Exponential Families

Last modified: 2026-11-12 22:02:32 (PST)

Many of the distributions in this course, such as the Bernoulli, Poisson and Normal distributions, share one algebraic form for their probability mass function or density. Distributions with that form are called an exponential family, and they share properties that make estimation easier. This page defines the form and its natural parameter, and checks the form for the Poisson distribution on data.

The exponential family form

Definition 0.1 (Exponential family)**.** Let X be a random variable whose distribution depends on one parameter η, and let $\mathbb{f}(x;\eta)$ be its probability mass function or density. The distributions of X form an **exponential family** if there are a nonnegative function h and a function T of x alone, and a function A of η alone, such that

$$\mathbb{f}(x;\eta) = h(x)\exp\{\eta\, T(x) - A(\eta)\} \tag{1}$$

for every value x and every allowed η.

Remark (Source: Exponential family)

Remark 0.1 (Source: Exponential family). Casella and Berger [1, Sec. 3.4], Eq. 3.4.1 write the family as a sum of k terms inside the exponential, each a function of the parameter times a function of x; Dobson and Barnett [2, Sec. 3.2], Eq. 3.3 write it in the form $\exp\{a(y)b(\theta) + c(\theta) + d(y)\}$. Equation 1 is the case of one term. It is Dobson and Barnett's canonical form when $T(x) = x$, and the natural parameter η plays the role of their $b(\theta)$. See the notation choices for how the symbols line up.

Example 0.1 (The Poisson distribution is an exponential family)**.** Let $X \sim \text{Pois}(\lambda)$, the Poisson distribution with rate $\lambda > 0$. Its probability mass function has one factor free of λ, and the factors λ^x and $\exp\{-\lambda\}$. We rewrite it in the form of Equation 1 by writing λ^x as an exponential of a logarithm:

$$
\begin{aligned}
\mathbb{f}(x;\lambda) &= \frac{\lambda^x \exp\{-\lambda\}}{x!} && \text{(Poisson mass function)} \\
&= \frac{1}{x!} \exp\{\log\{\lambda^x\}\} \exp\{-\lambda\} && \text{(a positive number is the exponential of its logarithm)} \\
&= \frac{1}{x!} \exp\{x \log\{\lambda\}\} \exp\{-\lambda\} && \text{(the logarithm of a power is the exponent times the logarithm)} \\
&= \frac{1}{x!} \exp\{x \log\{\lambda\} - \lambda\} && \text{(the product of exponentials is the exponential of the sum)}
\end{aligned}
$$

So the Poisson distributions are an exponential family with $h(x) = 1/x!$, $T(x) = x$, natural parameter $\eta = \log\{\lambda\}$, and $A(\eta) = \lambda = \exp\{\eta\}$.
To check the algebra on data, we use the number of tropical cyclones in northeastern Australia in each of 13 seasons [2, Sec. 1.6.5], with the rate set to the sample mean. We evaluate both sides of Equation 1 at every observed count.

R

```r
cyclones <- c(6, 5, 4, 6, 6, 3, 12, 7, 4, 2, 6, 7, 4)
rate <- mean(cyclones)

base_measure <- function(x) 1 / factorial(x)
natural_parameter <- log(rate)
log_partition <- function(eta) exp(eta)

exponential_form <- base_measure(cyclones) *
  exp(natural_parameter * cyclones - log_partition(natural_parameter))

data.frame(
  count = cyclones,
  poisson = dpois(cyclones, rate),
  exponential_form = exponential_form
)
#>    count    poisson exponential_form
#> 1      6 0.15764459       0.15764459
#> 2      5 0.17078164       0.17078164
#> 3      4 0.15417787       0.15417787
#> 4      6 0.15764459       0.15764459
#> 5      6 0.15764459       0.15764459
#> 6      3 0.11135069       0.11135069
#> 7     12 0.00683927       0.00683927
#> 8      7 0.12472979       0.12472979
#> 9      4 0.15417787       0.15417787
#> 10     2 0.06031496       0.06031496
#> 11     6 0.15764459       0.15764459
#> 12     7 0.12472979       0.12472979
#> 13     4 0.15417787       0.15417787
max(abs(dpois(cyclones, rate) - exponential_form))
#> [1] 1.11022e-16
```

Python

```python
import numpy as np
import pandas as pd
from scipy.special import factorial
from scipy.stats import poisson

cyclones = np.array([6, 5, 4, 6, 6, 3, 12, 7, 4, 2, 6, 7, 4])
rate = cyclones.mean()

natural_parameter = np.log(rate)
exponential_form = (1 / factorial(cyclones)) * np.exp(
    natural_parameter * cyclones - np.exp(natural_parameter)
)

table = pd.DataFrame(
    {
        "count": cyclones,
        "poisson": poisson.pmf(cyclones, rate),
        "exponential_form": exponential_form,
    }
)
print(table)
#>     count   poisson  exponential_form
#> 0       6  0.157645          0.157645
#> 1       5  0.170782          0.170782
#> 2       4  0.154178          0.154178
#> 3       6  0.157645          0.157645
#> 4       6  0.157645          0.157645
#> 5       3  0.111351          0.111351
#> 6      12  0.006839          0.006839
#> 7       7  0.124730          0.124730
#> 8       4  0.154178          0.154178
#> 9       2  0.060315          0.060315
#> 10      6  0.157645          0.157645
#> 11      7  0.124730          0.124730
#> 12      4  0.154178          0.154178
print(np.max(np.abs(poisson.pmf(cyclones, rate) - exponential_form)))
#> 5.551115123125783e-17
```

The two columns agree to within rounding error: the largest difference is 1.1e-16.

Definition 0.2 (Natural parameter)**.** In the form of Equation 1 of an exponential family, the parameter η that multiplies the statistic $T(x)$ inside the exponential is the **natural parameter**, also called the **canonical parameter** [2, Sec. 3.2].

Example 0.2 (The natural parameter of the Poisson distribution is the log rate)**.** In Example 0.1, the Poisson rate λ and the natural parameter η are related by

$$\eta = \log\{\lambda\} \qquad \text{and} \qquad \lambda = \exp\{\eta\}$$

For the cyclone data, the rate set to the sample mean is 5.538 cyclones per season, so the natural parameter is $\eta = 1.712$. The natural parameter can be any real number, while the rate must be greater than zero.

Example 0.3 (The Bernoulli distribution is an exponential family)**.**
Let $X \sim \text{Bernoulli}(\pi)$, the Bernoulli distribution with success probability $\pi \in (0, 1)$. Its probability mass function is $\pi^x(1-\pi)^{1-x}$ for $x \in \{0, 1\}$. We rewrite it in the form of Equation 1 by collecting the factors that have the exponent x:

$$\begin{aligned}
\mathbb{f}(x;\pi) &= \pi^x(1-\pi)^{1-x} && \text{(Bernoulli mass function)} \\
&= \pi^x(1-\pi)^{-x}(1-\pi) && \text{(a power with a sum of exponents is the product of the powe} \\
&= \left(\frac{\pi}{1-\pi}\right)^x(1-\pi) && \text{(powers with the same exponent combine)} \\
&= \exp\left\{\log\left\{\left(\frac{\pi}{1-\pi}\right)^x\right\}\right\}(1-\pi) && \text{(a positive number is the exponential of its logarithm)} \\
&= \exp\left\{x\log\left\{\frac{\pi}{1-\pi}\right\}\right\}(1-\pi) && \text{(the logarithm of a power is the exponent times the logarithı} \\
&= \exp\left\{x\log\left\{\frac{\pi}{1-\pi}\right\}\right\}\exp\{\log\{1-\pi\}\} && \text{(a positive number is the exponential of its logarithm)} \\
&= \exp\left\{x\log\left\{\frac{\pi}{1-\pi}\right\} + \log\{1-\pi\}\right\} && \text{(the product of exponentials is the exponential of the sum)}
\end{aligned}$$

The statistic is $T(x) = x$, and $h(x) = 1$. The natural parameter is the log odds, $\eta = \log\{\pi/(1-\pi)\}$. Inverting it gives $\pi = \exp\{\eta\}/(1+\exp\{\eta\})$, so $1-\pi = 1/(1+\exp\{\eta\})$, and the last term in the exponent is $\log\{1-\pi\} = -\log\{1+\exp\{\eta\}\}$. Hence the Bernoulli distributions are an exponential family with $A(\eta) = \log\{1+\exp\{\eta\}\}$.
To check the algebra on data, we use the `birthwt` data, which record the birth weight of babies born at Baystate Medical Center in Springfield, Massachusetts [3], as distributed in R's `MASS` package [4]. The variable `low` is 1 for a baby with low birth weight and 0 otherwise. We set the success probability to the sample mean of `low`, and evaluate both sides of Equation 1 at the two possible values of x.

R

```r
low <- MASS::birthwt$low
prob <- mean(low)

natural_parameter <- log(prob / (1 - prob))
exp_natural_parameter <- exp(natural_parameter)
log_partition <- log(1 + exp_natural_parameter)

x <- c(0, 1)
exponent <- natural_parameter * x - log_partition
exponential_form <- exp(exponent)
bernoulli <- dbinom(x, size = 1, prob = prob)

abs_difference <- abs(bernoulli - exponential_form)
max_difference <- max(abs_difference)

tibble::tibble(x, bernoulli, exponential_form)
#> # A tibble: 2 x 3
#>       x bernoulli exponential_form
#>   <dbl>     <dbl>            <dbl>
#> 1     0     0.688            0.688
#> 2     1     0.312            0.312
max_difference
#> [1] 1.11022e-16
```

Python

```
import numpy as np
import pandas as pd
from scipy.stats import bernoulli

birthwt = pd.read_csv("data/birthwt.csv")
prob = birthwt["low"].mean()

natural_parameter = np.log(prob / (1 - prob))
log_partition = np.log(1 + np.exp(natural_parameter))

x = np.array([0, 1])
exponential_form = np.exp(natural_parameter * x - log_partition)

table = pd.DataFrame(
    {
        "x": x,
        "bernoulli": bernoulli.pmf(x, prob),
        "exponential_form": exponential_form,
    }
)
print(table)
#>    x  bernoulli  exponential_form
#> 0  0   0.687831          0.687831
#> 1  1   0.312169          0.312169
print(np.max(np.abs(bernoulli.pmf(x, prob) - exponential_form)))
#> 2.220446049250313e-16
```

The value of `prob` is 0.312, so the natural parameter is $\eta = -0.79$. The two columns agree to within rounding error: the largest difference is 1.1e-16.

Example 0.4 (The Gaussian distribution with known variance is an exponential family)**.**

Let $X \sim \mathrm{N}(\mu, \sigma^2)$, the Gaussian distribution with mean $\mu \in \mathbb{R}$ and a variance $\sigma^2 > 0$ that we treat as known. Its density is $(2\pi\sigma^2)^{-1/2}\exp\{-(x-\mu)^2/(2\sigma^2)\}$. We rewrite it in the form of Equation 1 by expanding the square in the exponent and separating the factors that involve μ:

$$
\begin{aligned}
\mathbb{f}(x;\mu) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} && \text{(Gaussian density)} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{x^2 - 2x\mu + \mu^2}{2\sigma^2}\right\} && \text{(the square of a difference)} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{x^2}{2\sigma^2} + \frac{x\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2}\right\} && \text{(divide each term by the denominator)} \\
&= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{x^2}{2\sigma^2}\right\} \exp\left\{\frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2}\right\} && \text{(the exponential of a sum is the product of the exponentia}
\end{aligned}
$$

The statistic is $T(x) = x$, and $h(x) = \exp\{-x^2/(2\sigma^2)\}/\sqrt{2\pi\sigma^2}$, which does not involve μ. The natural parameter is $\eta = \mu/\sigma^2$. Inverting it gives $\mu = \sigma^2\eta$, so $\mu^2/(2\sigma^2) = \sigma^2\eta^2/2$. Hence the Gaussian distributions with a known variance are an exponential family with $A(\eta) = \sigma^2\eta^2/2$.

To check the algebra on data, we use the birth weights in grams from the `birthwt` data [3], as distributed in R's `MASS` package [4]. We set μ to the sample mean and σ^2 to the sample variance of the birth weights, and evaluate both sides of Equation 1 at every observed birth weight.

R

```r
bwt <- MASS::birthwt$bwt
mean_bwt <- mean(bwt)
sd_bwt <- sd(bwt)
variance <- sd_bwt^2

natural_parameter <- mean_bwt / variance
log_partition <- variance * natural_parameter^2 / 2

base_exponent <- -bwt^2 / (2 * variance)
base_scale <- sqrt(2 * pi * variance)
base_measure <- exp(base_exponent) / base_scale

exponent <- natural_parameter * bwt - log_partition
exponential_form <- base_measure * exp(exponent)
normal <- dnorm(bwt, mean = mean_bwt, sd = sd_bwt)

relative_difference <- abs(normal - exponential_form) / normal
max_relative_difference <- max(relative_difference)

tibble::tibble(bwt, normal, exponential_form)
#> # A tibble: 189 x 3
#>      bwt   normal exponential_form
#>    <int>    <dbl>            <dbl>
#>  1  2523 0.000463         0.000463
#>  2  2551 0.000473         0.000473
#>  3  2557 0.000475         0.000475
#>  4  2594 0.000487         0.000487
#>  5  2600 0.000489         0.000489
#>  6  2622 0.000496         0.000496
#>  7  2637 0.000501         0.000501
#>  8  2637 0.000501         0.000501
#>  9  2663 0.000508         0.000508
#> 10  2665 0.000508         0.000508
#> # i 179 more rows
max_relative_difference
#> [1] 3.52031e-15
```

Python

```python
import numpy as np
import pandas as pd
from scipy.stats import norm

birthwt = pd.read_csv("data/birthwt.csv")
bwt = birthwt["bwt"]
mean_bwt = bwt.mean()
variance = bwt.var(ddof=1)

natural_parameter = mean_bwt / variance
log_partition = variance * natural_parameter**2 / 2

base_measure = np.exp(-bwt**2 / (2 * variance)) / np.sqrt(2 * np.pi * variance)
exponential_form = base_measure * np.exp(natural_parameter * bwt - log_partition)
normal = norm.pdf(bwt, loc=mean_bwt, scale=np.sqrt(variance))

table = pd.DataFrame(
    {"bwt": bwt, "normal": normal, "exponential_form": exponential_form}
)
print(table.head())
#>     bwt    normal  exponential_form
#> 0  2523  0.000463          0.000463
#> 1  2551  0.000473          0.000473
#> 2  2557  0.000475          0.000475
#> 3  2594  0.000487          0.000487
#> 4  2600  0.000489          0.000489
print(np.max(np.abs(normal - exponential_form) / normal))
#> 3.324739754359985e-15
```

For these data, $\eta = 0.00554$ per gram. The two columns agree to within rounding error: the largest relative difference is 3.5e-15.

Example 0.5 (The uniform distribution on $[0, \beta]$ is not an exponential family)**.** Let $X \sim \text{Uniform}(0, \beta)$, the uniform distribution on the interval from 0 to $\beta > 0$. Its density is

$$\mathbb{f}(x; \beta) = \begin{cases} \dfrac{1}{\beta}, & 0 \le x \le \beta \\ 0, & \text{otherwise.} \end{cases}$$

Suppose these densities were those of an exponential family, indexed by some parameter η that determines β. Then Equation 1 would give

$$\mathbb{f}(x;\eta) = h(x)\exp\{\eta\,T(x) - A(\eta)\}.$$

The exponential function is positive for every η and every x, so $\mathbb{f}(x;\eta) > 0$ exactly when $h(x) > 0$. The set of values x with a positive density would then be the same for every parameter value.

The uniform density does depend on it. At $x = 3/2$, the density is $1/2$ when $\beta = 2$ and 0 when $\beta = 1$. So the uniform densities on $[0, \beta]$ are not an exponential family. The set of values x with a positive density depends on the parameter, which the form of Equation 1 does not allow.

The log-partition function

Exercise 0.1 (The exponential of the log-partition function is a normalizing integral)**.** Let X be a continuous random variable whose density has the form of Equation 1. Use the fact that a density integrates to one to show that

$$\exp\{A(\eta)\} = \int h(x)\exp\{\eta\,T(x)\}\,dx.$$

For a discrete random variable, replace the integral with a sum over the possible values.

Solution

Solution 0.1. We start from the fact that the density integrates to one, and isolate the factor $\exp\{-A(\eta)\}$ that does not depend on x:

$$\begin{aligned}
1 &= \int \mathbb{f}(x;\eta)\,dx && \text{(a density integrates to one)}\\
&= \int h(x)\exp\{\eta\,T(x) - A(\eta)\}\,dx && \text{(definition of an exponential family)}\\
&= \int h(x)\exp\{\eta\,T(x)\}\exp\{-A(\eta)\}\,dx && \text{(the exponential of a sum is the product of the exponentials)}\\
&= \exp\{-A(\eta)\}\int h(x)\exp\{\eta\,T(x)\}\,dx && \text{(a factor free of the integration variable moves out of the integral)}
\end{aligned}$$

Multiplying both sides by $\exp\{A(\eta)\}$ gives

$$\begin{aligned}
\exp\{A(\eta)\} &= \exp\{A(\eta)\}\exp\{-A(\eta)\}\int h(x)\exp\{\eta T(x)\}\,dx && \text{(multiply both sides by } \exp\{A(\eta)\}\text{)} \\
&= \exp\{A(\eta)-A(\eta)\}\int h(x)\exp\{\eta T(x)\}\,dx && \text{(the product of exponentials is the exponential} \\
&= \exp\{0\}\int h(x)\exp\{\eta T(x)\}\,dx && \text{(a number minus itself is zero)} \\
&= \int h(x)\exp\{\eta T(x)\}\,dx && \text{(the exponential of zero is one)}
\end{aligned}$$

Exercise 0.2 (The derivative of the log-partition function is the mean of the statistic)**.** In the setting of Exercise 0.1, differentiate both sides of

$$\exp\{A(\eta)\} = \int h(x)\exp\{\eta T(x)\}\,dx$$

with respect to η. Assume that the derivative of the integral is the integral of the derivative. Show that

$$\frac{dA(\eta)}{d\eta} = \mathbb{E}[T(X)].$$

Solution

Solution 0.2. The left side of Exercise 0.1 is a function of η alone, and its derivative is

$$\frac{d\exp\{A(\eta)\}}{d\eta} = \exp\{A(\eta)\}\,\frac{dA(\eta)}{d\eta} \quad \text{(chain rule)}$$

The derivative of the right side moves inside the integral, where only the exponential function depends on η:

$$\begin{aligned}
\frac{d}{d\eta}\int h(x)\exp\{\eta T(x)\}\,dx &= \int h(x)\,\frac{d\exp\{\eta T(x)\}}{d\eta}\,dx && \text{(the derivative of the integral is the integral of the} \\
&= \int h(x)\exp\{\eta T(x)\}\,T(x)\,dx && \text{(chain rule)} \\
&= \int T(x)\,h(x)\exp\{\eta T(x)\}\,dx && \text{(multiplication is commutative)}
\end{aligned}$$

The two derivatives are equal, because the two sides of Exercise 0.1 are equal:

$$\exp\{A(\eta)\}\,\frac{dA(\eta)}{d\eta} = \int T(x)\,h(x)\exp\{\eta T(x)\}\,dx.$$

Solving for the derivative of the log-partition function:

$$\begin{aligned}
\frac{dA(\eta)}{d\eta} &= \frac{1}{\exp\{A(\eta)\}} \int T(x)\, h(x) \exp\{\eta\, T(x)\}\, dx && \text{(divide both sides by } \exp\{A(\eta)\}\text{)} \\
&= \exp\{-A(\eta)\} \int T(x)\, h(x) \exp\{\eta\, T(x)\}\, dx && \text{(the reciprocal of an exponential is the exponential of th} \\
&= \int \exp\{-A(\eta)\}\, T(x)\, h(x) \exp\{\eta\, T(x)\}\, dx && \text{(a factor free of the integration variable moves into the i} \\
&= \int T(x)\, h(x) \exp\{\eta\, T(x)\} \exp\{-A(\eta)\}\, dx && \text{(multiplication is commutative)} \\
&= \int T(x)\, h(x) \exp\{\eta\, T(x) - A(\eta)\}\, dx && \text{(the product of exponentials is the exponential of the su} \\
&= \int T(x)\, \mathbb{f}(x;\eta)\, dx && \text{(definition of an exponential family)} \\
&= \mathbb{E}[T(X)] && \text{(definition of the expectation)}
\end{aligned}$$

Theorem 0.1 (The mean of the statistic is the derivative of the log-partition function)**.** *Let X have a distribution in an exponential family with natural parameter η, statistic T, and log-partition function A, and let η be in the interior of the allowed values. Then*

$$\mathbb{E}[T(X)] = \frac{dA(\eta)}{d\eta}.$$

Proof

Proof. The solution to Exercise 0.2, worked in Solution 0.2, proves the statement. It uses the solution to Exercise 0.1, worked in Solution 0.1. □

Example 0.1 (The mean of the Poisson distribution from its log-partition function)**.** In Example 0.1, the statistic is $T(x) = x$ and the log-partition function is $A(\eta) = \exp\{\eta\}$. By Theorem 0.1,

$$\mathbb{E}[X] = \mathbb{E}[T(X)] = \frac{dA(\eta)}{d\eta} = \exp\{\eta\}.$$

Since $\exp\{\eta\} = \lambda$, we have $\mathbb{E}[X] = \lambda$, the mean of the Poisson distribution.
To check Theorem 0.1 on data, we compute its two sides separately for the cyclone counts of Example 0.1, with λ set to the sample mean. The left side adds $x\, \mathbb{f}(x;\lambda)$ over the counts from 0 to 100, and the right side is a central difference of $A(\eta)$ at $\eta = \log\{\lambda\}$.

R

```
cyclones <- c(6, 5, 4, 6, 6, 3, 12, 7, 4, 2, 6, 7, 4)
rate <- mean(cyclones)
natural_parameter <- log(rate)

counts <- 0:100
weighted_masses <- counts * dpois(counts, rate)
mean_statistic <- sum(weighted_masses)

log_partition <- exp
step <- 1e-6
log_partition_up <- log_partition(natural_parameter + step)
log_partition_down <- log_partition(natural_parameter - step)
derivative <- (log_partition_up - log_partition_down) / (2 * step)

tibble::tibble(mean_statistic, derivative)
#> # A tibble: 1 x 2
#>   mean_statistic derivative
#>            <dbl>      <dbl>
#> 1           5.54       5.54
```

Python

```python
import numpy as np
import pandas as pd
from scipy.stats import poisson

cyclones = np.array([6, 5, 4, 6, 6, 3, 12, 7, 4, 2, 6, 7, 4])
rate = cyclones.mean()
natural_parameter = np.log(rate)

counts = np.arange(0, 101)
mean_statistic = np.sum(counts * poisson.pmf(counts, rate))

log_partition = np.exp
step = 1e-6
log_partition_up = log_partition(natural_parameter + step)
log_partition_down = log_partition(natural_parameter - step)
derivative = (log_partition_up - log_partition_down) / (2 * step)

pd.DataFrame({"mean_statistic": [mean_statistic], "derivative": [derivative]})
#>    mean_statistic  derivative
#> 0        5.538462    5.538462
```

The two sides agree at 5.538, and differ by 1.8e-10. That difference is numerical error.

Exercise 0.3 (The second derivative of the log-partition function is the variance of the statistic)**.** In the setting of Exercise 0.2, differentiate both sides of

$$\frac{dA(\eta)}{d\eta} = \int T(x)\, \mathbb{f}(x;\eta)\, dx$$

with respect to η. Assume that the derivative of the integral is the integral of the derivative. Use Theorem 0.1 and the second moment to show that

$$\frac{d^2 A(\eta)}{d\eta^2} = \mathrm{Var}(T(X)).$$

Solution

Solution 0.3. The left side of Exercise 0.3 differentiates to the second derivative of $A(\eta)$. On the right side, the derivative moves inside the integral, where only the density depends on η:

$$\frac{d}{d\eta}\int T(x)\,\mathbb{f}(x;\eta)\,dx = \int T(x)\,\frac{d\mathbb{f}(x;\eta)}{d\eta}\,dx \quad \text{(the derivative of the integral is the integral of the derivative)}$$

The derivative of the density follows from its form in Equation 1:

$$\begin{aligned}
\frac{d\mathbb{f}(x;\eta)}{d\eta} &= \frac{d}{d\eta}h(x)\exp\{\eta\,T(x) - A(\eta)\} && \text{(definition of an exponential family)}\\
&= h(x)\exp\{\eta\,T(x) - A(\eta)\}\,\frac{d(\eta\,T(x) - A(\eta))}{d\eta} && \text{(chain rule)}\\
&= \mathbb{f}(x;\eta)\,\frac{d(\eta\,T(x) - A(\eta))}{d\eta} && \text{(definition of an exponential family)}\\
&= \mathbb{f}(x;\eta)\left(T(x) - \frac{dA(\eta)}{d\eta}\right) && \text{(the derivative of a difference is the difference of the}
\end{aligned}$$

Substituting this derivative into the integral gives

$$\begin{aligned}
\int T(x)\,\frac{d\mathbb{f}(x;\eta)}{d\eta}\,dx &= \int T(x)\,\mathbb{f}(x;\eta)\left(T(x) - \frac{dA(\eta)}{d\eta}\right)dx && \text{(substitute the derivative of the densit}\\
&= \int\left((T(x))^2\,\mathbb{f}(x;\eta) - \frac{dA(\eta)}{d\eta}\,T(x)\,\mathbb{f}(x;\eta)\right)dx && \text{(distribute the product over the differe}\\
&= \int (T(x))^2\,\mathbb{f}(x;\eta)\,dx - \int \frac{dA(\eta)}{d\eta}\,T(x)\,\mathbb{f}(x;\eta)\,dx && \text{(the integral of a difference is the diffe}\\
&= \int (T(x))^2\,\mathbb{f}(x;\eta)\,dx - \frac{dA(\eta)}{d\eta}\int T(x)\,\mathbb{f}(x;\eta)\,dx && \text{(a factor free of the integration variab}\\
&= \mathbb{E}\left[(T(X))^2\right] - \frac{dA(\eta)}{d\eta}\,\mathbb{E}[T(X)] && \text{(definition of the expectation)}\\
&= \mathbb{E}\left[(T(X))^2\right] - (\mathbb{E}[T(X)])^2 && \text{(the derivative of the log-partition fun}\\
&= \mathrm{Var}(T(X)) && \text{(the second moment is the variance pl}
\end{aligned}$$

The last line rearranges the second moment of $T(X)$, which is its variance plus the square of its mean. Equating the two sides of Exercise 0.3 gives the result.

Theorem 0.2 (The variance of the statistic is the second derivative of the log-partition function)**.** *Let X have a distribution in an exponential family with natural parameter η, statistic T, and log-partition function A, and let η be in the interior of the allowed values. Then*

$$\mathrm{Var}(T(X)) = \frac{d^2 A(\eta)}{d\eta^2}.$$

Proof

Proof. The solution to Exercise 0.3, worked in Solution 0.3, proves the statement. It uses Theorem 0.1. □

Example 0.2 (The variance of the Poisson distribution from its log-partition function)**.**

In Example 0.1, the log-partition function is $A(\eta) = \exp\{\eta\}$, and the statistic is $T(x) = x$. By Theorem 0.2,

$$\mathrm{Var}(X) = \mathrm{Var}(T(X)) = \frac{d^2 A(\eta)}{d\eta^2} = \exp\{\eta\}.$$

Since $\exp\{\eta\} = \lambda$, we have $\mathrm{Var}(X) = \lambda$, the variance of the Poisson distribution.
To check Theorem 0.2 on data, we compute its two sides separately for the cyclone counts of Example 0.1, with λ set to the sample mean. The left side adds $(x-\lambda)^2\, \mathbb{f}(x;\lambda)$ over the counts from 0 to 100, and the right side is a second central difference of $A(\eta)$ at $\eta = \log\{\lambda\}$.

R

```r
cyclones <- c(6, 5, 4, 6, 6, 3, 12, 7, 4, 2, 6, 7, 4)
rate <- mean(cyclones)
natural_parameter <- log(rate)

counts <- 0:100
squared_deviations <- (counts - rate)^2
weighted_deviations <- squared_deviations * dpois(counts, rate)
variance_statistic <- sum(weighted_deviations)

log_partition <- exp
step <- 1e-4
log_partition_up <- log_partition(natural_parameter + step)
log_partition_mid <- log_partition(natural_parameter)
log_partition_down <- log_partition(natural_parameter - step)
second_diff <- log_partition_up - 2 * log_partition_mid + log_partition_down
second_derivative <- second_diff / step^2

tibble::tibble(variance_statistic, second_derivative)
#> # A tibble: 1 x 2
#>   variance_statistic second_derivative
#>                <dbl>             <dbl>
#> 1               5.54              5.54
```

Python

```
import numpy as np
import pandas as pd
from scipy.stats import poisson

cyclones = np.array([6, 5, 4, 6, 6, 3, 12, 7, 4, 2, 6, 7, 4])
rate = cyclones.mean()
natural_parameter = np.log(rate)

counts = np.arange(0, 101)
squared_deviations = (counts - rate) ** 2
variance_statistic = np.sum(squared_deviations * poisson.pmf(counts, rate))

log_partition = np.exp
step = 1e-4
log_partition_up = log_partition(natural_parameter + step)
log_partition_mid = log_partition(natural_parameter)
log_partition_down = log_partition(natural_parameter - step)
second_diff = log_partition_up - 2 * log_partition_mid + log_partition_down
second_derivative = second_diff / step**2

pd.DataFrame(
    {
        "variance_statistic": [variance_statistic],
        "second_derivative": [second_derivative],
    }
)
#>    variance_statistic  second_derivative
#> 0            5.538462           5.538462
```

The two sides agree at 5.538, and differ by 6.2e-08. That difference is numerical error, from stopping the counts at 100 and from the finite step of the second difference.

Sufficiency in an exponential family

Exercise 0.1 (The joint mass function of an exponential family sample factors through the sum of the statistic)**.** Let $X_1, \ldots, X_n \sim_{\text{iid}} \mathbb{f}(x; \eta)$ be a sample of a discrete random variable whose probability mass function has the form of Equation 1. Show that the joint mass function of the sample is

$$\mathbb{f}(\underset{\sim}{x};\eta) = \exp\left\{\eta \sum_{i=1}^{n} T(x_i) - n\,A(\eta)\right\} \prod_{i=1}^{n} h(x_i).$$

Solution

Solution 0.1. The observations are independent, so the joint mass function multiplies the individual mass functions. Each factor has the form of Equation 1:

$$\begin{aligned}
\mathbb{f}(\underset{\sim}{x};\eta) &= \prod_{i=1}^{n} \mathbb{f}(x_i;\eta) && \text{(independent observations)} \\
&= \prod_{i=1}^{n} h(x_i) \exp\{\eta\, T(x_i) - A(\eta)\} && \text{(definition of an exponential family)} \\
&= \prod_{i=1}^{n} h(x_i) \prod_{i=1}^{n} \exp\{\eta\, T(x_i) - A(\eta)\} && \text{(a product of two-factor terms is the product of the first fa} \\
&= \prod_{i=1}^{n} h(x_i) \exp\left\{\sum_{i=1}^{n} (\eta\, T(x_i) - A(\eta))\right\} && \text{(the product of exponentials is the exponential of the sum)} \\
&= \prod_{i=1}^{n} h(x_i) \exp\left\{\sum_{i=1}^{n} \eta\, T(x_i) - \sum_{i=1}^{n} A(\eta)\right\} && \text{(the sum of a difference is the difference of the sums)} \\
&= \prod_{i=1}^{n} h(x_i) \exp\left\{\eta \sum_{i=1}^{n} T(x_i) - \sum_{i=1}^{n} A(\eta)\right\} && \text{(a factor free of the index moves out of the sum)} \\
&= \prod_{i=1}^{n} h(x_i) \exp\left\{\eta \sum_{i=1}^{n} T(x_i) - n\,A(\eta)\right\} && \text{(a sum of } n \text{ equal terms is } n \text{ times the term)} \\
&= \exp\left\{\eta \sum_{i=1}^{n} T(x_i) - n\,A(\eta)\right\} \prod_{i=1}^{n} h(x_i) && \text{(multiplication is commutative)}
\end{aligned}$$

Theorem 0.1 (The sum of the statistic is sufficient in an exponential family)**.** *Let* $X_1, \ldots, X_n \sim_{\text{iid}} \mathbb{f}(x;\eta)$ *be a sample of a discrete random variable whose distribution is in an exponential family with natural parameter* η *and statistic* T*. Then the sum* $\sum_{i=1}^{n} T(X_i)$ *is sufficient for* η*.*

Proof

Proof. The solution to Exercise 0.1, worked in Solution 0.1, writes the joint mass function as a function of η and the sum, times a function of the data alone. That form is the one in the factorization theorem, which proves the statement. □

Example 0.1 (Two sets of Poisson counts with the same total have a mass function quotient free of the parameter)**.**
By Theorem 0.1, the sum of the counts is sufficient for the natural parameter of the Poisson distribution. So two sets of counts with the same sum differ in their joint mass functions only by a factor that does not depend on the parameter. To check the result on data, we take the first three cyclone counts of Example 0.1, 6, 5 and 4, and the counts 5, 5 and 5. Each set adds up to 15. We divide the joint mass function of the first set by that of the second at four values of λ.

R

```r
sample_a <- c(6, 5, 4)
sample_b <- c(5, 5, 5)

rates <- c(2, 4, 6, 8)
n_obs <- length(sample_a)
n_rates <- length(rates)

counts_a <- rep(sample_a, times = n_rates)
counts_b <- rep(sample_b, times = n_rates)
rates_rep <- rep(rates, each = n_obs)

log_mass_a <- dpois(counts_a, rates_rep, log = TRUE)
log_mass_b <- dpois(counts_b, rates_rep, log = TRUE)
log_matrix_a <- matrix(log_mass_a, nrow = n_obs)
log_matrix_b <- matrix(log_mass_b, nrow = n_obs)

log_joint_a <- colSums(log_matrix_a)
log_joint_b <- colSums(log_matrix_b)
masses_a <- exp(log_joint_a)
masses_b <- exp(log_joint_b)
ratio <- masses_a / masses_b

tibble::tibble(rates, masses_a, masses_b, ratio)
#> # A tibble: 4 x 4
#>   rates  masses_a  masses_b ratio
#>   <dbl>     <dbl>     <dbl> <dbl>
#> 1     2 0.0000392 0.0000470 0.833
#> 2     4 0.00318   0.00382   0.833
#> 3     6 0.00345   0.00414   0.833
#> 4     8 0.000641  0.000769  0.833
```

Python

```python
import numpy as np
import pandas as pd
from scipy.stats import poisson

sample_a = np.array([6, 5, 4])
sample_b = np.array([5, 5, 5])

rates = np.array([2, 4, 6, 8])
masses_a = [np.prod(poisson.pmf(sample_a, r)) for r in rates]
masses_b = [np.prod(poisson.pmf(sample_b, r)) for r in rates]
ratio = np.array(masses_a) / np.array(masses_b)

pd.DataFrame(
    {
        "rates": rates,
        "masses_a": masses_a,
        "masses_b": masses_b,
        "ratio": ratio,
    }
)
#>    rates  masses_a  masses_b     ratio
#> 0      2  0.000039  0.000047  0.833333
#> 1      4  0.003182  0.003818  0.833333
#> 2      6  0.003453  0.004144  0.833333
#> 3      8  0.000641  0.000769  0.833333
```

At every value of λ, the first mass function is 0.8333 times the second. That factor is $(5!)^3/(6!\,5!\,4!)$, which is free of λ.

The canonical link function

Definition 0.1 (Canonical link function)**.** Let X have a distribution in an exponential family with natural parameter η, statistic $T(x) = x$, and log-partition function A. By Theorem 0.1, the mean of X is $\mu = \frac{dA(\eta)}{d\eta}$. The **canonical link function** is the function that takes the mean to the natural parameter, the inverse function of that derivative:

$$\eta = \left(\frac{dA}{d\eta}\right)^{-1}(\mu).$$

Example 0.1 (The canonical links of the Poisson and Bernoulli distributions)**.**
In each case, Theorem 0.1 says the mean is the derivative of the log-partition function, so we differentiate it and then solve for η:

- **Poisson distribution.** In Example 0.1, $A(\eta) = \exp\{\eta\}$, so the mean is:

$$\begin{aligned} \mu &= \frac{dA(\eta)}{d\eta} && \text{(the mean is the derivative of the log-partition function)} \\ &= \exp\{\eta\} && \text{(the derivative of } \exp\{u\} \text{ is } \exp\{u\}) \end{aligned}$$

  The logarithm inverts the exponential function, so $\eta = \log\{\mu\}$. The canonical link function is the logarithm.

- **Bernoulli distribution.** In Example 0.3, $A(\eta) = \log\{1 + \exp\{\eta\}\}$, so the mean is:

$$\begin{aligned} \mu &= \frac{dA(\eta)}{d\eta} && \text{(the mean is the derivative of the log-partition function)} \\ &= \frac{\exp\{\eta\}}{1 + \exp\{\eta\}} && \text{(chain rule: the derivative of } \log\{u\} \text{ is } 1/u \text{ times the derivative of } u) \end{aligned}$$

  We solve this equation for η, one operation per line:

$$\begin{aligned} \mu(1 + \exp\{\eta\}) &= \exp\{\eta\} && \text{(multiply both sides by } 1 + \exp\{\eta\}) \\ \mu + \mu\exp\{\eta\} &= \exp\{\eta\} && \text{(distribute } \mu \text{ over the sum)} \\ \mu &= \exp\{\eta\} - \mu\exp\{\eta\} && \text{(subtract } \mu\exp\{\eta\} \text{ from both sides)} \\ \mu &= \exp\{\eta\}(1 - \mu) && \text{(factor out } \exp\{\eta\}) \\ \exp\{\eta\} &= \frac{\mu}{1 - \mu} && \text{(divide both sides by } 1 - \mu) \\ \eta &= \log\left\{\frac{\mu}{1 - \mu}\right\} && \text{(take the logarithm of both sides)} \end{aligned}$$

  The canonical link function is the logit, the function that takes a probability to its log odds.

Further reading

- The logistic regression chapter of the regression notes uses the log odds as its link. Linear models with a link function are called generalized linear models [2, Ch. 3].
- The Fall 2024 course site for UC Berkeley's Stat 210A, *Theoretical Statistics* (Will Fithian) has a course reader on exponential families, sufficiency, completeness and unbiased estimation.

References

[1] G. Casella and R. Berger, *Statistical inference*, 2nd ed. Cengage Learning, 2002. Available: https://www.cengage.com/c/statistical-inference-2e-casella-berger/9780534243128/

[2] A. J. Dobson and A. G. Barnett, *An introduction to generalized linear models*, 4th ed. CRC press, 2018. doi: 10.1201/9781315182780.

[3] D. W. Hosmer and S. Lemeshow, *Applied logistic regression*. New York: Wiley, 1989.

[4] W. N. Venables and B. D. Ripley, *Modern applied statistics with S*, 4th ed. New York: Springer, 2002. doi: 10.1007/978-0-387-21706-2.