1 Probabilities are sensitive to the form of the question that was used to generate the answer

1.1

Possible outcomes for X, the genders of the two children, are

$$X \in \{(B,B),(B,G),(G,B),(\cancel{G,G})\}$$

We have eliminated the possiblity of two girls. There are two remaining outcomes that include at least one girl, so the probability is $\frac{2}{3}$.

1.2

$X \in \{B,G\}$, so the probability is simply $\frac{1}{2}$.

2 Legal reasoning

2.1

Let G = defendant is guilty

Let B = defendant's blood type matches blood type at the scene

The probability that the defendant is guilty given that the defendant's blood type was found at the scene is then

$$P(G|B) = \frac{P(G \cap B)}{P(B)}$$

The prosecutor is confusing $P(B)$ for $P(G|B)$. The prosecutor is failing to account for $P(G)$, the prior probability that the defendant committed the crime, which is extremely low.

2.2

The defense attorney is ignoring all of the other evidence against the defendant. Not all of the 8000 people with the matching blood type are equally likely to have committed the crime.

3 Variance of a sum

Prove: $var[X+Y] = var[X] + var[Y] + 2cov[X,Y]$

$$var[X] = E[X^2] - \mu^2 \tag{1}$$
$$cov[X,Y] = E[XY] - E[X]E[Y] \tag{2}$$

$$
\begin{aligned}
var[X+Y] &= E[(X+Y)(X+Y)] - E^2[X+Y] \\
&= E[X^2 + 2XY + Y^2] - E^2[X+Y] \\
&= E[X^2 + 2XY + Y^2] - (\mu_x + \mu_y)^2 \\
&= E[X^2 + 2XY + Y^2] - (\mu_x^2 + 2\mu_x\mu_y + \mu_y^2) \\
&= E[X^2 + 2XY + Y^2] - \mu_x^2 - 2\mu_x\mu_y - \mu_y^2 \\
&= E[X^2] - \mu_x^2 + E[Y^2] - \mu_y^2 + 2(E[XY] - \mu_x\mu_y) \qquad \text{(By linearity of expectations)} \\
&= var[X] + var[Y] + 2cov[X,Y] \\
&= var[X+Y] \quad \square
\end{aligned}
$$

4 Bayes rule for medical diagnosis

Let D = You have the disease

Let T = You test positive

$$
\begin{aligned}
P(D) &= \frac{1}{10000} & P(\bar{D}) &= \frac{9999}{10000} \\
P(T|D) &= \frac{99}{100} & P(T|\bar{D}) &= \frac{1}{100} \\
P(D|T) &= \frac{P(D \cap T)}{P(T)} = \frac{P(T \cap D)}{P(T)} \\
P(D|T) &= \frac{P(T|D)P(D)}{P(T)} \\
P(D|T) &= \frac{P(T|D)P(D)}{P(T|D)P(D) + P(T|\bar{D})P(\bar{D})} \\
P(D|T) &= \frac{(.99)(.0001)}{(.99)(.0001) + (.01)(.9999)} \\
P(D|T) &= .98\%
\end{aligned}
$$

5 The Monty Hall problem

Let A = the prize is behind door 1

Let B = the prize is behind door 2

Let C = the prize is behind door 3

Let c = the host opens door 3

$$
\begin{aligned}
P(A) = P(B) = P(C) &= \frac{1}{3} \\
P(c|A) &= \frac{1}{2} \\
P(c|B) &= 1 \\
P(c|C) &= 0
\end{aligned}
$$

$$
\begin{aligned}
P(B|c) &= \frac{P(c|B)P(B)}{P(c)} \\
&= \frac{P(c|B)P(B)}{P(c|A)P(A) + P(c|B)P(B) + P(c|C)p(C)} \\
&= \frac{\frac{1}{3}}{\frac{1}{6} + \frac{2}{6} + 0} \\
&= \frac{2}{3}
\end{aligned}
$$

There is a $\frac{2}{3}$ chance that the prize is behind door 2, so the contestant should definitely switch.

6 Conditional independence

6.1

$$
\begin{aligned}
P(H = k|e_1, e_2) &= \frac{P(H = k \cap e_1, e_2)}{P(e_1, e_2)} \\
&= \frac{P(e_1 \cap e_2|H = k)P(H = k)}{P(e_1, e_2)} \qquad \text{(ii.)}
\end{aligned}
$$

(ii.) is sufficient.

6.2

$$
E_1 \perp\!\!\!\perp E_2|H \implies P(e_1, e_2|H) = P(e_1|H) \cdot P(e_2|H)
$$

$$
\begin{aligned}
P(H=k|e_1,e_2) &= \frac{P(e_1 \cap e_2|H=k)P(H=k)}{P(e_1,e_2)} \\
&= \frac{P(e_1|H=k)P(e_2|H=k)P(H=k)}{P(e_1,e_2)} \qquad \text{(i.)} \\
&= \frac{P(e_1|H=k)P(e_2|H=k)P(H=k)}{\sum_{j=1}^{K} P(e_1,e_2|H=j)P(H=j)} \\
&= \frac{P(e_1|H=k)P(e_2|H=k)P(H=k)}{\sum_{j=1}^{K} P(e_1|H=j)P(e_2|H=j)P(H=j)} \qquad \text{(iii.)}
\end{aligned}
$$

(i.), (ii.), and (iii.) are each sufficient.

7 Pairwise independence does not imply mutual independence

Example: Let X and Y represent two coin tosses, with value 0 or 1 for heads or tails, respectively. Let Z = 1 if two coin tosses result in exactly one heads.

$$
(X,Y,Z) = \left\{ \begin{array}{lll} (0, & 0, & 0) \\ (0, & 1, & 1) \\ (1, & 0, & 1) \\ (1, & 1, & 0) \end{array} \right\}
$$

$$
\begin{gathered}
P(X=1) = P(Y=1) = P(Z=1) = \frac{1}{2} \\
P(X=1|Y=1) = P(X=1|Z=1) = P(Y=1|Z=1) = \frac{1}{2} \\
P(X=1,Y=1,Z=1) \stackrel{?}{=} P(X=1)P(Y=1)P(Z=1) \\
0 \neq \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2}
\end{gathered}
$$

The state of any two variables $\in (X,Y,Z)$ determines the third. This means that X, Y, and Z are pairwise independent, but **not** mutually independent.

8 Conditional independence iff joint factorizes

By definition, $X \perp\!\!\!\perp Y|Z$ iff

$$
p(x,y|z) = p(x|z)p(y|z) \tag{1}
$$

We wish to prove an alternate definition - $X \perp\!\!\!\perp Y|Z$ iff there exist functions g and h such that

$$p(x, y|z) = g(x, z)h(y, z) \tag{2}$$

for all x, y and z such that $p(z) > 0$.

To prove this we will show that given equation 2, equation 1 follows, and $X \perp\!\!\!\perp Y|Z$.

We begin by integrating out x.

$$\begin{aligned}
\int_x p(x, y|z)dx &= \int_x g(x, z)h(y, z)dx \\
p(y|z) &= h(y, z)g(z) \\
p(y|z)\frac{1}{g(z)} &= h(y, z)
\end{aligned}$$

By a symmetric argument, we have $p(x|z)\frac{1}{h(z)} = g(x, z)$. Plugging into our original equation, we have

$$\begin{aligned}
p(x, y|z) &= p(y|z)p(x|z)\frac{1}{g(z)}\frac{1}{h(z)} \\
\int_x \int_y p(x, y|z)dxdy &= \int_x \int_y p(y|z)p(x|z)\frac{1}{g(z)}\frac{1}{h(z)}dx\,dy \\
1 &= \frac{1}{g(z)}\frac{1}{h(z)} \\
p(x, y|z) &= p(y|z)p(x|z) \cdot 1 \quad \square
\end{aligned}$$

9 Conditional independence

9.1

$$\overbrace{(X \perp\!\!\!\perp W|Z, Y)}^{A} \wedge \overbrace{(X \perp\!\!\!\perp Y|Z)}^{B} \stackrel{?}{\Longrightarrow} (X \perp\!\!\!\perp Y, W|Z)$$

By decomposition of $(X \perp\!\!\!\perp Y, W|Z)$, we have

$$X \perp\!\!\!\perp Y|Z \tag{1}$$

$$X \perp\!\!\!\perp W|Z \tag{2}$$

(1) is trivally true by (B). Now we must examine (2).

$$\overbrace{(X \perp\!\!\!\perp W|Z, Y)}^{A} \wedge \overbrace{(X \perp\!\!\!\perp Y|Z)}^{B} \stackrel{?}{\Longrightarrow} (X \perp\!\!\!\perp W|Z)$$

$$
\begin{aligned}
P(X|WYZ) &= P(X|YZ) && \text{by (B)} \\
&= P(X|Z) && \text{by (A)} \\
P(X|WYZ) &= P(X|WZ) && \text{by (A)} \\
P(X|Z) &= P(X|WZ) \implies X \perp\!\!\!\perp W|Z
\end{aligned}
$$

Both (1) and (2) are true, so it is true that

$$(X \perp\!\!\!\perp W|Z,Y) \wedge (X \perp\!\!\!\perp Y|Z) \implies (X \perp\!\!\!\perp Y,W|Z)$$

9.2

$$(X \perp\!\!\!\perp Y|Z) \wedge (X \perp\!\!\!\perp Y|W) \stackrel{?}{\implies} (X \perp\!\!\!\perp Y|Z,W)$$

We disprove this with a counter example. Let X, Y, and Z be i.i.d., where

$$P(X) = \begin{cases} .5, & \text{X = 1} \\ .5, & \text{X = -1} \end{cases}$$

Let $W = XYZ$.
Given Z, knowledge of Y gives no additional information about X.
Given W, knowledge of Y gives no additional information about X.
Given W, Z however, X and Y are **not** independent, because $XY = WZ$.

10 Deriving the inverse gamma density

$$
\begin{aligned}
X &\sim \text{Ga}(a,b) \\
p(x) &= \text{Ga}(x|a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-xb} \\
P_x(x) &= \int f(x)\, dx \\
P_x(x) &= \int \frac{b^a}{\Gamma(a)} x^{a-1} e^{-xb}\, dx
\end{aligned}
$$

Now we change variables to find the cdf of Y, $P_y(y)$

$$\begin{aligned}
x &= \frac{1}{y} = y^{-1} \\
\frac{dx}{dy} &= -y^{-2} = -\frac{1}{y^2} \\
dx &= -\frac{1}{y^2} dy \\
P_y(y) &= \int \frac{b^a}{\Gamma(a)} \left(y^{-1}\right)^{a-1} e^{-\frac{b}{y}} y^{-2} \, dy \\
&= \int \frac{b^a}{\Gamma(a)} y^{-(a+1)} e^{-\frac{b}{y}} \, dy \\
p(y) &= \frac{b^a}{\Gamma(a)} y^{-(a+1)} e^{-\frac{b}{y}} \\
&= IG(y|a,b)
\end{aligned}$$

11 Normalization constant for a 1D Gaussian

After squaring Z and converting to polar coordinates, we have

$$\begin{aligned}
Z^2 &= \int_0^{2\pi} \int_0^{\infty} r e^{\frac{-r^2}{2\sigma^2}} \, dr d\theta \\
&= \left[\theta \int_0^{\infty} r e^{-\frac{r^2}{2\sigma^2}} \, dr\right]_0^{2\pi} \\
&= 2\pi \int_0^{\infty} r e^{-\frac{r^2}{2\sigma^2}} \, dr
\end{aligned}$$

Let $u = \phi(r) = e^{-\frac{r^2}{2\sigma^2}}$.

$$\begin{aligned}
\frac{du}{dr} &= -\frac{r}{\sigma^2} e^{-\frac{r^2}{2\sigma^2}} \\
du &= -\frac{r}{\sigma^2} e^{-\frac{r^2}{2\sigma^2}} dr \\
\frac{-\sigma^2}{u} du &= r \, dr \\
\phi(0) &= 1 \\
\phi(\infty) &= e^{-\infty} = 0
\end{aligned}$$

Now we plug u and $r\ dr$ back into our equation for Z^2.

$$\begin{aligned} Z^2 &= 2\pi \int_1^0 u - \frac{\sigma^2}{u}\ du \\ &= -2\pi \int_1^0 \sigma^2\ du \\ &= -2\pi\sigma^2 \cdot \Big[u\Big]_1^0 \\ &= 2\pi\sigma^2 \\ Z &= \pm\sqrt{2\pi\sigma} \\ &= \sqrt{2\pi\sigma} \end{aligned}$$

We take only the positive value, since we arbitrarily squared Z at the beginning of our derivation.

12 Expressing mutual information in terms of entropies

$$
\begin{aligned}
I\left(X,Y\right) &= KL\left(p\left(X,Y\right)||p\left(X\right)p\left(Y\right)\right) \\
&= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
H(X) &= -\sum_{k=1}^{K} p(X=k) \log p(X=k) \\
&= -\sum_x p(x) \log p(x) \\
H(X|Y) &= \sum_y p(y) H(X|Y=y) \\
&= \sum_y p(y) \left(-\sum_x p(x) \log p(x) | Y=y \right) \\
&= -\sum_y \sum_x p(x,y) \log \frac{p(x,y)}{p(y)} \\
H(X) - H(X|Y) &= -\sum_x p(x) \log p(x) + \sum_y \sum_x p(x,y) \log \frac{p(x,y)}{p(y)} \\
&= -\sum_y p(y) \sum_x p(x) \log p(x) + \sum_y \sum_x p(x,y) \log \frac{p(x,y)}{p(y)} \\
&= -\sum_y \sum_x p(x,y) \log p(x) + \sum_y \sum_x p(x,y) \log \frac{p(x,y)}{p(y)} \\
&= \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\
&= I(X,Y) \quad \square
\end{aligned}
$$

13 Mutual information for correlated normals

$$
\begin{aligned}
I(X,Y) &= h(X) - h(X|Y) \\
&= h(X) - \big(h(X,Y) - h(Y)\big) \\
&= h(X) + h(Y) - h(X,Y)
\end{aligned}
$$

For a 1-d Gaussian,

$$h(X) = \frac{1}{2} log_2 \left[2\pi e \sigma^2 \right]$$

For a 2-d Gaussian,

$$h(X) = \frac{1}{2} log_2 \left[(2\pi e)^2 \sigma^4 \left(1 - \rho^2 \right) \right]$$

X_1 and X_2 are individually normal with variance σ^2, so $h(X_1) = h(X_2)$.

$$\begin{aligned} I(X_1, X_2) &= h(X_1) + h(X_2) - h(X_1, X_2) \\ &= \log_2 (2\pi e)^2 \sigma^4 - \frac{1}{2} \log_2 (2\pi e)^2 \sigma^4 (1 - \rho^2) \\ &= -\frac{1}{2} \log_2 (1 - \rho^2) \end{aligned}$$

1. $\rho = 1;\ I(X_1, X_2) = \infty$ (perfect correlation, $X_1 = X_2$)

2. $\rho = 0;\ I(X_1, X_2) = 0$ (independent)

3. $\rho = -1;\ I(X_1, X_2) = \infty$ (perfect correlation, $X_1 = -X_2$)

14 A measure of correlation (normalized mutual information)

$$r = 1 - \frac{H(Y|X)}{H(X)}$$

14.1

$$\begin{aligned} I(X, Y) &= H(X) - H(Y|X) \\ r &= \frac{H(X)}{H(X)} - \frac{H(Y|X)}{H(X)} \\ &= \frac{H(X)}{H(X)} - \frac{H(X|Y)}{H(X)} \qquad \text{(because } H(X) = H(Y)\text{)} \\ &= \frac{I(X, Y)}{H(X)} \end{aligned}$$

14.2

$$\begin{array}{lcl} 0 \leq & H(X|Y) & \leq H(X) \\ 0 \leq & \frac{H(X|Y)}{H(X)} & \leq 1 \\ 0 \leq & \frac{H(Y|X)}{H(X)} & \leq 1 \\ 0 \leq & 1 - \frac{H(Y|X)}{H(X)} & \leq 1 \\ 0 \leq & r & \leq 1 \end{array}$$

14.3

$r = \frac{I(X,Y)}{H(X)}$, so $r = 0$ when $I(X,Y) = 0$. This occurs when X and Y are independent.

14.4

$r = 1 - \frac{H(X|Y)}{H(X)}$, so $r = 1$ when $H(X|Y) = 0$. This occurs when X and Y are perfectly correlated.

15 MLE minimizes KL divergence to the empirical distribution

$$0 \leq KL(p||q) = \int_x p(x) \log \left(\frac{p(x)}{q(x)} \right) dx$$

$KL(p||q)$ is non-negative, so it is minimized at 0, or when

$$\begin{gathered} p(x) \log p(x) - p(x) \log q(x) = 0 \\ p(x) \log p(x) = p(x) \log q(x) \\ \log p(x) = \log q(x) \end{gathered}$$

$p(x)$ is the empirical distribution, so we have

$$p_{emp}(x) = q(x; \theta)$$

Therefore, θ must be the MLE.

$$p_{emp}(x) = q(x; \hat{\boldsymbol{\theta}})$$

16 Mean, mode, variance for the beta distribution

$$\text{Beta}(x|a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} \tag{1}$$

$$B(a,b) \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \tag{2}$$

$$\begin{aligned}
\text{E}[X] &= \int_0^1 x \cdot \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1}\, dx \\
&= \frac{1}{B(a,b)} \int_0^1 x^a (1-x)^{b-1}\, dx \\
&= \frac{1}{B(a,b)} \cdot B(a+1,b) \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \\
&= \frac{\Gamma(a+1)\Gamma(a+b)}{\Gamma(a)\Gamma(a+b+1)} \\
&= \frac{a\Gamma(a)\Gamma(a+b)}{\Gamma(a)(a+b)\Gamma(a+b)} \\
&= \frac{a}{a+b} \qquad \text{(mean)}
\end{aligned}$$

$$\begin{aligned}
\text{E}[X^2] &= \int_0^1 \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} x^2\, dx \\
&= \frac{1}{B(a,b)} \int_0^1 x^{a+1}(1-x)^{b-1}\, dx \\
&= \frac{1}{B(a,b)} B(a+2,b) \\
&= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \cdot \frac{\Gamma(a+2)\Gamma(b)}{\Gamma(a+b+2)} \\
&= \frac{\Gamma(a+b)(a+1)\Gamma(a+1)}{\Gamma(a)(a+b+1)\Gamma(a+b+1)} \\
&= \frac{\Gamma(a+b)(a+1)a\Gamma(a)}{\Gamma(a)(a+b+1)(a+b)\Gamma(a+b)} \\
&= \frac{a(a+1)}{(a+b)(a+b+1)}
\end{aligned}$$

$$
\begin{aligned}
\mathrm{Var}[X] &= \mathrm{E}[X^2] - (\mathrm{E}[X])^2 \\
&= \frac{(a+1)(a)}{(a+b)(a+b+1)} - \left(\frac{a}{a+b}\right)^2 \\
&= \frac{(a+1)(a)}{(a+b)(a+b+1)} - \frac{a^2}{(a+b)^2} \\
&= \frac{(a+b)(a+1)a}{(a+b)^2(a+b+1)} - \frac{a^2(a+b+1)}{(a+b)^2(a+b+1)} \\
&= \frac{a(a+b)(a+1) - a^2(a+b+1)}{(a+b)^2(a+b+1)} \\
&= \frac{a(a^2+a+ab+b) - a^3 - ba^2 - a^2}{(a+b)^2(a+b+1)} \\
&= \frac{ab}{(a+b)^2(a+b+1)} \qquad \text{(variance)}
\end{aligned}
$$

The mode of the distribution occurs where the pdf is at a maximum. To maximize the pdf, we must find where $p'(x) = 0$.

$$
\begin{aligned}
p(x) &= \frac{x^{a-1}(1-x)^{b-1}}{B(a,b)} \\
\tfrac{\mathrm{d}}{\mathrm{d}x}p(x) &= \frac{(a-1)x^{a-2}(1-x)^{b-1}}{B(a,b)} - \frac{x^{a-1}(b-1)(1-x)^{b-2}}{B(a,b)} = 0
\end{aligned}
$$

We can ignore the denominator.

$$
\begin{gathered}
(a-1)x^{a-2}(1-x)^{b-1} - x^{a-1}(b-1)(1-x)^{b-2} = 0 \\
x^{a-2} \cdot \Big[(a-1)(1-x)^{b-1} - x(b-1)(1-x)^{b-2}\Big] = 0 \\
x^{a-2}(1-x)^{b-2}\left[(a-1)(1-x) - x(b-1)\right] = 0 \\
a - ax - 1 + x - bx + x = 0 \\
-x(a+b-2) + a - 1 = 0 \\
x = \frac{a-1}{a+b-2} \quad \text{(mode)}
\end{gathered}
$$

17 Expected value of the minimum

Let $T = min(X_1, X_2, ..., X_N)$. We are interested in $\mathrm{E}[T]$, for which we will need the pdf of t. Let N represent the number of samples.

$$
\begin{aligned}
F(t) &= P(T \le t) \\
&= \begin{cases} 0 & \text{if } t < a \\ 1 - P(T > t), & \text{for } a \le t \le b \\ 1 & \text{if } t > b \end{cases} \\
&= 1 - P(\forall_i X_i > t) \quad \text{for } a \le t \le b \\
&= 1 - \prod_i P(X_i > t) \quad \text{for } a \le t \le b \quad \text{(because samples are indpendent)} \\
&= 1 - P(X_1 > t)^N \quad \text{for } a \le t \le b \\
&= 1 - \left(\frac{b-t}{b-a}\right)^N \quad \text{for } a \le t \le b
\end{aligned}
$$

Now we take the derivative, $f(t)$, which is the pdf.

$$
\begin{aligned}
f(t) &= \tfrac{\mathrm{d}}{\mathrm{d}t} F(t) \\
&= \tfrac{\mathrm{d}}{\mathrm{d}t} \left(1 - \left(\frac{b-t}{b-a}\right)^N\right) \\
&= 0 - \tfrac{\mathrm{d}}{\mathrm{d}t} \left(\frac{1}{(b-a)^N} \cdot (b-t)^N\right) \\
&= \frac{N}{(b-a)^N} \cdot (b-t)^{N-1}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{E}[t] &= \int_a^b t \cdot f(t)\, dt \\
\mathrm{E}[t] &= \frac{N}{(b-a)^N} \int_a^b (b-t)^{N-1} \cdot t\, dt
\end{aligned}
$$

We can solve this integral using substitution. Let $u = b - t$. Then $du = -dt$ and $t = b - u$.

$$
\begin{aligned}
E[t] &= \frac{N}{(b-a)^N} \int_{b-a}^{0} (u-b)u^{N-1}\, du \\
&= \frac{N}{(b-a)^N} \int_{b-a}^{0} u^N - bu^{N-1}\, du \\
&= \frac{N}{(b-a)^N} \left[\left(\frac{u^{N+1}}{N+1} - \frac{bu^N}{N} \right) \right]_{b-a}^{0} \\
&= \frac{N}{(b-a)^N} \frac{b(b-a)^N(N+1) - N(b-a)^{N+1}}{N(N+1)} \\
&= \frac{1}{(b-a)^N} \frac{(b-a)^N \left[b(N+1) - N(b-a) \right]}{N+1} \\
&= \frac{bN + b - bN + Na}{N+1} \\
&= \frac{Na+b}{N+1}
\end{aligned}
$$

For $N = 2, a = 0,$ and $b = 1$.

$$
\begin{aligned}
\mathrm{E}[t] &= \frac{(2 \cdot 0) + 1}{2+1} \\
&= \frac{1}{3}
\end{aligned}
$$