Stats Methods Week 2 example exam questions

1. Fig. 1 shows data for the masses and diameters of a large sample of widgets, which were published in a paper. The accompanying explanation states: "The spearman ρ correlation coefficient is 0.77 for 769 data points, so that the correlation between mass and diameter is highly significant ($p \approx 10^{-100}$)". Comment (with explanation) on whether this interpretation is correct or not. **[1 point]**

   Solution: The interpretation is incorrect because the data are clearly clustered into three groups, indicating 3 separate populations. Thus the assumption of independent and identically distributed data is clearly incorrect since the data are not drawn from a single population.

2. In a particular star cluster, 2 per cent of stars in the cluster are massive stars (stars with mass > 8 solar masses, though that is not relevant to this question) with the remainder non-massive. 70 per cent of the massive stars in the cluster are in binary star systems, while 50 per cent of non-massive stars in the cluster are in binary star systems. What is the probability that a star is massive if it is in a binary star system? **[1 point]**

   Solution: Define probabilities, with M used for massive stars, M^C non-massive, B for binary: $P(M) = 0.02$, $P(M^C) = 0.98$, $P(B|M) = 0.7$, $P(B|M^C) = 0.5$. Using Bayes' formula to swap conditionals and law of total probability for $P(B)$:

   $$P(M|B) = \frac{P(B|M)P(M)}{P(B)} = \frac{P(B|M)P(M)}{P(B|M)P(M)+P(B|M^C)P(M^C)}$$

   $$P(M|B) = \frac{0.7\times0.02}{0.7\times0.02+0.5\times0.98} = 0.0277\ldots \text{ recurring.}$$

3. After a search of DNA records for 10^5 people, a suspect has been identified for a recent crime (based solely on the DNA match). The probability of a false positive DNA match from a random (non-guilty) member of the population is estimated to be 10^{-6}. The DNA match is very reliable, so the probability of a match if the suspect is guilty can be assumed to be 1.

   Use Bayes' theorem to calculate the probability that the suspect is guilty given the DNA match. You will need to make your own reasonable assumption for the prior! **[2 points]**

   Solution: Bayes' theorem tells us that the probability of guilt (G) given the DNA match (M) is given by:

   $$P(G|M) = \frac{P(M|G)P(G)}{P(M)}$$

   and we can make the assumption that since only one person should be guilty in the sample of 10^5 people (though of course this could be less, since there may be a larger underlying population), the prior probability of guilt is $P(G) = 10^{-5}$ (this choice of prior is not required for the marks, just that

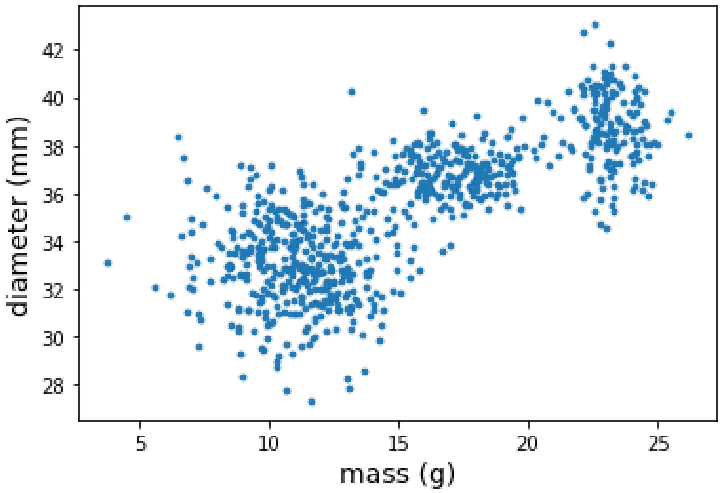

Figure 1: Masses and diameters for widgets.

it the reasoning makes sense and is explained): Then we can use the law of total probability to solve for P (M) (probability of a match) including the complement (not guilty) G^C to get:

$$\begin{aligned} P(G|M) &= \frac{P(M|G)P(G)}{P(M|G)P(G) + P(M|G^C)P(G^C)} \\ &= \frac{1 \times 0.00001}{1 \times 0.00001\ +\ 0.000001 \times 0.99999} \\ &= \frac{0.00001}{0.00001099999} = 0.9091 \text{ to 4 s.f.} \end{aligned}$$

4. A joint distribution has a covariance matrix

$$\begin{pmatrix} 12 & -3 \\ -3 & 5 \end{pmatrix}$$

Calculate the correlation coefficient for this distribution. [**1 point**]

Solution: In the covariance matrix the variances σ_x^2, σ_y^2 are the diagonal terms and the off-diagonal is the covariance σ_{xy}, so the correlation coefficient is:

$\rho(x,y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{-3}{\sqrt{12 \times 5}} = -0.387$ to 3 s.f.

5. Consider a model with two parameters, θ, ϕ, which you are using to explain your data D. Given the likelihood $p(D|\phi,\theta)$ and priors for θ and ϕ, show how to calculate the marginal posterior probability distribution for ϕ. You can assume that the priors for θ and ϕ are independent of one another. [**2 points**]

Solution: *A full calculation is not possible here, instead we need to show* **how** *to do the calculation given the information we already have (the likelihood and the priors). See Ep. 4 for how to work with joint pdfs, and remember to keep the variables on the same side of the conditional term.*

Start by writing Bayes' theorem in terms of the joint probability distribution for ϕ, θ:

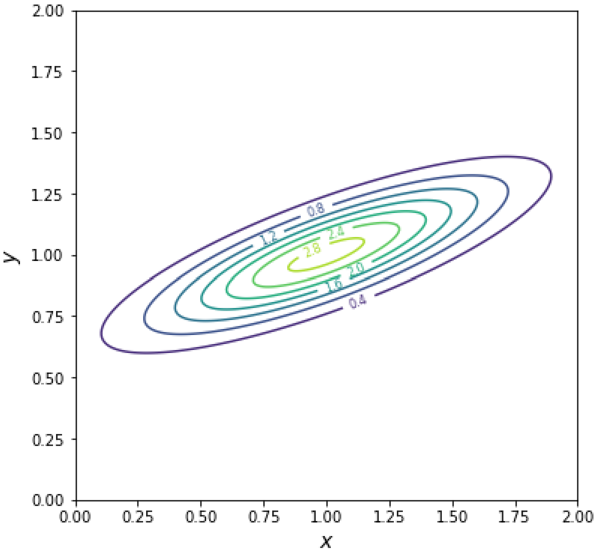

Figure 2: Joint pdf of x and y.

$p(\phi, \theta|D) = \frac{p(D|\phi,\theta)p(\phi,\theta)}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(D|\phi,\theta)p(\phi,\theta)\mathrm{d}\phi\mathrm{d}\theta}$

Then replace the joint prior with the product of $p(\phi)$, $p(\theta)$ (since the priors are independent of one another), and integrate both sides w.r.t. θ to get the final result:

$p(\phi|D) = \int_{-\infty}^{\infty} p(\phi, \theta|D)\mathrm{d}\theta = \frac{\int_{-\infty}^{\infty} p(D|\phi,\theta)p(\phi)p(\theta)\mathrm{d}\theta}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} p(D|\phi,\theta)p(\phi)p(\theta)\mathrm{d}\phi\mathrm{d}\theta}$

6. Fig. 2 shows the joint pdf of two random variables $p(x, y)$. Sketch (with approximate values on each axis) the marginal distribution of x and the marginal distribution of y. [**2 points**]

   Solution: See Fig. 3

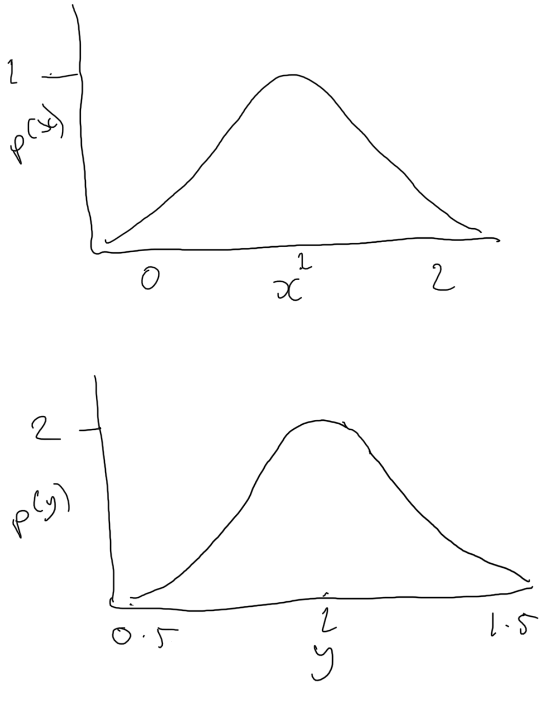

Figure 3: Marginal pdfs sketch for given joint pdf contours.