Solution Probability and Statistics Exam June 2016

1. $A : \{HHH, HTT, THT, TTH\}$, $B : \{HHH, HHT, HTH, THH\}$.

(i)

$$\Pr(A) = \frac{4}{8} = \frac{1}{2} = 0.5$$

$$\Pr(B) = \frac{4}{8} = \frac{1}{2} = 0.5$$

$$\Pr(A \cup B) = \frac{7}{8} = 0.875$$

[1]

(ii)

$$\Pr(B|A) = \frac{\Pr(A \cap B)}{\Pr(A)} \qquad [1]$$

$$= \frac{1/8}{1/2} = \frac{1}{4} = 0.25 \qquad [1]$$

(iii) $\Pr(B|A) \neq \Pr(B) \implies A$ and B are not mutually independent. [2]
Or, $\Pr(A \cap B) \neq \Pr(A)\Pr(B) \implies A$ and B are not mutually independent.

2. (i) $X \sim Bin(10, 0.8)$. [2]

(ii) $\Pr(X \geq 4) = \Pr(Y \leq 10 - 4 - 1) = \Pr(Y \leq 5) = 0.9936$. Where $Y \sim Bin(0.2, 10)$ models the number of bulbs that does not bloom. [3]

3. X is the number of items sold in 7 days, and follows a Poisson with mean $= 2 \times 7 = 14$. [1]

We have to find the smallest k such that $\Pr(X \leq k) \geq 0.95$. Using the tables $k = 20$ $\Pr(X \leq 20) = 0.9521$. [2]

The salesman have to buy at least 20 items. [2]

4. (i)

$$\begin{aligned}
\Pr(X \geq 75) &= 1 - \Pr(X < 75) && [1] \\
&= 1 - \Pr(Z < \frac{75-68}{10}) && [1] \\
&= 1 - \Pr(Z < \frac{7}{10}) \\
&= 1 - \Pr(Z < 0.7) \\
&= 1 - 0.7580 && [1] \\
&= 0.242
\end{aligned}$$

(ii)

$$\begin{aligned}
\Pr(70 \leq X \leq 75) &= \Pr(X \leq 75) - \Pr(X \leq 70) && [1] \\
&= \Pr(Z \leq \frac{75-68}{10}) - \Pr(Z \leq \frac{70-68}{10}) \\
&= \Pr(Z \leq \frac{7}{10}) - \Pr(Z \leq \frac{2}{10}) \\
&= \Pr(Z < 0.7) - \Pr(Z < 0.2) \\
&= 0.7580 - 0.5793 \\
&= 0.1787 && [1]
\end{aligned}$$

5. (i) We use a two-sample t-tests for independent samples to test the null hypothesis $H_0 : \mu_0 = \mu_1$ against the one-sided-greater alternative $H_1 : \mu_1 > \mu_0$ [or one-sided-less alternative $H_1 : \mu_0 < \mu_=1$]. [0.5] We assume that $x_{(1)1}, x_{(1)2}, \ldots, x_{(1)8}$ is a random sample from a $N(\mu_1, \sigma^2)$ distribution of the change in the amount of Carbon monoxide transfer for the people in Group 1. And, independently of the first sample, $x_{(0)1}, x_{(0)2}, \ldots, x_{(0)8}$ is a random sample from a $N(\mu_0, \sigma^2)$ distribution of the change in the amount of Carbon monoxide transfer for the people in Group 0. μ_1, μ_0 and $, \sigma^2$ are unknown. [0.5]

In the two-sample t-test the test statistics is:

$$t = \frac{\bar{x}_1 - \bar{x}_0}{s_P\sqrt{\frac{1}{n_1} + \frac{1}{n_0}}} \sim t_{n_1+n_0-2}$$

[1]

(ii) s_P^2 is the pooled estimate of the common variance:

$$s_P^2 = \frac{(n_1 - 1)s_1^2 + (n_0 - 1)s_0^2}{n_1 + n_0 - 2} = \frac{8 \times 4.630^2 + 6 \times 4.101^2}{9 + 7 - 2} = 19.45746 = 4.411^2$$

[1]

The test statistic is given by

$$t = \frac{3.953 - (-0.208)}{4.411\sqrt{\frac{1}{9} + \frac{1}{7}}} = 1.87185 \approx 1.9 \sim t_{14}$$

[1]

Table 9.

$$p\text{−value} = Pr(T > t) = 1 - F_{14}(1.9) = 1 - 0.9609 = 0.0391$$

.

[The p-value = 0.0411 if they use interpolation].

At 1% significance level, we fail to reject H_0. [1]

6. (i) If p denotes the proportion of individuals who are smokers, we test the null hypothesis $H_0 : p = 0.2$ against the alternative $H_1 : p \neq 0.2$. The sample proportion is $\hat{p} = 30/120 = 0.25$. [1]

The test statistic is

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0 q_0 / n}} = \frac{0.25 - 0.2}{\sqrt{(0.2)(0.8)/120}} \approx \frac{0.05}{0.0365} \approx 1.37$$

[1]

The p-value for the corresponding two-tail test is

$$p = 2(1 - \Phi(1.37)) = 2(1 - 0.9147) = 0.1706.$$

[1]

(ii) The p-value is > 0.05. [1]

There is no evidence to reject the claim that the proportion of smokers in the bar is the same as the one from the national statistics. [1]

7. (i) Chi-square test of goodness of fit. We want to test the null hypothesis is that the die is fair. [2]

(ii) $X^2 = 19.184$ is obtained from: [1]

O_r	E_r	$(O_r - E_r)^2/E_r$
212	166.667	12.331
140	166.667	4.267
156	166.667	0.683
170	166.667	0.067
172	166.667	0.171
150	166.667	1.667

[1]

Look Table 9, χ^2 with 5 d.f.

$$p \approx 1 - F(19) = 1 - 0.9981 = 0.0019$$

$p < 0.05$, we reject the null hypothesis that the die is fair. [1]

8. (i) We perform a chi-squared test for contingency tables. [1]

(ii) The expected values are:

2184.18	1399.82
867.82	556.18

[1]

$$\begin{aligned} X^2 &= \frac{(2524-2184.18)^2}{2184.18}+\frac{(1060-1399.82)^2}{1399.82}+\frac{(528-867.82)^2}{867.82}+\frac{(896-556.18)^2}{556.18} \\ &= 52.870+82.495+133.067+207.628 \\ &= 476.061 \end{aligned}$$

[2]

From Table 8 of *Lindley and Scott*, $\chi_1^2(5) = 3.841$.

We reject the null hypothesis of independence between the eye color of the parents and of the children. [1]

9. (a) i.

$$\begin{aligned}\Pr(X \leq 1) &= \Pr(X=0)+\Pr(X=1) & [1]\\ &= \binom{3000}{0}(0.002)^0(1-0.002)^{3000}+\binom{3000}{1}(0.002)^1(1-0.002)^{2999} & [1]\\ &= 1\times 1\times 0.002463904+3000\times 0.002\times 0.002468842 & [1]\\ &= 0.002463904+0.01481305\\ &= 0.01727696\\ &\approx 0.0173 & [1]\end{aligned}$$

ii. $\mu = n \times p = 3000 \times 0.002 = 6$ $Y \sim Poisson(6)$ use Table 2. [2]

$$\begin{aligned}\Pr(X \leq 1) &\approx \Pr(Y \leq 1)\\ &= 0.0174\end{aligned}$$

[2]

iii. $\mu = n \times p = 3000 \times 0.002 = 6$, $\sigma = n \times p \times (1-p) = 3000 \times 0.002 \times 0.998 = 5.988$, $W \sim N(6, 5.988)$ use Table 4. [1]

$$\begin{aligned}\Pr(X \leq 1) &\approx \Pr(W \leq 1)\\ &= \Pr\left(Z \leq \frac{1-6}{\sqrt{5.988}}\right) & [1]\\ &= \Pr\left(Z \leq \frac{-5}{2.447039}\right)\\ &= \Pr(Z \leq -2.043286)\\ &= 1-\Pr(Z \leq 2.043286) & [1]\\ &\approx 1-0.97932\\ &= 0.02068 & [1]\end{aligned}$$

(b) (i) Let's indicate $\Pr(D) = 0.002$ the probability that a dog has the disease, and $\Pr(D^c) = 0.998$ is the probability that a dog does not have the disease.
Let's indicate $\Pr(T|D) = 0.99$ the probability that the test reports correctly that the dog has the disease.
Let's indicate $\Pr(T^c|D^c) = 0.95$ the probability that the test reports correctly that the dog does not have the disease.
Let's indicate $\Pr(T|D^c)$ the probability that the test results come back positive given that the dog does not have the disease.

$$\Pr(T|D^c) = 1 - \Pr(T^c|D^c) = 1 - 0.95 = 0.05$$

[3]

(ii) We are interested in $\Pr(D|T)$, the probability that the dog is affected by the disease given that the test results come back positive.
Use the Bayes' Theorem:

$$
\begin{aligned}
\Pr(D|T) &= \frac{\Pr(D)\Pr(T|D)}{\Pr(T)} && [1] \\
&= \frac{\Pr(D)\Pr(T|D)}{\Pr(D)\Pr(T|D) + \Pr(D^c)\Pr(T|D^c)} && [1] \\
&= \frac{0.002 \times 0.99}{0.002 \times 0.99 + 0.998 \times 0.05} && [1] \\
&= \frac{0.00198}{0.05188} && [1] \\
&= 0.038165 \\
&\approx 4\% && [1]
\end{aligned}
$$

10. (a) We have to check that $\int_{-\infty}^{\infty} f(x)\, dx = 1$: [1]

$$\begin{aligned}
\int_{-\infty}^{\infty} f(x)\, dx &= \int_0^1 \theta x^{\theta-1}\, dx && [1]\\
&= \theta \frac{x^\theta}{\theta}\Big|_0^1 && [1]\\
&= x^\theta\Big|_0^1\\
&= 1 - 0\\
&= 1 && [1]
\end{aligned}$$

(b)

$$\begin{aligned}
\int_{-\infty}^{x} f(u)\, du &= \int_0^x \theta u^{\theta-1}\, du && [1]\\
&= u^\theta\Big|_0^x && [1]\\
&= x^\theta && [1]
\end{aligned}$$

So, $F(x)$:

$$F(x) = \begin{cases} 0 & \text{if } x \leq 0 \\ x^\theta & \text{if } 0 < x < 1 \\ 1 & \text{if } x \geq 1 \end{cases}$$

[1]

(c) $E[X]$:

$$\begin{aligned}
\mu \equiv E[X] &= \int_{-\infty}^{\infty} x f(x)\, dx && [1]\\
&= \int_0^1 x\ \theta x^{\theta-1}\, dx && [1]\\
&= \int_0^1 \theta x^\theta\, dx\\
&= \theta \frac{x^{\theta+1}}{\theta+1}\Big|_0^1 && [1]\\
&= \frac{\theta}{\theta+1}(1-0)\\
&= \frac{\theta}{\theta+1} && [1]
\end{aligned}$$

(d)

$$\begin{aligned}
\sigma^2 \equiv \text{Var}[X] &= E[X^2] - \mu^2 && [1]\\
&= \int_{-\infty}^{\infty} x^2 f(x)\, dx - \mu^2 && [1]\\
&= \int_0^1 x^2\, \theta x^{\theta-1}\, dx - \left(\frac{\theta}{\theta+1}\right)^2 && [1]\\
&= \int_0^1 \theta x^{\theta+1}\, dx - \left(\frac{\theta}{\theta+1}\right)^2 \\
&= \theta \frac{x^{\theta+2}}{\theta+2}\Big|_0^1 - \left(\frac{\theta}{\theta+1}\right)^2 && [1]\\
&= \frac{\theta}{\theta+2} - \left(\frac{\theta}{\theta+1}\right)^2 \\
&= \frac{\theta}{(\theta+2)(\theta+1)^2} && [1]
\end{aligned}$$

(e) $\Pr(X \geq 0.5) = 1 - \Pr(X < 0.5) = 1 - 0.5^2 = 0.75$ [3]

11. (a)

(i) $n = 20$

$$\bar{x} = \frac{4013}{20} = 200.65$$

[3]

(ii)

$$s^2 = \frac{903279 - 20 \times 200.65^2}{19} = \frac{903279 - 805208.4}{19} = \frac{98070.55}{19} = 5161.608$$

[4]

(iii)

$$\left(\bar{x} - t_{n-1}(50\alpha)\frac{s}{\sqrt{n}} \ , \ \bar{x} + t_{n-1}(50\alpha)\frac{s}{\sqrt{n}}\right) \ .$$

[1]

$se = \dfrac{s}{\sqrt{n}} = \dfrac{71.844}{4.359} = 16.06.$ $t_{19}(2.5) = 2.093$ [1]

$$(200.65 - (2.093)(16.06), 200.65 + (2.093)(16.06))$$

[1]

$$(167.03, 234.26)$$

[1]

(iv)

$$\left(\frac{(n-1)s^2}{\chi_9^2(2.5)} \ , \ \frac{(n-1)s^2}{\chi_9^2(97.5)}\right)$$

[2]

$$(19 \times \frac{5161.608}{32.85}, 19 \times \frac{5161.608}{8.907})$$

[2]

$$(2982.513, 10999.84)$$

[1]

(b) Let $X_1, X_2, \ldots, X_n, \ldots$ be a sequence of independently and identically distributed r.v.s, each having a distribution with mean μ and variance σ^2. Define

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad (n = 1, 2, 3, \ldots).$$

As $n \to \infty$,

$$\frac{(\bar{X} - \mu)}{\left(\dfrac{\sigma}{\sqrt{n}}\right)} \overset{D}{\to} Z,$$

where $Z \sim N(0, 1)$. [4]

12. (a) i. Two-sample t-test for paired observations. [1]

$x_1, x_2, \ldots, x_n$ and $y_1, y_2, \ldots, y_n$ are paired (not independent) samples from Normal distributions with means μ_{NoDrug} and μ_{Drug} (both unknown) respectively.

The analysis will be based on consideration of the differences $d_i = x_i - y_i$:

$$d_i \overset{iid}{\sim} N\left(\mu_D, \sigma_D^2\right)$$

where

$$\mu_D = \mu_{\text{NoDrug}} - \mu_{\text{Drug}} \ ,$$

and σ_D^2 are unknown. [2]

$H_0 : \mu_{\text{NoDrug}} = \mu_{\text{Drug}}$, $H_1 : \mu_{\text{NoDrug}} > \mu_{\text{Drug}}$ [or $H_0 : \mu_D = 0$, $H_0 : \mu_D > 0$] [1]

ii. Let $\bar{d}$ the sample mean of the differences and s_D the sample standard deviation of the differences:

$$t = \frac{\bar{d}}{\frac{s_D}{\sqrt{n}}} \sim t_{n-1}$$

[3]

iii. From the data there is no evidence that the new drugs decrease the cholesterol levels (p-value $>$ 0.05 and 0.01). [3]

(b) i. Non-parametric Wilcoxon signed-rank test. [1]

We assume that the random sample, $d_1, \ldots, d_{11}$ is symmetrically distributed about their median value η. This implies that we assume that the median is equal to the mean. $H_0 : \eta_{\text{NoDrug}} = \eta_{\text{Drug}}$, $H_1 : \eta_{\text{NoDrug}} > \eta_{\text{Drug}}$ [or $H_0 : \eta_D = 0$, $H_0 : \eta_D > 0$] [3]

ii. From the data there is no evidence that the new drugs decrease the cholesterol levels (p-value $>$ 0.05 and 0.01). [3]

iii. Wilcoxon signed test is a test on the median, while the paired sample t-test is a test on the mean. [1]

but since in both the tests we assume that the data are coming from a distribution that is symmetric about the mean, the median and the mean are the same. [1]

The two tests are both useful to test if the drug decreases the level of cholesterol, but the Wilcoxon signed rank test is more appropriate, since the dataset is small, and is hard to test the normality assumption. [1]