FIT5197 2018 S1 Assignment 3 Solutions

Kelvin Li and Wray Buntine

June 4, 2018

1 Week 2

Question: Develop the corresponding formula for $p(A \cup B \cup C)$ and $p(A \cup B \cup C \cup D)$ using $p(A)$, $p(B)$ etc. and the various 2, 3 and 4 place conjunctions $p(A \cap C)$, $p(A \cap B \cap C)$, etc.

Solution: You can do it by evaluating formula or by just working out pluses and minuses.

$$p(A \cup B \cup C) = p(A) + p(B) + p(C) + ...$$

so $p(A \cap B)$ occurs in both $p(A)$ and $p(B)$, so its double counted. So subtract one copy out, and likewise for the other doubles.

$$p(A \cup B \cup C) = p(A) + p(B) + p(C) - p(A \cap B) - p(A \cap C) - p(B \cap C) + ...$$

Now $p(A \cap B \cap C)$ occurs in all 6 terms so its cancelled out, so put it back.

$$p(A \cup B \cup C) = p(A) + p(B) + p(C) - p(A \cap B) - p(A \cap C) - p(B \cap C) + p(A \cap B \cap C)$$

Likewise for $p(A \cup B \cup C \cup D)$,

$$\begin{aligned} p(A \cup B \cup C \cup D) \quad = \quad & p(A) + p(B) + p(C) + p(D) \\ & - p(A \cap B) - p(A \cap C) - p(A \cap D) - p(B \cap C) - p(B \cap D) - p(C \cap D) \\ & + p(A \cap B \cap C) + p(A \cap B \cap D) + p(A \cap C \cap D) + p(B \cap C \cap D) \\ & - p(A \cap B \cap C \cap D) \end{aligned}$$

The other versions is to apply the formula twice. Consider the two place formula:

$$p(A \cup \mathcal{B}) = p(A) + p(\mathcal{B}) - p(A \cap \mathcal{B})$$

Replace $\mathcal{B}$ in this by $B \cup C$, so

$$p(A \cup B \cup C) = p(A) + p(B \cup C) - p(A \cap (B \cup C))$$

Now expand $p(B \cup C)$ with the two place formula and rearrange $p(A \cap (B \cup C))$

$$= p(A) + (p(B) + p(C) - p(B \cap C)) - (p((A \cap B) \cup (A \cap C)))$$

now apply the two place formula to $p((A \cap B) \cup (A \cap C))$

$$= p(A)+p(B)+p(C)-p(B \cup C)-(p(A \cap B) + p(A \cap C) - p((A \cap B) \cap (A \cap C)))$$

hence the result.

Question: Consider the triangular PDF defined for [-1,1]

$$p(x) = 1 - |x| \qquad \text{for } |x| < 1$$

and 0 otherwise. Find the functional form of the inverse CDF, $Q(p)$ What are the 10% and 75% quantiles.

Answer: first rewrite the function as

$$p(x) = \begin{cases} 1 + x & -1 < x < 0 \\ 1 - x & 0 \leq x < 1 \end{cases}$$

so do the integral to get the PDF

$$P(x) = \begin{cases} \frac{1}{2} + x + \frac{1}{2}x^2 & -1 < x < 0 \\ \frac{1}{2} + x - \frac{1}{2}x^2 & 0 \leq x < 1 \end{cases}$$

Set $p = P(x)$ and therefore solve $x^2 + 2x + (1 - 2p) = 0$ and $-x^2 + 2x + (1 - 2p) = 0$ using the quadratic formula to compute the inverse function

$$Q(p) = \begin{cases} -1 + \sqrt{2p} & 0 < p < \frac{1}{2} \\ 1 - \sqrt{2(1-p)} & \frac{1}{2} \leq p < 1 \end{cases}$$

The second half can be got using symmetry. Therefore $Q(0.1) = -1 + \sqrt{0.2}$ and $Q(0.75) = 1 - \sqrt{0.5}$.

2 Week 3

Question: You have K groups of items. Each group is equally likely, so $p(G = g) = 1/K$. The k-th group has n_k items that are equally likely. So $p(I|G = k) = 1/n_k$ for $k = 1, .., K$. Let G be the group you are in and let I be the item in the group. Compute $H(G)$ and $H(I|G)$ and hence $H(G, I)$.

Answer: $H(G) = \log_2 K$ and $H(I|G = k) = \log_2 n_k$. Therefore

$$H(I|G) = \sum_{k=1}^{K} p(G = k)H(I|G = k) = \frac{1}{K}\sum_{k=1}^{K} \log_2 n_k$$

Therefore

$$H(I, G) = \log_2 K + \frac{1}{K}\sum_{k=1}^{K} \log_2 n_k$$

Note some people might not realise n_k is indexed by k and think its a constant. In this interpretation

$$H(I, G) = \log_2 K + \log_2 n_k$$

Question: You have two Gaussians with identical variances s and diferent means m_1 and m_2. Let $p(x|N(m, s))$ denote the probability density function of the Gaussian $N(m, s)$. Define the probability density function

$$p(x|m_1, m_2, s, q) = qp(x|N(m_1, s)) + (1 - q)p(x|N(m_2, s))$$

Answer: Compute the mean and variance of x using expected value formula.

$$\begin{aligned}
\mathbb{E}[x] &= \int x\, p(x|m_1, m_2, s, q)\mathrm{d}\, x \\
&= \int x\, (qp(x|N(m_1, s)) + (1 - q)p(x|N(m_2, s)))\, \mathrm{d}\, x \\
&= qm_1 + (1 - q)m_2
\end{aligned}$$

Now for the simple Gaussian, $\mathbb{E}[x] = m^2 + s^2$, so

$$\begin{aligned}
\mathbb{E}\left[x^2\right] &= \int x^2\, p(x|m_1, m_2, s, q)\mathrm{d}\, x \\
&= \int x^2\, (qp(x|N(m_1, s)) + (1 - q)p(x|N(m_2, s)))\, \mathrm{d}\, x \\
&= q(m_1^2 + s^s) + (1 - q)(m_2^2 + s^2) \\
&= s^2 + qm_1^2 + (1 - q)m_2^2
\end{aligned}$$

Therefore the variance is

$$\begin{aligned}
\mathbb{E}\left[x^2\right] - \mathbb{E}\left[x\right]^2 &= s^2 + qm_1^2 + (1-q)m_2^2 - (qm_1 + (1-q)m_2)^2 \\
&= s^2 + qm_1^2 + (1-q)m_2^2 - q^2m_1^2 - (1-q)^2m_2^2 - 2q(1-q)m_1m_2 \\
&= s^2 + q(1-q)m_1^2 + q(1-q)m_2^2 - 2q(1-q)m_1m_2 \\
&= s^2 + q(1-q)(m_1 - m_2)^2
\end{aligned}$$

3 Week 4

Question: Consider a binomial distributon with $n = 500$ and $\theta = 0.001$. Use the appropriate CDF functions in R to compute $p(k < 10|n = 500, \theta = 0.001)$:

1. the exact value
2. the value according to the Gaussian approximation in lectures
3. the value according to the Poisson approximation in lectures
4. write down a consise formula for the exact value.

(NB. Wray made a mistake, meant to use $\theta = 0.01$)

Answer:

1. the exact value is `pbinom(9, size=500, prob=0.001)` which gives 1
2. so its Gaussian with mean $n\theta = 0.5$ and variance $n\theta(1-\theta) = 0.4995$, so use $Z = \frac{9.5-0.5}{\sqrt{0.4995}} = 12.73$, which is very high, so probability of 1 (**NB.** shouldn't use Gaussian approximation because $n\theta < 5$)
3. so its Poisson with mean $n\theta = 0.5$, so in R use `ppois(9,lambda=0.5)` which gives 1
4. the formula is
$$\sum_{k=0}^{9} \binom{500}{k} \theta^k (1-\theta)^{500-k}$$

Question: IQ is supposed to Gaussian with a mean of 100 and a standard deviaton of 15. At a high-school reunion, where everyone atends, 2 of your classmates out of 40 claim to have IQs greater than 150. What is the probability that 2 or more would have an IQ greater than 140. Represent your

solution as an expression of $\theta = p(IQ > 140)$ and give θ.

Answer: So $p(IQ > 140)$ using $IQ \sim N(100, 15)$ uses $Z = \frac{140-100}{15} = 2.666$ and gets (from `1-pnorm(2.66)`) which is $\theta = 0.003907033$ or near enough 0.004. The probability is

$$1-(1-\theta)^{40}-40(1-\theta)^{39}\theta = 1-0.8550579-0.1341536 = 0.0107885 \approx 0.011$$

4 Week 5

Question: You take measurements for a location in 3D space. Your measurement errors in each dimension, e_X, e_Y and e_Z are all Gaussian with zero mean and variance ϵ^2. The total squared error is

$$E = e_X^2 + e_Y^2 + e_Z^2$$

Describe the distribution of E. Hence, or otherwise, find the mean and variance of E? What is a 90% confdence interval for E as a factor of ϵ^2.

Answer: so e_X/ϵ, e_X/ϵ, and e_X/ϵ are all standard Normal and independent, so therefore E/ϵ^2 is χ^2 with 3 degrees of freedom. So to get the 90% confidence interval, we get the lower and upper 5% and 95% quantile of the χ^2 which for $df = 2$ are [0.10258,5.991465]. So the answer is $\epsilon^2[0.10258, 5.991465]$ or otherwise $[\epsilon^2\chi^2_{0.05,2}, \epsilon^2\chi^2_{0.95,2}]$.

One could also do a Gaussian approximation for the chisquared but it won't be very accurate.

5 Week 7

Question: In a famous experiment to determine the efficacy of aspirin in preventing heart attacks, 22,000 healthy middle-aged men were randomly divided into two equal groups, one of which was given a daily dose of aspirin and the other a placebo that looked and tasted identical to the aspirin. The experiment was halted at a time when 104 men in the aspirin group and 189 in the control group had had heart attacks. Use these data to test the hypothesis that the taking of aspirin does not change the probability of having a heart attack.

Answer: Group A has 104 failures out of 11000 and group B has 189 failures out of 11000. Let p_A and p_B be the population failure rates, with $n_A = n_B = 11000$. Null hypothesis has $p_A = p_B$ and alternative hypothesis is $p_A \neq p_B$. Since it is binomial data with difference between means, we will use the CLT. Let $\hat{p}_A = \frac{104}{11000}$ and $\hat{p}_A = \frac{189}{11000}$. So apply the CLT to get a hypothesis test. The estimated mean and variance of group A for large sample by CLT are $\hat{p}_A$ and $\hat{p}_A(1-\hat{p}_A)/n_A$ and for group B for are $\hat{p}_B$ and $\hat{p}_B(1-\hat{p}_B)/n_B$. Under the assumption $p_A = p_B$ we can pool the counts to get a pooled estimate of mean of $\hat{p}_{AB} = \frac{104+189}{22000}$ which means the Bernoulli has variance $\hat{p}_{AB}(1-\hat{p}_{AB})$. The difference between means hypothesis test for the Gaussian with known variance gives a Z score of

$$Z = \frac{(\hat{p}_A - \hat{p}_B) - 0}{\sqrt{\hat{p}_{AB}(1-\hat{p}_{AB})\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}}$$

The evaluates to $Z = -4.98$ which is clearly significant at a high level, so reject the null hypothesis. Note, if you didn't pool the variances, you would get

$$Z = \frac{(\hat{p}_A - \hat{p}_B) - 0}{\sqrt{\left(\frac{\hat{p}_A(1-\hat{p}_A)}{n_A} + \frac{\hat{p}_B(1-\hat{p}_B)}{n_B}\right)}}$$

and the results are similar.

Question: A sample of 10 fish were caught at lake A and their PCB concentrations were measured using a certain technique. The resulting data in parts per million were Lake A:

11.5, 10.8, 11.6, 9.4, 12.4, 11.4, 12.2, 11, 10.6, 10.8

In addition, a sample of 8 fish were caught at lake B and their levels of PCB were measured by a different technique than that used at lake A. The resultant data were Lake B:

11.8, 12.6, 12.2, 12.5, 11.7, 12.1, 10.4, 12.6

If it is known that the measuring technique used at lake A has a variance of .09 whereas the one used at lake B has a variance of .16, could you reject (at the 5 percent level of significance) a claim that the two lakes are equally contaminated?

Answer: Have μ_A and μ_B, and $\sigma_A^2 = 0.09$ and $\sigma_B^2 = 0.16$ are different. Null hypothesis has $\mu_A = \mu_B$. Hypothesis testing is done with difference between means with known variances case. Various estimates are $\bar{X}_A = 11.17$, $\bar{X}_B = 11.988$. By null hypothesis, Z given by

$$Z = \frac{(\bar{X}_A - \bar{X}_B) - 0}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}}$$

which evaluates to $Z = 4.803$. This is clearly significant at the 95% level.

6 Week 8

Question: 12 law students have heights (H) and salary (S) given. The means are: $\overline{H} = 70.33333$, $\overline{S} = 97.5$, $\overline{H^2} = 4962.333$, $\overline{S^2} = 9647.333$, $\overline{HS} = 6880.167$. Using the formulas for simple linear regression in the formula sheet we get regression co-efficients $b_0 = -4.827431$ and $b_1 = 1.4549$ giving the estimated formula

$$S = -4.827431 + 1.4549 * H$$

Now the RSS is computed as (for $n = 12$)

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} (S_i - \beta_0 - \beta_1 H_i)^2$$

This can be expanded as

$$RSS(\beta_0, \beta_1) = n\left(\overline{S^2} + \beta_0^2 + \beta_1^2\overline{H^2} - 2\beta_0\overline{S} + 2\beta_0\beta_1\overline{H} - 2\beta_1\overline{HS}\right)$$

So we get that $RSS = 12 * 108.0548$. We can do a hypothesis test that $b_1 = 0$. According to the formula sheet, we get the test statistic

$$t_{n-2} = \frac{1}{\sqrt{\frac{RSS}{n(n-2)}\frac{1}{\overline{X^2} - \overline{X}^2}}}\hat{\beta}_1$$

which is computed as $t_{10} = 1.745636$. So we accept the hypothesis that $b_1 = 0$ at any reasonable confidence level so H and S are not correlated.

Note, if your statistics was better, you could have done a simpler test directly on the sample correlation. This would be accepted as well and should lead to a similar t statistic.

7 Week 10

Question: You have a discrete distribution:

$$\begin{aligned} p(X{=}0) &= 0.5 \\ p(X{=}1) &= 0.3 \\ P(X{=}2) &= 0.2 \end{aligned}$$

- Develop an inverse transform sampler for this.
- Develop a rejection sampler for this using a uniform distribution as your proposal distribution.
- What proportion of proposed samples will be rejected?

Solution: The cummulative distribution is:

$$\begin{aligned} p(X{=}0) &= 0.5 \\ p(X{\leq}1) &= 0.8 \\ P(X{\leq}2) &= 1.0 \end{aligned}$$

The quantile distribution

$$q(p) = \begin{cases} 0 & p < 0.5 \\ 1 & 0.5 \leq p < 0.8 \\ 2 & 0.8 \leq p \leq 1.0 \end{cases}$$

The inverse sampler is defined as:

1. sample $U \sim \text{uniform}(0, 1)$
2. return $q(U)$

The uniform distribution is:

$$\begin{aligned} p_{prop}(X{=}0) &= \frac{1}{3} \\ p_{prop}(X{=}1) &= \frac{1}{3} \\ P_{prop}(X{=}2) &= \frac{1}{3} \end{aligned}$$

In the formula use $q(X) = p(X)$ and $C = \frac{2}{3}$. Then the condition is satisfied ($Cq(X) \leq p_{prop}(X)$). Thus the algorithm becomes

1. sample X from $1, 2, 3$ uniformly
2. sample $U \sim \text{uniform}(0, 1)$
3. if $X = 1$ accept; if $X = 2$ reject if $U > 0.6$; if $X = 3$ reject if $U > 0.4$.

8 Week 11

Question: First, complete the practical building a tree with the diabetes data. Now, for "diabetes_train.csv" compute the counts table for the Booleanised Y against Z for each of the cases:

- $Z1 \equiv S5 < 4.88275$
- $Z2 \equiv BP < 102.5$
- $Z3 \equiv BMI < 31.55$

Compute their entropy, $H(Y|Z)$ for the three cases. Hence, which is the best test to use at the root of the a tree?

Solution: The three tables are given below:

Y	$Z1$=0	$Z1$=1
Y=0	49	213
Y=1	64	28

$Z2$=0	$Z2$=1
45	217
53	39

$Z3$=0	$Z3$=1
15	247
37	55

The entropies (to base 2) required are

- $H(Y|Z1) = 0.667985$
- $H(Y|Z2) = 0.7207375$
- $H(Y|Z3) = 0.7114307$

These can be computed most easily using the difference formula $H(Y|Z) = H(Y, Z) - H(Z)$. So use $Z3$ as the root.