7.8 Consistency of Observations

Since Press' paper is part of a book that isn't necessarily widely available, here are are some details of the solution. The notation in our Chapter 7 is used.
The key notion is that observations are either "good" (are drawn from a Gaussian which has the quoted standard deviation) or "bad" (drawn from a Gaussian which has some other, bigger, standard deviation than the nominal or quoted one). All of these Gaussians have the same mean μ_w, which is what we want to know. The probability of an observation being good is p and this has some prior probability distribution $\text{prob}(p)$.
We proceed in the standard way using Bayes' theorem. The posterior probability distribution of μ_w and p is (assuming we have N observations in total):

probability of the data, given μ_w and p and that a specific set n of them are are "good" and $N-n$ are "bad"
$\times$
probability that n are "good" and $N-n$ are "bad",
$\times$
the priors on μ_w and p,
summed over all the possible permutations.

The probability that n are "good" and $N-n$ are "bad" is just

$$p^n(1-p)^{N-n}$$

There is no binomial coefficient here because we are dealing with distinguishable cases – it matters which of the observations are labelled "good" or "bad". So we have to deal with all of the permutations of possibly "good" or "bad" observations.
If we simply write down the general likelihood term it is pretty obscure what is going on, because of the multitude of distinguishable possibilities. Let's take a simpler case, $N=3$, and for brevity let's denote the probability of the i-th datum (given that it is good) by g_i. So g is an abbreviation for

$$\text{prob}(X_i|\mu_w,\sigma)$$

as in Equation 7.41 in the book.
Similarly, g_i' the probability of the i-th datum in the case where the measurement is "bad" and a bigger standard deviation applies than the nominal one.
The term in the posterior involving all the permutations is this one:

probability of the data, given μ_w and p and that n are "good" and $N-n$ are "bad" $\times$ probability that n are "good" and $N-n$ are "bad" and spelling out all the permutations, it looks like this.

$$
\begin{aligned}
g_1 g_2 g_3 p^3 \\
+ \\
g_1' g_2 g_3 p^2(1-p) + g_1 g_2' g_3 p^2(1-p) + g_1 g_2 g_3' p^2(1-p) \\
+ \\
g_1' g_2' g_3 p(1-p)^2 + g_1 g_2' g_3' p(1-p)^2 + g_1' g_2 g_3' p(1-p)^2 \\
+ \\
g_1' g_2' g_3' (1-p)^3.
\end{aligned}
\tag{1}
$$

Now this definitely looks like $(Ap + B(1-p))^3$ for some A and B, and playing with it for a bit it's not hard to see that Equation (1) factorizes to

$$(g_1 p + g_1'(1-p))(g_2 p + g_2'(1-p))(g_3 p + g_3'(1-p)).$$

This would work for any N, although getting increasingly messy, so the book's 7.41 is established; we just need to remember that we also have to multiply by the priors on p and μ_w and we have 7.41.
If this factorization seems non-obvious, take heart: Press confesses in his article that he didn't see it at first either.

Now to the question, the probability that a *particular* observation, say the kth, is "good". This means we want

probability that k is good, others unspecified, given the data, p, and μ_w – which from Bayes' Theorem is going to be proportional to

(probability of the data, given μ_w and p and that the kth is good, others unspecified
$\times$
probability the kth is good, others unspecified
$\times$
the priors on μ_w and p)
summed over all the possible permutations in which the kth is good, others unspecified marginalized (integrated) over p and μ_w.

For our example, $N = 3$, let's take $k = 1$. Then the sum over the various permutations looks like this:

$$(g_1 p) g_2 g_3 p^2 \quad +$$

$$
\begin{aligned}
(g_1 p)(g_2' g_3 p(1-p) + g_2 g_3' p(1-p)) &\quad + \\
(g_1 p)((g_2' g_3' p^3)(1-p)^2). & \qquad (2)
\end{aligned}
$$

The $(g_1 p)$ term turns up in each case because it is the probability of the first datum being drawn from a "good" distribution, multiplied by the probability (recall this is a parameter, with a prior) that an observation is good. This factorizes as before:

$$
(g_1 p)(g_2 p + g_2'(1-p))(g_3 p + g_3'(1-p)) \qquad (3)
$$

which is the "ancillary likelihood" $\mathcal{L}_1$ defined just after Equation 7.43 of Chapter 7. We therefore have, for the case we are considering,

$$
\begin{aligned}
\text{prob(observation 1 is good, others unspecified)} &\propto \\
\int dp\, d\mu_w\, \mathcal{L}_1 \times \text{priors on } p, \mu_w &
\end{aligned}
$$

We get the required normalizing factor from noting that

$$
\begin{aligned}
\text{prob(Observation 1 is good, others unspecified)} &\quad + \\
\text{prob(Observation 1 is bad, others unspecified)} &\quad = 1
\end{aligned}
$$

In our $N = 3$ example, the first term in this will contain

$$
(g_1 p)(g_2 p + g_2'(1-p))(g_3 p + g_3'(1-p))
$$

and the second will contain

$$
(g_1'(1-p))(g_2 p + g_2'(1-p))(g_3 p + g_3'(1-p))
$$

so that the sum is just the quantity $\mathcal{L}$ defined in 7.41: the result we want, 7.43, follows immediately. While the derivation has proceeded for specific N, this is just to make the key factorization, the step from Equation (2) to Equation (3), more obvious.