Math 318 – homework 6 solutions

Problem 1. Assume X_i are iid samples with $N(\mu, \sigma^2)$ distribution, with unknown μ, σ. We would like to test the null hypothesis $\mu = 200$. A sample of size 9 has sample mean $X = 205$.

(a) Find the p-value (probability of deviating at least that much from the mean)if the standard deviation is known to be
  (i) $\sigma = 5$
  (ii) $\sigma = 10$
  (iii) $\sigma = 15$

(b) In which of the three cases would the null hypothesis be rejected at the 5% level of significance?

(c) In which of the three cases would the null hypothesis be rejected at the 1% level of significance?

Solution.

(a) The distribution of the sample mean is $N(\mu, \sigma^2/n)$. If $\mu = 200$ then

$$\mathbb{P}(|\bar{X} - 200| \geq 5) = \mathbb{P}(|Z| \geq 5\sqrt{n}/\sigma) = 2\Phi(-5 \cdot 3/\sigma).$$

For the three cases this is $2\Phi(-3) = 0.0026$, $2\Phi(-1.5) = 0.1336$ and $2\Phi(1) = 0.617$.

(b) We reject the hypothesis with 95% confidence in the first case but not the others.

(c) Only in the first case.

Problem 2. ESP cards have one of five shapes on them (cross, circle, square, star and waves) with equal probability. A psychic claims he can guess a card correctly with probability 0.5. James Randy designs an experiment, where the psychic tries to guess N cards, and is considered successful if he made at least $N/3$ correct guesses.

(a) How large does N need to be so that the probability of success is at most 1/1000, if the psychic has no special ability (probability 1/5 at each guess)?

(b) For that N, what is the probability of failure if the psychic's claim is true?

Solution.

(a) The number X of correct guesses is $\text{Bin}(n, 1/5)$. Using the CLT, for a standard normal Z,

$$\mathbb{P}(X > n/3) \approx \mathbb{P}(Z > \frac{n/3 - n/5}{\sqrt{4/25n}}) = \mathbb{P}(Z > \frac{1}{3}\sqrt{n}).$$

We want this to be under 0.001. Since $\Phi(3.09) = 1 - 0.001$, we need $\sqrt{n}/3 \geq 3.09$, or $N \geq 86$. (Note: Requiring 3 standard deviations, giving $N = 81$ is reasonably accurate. Using binomials, the exact answer is 90.)

(b) Assuming the claim is true, X is $\text{Bin}(86, 1/2)$, so

$$\mathbb{P}(X < n/3) \approx \mathbb{P}(Z < \frac{n/3 - n/2}{\sqrt{n/4}}) = \mathbb{P}(Z < -\sqrt{n}/3) \approx 1/1000.$$

In fact, for every n, we get from the CLT that

$$\mathbb{P}(\text{Bin}(n, 1/5) > n/3) \approx \mathbb{P}(\text{Bin}(n, 1/2) < n/3).$$

(This is a coincidence for the numbers (1/5,1/3,1/2) used here.)

Problem 3. A bakery claims their bread weighs on average 1Kg.

(a) An investigator weighed 1000 loaves over a year, and summarized his results in a file `bread_data` (available on the course website). Based on this data, find a 95% confidence interval for μ.

(b) Based on this data, can we rule out the bakery's claim at a 5% level?

(c) If the mean weight actually is 1Kg, what is the likelyhood of us reaching that decision?

(d) If 50 investigators repeated this procedure, how many of them on average would reject the bakery's claim?

Solution.

(a) The file has 1000 samples with sample mean 0.9993 and sample variance $S^2 \approx 0.01^2$. Therefore the 95% confidence interval is $\bar{X} \pm 1.96S/\sqrt{n} = (0.9987, 0.99995)$.

(b) This interval does not contain 1, so we reject the claim.

(c) The likelyhood of rejecting a true claim when using a 95% confidence interval is 5% by definition.

(d) On average 5% of the investigators will claim the Bakery is cheating with 95% confidence (so $EX = 2.5$).

Problem 4. Let (X, Y) be a point in the rectangle $0 \le x \le 2$ and $0 \le y \le 1$, with joint pdf $f(x, y) = \frac{xy+1}{3}$.

(a) Find the conditional pmf $f_{X|Y}(x|y)$ and $f_{Y|X}(y|x)$.

(b) Find the conditional expectations $E[X|Y]$ and $E[Y|X]$.

Solution. We first calculate for $x \in [0, 2]$ and $y \in [0, 1]$:

$$f_x(x) = \int_0^1 \frac{xy+1}{3} dy = \tfrac{x+2}{6} \qquad\qquad f_y(y) = \int_0^2 \frac{xy+1}{3} dx = \frac{2y+2}{3}$$

(a) We have

$$f_{X|Y}(x|y) = \frac{f(x,y)}{f_Y(y)} = \frac{(xy+1)/3}{(2y+2)/3)} = \frac{xy+1}{2y+2},$$
$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)} = \frac{(xy+1)/3}{(x+2)/6)} = \frac{2xy+2}{x+2}.$$

(b) We have

$$E[X|Y] = \int_0^2 x f(x|y) dx = \int_0^2 \frac{x(xy+1)}{2y+2} dx = \frac{4y+3}{3y+3}. f_{Y|X}(y|x) \quad = \int_0^1 y f(y|x) dy = \int_0^1 \frac{2xy^2+2y}{x+2} = \frac{2x+3}{6x+12}.$$

Problem 5. Let X be uniform from the set $\{4, 6, 8, 12\}$. Given X, we take an X-sided die, and toss it X times. Let Y be the sum of these dice.

(a) Find $E[Y|X]$. (Hint: the expected result on an n-sided die is $\frac{n+1}{2}$.)

(b) Use that to find $E[Y]$.

Solution.

(a) Given X, each of the dice is uniform on $\{1, \ldots, X\}$, so has expected value $\frac{X+1}{2}$. Since there are X of them, the expected total is $\mathbb{E}[Y|X] = X\frac{X+1}{2}$.

(b) We have

$$E[Y] = \sum p(x) E[Y|X = x] = \sum_x \frac{1}{4} \frac{x^2+x}{2},$$

where the sum is over x in $\{4, 6, 8, 12\}$. This is

$$E[Y] = \frac{1}{4}(10 + 21 + 36 + 78).$$

Problem 6. Consider the standard normal probability density function $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$. It is known that there is no closed form for the antiderivative of this function, i.e., for the c.d.f. Φ of the standard normal. However, the c.d.f. can be approximated accurately, and the tables give

$$\int_0^1 f(x)\, dx = \Phi(1) - \Phi(0) \approx 0.3413.$$

(a) To demonstrate the method of Monte Carlo integration, approximate the integral $\int_0^1 f(x)\,dx$ by generating 10000 i.i.d. uniform random numbers on $[0,1]$ and computing the approximation

$$I_{10000} = \frac{f(U_1) + f(U_2) + \cdots + f(U_{10000})}{10000}.$$

Do this 100 times and record the results from each run.
Compute the average result (sample mean) and the sample variance of the 100 results, and compare thpare them to their expectation. Submit your code and your output.

(b) Another method for approximating this integral is to recall that for $f \geq 0$, the integral $\int_0^1 f(x)\,dx$ represents the area underneath the graph of f from $x = 0$ to $x = 1$. To estimate this area, one could simulate a large number of uniform points in the *square* with corners at $(0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$; then, find the proportion of points that lie underneath the curve $y = f(x)$. Give an argument (a formal proof is not required, just a motivation) as to why this is a reasonable way to approximate this area.

(c) Perform the approximation in (b) by writing code to simulate 10000 i.i.d. uniform random numbers in the square $[0,1] \times [0,1]$ and determine the proportion of them falling in the region $y \leq f(x)$. Do this 100 times and record the sample mean and sample variance of the results. Submit your code and your output.

(d) Is one of the two methods better? Which and why?

Solution.

(a) See notebook. Note that for the sample variance we divide by $N - 1$, not by N (make sure to do this is using `np.var`.

(b) The probability that a point falls under the curve is the area. The law of large numbers says in the long run the fraction will be close to the area.

(c) See notebook.

(d) The first method has a much smaller sample variance (by a factor of 100), meaning it is closer to the actual number. This is since the second method adds extra randomness in the y-coordinate which is not needed if the unction is known.