A. Starting from the "standard" form of each PDF/PMF, show that the following distributions are in an exponential family, and find the corresponding b, c, θ, and $a(\phi)$.

(i) $Y \sim \mathbf{N}(\mu, \sigma^2)$ **for known** σ^2

Let's begin by writing the PDF for the normal distribution:

$$\begin{aligned}
f(y|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y-\mu)^2\right\} \\
&= \exp\left\{-\frac{1}{2\sigma^2}(y^2 - 2y\mu + \mu^2)\right\} \cdot \exp\left\{\log\left((2\pi\sigma^2)^{-1/2}\right)\right\} \\
&= \exp\left\{\frac{y^2}{-2\sigma^2} + y\frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right\} \\
&= \exp\left\{\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} + \left(-\frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right)\right\}
\end{aligned}$$

By letting $\theta = \mu, a(\phi) = \sigma^2, b(\theta) = \frac{1}{2}\mu^2$, and $c(y|\phi) = -\frac{y^2}{2\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)$, we can write the normal distribution with fixed variance in the form of an exponential family:

$$f(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y|\phi)\right\}$$

(ii) $Y = Z/N$**, where** $Z \sim \mathbf{Binom}(N, P)$ **for known** N

We can easily obtain the PMF of Y through a transformation of random variables:

$$\begin{aligned}
P\left(Y = \frac{Z}{N}\right) &= P\left(Z = NY\right) \cdot \left|\frac{1}{N}\right| \\
&= \binom{N}{NY} P^{NY}(1-P)^{N-NY} \cdot \frac{1}{N} \\
&= \exp\left\{\log\left[\binom{N}{NY} P^{NY}(1-P)^{N-NY} \cdot \frac{1}{N}\right]\right\} \\
&= \exp\left\{\log\left[\binom{N}{NY}\right] + NY\log(P) + N\log(1-P) - NY\log(1-P) - \log(N)\right\} \\
&= \exp\left\{Y\left[N\log\left(\frac{P}{1-P}\right)\right] - N\log\left(\frac{1}{1-P}\right) + \log\left[\binom{N}{NY}\right] - \log(N)\right\}
\end{aligned}$$

Let $\theta = N\log\left(\frac{P}{1-P}\right), b(\theta) = N\log\left(\frac{1}{1-P}\right), a(\phi) = 1$, and $c(y|\phi) = \log\left[\binom{N}{NY}\right] - \log(N)$ to get the form of an exponential family.

(iii) $Y \sim \textbf{Pois}(\lambda)$
You know the drill!

$$\begin{aligned}
f(y|\lambda) &= \frac{e^{-\lambda}\lambda^y}{y!} \\
&= \exp\left\{\log\left[\frac{e^{-\lambda}\lambda^y}{y!}\right]\right\} \\
&= \exp\left\{-\lambda + y\log(\lambda) - \log(y!)\right\} \\
&= \exp\left\{y\log(\lambda) - \lambda + (-\log(y!))\right\}
\end{aligned}$$

The above is in the desired exponential family form since we can let $\theta = \log(\lambda), a(\phi) = 1, b(\theta) = \lambda$, and $c(y|\phi) = -\log(y!)$. Therefore, we have shown that we can write the PDF of Y in the desired exponential family form.

B. We want to characterize the mean and variance of a distribution in the exponential family. To do this, we'll take an unfamiliar route, involving a preliminary lemma. Define the score $s(\theta)$ as the gradient of the log-likelihood with respect to θ:

$$s(\theta) = \frac{\partial}{\partial\theta} \log L(\theta).$$

While we think of the score as a function of θ, clearly the score also depends on the data. So a natural question is: what can we say about the distribution of the score over different random realizations of the data under the true data-generating process, i.e., at the true θ? It turns out we can say the following, sometimes referred to as the score equations:

$$\begin{aligned} E\left[s(\theta)\right] &= 0 \\ \mathcal{I}(\theta) \equiv \mathbf{var}(s(\theta)) &= -E\left[H(\theta)\right], \end{aligned}$$

where the mean and variance are taken under the true θ. Prove these score equations.

First, we prove $E\left[s(\theta)\right] = 0$. Note that while the score is a function of θ, it's also dependent on the data y. Therefore, we can take the expected value of the score over the sample space $\mathcal{Y}$. Let's write out the form of $E\left[s(\theta)\right]$:

$$\begin{aligned} E\left[s(\theta)\right] &= \int_{\mathcal{Y}} s(\theta) f(y|\theta) dy \\ &= \int_{\mathcal{Y}} \frac{\partial}{\partial\theta} \log L(\theta) \cdot f(y|\theta) dy \\ &= \int_{\mathcal{Y}} \frac{\frac{\partial}{\partial\theta} L(\theta)}{L(\theta)} \cdot f(y|\theta) dy \end{aligned}$$

Now, we use the statistical trick that we can rewrite the likelihood function as a PDF, since we integrate over $\mathcal{Y}$ with PDF $f(y|\theta)$:

$$\begin{aligned} E\left[s(\theta)\right] &= \int_{\mathcal{Y}} \frac{\frac{\partial}{\partial\theta} f(y|\theta)}{f(y|\theta)} \cdot f(y|\theta) dy \\ &= \int_{\mathcal{Y}} \frac{\partial}{\partial\theta} f(y|\theta) dy \\ &= \frac{\partial}{\partial\theta} \int_{\mathcal{Y}} f(y|\theta) dy \\ &= \frac{\partial}{\partial\theta} (1) = 0, \end{aligned}$$

where we assume that any necessary technical conditions are met to be able to switch the order of integration and differentiation.

Now, we prove that $\text{var}(s(\theta)) = -E\left[H(\theta)\right]$. Using the provided hint, suppose we differentiate the first equation with respect to θ^T:

$$\begin{aligned}
\frac{\partial}{\partial \theta^T} E(s(\theta)) &= \frac{\partial}{\partial \theta^T} \int_{\mathcal{Y}} \frac{\partial}{\partial \theta} \log L(\theta) f(y|\theta) dy \\
&= \int_{\mathcal{Y}} \frac{\partial}{\partial \theta^T} \left[\frac{\partial}{\partial \theta} \log L(\theta) f(y|\theta) \right] dy \\
&= \int_{\mathcal{Y}} \frac{\partial}{\partial \theta} \log L(\theta) \cdot \frac{\partial}{\partial \theta^T} f(y|\theta) + f(y|\theta) \frac{\partial^2}{\partial \theta^T \theta} \log L(\theta) dy \\
&= \int_{\mathcal{Y}} \frac{\partial}{\partial \theta} \log L(\theta) \cdot \frac{\partial}{\partial \theta^T} L(\theta) dy + \int_{\mathcal{Y}} f(y|\theta) \frac{\partial^2}{\partial \theta^T \theta} \log L(\theta) dy \\
&= \int_{\mathcal{Y}} \frac{\partial}{\partial \theta} \log L(\theta) \cdot \frac{\partial}{\partial \theta^T} \log L(\theta) \cdot f(y|\theta) dy + E\left[\frac{\partial^2}{\partial \theta^T \theta} \log L(\theta) \right] \\
&= E\left[\frac{\partial}{\partial \theta} \log L(\theta) \cdot \frac{\partial}{\partial \theta^T} \log L(\theta) \right] + E\left[H(\theta)\right] \\
&= E\left[s(\theta) s(\theta)^T \right] + E\left[H(\theta)\right] \\
&\overset{set}{=} \frac{\partial}{\partial \theta^T}(0) = 0
\end{aligned}$$

Before we get to the big reveal, let's acknowledge the nice property that we used to obtain the fifth equality:

$$\begin{aligned}
\frac{\partial}{\partial \theta} \log L(\theta) &= \frac{\frac{\partial}{\partial \theta} L(\theta)}{f(y|\theta)} \\
\implies \frac{\partial}{\partial \theta} L(\theta) &= \frac{\partial}{\partial \theta} \log L(\theta) \cdot f(y|\theta)
\end{aligned}$$

Now, we see from the above that

$$\begin{aligned}
\text{var}\left[s(\theta)\right] &= E[s(\theta)s(\theta)^T] - (E[s(\theta)])^2 \\
&= E[s(\theta)s(\theta)^T] \\
&= -E[H(\theta)]
\end{aligned}$$

C. Use the score equations you just proved to show that, if $Y \sim f(y|\theta, \phi)$ is an exponential family, then

$$E(Y) = b'(\theta)$$
$$\textbf{var}(Y) = a(\phi)b''(\theta)$$

Thus, the variance of Y is a product of two terms: $b''(\theta)$ depends only on the canonical parameter θ, and hence on the mean, since we showed that $E(Y) = b'(\theta)$; $a(\phi)$ is independent of θ. Note that the most common form of a is $a(\phi) = \phi/w$ where ϕ is called a dispersion parameter and w is a known prior weight that can vary from one observation to another.

Recall that $E[s(\theta)] = E\left[\frac{\partial}{\partial\theta}\log L(\theta)\right] = 0$. Assume without loss of generality that there are n observations. For exponential families, we know that the log-likelihood is

$$\begin{aligned}
\log L(\theta) &= \log\left[\prod_{i=1}^{n}\exp\left\{\frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i|\phi)\right\}\right] \\
&= \sum_{i=1}^{n}\left[\frac{y_i\theta - b(\theta)}{a(\phi)} + c(y_i|\phi)\right] \\
&= \frac{\theta}{a(\phi)}\sum_{i=1}^{n} y_i - \frac{nb(\theta)}{a(\phi)} + \sum_{i=1}^{n} c(y_i|\phi)
\end{aligned}$$

By taking the expectation of the gradient of the log-likelihood with respect to θ, we obtain the following:

$$\begin{aligned}
E[s(\theta)] &= \int_{\mathcal{Y}} \frac{\partial}{\partial\theta}\left[\frac{\theta}{a(\phi)}\sum_{i=1}^{n} y_i - \frac{nb(\theta)}{a(\phi)} + \sum_{i=1}^{n} c(y_i|\phi)\right] f(y|\theta)dy \\
&= \int_{\mathcal{Y}} \left[\frac{1}{a(\phi)}\sum_{i=1}^{n} y_i - \frac{nb'(\theta)}{a(\phi)}\right] f(y|\theta)dy \\
&= E\left[\frac{\sum_{i=1}^{n} y_i}{a(\phi)} - \frac{nb'(\theta)}{a(\phi)}\right] \\
&= \frac{1}{a(\phi)}\sum_{i=1}^{n} E(Y) - \frac{nb'(\theta)}{a(\phi)} \\
&\overset{set}{=} 0
\end{aligned}$$

By manipulating the above equations, we get

$$E(Y) = b'(\theta)$$

Now, let's obtain the variance of Y:

$$\begin{aligned}
\text{var}(s(\theta)) &= \text{var}\left[\frac{1}{a(\phi)}\sum_{i=1}^{n} y_i - \frac{nb'(\theta)}{a(\phi)}\right] \\
&= \frac{1}{a(\phi)^2}\sum_{i=1}^{n}\text{var}(Y) \\
&\stackrel{set}{=} -E\left[H(\theta)\right]
\end{aligned}$$

Note that

$$\begin{aligned}
-E[H(\theta)] &= -E\left[\frac{\partial}{\partial\theta^T}\left(\frac{1}{a(\phi)}\sum_{i=1}^{n} y_i - \frac{nb'(\theta)}{a(\phi)}\right)\right] \\
&= -\int_{\mathcal{Y}}\frac{\partial}{\partial\theta^T}\left(\frac{1}{a(\phi)}\sum_{i=1}^{n} y_i - \frac{nb'(\theta)}{a(\phi)}\right) f(y|\theta)dy \\
&= \int_{\mathcal{Y}}\frac{nb''(\theta)}{a(\phi)} f(y|\theta)dy \\
&= E\left[\frac{nb''(\theta)}{a(\phi)}\right] \\
&= \frac{nb''(\theta)}{a(\phi)}
\end{aligned}$$

By combining the two above derivations, we see that

$$\begin{aligned}
\frac{1}{a(\phi)^2}\sum_{i=1}^{n}\text{var}(Y) &= \frac{nb''(\theta)}{a(\phi)} \\
\implies \text{var}(Y) &= a(\phi)b''(\theta)
\end{aligned}$$

D. To convince yourself that your result in (C) is correct, use these results to compute the mean and variance of the N(μ, σ^2) distribution.

Recall from (a) that $\theta = \mu, a(\phi) = \sigma^2$, and $b(\theta) = \frac{1}{2}\mu^2$. While (a) *did* assume that σ^2 was known, we see that the result still holds! From (c), we found that

$$\begin{aligned}
E(Y) &= b'(\theta) \\
&= \frac{\partial}{\partial \mu}\left(\frac{1}{2}\mu^2\right) \\
&= \mu, \\
\mathrm{var}(Y) &= a(\phi)b''(\theta) \\
&= \sigma^2 \frac{\partial^2}{\partial \mu^2}\left(\frac{1}{2}\mu^2\right) \\
&= \sigma^2 \frac{\partial}{\partial \mu}(\mu) \\
&= \sigma^2,
\end{aligned}$$

which are certainly the mean and variance of a N(μ, σ^2) distribution.

A. Deduce from your results above that, in a GLM,

$$\theta_i = (b')^{-1}\left(g^{-1}(x_i^T\beta)\right),$$
$$\mathbf{var}(Y_i) = \frac{\phi}{w_i}V(\mu_i)$$

for some function V that you should specify in terms of the building blocks of the exponential family model. V is often referred to as the variance function, since it explicitly relates the mean and the variance in a GLM.

Let's start with proving the first equation. Recall that $E(Y_i) = b'(\theta_i)$ and, by definition of GLM, $E(Y_i) = \mu_i$. Additionally, by definition, $g(\mu_i) = x_i^T\beta$. Therefore, we can simply equate these equations in the following way:

$$\begin{aligned} E(Y_i) &= b'(\theta_i) = \mu_i \\ \mu_i &= g^{-1}(x_i^T\beta), \\ \implies b'(\theta_i) &= g^{-1}(x_i^T\beta), \\ \implies \theta_i &= (b')^{-1}\left(g^{-1}(x_i^T\beta)\right), \end{aligned}$$

which is the first equation.

Now, we want to prove the second equation, containing the variance of Y_i. Recall that $\text{var}(Y_i) = a(\phi)b''(\theta)$. In the formulation of the GLM, we see that $a(\phi) = \frac{\phi}{w_i}$. Therefore,

$$\begin{aligned} \text{var}(Y_i) &= \frac{\phi}{w_i}b''(\theta_i) \\ &= \frac{\phi}{w_i}b''\left((b')^{-1}\left(g^{-1}(x_i^T\beta)\right)\right) \\ &= \frac{\phi}{w_i}b''\left((b')^{-1}(\mu_i)\right) \end{aligned}$$

By letting $V(\mu_i) = b''\left((b')^{-1}(\mu_i)\right)$, we get the second equation. Notice how we wrote $V(\mu_i)$ as a function of μ_i using functions of the exponential family (i.e., function $b(\cdot)$).

B. Take two special cases. (1) Suppose that Y is a Poisson GLM, i.e., that the stochastic component of the model is a Poisson distribution. Show that $V(\mu) = \mu$. (2) Suppose that $Y = Z/N$ is a Binomial GLM, i.e., that the stochastic component of the model is a Binomial distribution $Z \sim$ Binom(N, P) and that Y is the fraction of yes outcomes. Show that $V(\mu) = \mu(1-\mu)$.

First special case:

We can always write the stochastic component of the model in the following form:

$$f(y_i|\theta_i, \phi_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi/w_i} + c(y_i|\phi/w_i)\right\},$$

which resembles an exponential family where $a(\phi) = \frac{\phi}{w_i}$. Note that we can write the Poisson PMF as

$$\begin{aligned}
f(y_i|\lambda_i) &= \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} \\
&= \exp\left\{\log\left[\frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}\right]\right\} \\
&= \exp\left\{-\lambda_i + y_i\log(\lambda_i) - \log(y_i!)\right\},
\end{aligned}$$

which is an exponential family where $\theta_i = \log(\lambda_i)$, $b(\theta_i) = \lambda_i = e^{\theta_i}$, $a(\phi) = \frac{\phi}{w_i} = 1$, and $c(y_i|\phi/w_i) = -\log(y_i!)$. Recall that, for an exponential family,

$$\begin{aligned}
E(Y_i) &= b'(\theta_i), \\
\text{var}(Y_i) &= a(\phi)b''(\theta_i),
\end{aligned}$$

so we easily obtain

$$\begin{aligned}
E(Y_i) &= b'(\theta_i) \\
&= e^{\theta_i} \\
&= \lambda_i, \\
\text{var}(Y_i) &= a(\phi)b''(\theta) \\
&= \frac{\partial}{\partial\theta_i}(e^{\theta_i}) \\
&= \lambda_i,
\end{aligned}$$

By definition, $\mu_i = E(Y_i)$, so $\mu_i = \lambda_i$. Now, we set $V(\mu_i) = b''(\theta)$ to satisfy $\text{var}(Y_i) = \frac{\phi}{w_i}V(\mu_i)$. Therefore, $V(\mu_i) = \mu_i$.

Second special case:

If the stochastic component is $Z \sim \text{Binom}(N, P)$, then we can write

$$f(z_i|\theta_i, \phi_i) = \exp\left\{\frac{z_i\theta_i - b(\theta_i)}{\phi/w_i} + c(z_i|\phi/w_i)\right\}$$

Note that since Z follows a binomial distribution, we can write its PMF as

$$\begin{aligned}
f(z_i|\theta_i, \phi_i) &= \binom{N}{z_i} P^{z_i}(1-P)^{N-z_i} \\
&= \exp\left\{\log\left[\binom{N}{z_i} P^{z_i}(1-P)^{N-z_i}\right]\right\} \\
&= \exp\left\{z_i \log(P) - z_i \log(1-P) + N\log(1-P) + \log\left[\binom{N}{z_i}\right]\right\}
\end{aligned}$$

Note that we can only write theis PMF in the form of an exponential family for the GLM when N is known. So, let's assume N is known to show the PMF of z_i can be written in the desired exponential family form:

$$\begin{aligned}
f(z_i|\theta_i, \phi_i) &= \exp\left\{z_i \log(P) - z_i \log(1-P) + N\log(1-P) + \log\left[\binom{N}{z_i}\right]\right\} \\
&= \exp\left\{z_i \log\left(\frac{P}{1-P}\right) - N\log\left(\frac{1}{1-P}\right) + \log\left[\binom{N}{z_i}\right]\right\},
\end{aligned}$$

where $a(\phi) = \frac{\phi}{w_i} = 1$, $\theta_i = \log\left(\frac{P}{1-P}\right)$, $b(\theta_i) = N\log\left(\frac{1}{1-P}\right)$, and $c(z_i|\phi/w_i) = \log\left[\binom{N}{z_i}\right]$. Once again, we find the form of $V(\mu_i)$ by setting it equal to $b''(\theta_i)$. Note that

$$\begin{aligned}
\theta_i &= \log\left(\frac{P}{1-P}\right) \\
\implies e^{\theta_i} &= \frac{P}{1-P} \\
\implies e^{\theta_i} &= P + Pe^{\theta_i} \\
\implies P(1+e^{\theta_i}) &= e^{\theta_i} \\
\implies P &= \frac{e^{\theta_i}}{1+e^{\theta_i}}
\end{aligned}$$

Knowing P, we can obtain the proper form for $b(\theta_i)$:

$$b(\theta_i) = N\log\left(1 + e^{\theta_i}\right)$$

Now, we can obtain $b''(\theta_i)$:

$$\begin{aligned}
b'(\theta_i) &= \frac{N}{1+e^{\theta_i}} \cdot e^{\theta_i} \\
b''(\theta_i) &= \frac{N}{1+e^{\theta_i}} \cdot e^{\theta_i} - \left(e^{\theta_i}\right)^2 \frac{N}{(1+e^{\theta_i})^2}
\end{aligned}$$

We can find μ_i with

$$\begin{aligned}
E(Z_i) &= b'(\theta_i) \\
&= \frac{N}{1 + \frac{P}{1-P}} \cdot \frac{P}{1-P} \\
&= (1-P)N \cdot \frac{P}{1-P} \\
&= NP \\
&\stackrel{set}{=} \mu,
\end{aligned}$$

which we set to μ by definition of the GLM. Now, we can find $b''(\theta_i)$ in terms of $\mu = NP$:

$$\begin{aligned}
b''(\theta_i) &= NP - \left(\frac{P}{1-P}\right)^2 (1-P)^2 N \\
&= NP - NP^2 \\
&= NP(1-P)
\end{aligned}$$

Now, notice that all of the above work was in terms of $z_i \sim \text{Binom}(N, P)$; however, this question asks us to show that $V(\mu) = b''(\theta_i) = \mu(1-\mu)$ for $Y = \frac{Z}{N}$. We can simply divide $b''(\theta_i)$ by N to obtain the corresponding $b''(\theta_i)$ for Y. Therefore, $V(\mu) = P(1-P) = \mu(1-\mu)$.

C. To specify a GLM, we must choose the link function $g(\mu_i)$. Recall that g links the predictors with the mean of the response: $g(\mu_i) = x_i^T \beta$. Since you've shown that

$$\theta_i = (b')^{-1} \left\{ g^{-1}(x_i^T \beta) \right\},$$

a "simple" choice of link function is one where $g^{-1} = b'$. This is known as the canonical link, in which case the canonical parameter simplifies to $\theta_i = x_i^T \beta$. So under the canonical link $g(\mu) = (b')^{-1}(\mu)$, we have the model

$$f(y_i|\beta, \phi) = \exp\left\{ \frac{y_i x_i^T \beta - b(x_i^T \beta)}{\phi / w_i} + c(y_i|\phi/w_i) \right\}.$$

Now return to the two special cases from the previous problem and find the canonical link $g(\mu)$.

First special case:
Recall that $b(\theta_i) = e^{\theta_i}$. We want to find $g(\mu)$ that satisfies $g(\mu) = (b')^{-1}(\mu)$. Note that $b'(\theta_i) = e^{\theta_i}$, so

$$\begin{aligned} g(\mu) &= (b')^{-1}(\mu) \\ \implies b'(g(\mu)) &= \mu \\ \implies e^{g(\mu)} &= \mu \\ \implies g(\mu) &= \log(\mu) \end{aligned}$$

Therefore, the canonical link is the log link, i.e., $g(\mu) = \log(\mu)$.
Second special case:
In this case, we found that

$$\begin{aligned} b(\theta_i) &= \log\left(1 + e^{\theta_i}\right), \\ b'(\theta_i) &= \frac{1}{1 + e^{\theta_i}} \cdot e^{\theta_i}. \end{aligned}$$

We want to find $g(\mu)$ that satisfies $g(\mu) = (b')^{-1}(\mu)$:

$$\begin{aligned} b'(g(\mu)) &= \mu \\ \implies \frac{e^{g(\mu)}}{1 + e^{g(\mu)}} &= \mu \\ \implies e^{g(\mu)} &= \mu + \mu e^{g(\mu)} \\ \implies e^{g(\mu)} &= \frac{\mu}{1 - \mu} \end{aligned}$$

Therefore, our canonical link is $g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$.

A.