For extra practice, several additional review problems are printed below. Solutions to these problems can be found on the exams page of the course website. I've tried to sort the problems by whether I think they are more appropriate as an in-class question or as a take-home question. While these questions are representative of the typical scope and difficulty of individual exam questions, this review is not comprehensive, nor does it necessarily represent the total amount of time available for the exam.

In-class Exam.

(1) Surprise eggs are type of collectible popular among children; each egg contains a randomized toy that is revealed once the egg is purchased and opened. Suppose there are a total of n unique toys types, and that each egg has equal chance of containing each type. If you purchase t eggs, what is the expected number of distinct toys you will collect?

Solution. For each toy type j, let A_j be the event that the toy was obtained at least once in t eggs. Since each egg has a $\frac{1}{n}$ chance of containing a specific toy, and the toys in different eggs are independent, then $P(A_j) = 1 - (\frac{n-1}{n})^t$. Let I_j be the indicator for A_j and let N be the total number of distinct toys collected. Then $N = \sum_j I_j$ and by the fundamental bridge and linearity of expectation,

$$E[N] = \sum_{j=1}^{n} E[I_j] = \sum_{j=1}^{n} P(A_j) = n\left(1 - \left(\frac{n-1}{n}\right)^t\right)$$

(2) Let $Z \sim N(0,1)$ and $Y = |Z|$. We say that Y has the *folded Normal* distribution. Find **two** expressions for the MGF of Y as unsimplified integrals, one integral based on the PDF of Y and one based on the PDF of Z.

Solution. To find the PDF of Y, we first find the CDF of Y. Let $y \geq 0$, then

$$F_Y(y) = P(Y \leq y) = P(|Z| \leq y) = P(-y \leq Z \leq y) = \Phi(y) - \Phi(-y)$$

where Φ is the CDF of the standard Normal. To get the PDF, we differentiate:

$$f_Y(y) = \frac{d}{dy}F_Y(y) = \varphi(y) + \varphi(-y) = 2\varphi(y) \qquad y \geq 0$$

since φ is symmetric around 0. By LOTUS, the MGF of Y is then

$$M(t) = E[e^{tY}] = \int_{-\infty}^{\infty} e^{ty} f_Y(y)\,dy = \int_0^{\infty} 2e^{ty}\varphi(y)\,dy$$

On the other hand, since $Y = |Z|$, then

$$M(t) = E[e^{tY}] = E[e^{t|Y|}] = \int_{-\infty}^{\infty} e^{t|y|}\varphi(y)\,dy.$$

(3) Find the probability that the quadratic polynomial $Ax^2 + Bx + 1$ has at least one real root, if A, B are iid $\mathrm{Unif}(0,1)$. What is the expected number of real roots of the polynomial?

Solution. By the quadratic formula, the polynomial has at least one real root when

$$B^2 - 4A \geq 0 \iff A \leq \frac{B^2}{4}$$

Since (A, B) are iid uniform on $[0,1]$, then their joint density is $f(a,b) = 1$ for $0 \leq a, b \leq 1$ and

$$P(A \leq \frac{B^2}{4}) = \int_0^1 \int_0^{b^2/4} 1\,da\,db = \int_0^1 b^2/4\,db = \frac{1}{12}$$

The polynomial has exactly 1 real root only when $B^2 = 4A$, which happens with probability 0, since both A and B are continuous variables. By the preceding calculation, the polynomial has two real roots with probability $\frac{1}{12}$ and therefore has no real roots with probability $\frac{11}{12}$. Therefore, the expected number of roots is

$$2 \cdot \frac{1}{12} + 1 \cdot 0 + 0 \cdot \frac{11}{12} = \frac{1}{6}.$$

(4) A variable X is said to have the *arcsine* distribution if its CDF F is given by

$$F(x) = \frac{2}{\pi} \sin^{-1}(\sqrt{x}) \qquad \text{for } 0 < x < 1$$

and $F(x) = 0$ for $x \leq 0$ and $F(x) = 1$ for $x \geq 1$.
(a) Check that F is indeed a valid CDF.
(b) Find the corresponding PDF f. (Recall that the derivative of $\sin^{-1}(x)$ is $\frac{1}{\sqrt{1-x^2}}$)
(c) Find a formula for the quantile function F^{-1} of X.
(d) Suppose $U \sim \text{Unif}(0,1)$ and let F^{-1} be the quantile function of X. What is the name of the distribution of $F^{-1}(U)$?

Solution. (a) In order for F to be a valid CDF, it must:
(i) Be increasing; which it is, since the the derivative of $\frac{2}{\pi}\sin^{-1}(\sqrt{x})$ is $\frac{1}{\pi\sqrt{x(1-x)}}$ which is non-negative.
(ii) Be right-continuous; which it is, since it is continuous (note that the limit of $\frac{2}{\pi}\sin^{-1}(\sqrt{x})$ as $x \to 0^+$ is 0 and that the limit as $x \to 1^-1$ is 1).
(iii) Approach 0 as $x \to -\infty$, which it does, since $F(x) = 1$ if $x \geq 1$ and $F(x) = 0$ if $x \leq 0$.
(b) To find the PDF, we differentiate using the chain rule:

$$f(x) = \frac{d}{dx}F(x) = \frac{1}{\pi\sqrt{x(1-x)}}$$

(c) The quantile function is the inverse of F. Let $y = F(x) = \frac{2}{\pi}\sin^{-1}(\sqrt{x})$. Solving for y gives

$$x = F^{-1}(y) = \sin^2\left(\frac{\pi y}{2}\right)$$

(d) By the universality of the uniform, the variable $F^{-1}(U) = \sin^2\left(\frac{\pi U}{2}\right)$ has the arcsin distribution.

(5) Let $X \sim \text{Pois}(\lambda)$. (While proofs of each of the following are in the text, it's good practice with Taylor Series to compute them yourself).
(a) Show directly that the PMF for X,

$$p(k) = \frac{e^{-\lambda}\lambda^k}{k!} \qquad \text{for} \qquad k \geq 0$$

is a valid PMF.
(b) Show using the PMF of X that the mean and variance of X are both λ.
(c) Give an alternative proof of the preceding fact by finding the MGF of X using LOTUS and then computing moments.

Solution. (a) In order for p to be a valid PMF, it must be non-negative and must sum to 1 across the support of X. The first property is true, since $e^{-\lambda}$, λ^k and $k!$ are all non-negative for $k \geq 0$ and $\lambda \geq 0$. For the second property,

$$\sum_{k=0}^{\infty} p(k) = \sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = e^{-\lambda}e^{\lambda} = 1$$

using the Taylor Series for e^{λ}.
(b) The mean of X is

$$E[X] = \sum_{k=0}^{\infty} k\frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}\lambda\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} = e^{-\lambda}\lambda\sum_{k=0}^{\infty}\frac{\lambda^k}{(k-1)!} = e^{-\lambda}\lambda\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = e^{-\lambda}\lambda e^{\lambda} = \lambda$$

The second moment of X is

$$\begin{aligned} E[X^2] = \sum_{k=0}^{\infty} k^2 \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda}\lambda \sum_{k=1}^{\infty} k \frac{\lambda^{k-1}}{(k-1)!} &= e^{-\lambda}\lambda \sum_{k=1}^{\infty} \frac{d}{d\lambda} \frac{\lambda^k}{(k-1)!} \\ &= e^{-\lambda}\lambda \frac{d}{d\lambda} \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} \\ &= e^{-\lambda}\lambda \frac{d}{d\lambda} \lambda \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \\ &= e^{-\lambda}\lambda \frac{d}{d\lambda} \lambda \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \\ &= e^{-\lambda}\lambda \frac{d}{d\lambda} \lambda e^{\lambda} \\ &= e^{-\lambda}\lambda (e^{\lambda} + \lambda e^{\lambda}) \\ &= \lambda + \lambda^2 \end{aligned}$$

The variance of X is therefore,

$$\mathrm{Var}(X) = E[X^2] - (E[X])^2 = \lambda + \lambda^2 - \lambda^2 = \lambda$$

(c) Using the definition of the MGF and LOTUS,

$$M(t) = E[e^{tX}] = \sum_k e^{tk} \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda} \sum_k \frac{(e^t\lambda)^k}{k!} = e^{-\lambda} e^{e^t \lambda} = e^{\lambda(e^t - 1)}$$

To calculate moments, we need to find the first and second derivatives of M:

$$M'(t) = \lambda e^t e^{\lambda(e^t-1)} = \lambda e^{t+\lambda(e^t-1)}$$

and

$$M''(t) = \lambda(1 + \lambda e^t) e^{t+\lambda(e^t-1)}$$

Therefore,

$$E[X] = M'(0) = \lambda e^{0+\lambda(e^0-1)} = \lambda$$

and

$$E[X^2] = M''(0) = \lambda(1 + \lambda e^0) e^{0+\lambda(e^0-1)} = \lambda + \lambda^2$$

These values agree with those calculated in previous parts.

(6) Suppose X and Y are independent variables with $X \sim \mathrm{Geom}(p_1)$ and $Y \sim \mathrm{Geom}(p_2)$. Let $q_1 = 1 - p_1$ and $q_2 = 1 - p_2$.

(a) Find $P(X < Y)$. *Hint: recall that the CDF of* $Z \sim \mathrm{Geom}(p)$ *is* $F_Z(z) = 1 - q^{z+1}$ *for nonnegative integers* z.

(b) What does your answer simplify to in the case when $p_1 = p_2$?

Solution. (a) Using LoTP, and the fact that X and Y are independent,

$$\begin{aligned} P(X<Y) = \sum_{k=0}^{\infty} P(X<Y|Y=k)P(Y=k) &= \sum_{k=0}^{\infty} P(X<k)P(Y=k) \\ &= \sum_{k=0}^{\infty} (1-q_1^k)q_2^k p_2 \\ &= p_2 \sum_{k=0}^{\infty} q_1^k - (q_1q_2)^k \\ &= p_2 \left(\frac{1}{1-q_1} - \frac{1}{1-q_1q_2} \right) \end{aligned}$$

(b) When $p_1 = p_2 = p$ and $q_1 = q_2 = q$, this simplifies to

$$P(X<Y) = p\left(\frac{1}{p} - \frac{1}{1-q^2}\right) = 1 - \frac{p}{1-q^2} = 1 - \frac{1}{1+q} = \frac{q}{1+q}$$

Take-home Exam.

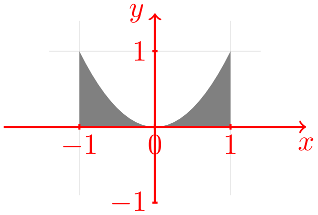
(1) Suppose (X, Y) are random variables with joint PDF given by

$$f(x, y) = c, \ \text{ for } \ -1 \leq x \leq 1, \, 0 \leq y \leq x^2$$

for some constant c.

(a) Sketch the region in the xy-plane where the joint PDF is non-zero.
(b) Find the value of the constant c that makes f a valid joint PDF.
(c) Find the marginal distribution of X, as well as the marginal distribution of Y.
(d) Compute the mean of X and the mean of Y.
(e) Calculate $\text{Cov}(X, Y)$.
(f) Are X and Y independent? Explain.

Solution. (a) The region corresponds to the area under the parabola $y = x^2$ above the x-axis, from $x = -1$ to $x = 1$.



(b) Since $f(x, y)$ is a joint PDF, then the volume under the surface of $z = f(x, y)$ must be 1. That is,

$$1 = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) \, dx dy = \int_{-1}^{1} \int_{0}^{x^2} c \, dy \, dx = c \int_{-1}^{1} x^2 \, dx = c \frac{2}{3}$$

and so $c = \frac{3}{2}$.

(c) The marginal distribution of X can be obtained by integrating out y from the joint PDF:

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) \, dy = \int_{0}^{x^2} \frac{3}{2} \, dy = \frac{3}{2} x^2 \qquad -1 \leq x \leq 1$$

Similarly, the marginal distribution of Y can be found by integrating out x from the joint PDF:

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) \, dx = \int_{\sqrt{y}}^{1} \frac{3}{2} \, dx + \int_{-1}^{-\sqrt{y}} \frac{3}{2} \, dx = \frac{3}{2}(1 - \sqrt{y} - \sqrt{y} + 1) = 3(1 - \sqrt{y}) \qquad 0 \leq y \leq 1$$

(d) The mean of X is

$$E[X] = \int_{-1}^{1} x \frac{3}{2} x^2 \, dx = \frac{3}{2} \frac{1}{4} x^4 \Big|_{-1}^{1} = 0$$

while the mean of Y is

$$E[Y] = \int_{0}^{1} y \cdot 3(1 - \sqrt{y}) \, dy = 3 \left(\frac{y}{2} - \frac{2y^{5/2}}{5} \right) \Big|_{0}^{1} = \frac{3}{10}$$

(e) We first calculate $E[XY]$ using 2D Lotus:

$$E[XY] = \int_{-1}^{1} \int_{0}^{x^2} xy \frac{3}{2} \, dy dx = \frac{3}{2} \int_{-1}^{1} x \int_{0}^{x^2} y \, dy dx = \frac{3}{2} \int_{-1}^{1} x \left(\frac{y^2}{2} \Big|_{0}^{x^2} \right) = \frac{3}{4} \int_{-1}^{1} x^5 \, dx = 0$$

where the final equality followed by symmetry and the fact that the integrand was an odd function.

(f) The variables X and Y are not independent. If X is close to 0, then Y must also be close to 0, for example.

(2) Suppose $m \geq 2$ and let $X_1, \ldots, X_m$ be iid $\text{Exp}(\lambda)$, let $S_m = \sum_{k=1}^{m} X_k$ and let $M = \min\{X_1, \ldots, X_m\}$. Is S_m exponentially distributed? Is M exponentially distributed? Justify your answer.

Solution. First, observe that S_m is not exponentially distributed. Not that any $\text{Expo}(\lambda)$ variable must have mean $\frac{1}{\lambda}$ and variance $\frac{1}{\lambda^2}$, and so the variance of an exponential is the square of its mean. But by linearity of expectation, and properties of variance,

$$E[S_m] = nE[X_1] = \sum E[X_i] = \frac{n}{\lambda} \qquad \text{Var}(S_m) = \sum \text{Var}(X_i) = \frac{n}{\lambda^2}$$

But this means $\text{Var}(S_m) \neq (E[S_m])^2$.

Second, we will see that M is exponentially distributed, by calculating its CDF and then PDF. Let $x \geq 0$, then

$$\begin{aligned} F_M(x) = P(M \leq x) = 1 - P(M \geq x) =& 1 - P(X_1 \geq x, X_2 \geq x, \ldots, X_n \geq x) \\ =& 1 - P(X_1 \geq x) \cdots P(X_n \geq x) \\ =& 1 - (e^{-\lambda x})^n \\ =& 1 - e^{-n\lambda x} \end{aligned}$$

which is the CDF of an $\text{Expo}(n\lambda)$ variable.

(3) Let $A_1, A_2, \ldots, A_m$ be an arbitrary collection of events (we do not assume that the A_i's are either disjoint or independent). Show that

$$P(A_1 \cap A_2 \cap \cdots \cap A_n) \geq \left(\sum_{i=1}^{m} P(A_i)\right) - m + 1$$

by first considering indicator variables and then using the fundamental bridge.

Solution. Let $I_1, \ldots, I_,$ be the indicators for $A_1, A_2, \ldots, A_m$. Observe that if the event $A_1, \ldots, A_n$ occurs, then each of the A_j must occur, and therefore, $\sum_j I_j = m$, and so

$$I_{A_1,\ldots,A_m} = 1 = (\sum I_j) - m + 1$$

On the other hand, if the event $A_1, \ldots, A_m$ does not occur, then at least one of the A_j does not occur, and so $\sum_j I_j < m$. Hence

$$I_{A_1,\ldots,A_m} = 0 \geq (\sum I_j) - m + 1$$

In both cases,

$$I_{A_1,\ldots,A_m} \geq (\sum I_j) - m + 1$$

Taking expected values of both sides of the inequality and using linearity and fundamental bridge, we get

$$P(A_1, \ldots, A_m) \geq \sum_j P(A_j) - m + 1$$

as desired.

(4) An immortal ant wanders along an infinitely long stick that bears striking resemblance to the x-axis. Suppose at the dawn of time, the ant starts at position 0, and that each second thereafter, the ant moves 1cm in either the positive or the negative direction, each with equal probability, independent of the ant's previous movement. Let X_n be the ant's position at time n.

(a) Compute $E[X_n]$.

(b) ~~Compute $E[|X_n|]$.~~

(c) Compute $\text{Var}(X_n)$.

(d) Compute $\text{Corr}(X_n, X_{n-1})$ for $n \geq 2$.

(e) Compute $\mathrm{Corr}(X_n, X_1)$ for $n \geq 2$.

Solution. (a) For each j, let $Y_j = 1$ if the ant moved in the positive direction at time j, and $Y_j = -1$ if the ant moved in the negative direction at time j. Since the ant moves in each direction with equal probability, independent of its previous motion, then $Y_j \sim \mathrm{DUnif}(-1, 1)$, the Y_j's are independent, and $E[Y_j] = 0$. Moreover, $X_n = \sum_{j=1}^n Y_j$ and so by linearity of expectation,

$$E[X_n] = \sum_{j=1}^{n} E[Y_j] = 0$$

(b) **This part is actually trickier than I intended it to be. I'm happy to provide a solution to those interested, but this part is more difficult than problems on the actual take-home exam.**

(c) Since the variance of a sum of independent variables is the sum of their variances, then

$$\mathrm{Var}(X_n) = \sum \mathrm{Var}(Y_j) = n\mathrm{Var}(Y_j) = n$$

since

$$\mathrm{Var}(Y_1) = E[Y_1^2] - (E[Y_1])^2 = 1$$

(d) To find $\mathrm{Corr}(X_n, X_{n-1})$, we first find $\mathrm{Cov}(X_n, X_{n-1})$. Using the Bilinearity property of Covariance:

$$\mathrm{Cov}(X_n, X_{n-1}) = \mathrm{Cov}(Y_n + \sum_{j=1}^{n-1} Y_j, \sum_{j=1}^{n-1} Y_j) = \mathrm{Cov}(Y_n, \sum_{j=1}^{n-1} Y_j) + \mathrm{Var}(X_{n-1}) = n - 1$$

as $\mathrm{Cov}(Y_n, \sum_{j=1}^{n-1} Y_j) = 0$ since Y_n is independent of the other Y_j's.
Then,

$$\mathrm{Corr}(X_n, X_{n-1}) = \frac{\mathrm{Cov}(X_n, X_{n-1})}{\mathrm{SD}(X_n)\mathrm{SD}(X_{n-1}} = \frac{n-1}{\sqrt{n}\sqrt{n-1}} = \frac{\sqrt{n-1}}{\sqrt{n}}$$

which approaches 1 as $n \to \infty$.

(e) Similar to the previous part, to find $\mathrm{Corr}(X_n, X_1)$, we first find $\mathrm{Cov}(X_n, X_1)$. Using the Bilinearity property of Covariance:

$$\mathrm{Cov}(X_n, X_{n-1}) = \mathrm{Cov}(Y_1 + \sum_{j=2}^{n} Y_j, Y_1) = \mathrm{Var}(Y_1) + \mathrm{Cov}(\sum_{j=2}^{n} Y_j, Y_1) = 1$$

since Y_1 is independent of the other Y_j. Therefore,

$$\mathrm{Corr}(X_n, X_1) = \frac{\mathrm{Cov}(X_n, X_1)}{\mathrm{SD}(X_n)\mathrm{SD}(X_1} = \frac{1}{\sqrt{n}}$$

which goes to 0 as $n \to \infty$.

(5) A variable X is said to have the *Gumbel Distribution* if $X = -\log Y$ where $Y \sim \mathrm{Expo}(1)$.

(a) Find the CDF of the Gumbel distribution.

(b) Let $X_1, X_2, \ldots$ be iid Expo(1) and let $M_n = \max\{X_1, \ldots, X_n\}$. Show that $M_n - \log n$ converges in distribution to the Gumbel distribution; that is, show that the CDF of $M_n - \log n$ converges to the Gumbel CDF as $n \to \infty$. *Hint: Recall that* $e^x = (1 + \frac{x}{n})^n$.

Solution. (a) Let $X = -\log Y$ where $Y \sim \mathrm{Expo}(1)$. Using the change-of-variables formula, if $x \geq 0$, then

$$F_X(x) = P(X \leq x) = P(-\log Y \leq x) = P(Y \geq e^{-x}) = e^{-e^{-x}}$$

(b) Let $Y_n = M_n - \log n$. Then

$$\begin{aligned}
F_{Y_n}(x) = P(M_n - \log n \le x) =& P(M_n \le x + \log n)\\
=& P(X_1 \le x + \log n, X_2 \le x + \log n, \ldots, X_n \le x + \log n)\\
=& P(X_1 \le x + \log n) P(X_1 \le x + \log n) \cdots P(X_n \le x + \log n)\\
=& [1 - e^{-(x+\log n)}]^n\\
=& [1 - e^{-x} e^{-\log n)}]^n\\
=& \left[1 - \frac{e^{-x}}{n}\right]^n
\end{aligned}$$

Taking the limit as $n \to \infty$,

$$\lim_{n\to\infty} F_{Y_n}(x) = \lim_{n\to\infty} \left[1 - \frac{e^{-x}}{n}\right]^n = e^{-e^{-x}}$$

as desired.

(6) Suppose $X \sim \mathrm{Bin}(20, 0.25)$.

(a) Use `dbinom` in R to show that the mean X is 5. *Hint: If v and w are vectors, then `sum(v*w)` is the sum of pairwise products of entries of v and w.*

(b) Use `qbinom` in R to show that 5 is a median of X.

(c) Use `dbinom` in R to show that the mode of X is 5. *Note that the function `max(v)` computes the largest value of the vector v and that the function `which.max(v)` finds the position of the maximal value of vector v.*

(d) Verify that X is **not** symmetric by creating a plot of the PMF in R, despite the fact the mean, median and mode of X are all equal.

Solution. (a) Recall that the expected value of X is

$$E[X] = \sum_x x P(X = x)$$

which in R can be computed as

```
sum(0:20 * dbinom(0:20, 20, 0.25 )
## 5
```

(b) The median of X can be found using the quantile function in R:

```
qbinom(0.5, 20, 0.25)
## 5
```

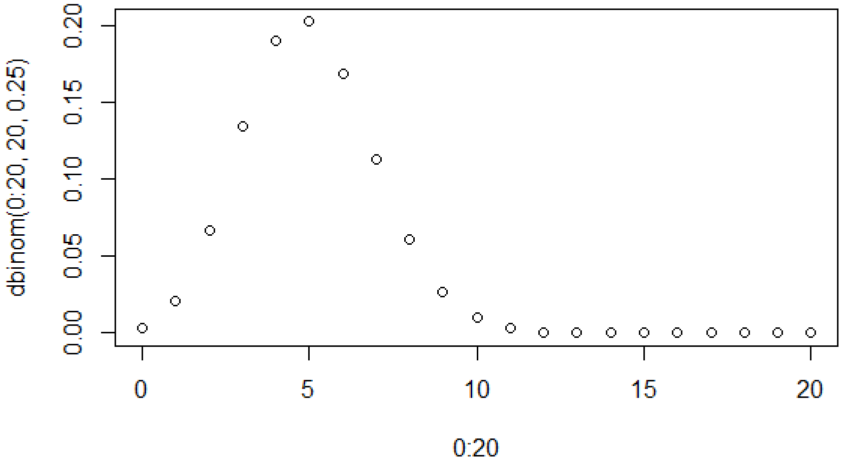
(c) To find the mode in R, we need to find the largest value in the vector `dbinom(0:20, 20, 0.25)`. The `which.max` function will identify the index of the largest component:

```
which.max(dbinom(0:20, 20, 0.25)
## 6
```

However, note that his vector computes probabilities starting at $k = 0$, so the 6th component of the vector is $k = 5$.

(d) We can graph the PMF in R using `plot`:

```
plot(0:20, dbinom(0:20, 20, 0.25))
```


dbinom(0:20, 20, 0.25)
0.00
0.05
0.10
0.15
0.20
0
5
10
15
20
0:20