Likelihood theory

Kristian Wichmann

July 12, 2017

1 Statistical models

A *statistical model* $\mathcal{P}$ is a family of probability distributions on a measurable space $(\mathcal{X}, \mathbb{E})$ indexed by parameters θ from a parameter space Θ. We can sum this up as:

$$\mathcal{P} = \{\nu_\theta | \theta \in \Theta\} \tag{1.1}$$

1.1 Dominated statistical models

We call such a model *dominated* if there exists a σ-finite measure μ on $(\mathcal{X}, \mathbb{E})$, such that all the distributions in the model has a density function f_θ with respect to μ. Or equivalently, that all the distributions is absolutely continuous with respect to μ:

$$\forall \nu \in \mathcal{P} : \nu \ll \mu \tag{1.2}$$

The Radon-Nikodym derivative $\frac{d\nu}{d\mu}$ is then a density function for ν with respect to μ. We call μ a *dominating measure* for the model.

This may all sound a little hairy, but in practice, the dominating measure will almost always be the Lebesgue measure (for continuous distributions), the counting measure (for discrete distributions), or some combination of the two.

2 The likelihood function

2.1 Definition

Let $\mathcal{P}$ be a dominated statistical model. The *likelihood* function for an outcome $x \in \mathcal{X}$ is a function $L_x : \Theta \to \mathbb{R}$ associates a number to every parameter configuration:

$$L_x(\theta) = f_\theta(x) \tag{2.1}$$

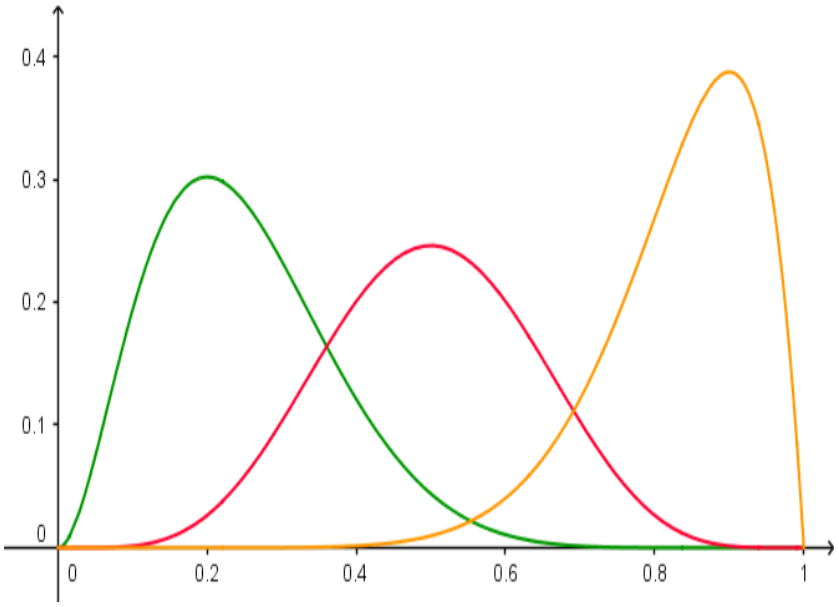

Figure 1: Likelihood function for $n = 10$ and $k = 2, 5, 9$ (green, red, orange) respectively.

The interpretation of the likelihood function is, that the higher its value, the more likely it seems that θ is the true parameters of the model. Hence, we will often seek out the set of parameters which maximize the likelihood function. This process is known as *maximum likelihood estimation* or MLE for short. Note that there's no mathematical justification of this process in itself.

2.1.1 Example: Coin tosses

We consider a repeated coin toss, each i.i.d. Bernoulli processes with parameter p - the probability that the outcome is heads. If the coin is tossed n times, the outcome space is $\mathcal{X} = \{0, 1, 2, \cdots, n\}$ where the number of the outcome heads is counted (the dominating measure is the counting measure). Given a specific outcome $k \in \mathcal{X}$, the likelihood function can be found by the binomial distribution:

$$L_k(p) = \binom{n}{k} p^k (1-p)^{n-k} \tag{2.2}$$

Here $p \in \Theta = [0, 1]$. Figure 1 shows examples of this function.

Now, we can perform MLE by finding the value of the parameter p which

maximizes L_k. We differentiate using the product rule:

$$\frac{\partial L_k}{\partial p} = \binom{n}{k} \left(k(p^{k-1}(1-p)^{n-k} - p^k(n-k)(1-p)^{n-k-1} \right) \tag{2.3}$$

For this to be zero, the binomial coefficient is irrelevant, so:

$$k(p^{k-1}(1-p)^{n-k} = p^k(n-k)(1-p)^{n-k-1} \Leftrightarrow \tag{2.4}$$

$$k(1-p) = (n-k)p \Leftrightarrow \tag{2.5}$$

$$k = np \tag{2.6}$$

In other words, $p_{\text{MLE}} = \frac{k}{n}$. This will probably not be much of a surprise to anyone.

However, this estimate might not always be sensible. Specifically, if you've made a very small amount of count throws. If $n = 1$, you will conclude that $p = 0$ or $p = 1$, which meshes badly with our intuition about coin throws. This may be modelled as a *prior distribution* of p, leading to a Bayesian analysis. Contrast with the case where $m \gg 1$: When we have a lot of repetitions, we will be more certain of the value of the parameter p. This idea of probability as a limit for a large number of repetitions is at the heart of the frequentist interpretation.

2.2 The log-likelihood function

When the density functions are nowhere zero, it often makes sense to deal with the logarithm of the likelihood function instead. Since the logarithm is a strictly monotonic function, this makes no difference for the purpose of MLE. Some presentations (this included), introduces a sign change as well:

$$l_x(\theta) = -\log f_\theta(x) \tag{2.7}$$

So MLE means one of the following, equivalent procedures:

- Maximizing the likelihood function L_x.
- Minimizing the log-likelihood function l_x.

2.2.1 Example: Fish weights

n adult fish of the same species are caught and weighed. The weights can be reasonably modelled by a normal distribution $N(\mu, \sigma^2)$ (and so the dominating measure is the Lebesgue measure). For simplicity, we will assume that the variance is known from historical data. The observations are:

$$x = (w_1, w_2, \cdots, w_n) \tag{2.8}$$

Now, the likelihood function is:

$$L_x(\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(w_i - \mu)^2}{2\sigma^2}\right] \tag{2.9}$$

Here we see the practicality of taking the logarithm to get the log-likelihood: It turns a product like this into a much more manageable sum:

$$l_x(\mu) = -\log L_x(\mu) = -n \log \frac{1}{\sqrt{2\pi\sigma^2}} + \sum_{i=1}^{n} \frac{(w_i - \mu)^2}{2\sigma^2} \tag{2.10}$$

Differentiating to find the minimum:

$$\frac{\partial l_x}{\partial \mu} = \frac{1}{2\sigma^2} \sum_{i=1}^{n} 2(w_i - \mu) = \frac{1}{\sigma^2} \left[\sum_{i=1}^{n} w_i - n\mu\right] \tag{2.11}$$

Setting this to zero we find:

$$\mu_{\text{MLE}} = \frac{1}{n} \sum_{i=1}^{n} w_i \tag{2.12}$$

Once again, hardy a surprising result.

3 Score function and observed information

When the parameter space is an open subset of $\mathbb{R}^k$ (usually the case), we define these as follows:

- The *score function* is the gradient of the log-likelihood:

$$V_x(\theta) = \nabla_\theta l_x(\theta) \tag{3.1}$$

- The *observed information function* is the Hessian matrix of the log-likelihood function:

$$\mathcal{J}_x(\theta) = H_\theta[l_x(\theta)] = \nabla_\theta(\nabla_\theta l_x(\theta))^t = \nabla_\theta V_x(\theta)^t \tag{3.2}$$

The *observed information* is the observed information function evaluated at the MLE estimator: $\mathcal{J}_x(\theta_{\text{MLE}})$.

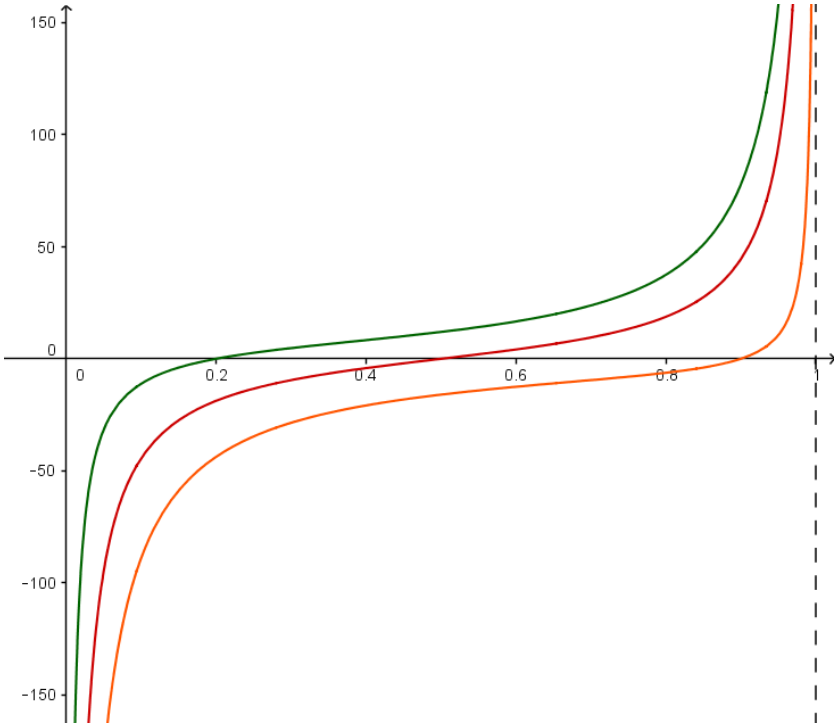

Figure 2: Score function for $n = 10$ and $k = 2, 5, 9$ (green, red, orange) respectively.

3.1 Example: Coin toss

As we saw above, here the likelihood is given by the binomial distribution:

$$L_k(p) = \binom{n}{k} p^k (1-p)^{n-k} \Rightarrow l_k(p) = -\log \binom{n}{k} - k \log p - (n-k) \log(1-p) \tag{3.3}$$

Now to get the score function, we differentiate with respect to p:

$$V_k(p) = \frac{\partial l_x}{\partial p} = -\frac{k}{p} + \frac{n-k}{1-p} = \frac{(n-k)p - k(1-p)}{p(1-p)} = \frac{n-k}{p(1-p)} \tag{3.4}$$

Three examples of the score function are shown in figure 2.

The observed information function, like the score function, is simply a scalar in this case:

$$\mathcal{J}_k(p) = \frac{\partial V_k}{\partial p} = \frac{k}{p^2} + \frac{n-k}{(1-p)^2} \tag{3.5}$$

Examples of these functions are shown in figure 3.

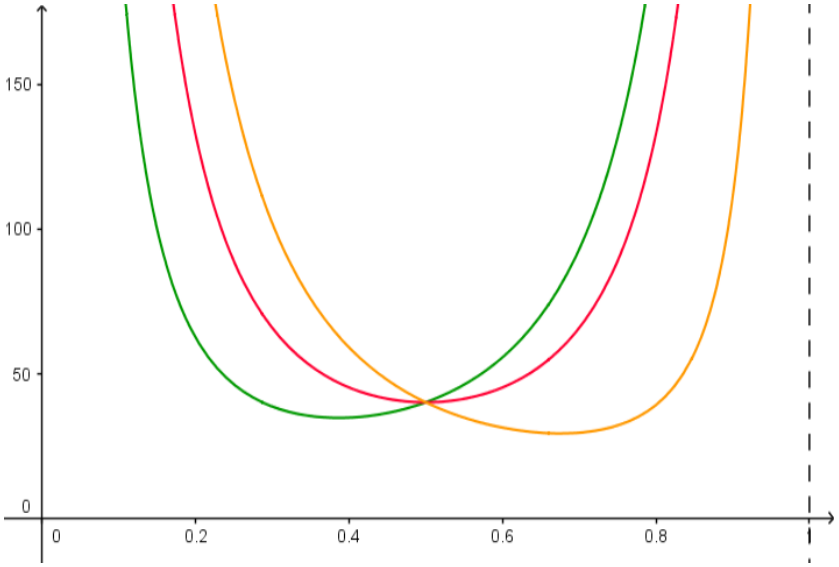

Figure 3: Observed information function for $n = 10$ and $k = 2, 5, 9$ (green, red, orange) respectively.

The observed information is found by setting $p = p_{\text{MLE}} = \frac{k}{n}$ in 3.5:

$$\frac{kn^2}{k^2} + \frac{n-k}{\frac{(n-k)^2}{n^2}} = \frac{n^2}{k} + \frac{n^2}{n-k} = \frac{n^2(n-k+k)}{k(n-k)} = \frac{n}{\frac{k}{n} \cdot \left(1 - \frac{k}{n}\right)} \tag{3.6}$$

Reinstating p_{MLE}, this means that the observed information is:

$$\frac{n}{p_{\text{MLE}}(1 - p_{\text{MLE}})} \tag{3.7}$$

3.2 Example: Fish weights

Again, consider the example with the normally distributed fish weights, this time with an unknown variance σ^2 as well. The likelihood function in this case is:

$$L_x(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right] \tag{3.8}$$

Here, x_i is the observed weight of the i'th fish. The log-likelihood is:

$$l_x(\mu, \sigma^2) = -\log L_x(\mu, \sigma^2) = \frac{n}{2}\log(2\pi) + \frac{n}{2}\log\sigma^2 + \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2} \tag{3.9}$$

The score function is then a two dimensional vector:

$$V_x(\mu, \sigma^2) = \begin{pmatrix} \frac{\partial l_x}{\partial \mu} & \frac{\partial l_x}{\partial(\sigma^2)} \end{pmatrix}^t \tag{3.10}$$

Note that it's σ^2 which is the parameter, not σ itself. Differentiating with respect to each parameter, we find:

$$\frac{\partial l_x}{\partial \mu} = \frac{1}{2\sigma^2} \sum_{i=1}^{n} 2(x_i - \mu)(-1) = -\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) \tag{3.11}$$

$$\frac{\partial l_x}{\partial(\sigma^2)} = \frac{n}{2}\frac{1}{\sigma^2} - \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu)^2 \tag{3.12}$$

That makes the score vector:

$$V_x(\mu, \sigma^2) = \begin{pmatrix} -\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) & \frac{n}{2\sigma^2} - \frac{1}{2\sigma^4} \sum_{i=1}^{n} (x_i - \mu)^2 \end{pmatrix}^t \tag{3.13}$$

The observed information function is a two by two matrix:

$$\mathcal{J}_x(\mu, \sigma^2) = \begin{pmatrix} \frac{\partial^2 l_x}{\partial \mu^2} & \frac{\partial^2 l_x}{\partial \mu \partial(\sigma^2)} \\ \frac{\partial^2 l_x}{\mu \partial(\sigma^2) \partial} & \frac{\partial^2 l_x}{\partial(\sigma^2)^2} \end{pmatrix} \tag{3.14}$$

By the usual rules of second partial derivatives, the matrix is symmetric, so there's three calculations to be done:

$$\frac{\partial^2 l_x}{\partial \mu^2} = \frac{\partial}{\partial \mu} \left[-\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) \right] = \frac{1}{\sigma^2} \sum_{i=1}^{n} 1 = \frac{n}{\sigma^2} \tag{3.15}$$

$$\frac{\partial^2 l_x}{\partial \mu \partial(\sigma^2)} = \frac{\partial}{\partial(\sigma^2)} \left[-\frac{1}{\sigma^2} \sum_{i=1}^{n} (x_i - \mu) \right] = \frac{1}{\sigma^4} \sum_{i=1}^{n} (x_i - \mu) \tag{3.16}$$

$$\frac{\partial^2 l_x}{\partial(\sigma^2)^2} = \frac{\partial}{\partial(\sigma^2)} \left[\frac{n}{2\sigma^2} - \frac{1}{2(\sigma^2)^2} \sum_{i=1}^{n} (x_i - \mu)^2 \right] = \tag{3.17}$$

$$-\frac{n}{2(\sigma^2)^2} + \frac{1}{(\sigma^2)^3} \sum_{i=1}^{n} (x_i - \mu)^2 \tag{3.18}$$

Putting it all together:

$$\mathcal{J}_x(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & \frac{1}{\sigma^4} \sum_{i=1}^{n} (x_i - \mu) \\ \frac{1}{\sigma^4} \sum_{i=1}^{n} (x_i - \mu) & -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6} \sum_{i=1}^{n} (x_i - \mu)^2 \end{pmatrix} \tag{3.19}$$

4 Fisher information

The *Fisher information* (sometimes simply called the information) is the expected value of the observed information function:

$$\mathcal{I}(\theta) = E[\mathcal{J}_x(\theta)] \tag{4.1}$$

Note that this is not a function of the actual outcome - The Fisher information is a property of the model, not the realized observations.

4.1 Example: Bernoulli distribution

The (log)likelihood for a random Bernoulli variable X is:

$$L_x(p) = p^x(1-p)^{1-x} \Rightarrow l_x(p) = -x\log p - (1-x)\log(1-p) \tag{4.2}$$

The score function is:

$$V_x(p) = -\frac{x}{p} + \frac{1-x}{1-p} \tag{4.3}$$

The observed information function is:

$$\mathcal{J}_x(p) = \frac{x}{p^2} + \frac{1-x}{(1-p)^2} \tag{4.4}$$

Now, to get the Fisher information we find the expectation value by summing over the two possible values of x:

$$\mathcal{I}(p) = E[\mathcal{J}_x(p)] = \sum_x P(X=x)\mathcal{J}_x(p) = p\frac{1}{p^2} + (1-p)\frac{1}{(1-p)^2} = \tag{4.5}$$

$$\frac{1}{p} + \frac{1}{1-p} = \frac{1-p+p}{p(1-p)} = \frac{1}{p(1-p)} \tag{4.6}$$

This is simply the reciprocal of the variance of X.

4.2 Example: Binomial distribution

From equation 3.5 we know that:

$$\mathcal{J}_k(p) = \frac{k}{p^2} + \frac{n-k}{(1-p)^2} = \frac{(1-p)^2k + p^2(n-k)}{p^2(1-p)^2} \tag{4.7}$$

The numerator is:

$$(1-2p+p^2)k + p(n-k) = k - 2pk + p^2k + pn - pk = k - 2pk + np^2 \tag{4.8}$$

So:

$$\mathcal{J}_k(p) = \frac{(1-2p)k + np^2}{p^2(1-p)^2} \tag{4.9}$$

Now, the Fisher information is found by taking the expectation value of this, again by summing over outcomes (here $n+1$ possibilities). This means that everything that does not involve k can be taken outside of the expectation:

$$\mathcal{I}(p) = E[\mathcal{J}_x(p)] = \frac{1}{p^2(1-p)^2}((1-2p)\ E[k] + np^2) \tag{4.10}$$

The expected value of k is np, so the parenthesis is equal to:

$$(1-2p)np + np^2 = np - 2np^2 + np^2 = np - np^2 = n(p-p^2) = np(1-p) \tag{4.11}$$

All in all, we get:

$$\mathcal{I}(p) = \frac{n}{p(1-p)} \tag{4.12}$$

5 Bartlett's identities

These identities holds when the log-likelihood is sufficiently nice that integration and differentiation with respect to θ are interchangeable. In that case, the following two theorems are true:

Theorem 5.1. *The expectation value of the score function is zero:*

$$E[V_x(\theta)] = 0 \tag{5.1}$$

Proof. To prove the identity, note that because $L_x(\theta)$ regarded as a function of x is a probability density function we have:

$$\int L_x(\theta)\ d\mu(x) = 1 \tag{5.2}$$

Differentiating with respect to θ on each side of the equation and using the regularity assumption above we get:

$$\frac{\partial}{\partial\theta}\int L_x(\theta)\ d\mu(x) = \int \frac{\partial}{\partial\theta} L_x(\theta)\ d\mu(x) = 0 \tag{5.3}$$

Now rewrite the integrand by multiplying by 1 in the form $\frac{L_x(\theta)}{L_x(\theta)}$ (we assume that the likelihood is positive everywhere):

$$\int \underbrace{\frac{\frac{\partial L_x}{\partial\theta}}{L_x(\theta)}} L_x(\theta)\ d\mu(x) = 0 \tag{5.4}$$

Now, the underbraced part is actually the score function except for a sign change, since:

$$V_x(\theta) = \frac{\partial l_x}{\partial \theta} = -\frac{\partial}{\partial \theta} \log L_x(\theta) = -\frac{1}{L_x(\theta)} \frac{\partial L_x}{\partial \theta} \tag{5.5}$$

Which means:

$$\int V_x(\theta) L_x(\theta) \; d\mu(x) = 0 \tag{5.6}$$

But this is just the expected value. □

Theorem 5.2. *The variance of the score function is equal to the Fisher information:*

$$Var(V_x(\theta)) = \mathcal{I}(\theta) \tag{5.7}$$

Proof. The proof uses the same basic idea as above, except that a second derivative is used:

$$\frac{\partial^2}{\partial \theta^2} \int L_x(\theta) \; d\mu(x) = \int \frac{\partial^2}{\partial \theta^2} L_x(\theta) \; d\mu(x) = 0 \tag{5.8}$$

Following the same steps as above, the integrand can be written as:

$$\frac{\partial}{\partial \theta} \left[V_x(\theta) L_x(\theta) \right] = \frac{\partial V_x}{\partial \theta} L_x(\theta) + V_x(\theta) \frac{\partial L_x}{\partial \theta} \tag{5.9}$$

$$\mathcal{J}_x(\theta) L_x(\theta) - \left(V_x(\theta) \right)^2 L_x(\theta) \tag{5.10}$$

Here, the second term has been rewritten in the same way we did above. Now, when integrating this yields:

$$\mathcal{I}(\theta) - \mathrm{Var}(V_x(\theta)) = 0 \tag{5.11}$$

Here we've used that the expectation value of the score is zero. The desired result immediately follows. □

These are known as Bartlett's first and second identity, respectively. In principle, one could continue this pattern to higher order derivatives for more identities.

6 Likelihood for exponential families

Recall, that an exponential family $\mathcal{P}$ on a measurable space $(\mathcal{X}, \mathbb{E})$ has parameter space $\Theta \subseteq \mathbb{R}^k$, a measurable $t : \mathbb{E} \to \mathbb{R}^k$ and a base measure μ on $(\mathcal{X}, \mathbb{E})$ such that:

$$\forall A \in \mathbb{E}, \theta \in \Theta : \; \nu_\theta(A) = \frac{1}{c(\theta)} \int_A \exp\left[\theta^t t(x) \right] \; d\mu(x) \tag{6.1}$$

Here $c(\theta)$ is a normalization constant. In other words, μ is a dominating measure for the model, and the associated likelihood function is:

$$L_x(\theta) = \frac{1}{c(\theta)} \exp[\theta^t t(x)] \tag{6.2}$$

6.1 Useful identities

These formulas from the general theory of exponentials families will come in handy in a second:

$$\nabla_\theta \log c(\theta) = E[t(X)] \equiv \tau(\theta) \tag{6.3}$$

$$\nabla_\theta (\nabla_\theta \log c(\theta))^t = \mathrm{Var}(t(X)) \equiv \kappa(\theta) \tag{6.4}$$

6.2 Log-likelihood

The log-likelihood function takes on a particularly simple form for such a model:

$$l_x(\theta) = -\log L_x(\theta) = \log c(\theta) - \theta^t t(x) \tag{6.5}$$

6.3 Score function

The score function is then:

$$V_x(\theta) = \nabla_\theta l_x(\theta) = \nabla_\theta \log c(\theta) - t(x) \tag{6.6}$$

Using equation 6.3 this becomes:

$$V_x(\theta) = \tau(\theta) - t(x) \tag{6.7}$$

6.4 Observed information function

The observed information function is:

$$\mathcal{J}_x(\theta) = \nabla_\theta (\nabla_\theta l_x(\theta))^t = \nabla_\theta (\nabla_\theta \log c(\theta) - t(x))^t = \kappa(\theta) \tag{6.8}$$

This means that - somewhat counter-intuitive - the observed information function does not depend on the observation x at all!

This has the immediate consequence that the observed information as well as the Fisher information is also equal to $\kappa(\theta)$.

But wait! We know that the binomial and normal distributions are both exponential families. Yet, above we found their observed information functions in equations 3.5 and 3.19 to explicitly depend on the observations. So what is happening? The problem is, that these distributions are not "naturally" formulated in the exponential form above: *Reparametrization* is needed to express them in the form of equation 6.1.

7 Reparametrization