# Probabilities and Statistics: 2017 (Solutions)

Victor Braun

## Exercise 1

(a) $\Omega = \{1,2,3,4,5,6\}^6$, the probability to obtain one combination $\omega \in \Omega$ is equal to $\frac{1}{|\Omega|} = \frac{1}{6^6}$.
The probability of having the same numbers is $6 * \frac{1}{6^6} = \frac{1}{6^5}$ while the probability of having all different numbers is $\frac{6!}{6^6} = \frac{5!}{6^5}$. Thus the latter is most likely to happen.

(b) $p(Y \leq y) = p(X^2 \leq y) = p(X \leq \sqrt{y})$, here $y > 0$.

$$p(X \leq \sqrt{y}) = \int_0^{\sqrt{y}} \lambda e^{-\lambda x}\, dx = \left[-e^{-\lambda x}\right]_0^{\sqrt{y}} = 1 - e^{-\lambda\sqrt{y}}$$

$$\Downarrow$$

$$f_Y(y) = \frac{\lambda}{2\sqrt{y}} e^{-\lambda\sqrt{y}}$$

(c) $p(X < Y) = \sum_{y=1}^{\infty} p(X < Y \mid Y = y)p(Y = y) = \sum_{y=1}^{\infty} \left(p(Y = y) \sum_{x=1}^{y-1} p(X = x)\right) =$
$p\sum_{y=1}^{\infty}(1-p)^{x-1}(1-(1-p)^{x-1}) = \frac{1-p}{2-p}$. Here we just distribute and use geometric series.

(d) $var(X) = b^2 var(Z_1) + c^2 var(Z_2)$, $var(Y) = e^2 var(Z_1) + f^2 var(Z_2)$
$cov(X,Y) = be \cdot var(Z_1) + cf \cdot var(Z_2) \Rightarrow corr(X,Y) = \frac{be+cf}{\sqrt{(b^2+c^2)(e^2+f^2)}}$

(e) $\mathbb{E}(X_1 + X_2 + 1) = 1 - 2 + 1 = 0$, $cov(X_1, X_2) = 1 \cdot 2 \cdot 3 = 6$, $var(X_1 + X_2 + 1) = 4 + 9 + 2 \cdot 6 = 25$
Thus $X_1 + X_2 + 1 \sim \mathcal{N}(0, 25)$

(f) $p(V = v \mid U = 1) = \frac{p(V=v,\, U=1)}{p(U=1)} = 2v$. By calculating the marginal density.

(g) As the distribution is continuous, by definition $p(X \leq x_p) = p \Rightarrow (X \geq x_p) = 1 - p$

(h) The median doesn't care about outliers, that's why it is robust. Nevertheless, to have a good approximation, the median needs way more data than the mean.

(i) By the delta method, $\frac{1}{\overline{X}} \sim \mathcal{N}\left(\frac{1}{\mu},\ \frac{\sigma^2}{\mu^4 n}\right)$.

(j) $p(I_a(1-I_b)=1) = p(I_a=1)p(I_b=0) = p(1-p) = \tilde{p}$, Thus $Y_i = I_{2i-1}(1-I_{2i}) \sim Ber(\tilde{p})$
$\mathbb{E}(T) = n^{-1}\sum_{j=1}^{n}\mathbb{E}(Y_i) = \frac{np(1-p)}{n} = p(1-p)$. Thus our bias is null.

The MSE of an unbiased estimator is just its variance, so we have:

$$MSE(T) = var(T) = n^{-2}\sum_{j=1}^{n} var(Y_i) \overset{\text{by ind.}}{=} n^{-1}(\tilde{p}(1-\tilde{p})) = n^{-1}(p - 2p^2 + 2p^3 - p^4)$$

# Exercise 2

Let F be the event "the transaction is fraudulent"

(a) $p(X>1) = p(X>1 \mid F)p(F) + p(X>1 \mid \overline{F})p(\overline{F}) = 0.01\int_1^\infty f(x)dx = 0.01\left[(1+9x)^{-1}\right]_\infty^1 = 0.001$

(b) $p(F \mid X=x) = \frac{p(X=x \mid F)p(F)}{p(X=x \mid F)p(F) \ + \ p(X=x \mid \overline{F})p(\overline{F})} = \frac{pf(x)}{pf(x) \ + \ 1-p} =$

$$p(F \mid X=x) = \begin{cases} 0 & x<0 \\ \frac{pf(x)}{pf(x) \ + \ 1-p} & 0 \le x \le 1 \\ 1 & x>1 \end{cases}$$

(c) Let L be the loss.

$$p(L=l) = \begin{cases} p(X=l,\ F) & x \le 1 \\ 0 & x>1 \end{cases}$$

$\mathbb{E}(L) = \int_0^1 xp(L=x)dx = p\int_0^1 xf(x)dx \approx 0.001$

# Exercise 3

Questions (a) and (b) are not corrected as QQ-plots weren't in the 2022 program.

(c) We know that for a big n, $\frac{\overline{X}-\mu}{\sqrt{s^2/n}} \sim t_{n-1}$, and thus a $(1-\alpha)\cdot 100\%$ confidence interval is given by:

$$\left[\overline{X} - t_{n-1}(1-\alpha/2)\frac{s}{\sqrt{n}},\ \overline{X} + t_{n-1}(1-\alpha/2)\frac{s}{\sqrt{n}}\right]$$

$$\Downarrow$$

$$[25.55,\ 29.33]$$

(d) ?

## Exercise 4

(a) As the number of goals during a game is random but not uniform, a good approximation would be a bell curve. The three distributions respect that, however hypergeometric and binomial distributions are used to count successes, here we have no notion of success. That's why the best distribution here would be a Poisson distribution.

(b) The likelihood for a sample of size n is equal to

$$L(\hat{\lambda}) = \prod_{i=1}^{n} \frac{\hat{\lambda}^{y_i}}{y_i!} e^{-\hat{\lambda}} \Rightarrow l(\hat{\lambda}) = \sum_{i=1}^{n} \Big( y_i \ln(\hat{\lambda}) - \hat{\lambda} - \ln(y_i!) \Big) = -\hat{\lambda} n + \ln(\hat{\lambda}) \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \ln(y_i!)$$

As our function is concave, the maximum is given by the null derivative of the likelihood

$$\frac{dl(\hat{\lambda})}{d\hat{\lambda}} = \frac{1}{\hat{\lambda}} \sum_{i=1}^{n} y_i - n = 0 \iff \hat{\lambda} = n^{-1} \sum_{i=1}^{n} y_i = \textit{sample mean}$$

The mean of our sample of size n = 47 is equal to $\overline{x} = 2.45$.

$$\textit{Also, } J(\hat{\lambda}) = \frac{-d^2 l(\hat{\lambda})}{d\hat{\lambda}^2} = \frac{1}{\hat{\lambda}^2} \sum_{i=1}^{n} y_i \Rightarrow \hat{\lambda} \sim \mathcal{N}(\lambda, J(\hat{\lambda})^{-1})$$

Our confidence interval of 95% ($\alpha = 0.05$)is then given by:

$$\left[ \bar{x} - z_{0.975} \sqrt{\frac{\bar{x}}{n}},\ \bar{x} + z_{0.975} \sqrt{\frac{\bar{x}}{n}} \right] = [2.003,\ 2.897]$$

(c) To give evidence against $H_0$, we use the P-value, where $p_{obs}$ = the probability to observe our data under the null hypothesis. As here $p_{obs} = 3\%$ which is a small level and thus a strong evidence against $H_0$.

(d) As the $H_1$ is more likely to be true, the number of goals during a game depends on the round, so it makes no sense to find a common mean to all the games.