sampling distribution of $\bar{Y}$. On average the sample mean equals the population mean. The average of sampling distribution of $\bar{Y}$ is $\mu$. The standard deviation of $\bar{Y}$ is equal the standard deviation of $Y$ devided by square root of the sample size. so,

$$\sigma_{\bar{Y}} = sd(\bar{Y}) = \frac{\sigma}{\sqrt{n}}$$

In our problem, $n = 16$, so $\sigma_{\bar{Y}} = \frac{22}{\sqrt{16}} = 5.5$

$$\begin{aligned} P(135 < \mu\bar{Y} < 155) &= P(\frac{135-145}{5.5} < Z < \frac{155-145}{5.5}) = P(-1.82 < Z < 1.82) \\ &= P(Z < 1.82) - P(Z < -1.82) = 0.9312 \end{aligned}$$

c) If we choose a random sample of 32 plants, what is the probability that the mean hight would be between 135cm and 155cm tall?

$$\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}} = \frac{22}{\sqrt{32}} = 3.889087$$

$$\begin{aligned} P(135 < \mu\bar{Y} < 155) &= P(\frac{135-145}{3.89} < Z < \frac{155-145}{3.89}) = P(-2.57 < Z < 2.57) \\ &= P(Z < 2.57) - P(Z < -2.57) = 0.9949151 - 0.005084926 = 0.99 \end{aligned}$$

---

## Problem 3

We know that two events $A$ and $B$ are independent if $P(A,B) = P(A)P(B)$. We have to calculate the product of $P(RightHand)$ and $P(RightFoot)$ and see if it is equal to $P(RightHand, RightFoot)$.

$$\begin{aligned} P(RightHand, RightFoot) &= \frac{2012}{2391} = 0.8415 \\ P(RightHand) &= \frac{2012+142}{2391} = 0.9 \\ P(RightFoot) &= \frac{2012+121}{2391} = 0.89 \end{aligned}$$

obviously, $P(RightHand, RightFoot)$ is not equal to the product of $P(RightHand)$ and $P(RightFoot)$. We conclude that hand and foot preferences are not independent.

---

## Exercises from K. Murphy:

### Exercise 2.1 from K.Murphy

a) Let $X$ be one child and $Y$ be the other child. Then the event space is:

| $XY$ | $Prob.$ |
|---|---|
| $GG$ | $1/4$ |
| $GB$ | $1/4$ |
| $BG$ | $1/4$ |
| $BB$ | $1/4$ |

Let $N_g$ be the number of girls and $N_b$ the number of boys. We have the constraint (side information) that $N_b + N_g = 2$ and $0 \leq N_b, N_g \leq 2$. We are told $N_b \geq 1$ and asked to compute the probability of the event $N_g = 1$ (i.e., one child is a girl). By Bayes rule we have,

$$\begin{aligned} p(N_g = 1 | N_b \geq 1) &= \frac{p(N_b \geq 1 | N_g = 1) p(N_g = 1)}{p(N_b \geq 1)} \\ &= \frac{1 \times 1/2}{3/4} = 2/3 \end{aligned}$$

b) Let $Y$ be the identity of the obserevd child and $X$ be the identity of the other child. We want $p(X = g | Y = b)$. By Bayes rule we have,

$$\begin{aligned} p(X = g | y = b) &= \frac{p(Y = b | X = g) p(X = g)}{p(Y = b)} \\ &= \frac{(1/2) \times (1/2)}{1/2} = 1/2 \end{aligned}$$

This seems like a paradox because it seems that in both cases we could condition on the fact that "at least one child is a boy". But that is not correct; you must condition on the event actually observed, not its

logical implications. In the first case, the event was "He said yes to my question." In the second case, the event was "One child appeared in front of me". The generating distribution is different for the two events. Probabilities reflect the number of possible ways an event can happen, like the number of roads to a town. Logical implications are further down the road and may be reached in more ways, through different towns. The different number of ways change the probability.

---

**Exercise 2.2 from K.Murphy** legal reasoning

Let $E$ be the evidence (the obsereved blood type), and $I$ be the event that the defendant is innocent, and $G = \neg I$ be the event that the defendant is guilty.

a) The prosecutor is confusing $p(E|I)$ with $p(I|E)$. We are told that $p(E|I) = 0.01$ but the relevant quantity is $p(I|E)$. By Bayes rule, this is,

$$p(I|E) = \frac{p(E|I)p(I)}{p(E|I)p(I) + p(E|G)p(G)} = \frac{0.01p(I)}{0.01p(I) + (1 - p(I))}$$

Since $p(E|G) = 1$ and $p(G) = 1 - p(I)$. So we cannot determine $p(I|E)$ withouth knowing the prior probability $p(I)$. So $p(E|I) = p(I|E)$ only if $p(G) = p(I) = 0.5$, which is hardly a presumption of innocence.

b) The defender is quoting $p(G|E)$ while ignoring $p(G)$. The prior odds are,

$$\frac{p(G)}{P(I)} = \frac{1}{799,999}$$

The posterior odds are,

$$\frac{G|E}{p(I|E)} = \frac{1}{7999}$$

So the evidence has increased the odds of guilt by a factor of 1000. This is clearly relevant, although perhaps still not enough to find the suspect guilty.

---

**Excercise 2.4 from K.Murphy** Bayes rule for medical diagnosis

Let $T = 1$ represents a positive test outcome, $T = 0$ represents a negative test outcome, $D = 1$ mean you have the disease, and $D = 0$ mean you don't have the disease. We are told,

$$\begin{aligned} P(T = 1|D = 1) &= 0.99 \\ P(T = 0|D = 0) &= 0.99 \\ P(D = 1) &= 0.0001 \end{aligned}$$

We are asked to compute $P(D = 1|T = 1)$, which we can do using Bayes' rule:

$$\begin{aligned} P(D = 1|T = 1) &= \frac{P(T = 1|D = 1)P(D = 1)}{P(T = 1|D = 1)P(D = 1) + P(T = 1|D = 0)P(D = 0)} \\ &= \frac{0.99 \times 0.0001}{0.99 \times 0.0001 + 0.01 \times 0.9999} = 0.009804 \end{aligned}$$

So although you are much more likely to have the disease (given that you have tested positive) than a random member of the population, you are still unlikely to have it.

---

**Exercise 2.5 from K.Murphy** The Monty Hall problem

Let $H_i$ denote the hypothesis that the prize is behind door $i$. We make the following assumption: the three hypotheses $H_1$, $H_2$ and $H_3$ are equiprobable a priori, i.e.

$$P(H_1) = P(H_2) = P(H_3) = \frac{1}{3}$$

The datum we receive, after choosing door 1, is one of $D = 3$ and $D = 2$ (meaning door 3 or 2 is opened, respectively). We assume that these two possible outcomes have the following probabilities. If the prize is behind door 1 then the host has a free choice; in this case we assume that the host selects at random between $D = 2$ and $D = 3$. Otherwise the choice of the host is forced and the probabilities are 0 and 1.

$$P(D = 2|H_1) = \frac{1}{2}, P(D = 2|H_2) = 0, P(D = 2|H_3) = 1$$
$$P(D = 3|H_1) = \frac{1}{2}, P(D = 3|H_2) = 1, P(D = 3|H_3) = 0$$

Now, using Bayes theorem, we evaluate the posterior probabilities of the hypotheses:

$$P(H_i|D = 3) = \frac{P(D = 3|H_i)P(H_i)}{P(D = 3)}$$
$$P(H_1|D = 3) = \frac{(1/2)(1/3)}{P(D = 3)}, P(H_2|D = 3) = \frac{(1)(1/3)}{P(D = 3)}, P(H_3|D = 3) = \frac{(0)(1/3)}{P(D = 3)}$$

The denominator $P(D = 3)$ is $1/2$ because it is the normalizing constant for this posterior distribution. So

$$P(H_1|D = 3) = \frac{1}{3}, P(H_2|D = 3) = \frac{2}{3}, P(H_3|D = 3) = 0$$

So the contestant should switch to door 2 in order to have the biggest chance of getting the prize.

---

**Exercise 2.12 from K.Murphy** Show that $I(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$

$$\begin{aligned}
I(X,Y) &= \sum_{x,y} p(x,y) log \frac{p(x,y)}{p(x)p(y)} \\
&= \sum_{x,y} p(x,y) log \frac{p(x|y)}{p(x)} \\
&= -\sum_{x,y} p(x,y) logp(x) + \sum_{x,y} p(x,y) logp(x|y) \\
&= -\sum_{x} p(x) logp(x) - (-\sum_{x,y} p(x,y) logp(x|y)) \\
&= -\sum_{x} p(x) logp(x) - (-\sum_{y} p(y) \sum_{x} p(x|y) logp(x|y))) \\
&= H(X) - H(X|Y)
\end{aligned}$$

We can show $I(X, Y) = H(Y) - H(Y|X)$ by symmetry.