Math Review Solutions.
Probability Problem Set 1.

1. Recall that the Borel σ-algebra on the interval $(0, 1)$ is the smallest σ-algebra containing all the countable unions and intersections of open intervals that lie in $(0, 1)$ (As mentioned in class, you can think of it as all the countable unions and intersections of open intervals in $(0, 1)$ although that is slightly inaccurate). The analogue of an open interval in $\mathbb{R}^4$ is an open hyper-rectangle of the form $(a_1, b_1) \times (a_2, b_2) \times (a_3, b_3) \times (a_4, b_4)$ where × denotes the Cartesian product and $\{a_n, b_n\}_{n=1}^{4} \subset (0, 1)$. The Borel σ-algebra, $\mathcal{B}$ on $(0, 1)^4$ is the smallest σ-algebra containing all the countable unions and intersections of open hyper-rectangles in $(0, 1)^4$. The choice $\mathcal{S} = \mathcal{B}$ is an appropriate choice. Choices for P can be constructed by integrating pdfs. If f is a pdf on $(0, 1)^4$, then one can take $P(B) = \int_B f(\mathbf{x})\, d\mathbf{x}$. Note that vector notation is used as shorthand to represent the integral; it is actually a quadruple integral.

   Food for thought. This won't be on the test! The normalized histograms of the Iris features are shown in Fig. C.4. Can you think of a good pdf model? (Feel free to look online, e.g. multi-variate Gaussian, Laplacian, exponential, mixtures,...).

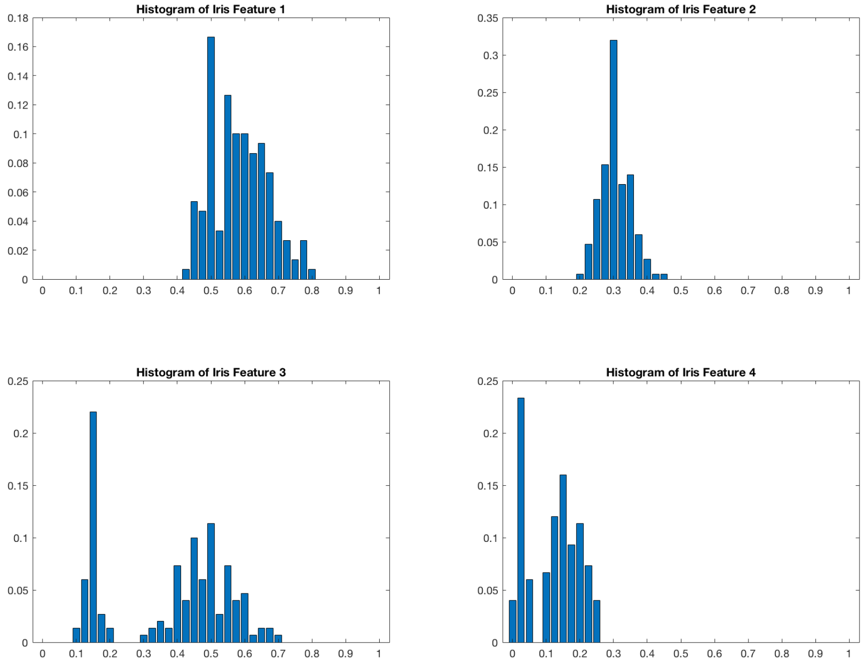

Figure C.4 Normalized Histograms of Individual Iris data features (divided by 10).

2. The experiment is straightforward.

   (a) Draw many sets of samples from a distribution P, $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_N$ where $\mathcal{X}_n = \{\mathbf{x}_{n,1}, \mathbf{x}_{n,2}, \ldots, \mathbf{x}_{n,M_n}\}$ is a collection of independently drawn samples from P, for $n = 1, 2, \ldots, N$.

   (b) Calculate the sample mean of each set of samples, $\bar{\mathbf{x}}_n = \frac{1}{M_n} \sum_{m=1}^{M_n} \mathbf{x}_{n,m}$.

   (c) Calculate the sample mean of the means, $\bar{\boldsymbol{\mu}}^* = \frac{1}{N} \sum_{n=1}^{N} \bar{\mathbf{x}}_n$.

   The vector $\bar{\boldsymbol{\mu}}^*$ is the estimate of the population mean.

   Food for thought. The experiment is illustrated with an example. Suppose P is a Ξ-squared distribution. The above experiment is conducted 4 times corresponding to N = 10, 100, 1000, 10,000 and with $M_n = 100$ for all values of n. Fig. C.5 shows 9 typical examples of normalized histograms obtained by drawing N = 1000 samples from Ξ-squared distributions 9 times. Fig. C.6 shows normalized histograms of the 10, 100, 1000, and 10000 means. Does Fig. C.6 give you any ideas about distributions of means?

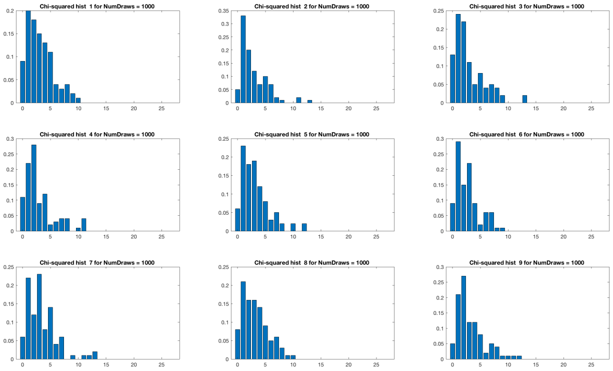

Figure C.5 Normalized Histograms of Samples from Ξ-squared distributions.

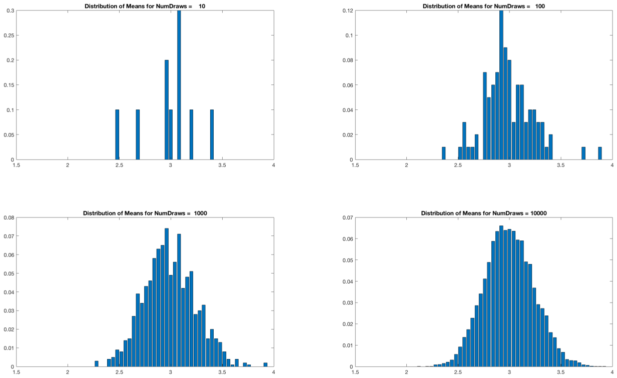

Figure C.6 Normalized Histograms of means of N Samples from Ξ-squared distributions.

3. (a) Let $x, y, z \in \{L, M, R\}$ with $x \neq y \neq z$ and let the triple (x, y, z) represent the event that T_x is the fastest route, that T_y is the initial route selected, and that T_z is not the fastest. Let $P(A)$ for $A \in \{L, M, R\}$ denote the probability that T_A is the smallest estimated travel time. Then $P(L)$ = $P(M)$ =$P(R)$ = $\frac{4}{12} = \frac{1}{3}$ since each fork appears as the first entry exactly 4 times.

There are several interpretations of this problem. First assume that one selects a route and is told that a specific different route is not the fastest route and a decisions is to be made about whether or not to switch routes.

Outcomes corresponding to not switching are

i. The good route is L and one doesn't switch corresponds to the outcomes $\{(L, L, M), (L, L, R)\}$.
ii. The good route is M and one doesn't switch corresponds to the outcomes $\{(M, M, L), (M, M, R)\}$.
iii. The good route is R and one doesn't switch corresponds to the outcomes $\{(R, R, L), (R, R, M)\}$.

Outcomes corresponding to switching are

i. The good route is L and one switches corresponds to the outcomes $\{(L, R, M), (L, M, R), (L, R, R), (L, M, M)\}$.
ii. The good route is M and one switches corresponds to the outcomes $\{(M, L, M), (M, R, L), (M, R, R), (M, L, L)\}$.
iii. The good route is R and one switches corresponds to the outcomes $\{(R, L, M), (R, M, L), (R, M, M), (R, L, L)\}$.

The set of outcomes of interest is the union of the all the sets. Therefore, the are more outcomes that find the best route by switch than by not switching. Specifically, the set of all possible winning outcomes contains 18 triples. 6 of these triples correspond to not switching and 12 correspond to switching. Therefore, in this case, the probability increases from $\frac{1}{3}$ to $\frac{2}{3}$.

On the other hand, if one has selected the right fork and is told that a specific different route is not the fastest route, there is a different set of possible outcomes:

Winning outcomes corresponding to not switching are

i. The good route is R and one doesn't switch corresponds to the outcomes $\{(R, R, L), (R, R, M)\}$.

Winning outcomes corresponding to switching are

i. The good route is M and one switches corresponds to the outcomes $\{(M, R, L)\}$.

In this case, there is no advantage to switching

(b) One could calculate the average of the true travel times, say μ_T and the mean absolute differences between the true travel times and the estimated travel times. The latter calculation can be used to construct a normalized histogram. The parameter s could be set so that the probability that the mean absolute difference between the true travel time and the estimated travel times is greater than $s\mu_T \leq \epsilon$ for some small number ϵ, e.g., $\epsilon = 0.05$.

4. Recall that $\mu_k = \sum_{x \in \mathcal{X}} x^k P(x)$ and that $\gamma_k = \sum_{x \in \mathcal{X}} (x - \mu_k)^k P(x)$. Therefore

$$\begin{aligned}
\mu_1 &= (1)(0.1) + (2)(0.3) + (3)(0.6) && = 2.30 \\
\mu_2 &= (1)(0.1) + (4)(0.3) + (9)(0.6) && = 6.30 \\
\gamma_1 &= (1-2.3)^2(0.1) + (2-2.3)^2(0.3) + (3-2.3)^2(.6) && = 0.48
\end{aligned}$$

5. Let $K = \frac{1}{\sqrt{2\pi}}$. The expected value of x is given by

$$\mathbf{E}[x] = \int_{-\infty}^{\infty} x f(x)\, dx = \frac{K}{\sigma} \int_{-\infty}^{\infty} x e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

Adding and subtracting $\frac{\mu}{\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$ yields

$$\mathbf{E}[x] = \frac{K}{\sigma} \int_{-\infty}^{\infty} (x-\mu)\, e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx + \mu \left[\frac{K}{\sigma} \int_{-\infty}^{\infty} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx \right]$$

Setting $z = \frac{x-\mu}{\sigma}$ so that $\sigma dz = dx$ and noting that the factor in the square brackets in the second term is 1 yields

$$\mathbf{E}[x] = K \int_{-\infty}^{\infty} z e^{-\frac{z}{2}} dz + \mu = 0 + \mu = \mu$$

since $ze^{-\frac{z}{2}}$ is an odd function.

6. The random vector $\mathbf{x}$ can be transformed to the random vector $\mathbf{y}$ using $\mathbf{y} = \mathbf{V}^t(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})$. In the transformed domain, the level curves have the form

$$f(\mathbf{y}) = \mathbf{y}^t \Lambda^{-1} \mathbf{y} = \frac{y_1^2}{16} + \frac{y_2^2}{1} = c^2.$$

The level curves are ellipses centered at $\boldsymbol{\mu}_\mathbf{x}$ with semi-major and minor axes directions defined by the eigenvectors (the columns of $\mathbf{V}$) and with magnitudes $\sqrt{\lambda_1}c$ and $\sqrt{\lambda_2}c$ where $\lambda_1 = 16$ and $\lambda_2 = 1$.

7. The mean vector $\boldsymbol{\mu}_\mathbf{y} = \mathbf{E}[\mathbf{y}] = \mathbf{V}^t \mathbf{E}[\mathbf{x} - \boldsymbol{\mu}_\mathbf{x}] = \mathbf{0}$. The covariance matrix of $\mathbf{y}$ is

$$\begin{aligned}
\mathbf{E}\left[\mathbf{y}\mathbf{y}^t\right] &= \mathbf{E}\left[\mathbf{V}^t(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})^t \mathbf{V}\right] \\
&= \mathbf{V}^t \mathbf{E}\left[(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})(\mathbf{x} - \boldsymbol{\mu}_\mathbf{x})^t\right] \mathbf{V} \\
&= \mathbf{V}^t \boldsymbol{\Sigma} \mathbf{V} \\
&= \boldsymbol{\Lambda}
\end{aligned}$$

8. (a) Let $\mathbf{W} = \mathbf{U} + \mathbf{V}$. The formula for the pdf of $\mathbf{W}$, $f_{\mathbf{W}}(w)$, is the convolution of $f_{\mathbf{U}}(u)$ and $f_{\mathbf{V}}(v)$ which is given by $f_{\mathbf{W}}(w) = \int_{-\infty}^{\infty} f_{\mathbf{U}}(u)\, f_{\mathbf{V}}(w-u)\, du$. The integrand is 0 unless $u \in [0,2]$ and $(w-u) \in [-1,3]$.

   As shown in Fig. C.7, the right edge of $f_{\mathbf{V}}(w-u)$ will partially overlap with $f_{\mathbf{U}}(u)$ while $(w-u) \in [-2,0]$ and the left edge of $f_{\mathbf{V}}(w-u)$ will partially overlap $f_{\mathbf{U}}(u)$ while $(w-u) \in [2,4]$. There will be maximum overlap when $(w-u) \in (0,2)$. Therefore,

$$\begin{aligned} \text{if } w \in [-2,0]\,,\ \text{then } f_{\mathbf{W}}(w) &= \int_{-2}^{w} \frac{1}{8} dw &&= \frac{w+2}{8} \\ \text{if } w \in [0,2]\,,\ \text{then } f_{\mathbf{W}}(w) &= \int_{0}^{2} \frac{1}{8} dw &&= \frac{1}{4} \\ \text{if } w \in [2,4]\,,\ \text{then } f_{\mathbf{W}}(w) &= \int_{w}^{4} \frac{1}{8} dw &&= \frac{4-w}{8} \end{aligned}$$

   The cumulative distribution is given by $F(w) = \int_{-\infty}^{w} f_{\mathbf{W}}(z)\, dz$. Therefore,

   $F(\infty) = \int_{-2}^{0} f_{\mathbf{W}}(w)\, dw + \int_{0}^{2} f_{\mathbf{W}}(w)\, dw + \int_{2}^{4} f_{\mathbf{W}}(w)\, dw = \frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1$

   which shows that $f_{\mathbf{W}}$ is a pdf.

   (b) The convolution of Gaussians, $\mathcal{N}(x|\mu_1, \sigma_1)$ and $\mathcal{N}(x|\mu_2, \sigma_2)$ is also a Gaussian given by $\mathcal{N}(x|\mu_{1,2}, \sigma_{1,2})$ where $\mu_{1,2} = \mu_1 + \mu_2$. If one assumes that the two random variables are uncorrelated, then and $\sigma_{1,2} = \sqrt{\sigma_1^2 + \sigma_2^2}$. If they are correlated, then $\sigma_{1,2} = \sqrt{\sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_y}$ where ρ is the correlation, cf. Wikipedia for the derivation. One can plug into these formulas.

9. Clearly $f(\mathbf{x}) \geq 0$ since $f_m(\mathbf{x}) \geq 0$ and $c_m \geq 0$ for $m = 1, 2, \ldots, M$. Furthermore,

$$\int f(\mathbf{x})\, d\mathbf{x} = f_m(\mathbf{x})\, d\mathbf{x} = \sum_{m=1}^{M} c_m \int f_m(\mathbf{x})\, d\mathbf{x} = \sum_{m=1}^{M} c_m = 1$$

   since $\int f_m(\mathbf{x})\, d\mathbf{x} = 1$ for $m = 1, 2, \ldots, M$.

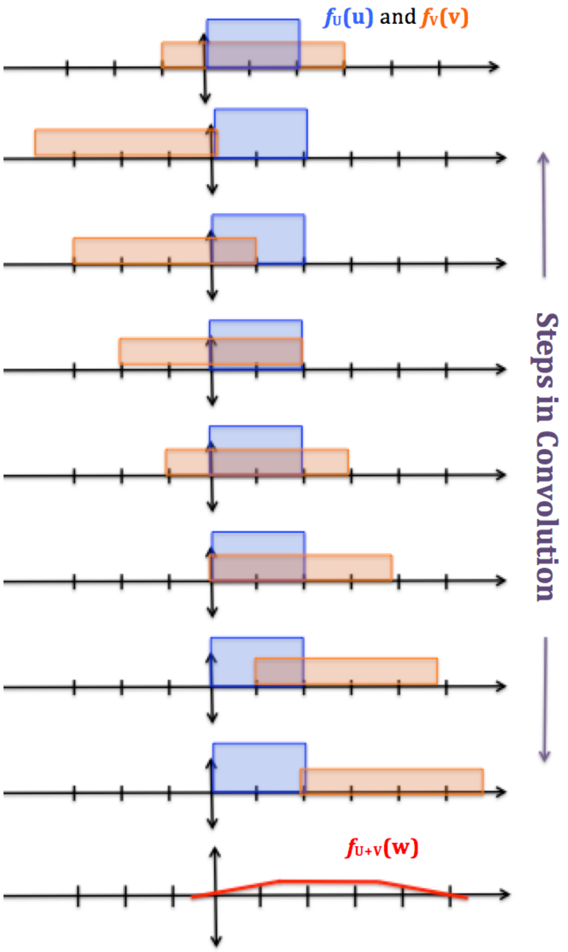

Figure C.7 Depiction of convolution of pdfs required to calculate the distribution of $U + V$. The top plot shows the original distributions of U and V, the bottom plot shows the distribution of $U + V$, and the intermediate plots show some of the steps in the convolution for w barely larger than -2, w = -1,0,1,2,3, and w barely smaller than 4.