Practice questions

1. The Department of Transportation reports the following numbers of accidents in different days of the week:

M	T	W	T	F	S	S
35	23	29	31	34	60	25

   (a) You suspect that the chance of an accident depends on the day of the week. State the null hypothesis and calculate the p-value for your (alternative) hypothesis.

   Solution: We model the distribution of accidents among days of the week as tosses of a 7-sided die with probability p_i for the accident occuring on day i. The null hypothesis is that the die is fair, i.e., $p_1 = \cdots = p_7 = 1/7$. The alternative hypothesis is that it is not. The total number of samples is $n = 35 + 23 + \cdots + 25 = 237$. The chi-squared statistic is

$$\mathcal{X}^2 = \frac{(35 - 237/7)^2}{237/7} + \cdots + \frac{(25 - 237/7)^2}{237/7} \approx 26.941,$$

   giving a p-value of $\mathrm{P}(\chi^2(6) \geq 26.941) \approx 1.5 \cdot 10^{-4}$, indicating strong support for the alternative hypothesis.

   (b) You suspect that the chance of an accident is different on weekdays and weekends. What is the p-value now?

   Solution: Out of the 237 cases, 152 occurred on weekdays and 85 occured on weekends. If the chance of accidents was equally likely we would have expected 5/7 of the accidents to occur on weekdays and 2/7 to occur on weekends. The chi-squared statistic is

$$X^2 = \frac{(152 - 237 \cdot \frac{5}{7})^2}{237 \cdot \frac{5}{7}} + \frac{(85 - 237 \cdot \frac{2}{7})^2}{237 \cdot \frac{2}{7}} \approx 6.178.$$

   The p-value is $\mathrm{P}(\chi^2(1) \geq 6.178) \approx 0.013$ so there is again strong evidence against the null hypothesis, but less so than in part (a).

2. You observe the following sorted sequence of samples of a continuous random variable, which is hypothesized by default to be Exponential(1).

0.013	0.018	0.066	0.086	0.136	0.138	0.172
0.311	0.321	0.654	0.828	1.060	1.326	1.373
1.682	1.860	2.232	3.191	3.715	3.720	5.780

   (a) How should you partition the range of an Exponential(1) random variable X into three intervals I_1, I_2, I_3 so that $\mathrm{P}(X \in I_1) = \mathrm{P}(X \in I_2) = \mathrm{P}(X \in I_3) = 1/3$?

   Solution: The CDF of an Exponential(1) random variable is $\mathrm{P}(X \leq x) = 1 - e^{-x}$ for $x \geq 0$. If we set $I_1 = [0, x_1), I_2 = [x_1, x_2), I_3 = [x_2, \infty)$ we need to choose x_1, x_2 so that $\mathrm{P}(X \leq x_1) = 1/3$ and $\mathrm{P}(X \leq x_2) = 2/3$, from where $x_1 = \ln \frac{3}{2} \approx 0.405$ and $x_2 = \ln 3 \approx 1.099$.

(b) What is the p-value for the chi-square test with respect to the partition in part (a)? Does it support an alternative hypothesis?

Solution: We are given $n = 21$ samples, out of which 9, 3, and 9 fall inside intervals I_1, I_2, and I_3, respectively. Under the null hypothesis we would expect $21/3 = 7$ samples to fall in each interval, giving a chi-squared statistic value

$$X^2 = \frac{(9-7)^2}{7} + \frac{(3-7)^2}{7} + \frac{(9-7)^2}{7} \approx 3.43$$

and p-value $\mathrm{P}(\chi^2(2) \geq 3.43) \approx 0.18$. This is not strong evidence against the null hypothesis.

3. A hospital is performing an experiment about the effect of different methods in administering a drug. Apply the chi-square test for independence to determine the p-value for the null hypothesis that the effect is independent of the administration method.

	Effective	Ineffective	Number
Oral	58	40	98
Injection	64	31	95
Sum	122	71	193

Solution: The null hypothesis H_0 is that the effect is independent of the administration method. The maximum likelihood estimates of the probabilities that a patient takes an oral vaccine and that the vaccine is effective are $\hat{p} = 98/193$ and $\hat{q} = 122/193$, respectively. If the method of administration was independent of the effect, the expected counts in each category for $n = 193$ patients would be

$$\begin{bmatrix} n\hat{p}\hat{q} & n\hat{p}(1-\hat{q}) \\ n(1-\hat{p})\hat{q} & n(1-\hat{p})(1-\hat{q}) \end{bmatrix} \approx \begin{bmatrix} 61.95 & 36.05 \\ 60.05 & 34.95 \end{bmatrix}$$

resulting in a chi-squared statistic value

$$X^2 = \frac{(58-61.95)^2}{61.95} + \frac{(40-36.05)^2}{36.05} + \frac{(64-60.05)^2}{60.05} + \frac{(31-34.95)^2}{34.95} = 1.391$$

As there is one degree of freedom, the p-value is about $\mathrm{P}(\chi^2(1) \geq 1.391) \approx 0.25$. This is not strong evidence in favor of an alternative hypothesis.

4. In this question you will prove the correctness of the chi-square test for discrete random variables that take two values.

(a) Show that the chi-square statistic X^2 for n samples of an Indicator(p) random variable, N of which come up positive, has value $(N-np)^2/(np(1-p))$.

Solution: The actual counts for the positive and negative outcomes are N and $n-N$, while the expected counts are np and $n(1-p)$. The value of the chi-squared statistic is

$$\begin{aligned} X^2 &= \frac{(N-np)^2}{np} + \frac{((n-N)-n(1-p))^2}{n(1-p)} \\ &= \frac{(N-np)^2 \cdot (1-p) + (N-np)^2 \cdot p}{np(1-p)} \\ &= \frac{(N-np)^2}{np(1-p)}. \end{aligned}$$

(b) Show that $X^2 \geq t^2$ if and only if $|N - \mu| \geq t\sigma$, where μ and σ are the Binomial(n, p) mean and standard deviation, respectively.

Solution: The Binomial(n, p) random variable has mean $\mu = np$ and variance $\sigma^2 = np(1 - p)$. By part (a),

$$\mathrm{P}(X^2 \geq t^2) = \mathrm{P}\left(\frac{(N - np)^2}{np(1 - p)} \geq t^2\right) = \mathrm{P}\left(\frac{(N - \mu)^2}{\sigma^2} \geq t^2\right) = \mathrm{P}(|N - \mu| \geq t\sigma).$$

(c) Using the central limit theorem, show that under the null hypothesis, as n gets large, $\mathrm{P}(X^2 \geq t^2)$ approaches $\mathrm{P}(Y \geq t^2)$, where Y is a $\chi^2(1)$ random variable.

Solution: As a binomial random variable is a sum of independent Indicator(p) random variables, by the Central Limit Theorem, when p is fixed and n approaches infinity, $\mathrm{P}(|N - \mu| \geq t\sigma)$ approaches $\mathrm{P}(|Z| \geq t)$ for a Normal$(0, 1)$ random variable Z. By part (b), $\mathrm{P}(X^2 \geq t^2)$ approaches $\mathrm{P}(|Z| \geq t) = \mathrm{P}(Z^2 \geq t^2)$. The random variable Z^2 is the square of a Normal$(0, 1)$ which is precisely $\chi^2(1)$.