Solution: chapter 2, problem 5, part a:

Let y be the observed value of a sampling from a normal distribution with mean μ and standard deviation 1. We'll reserve $\widehat{\mu}$ for the estimator for μ; the distinction will be important later. We might as well allow for a more general situation by letting $1-\alpha_u$ be the confidence level for the upper limit and $1-\alpha_t$ be the confidence level for the two sided interval. For the explicit case in the problem statement we have $\alpha_u = \alpha_t = 0.3173$. Note that we assume that the "68%" in the statement of the problem and below is approximate, and that what we want is the "1σ" interval.

The one-sided p-value for $H_0 : \mu = 0$ is given by

$$p = \int_y^\infty \mathsf{N}(x)dx, \tag{2.1}$$

where N denotes the standard normal [e.g., section 4.4 in Narsky and Porter (2014a)] for random variable Y. We shall also use the notation $\mathsf{N}(x;\mu,1)$ to denote the normal pdf with mean μ and variance 1, evaluated at x. Let $y_{\rm crit}$ be the "critical value" for y, that is the value such that if $y < y_{\rm crit}$ we will compute an upper limit, and if $y > y_{\rm crit}$ we will compute a two-sided interval. If we set $p = 0.1$, we find a critical value of $y_{\rm crit} = 1.2816$.

Note that the point is here that when we decide to quote, say, an upper limit at 68% confidence, we mean that we are going to use our "standard" prescription for doing so, without making any allowance for the fact that we might have quoted a two-sided interval if the observation had come out differently. In the present example, this means that the upper limit is going to be computed by first solving for u in

$$1-\alpha_u = \int_{-\infty}^u \mathsf{N}(x)dx. \tag{2.2}$$

Then the upper limit corresponding to observation y is just $y_u = y+u$. For the stated problem, $u = 0.4752$. For the two-sided interval, we compute t in

$$1-\alpha_t = \int_{-t}^t \mathsf{N}(x)dx. \tag{2.3}$$

Then the two-sided (symmeteric) interval is $y \pm t$. For our case, $t = 1$.

Now we can compute the coverage, $c(\mu)$ of our interval, for any given parameter value μ:

$$c(\mu) = P(Y+u > \mu \cap \mathrm{ul};\mu) + P(Y+t > \mu > Y-t \cap \mathrm{ts};\mu), \tag{2.4}$$

where "ul" refers to the choice of an upper limit and "ts" to the choice of a two-sided interval. Let $P(\mathrm{ul};\mu)$ be the probability to quote an upper limit, and $P(\mathrm{ts};\mu)$ be the probability to quote a two-sided interval. We have:

$$P(\mathrm{ts};\mu) = 1 - P(\mathrm{ul};\mu) = \int_{y_{\rm crit}}^\infty \frac{1}{\sqrt{2\pi}} e^{\frac{1}{2}(y-\mu)^2} dy. \tag{2.5}$$

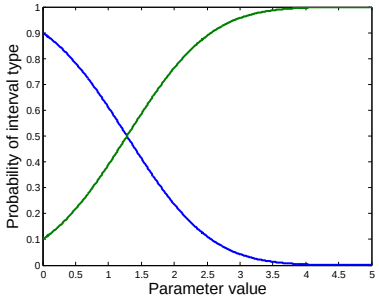

Figure 2.1 The probability to quote an upper limit (blue) or a two-sided interval (green) as a function of the parameter value.

These probabilities are shown in Fig. 2.1 for the stated problem.

It remains to compute the two components of the coverage in (2.4), based on whether or not $y > y_{\rm crit}$. For the upper limit term we have

$$c_{\rm ul}(\mu) \equiv P(Y + u > \mu \cap {\rm ul}; \mu) = \begin{cases} 0, & y_{\rm crit} < \mu - u \\ \int_{\mu-u}^{y_{\rm crit}} {\sf N}(y; \mu, 1) dy, & \text{otherwise.} \end{cases} \tag{2.6}$$

This is plotted as the blue curve in the left plot of Fig. 2.2. For the two-sided interval term the conditional probability is

$$c_{\rm ts}(\mu) \equiv P(Y - t < \mu < Y + t \cap {\rm ts}; \mu) = \begin{cases} 0, & y_{\rm crit} > \mu + t \\ \int_{\max(y_{\rm crit}, \mu - t)}^{\mu+t} {\sf N}(y; \mu, 1) dy, & \text{otherwise.} \end{cases} \tag{2.7}$$

This is plotted as the green curve in the left plot of Fig. 2.2. Finally, the overall coverage according to Eq. 2.4 is shown with the red curve in the left plot of Fig. 2.2. The MATLAB code listing for this calculation is provided in Narsky and Porter (2014b). Note that it is the overall coverage (red curves) that really matters; the other curves are shown so that it can be understood how the overall coverage comes about.

The detailed features of the coverage graph can be readily understood by examining the different cases leading to these features. For example, the coverage is minimum for parameter values where the coverage probabilities for both the upper limit and two-sided interval are non-zero. If an upper limit is quoted, its probability to cover is zero for parameter values greater than 1.7493. This is arrived at as follows: 68% of the area under the normal distribution occurs for $y > \mu - 0.4677$. Thus, the 68% CL upper limit given a sampling y is $y + 0.4677$. Our decision point for chosing to

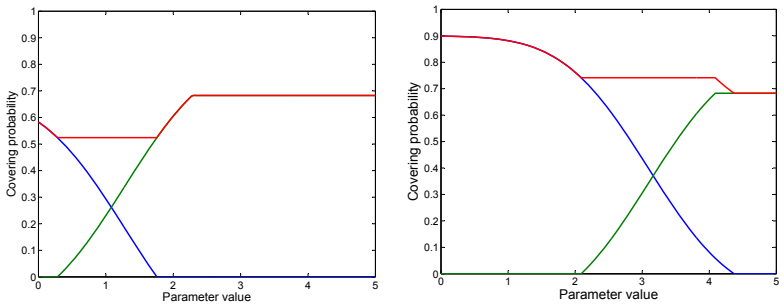

Figure 2.2 The coverage of the intervals obtained when we use the result to decide whether to quote a two-sided or upper limit. The red curves show the coverage. The blue curves show the contribution from the upper limit, and the green curves show the contribution from the two-sided interval. Left: Coverages computed according to the formulas. Right: Coverage of intervals, as in the left plot, except that the critical p-value is here 0.001, and the upper limits are computed at 90% confidence.

quote an upper limit is $y_{\rm crit} = 1.2816$. Thus, the largest upper limit that will ever be quoted is $1.2816 + 0.4677 = 1.7493$, and if μ is greater than this value it will never be included in the upper limit.

It is observed that the overall coverage is not 68% until large values of the parameter μ are reached. For small values of μ the algorithm undercovers. This is undesirable, and hence this practice of deciding how to quote a result based on the result (popularly referred to as the "flip-flop") should be avoided if possible.

It may be suggested that the problem does not deal with what people really do. That is, sometimes they make the decision based on a much smaller p-value, and quote a 90% upper limit when an upper limit is chosen, instead of keeping the 68% used for the two-sided interval. Unfortunately, this doesn't help. For a p-value of 0.001, corresponding to approximately 3σ significance (one-sided), and 90% upper limits and 68% two-sided intervals, we obtain the coverage shown in the right plot of Fig. 2.2. The coverage approaches 90% in the limit $\mu \to 0$ and approaches 68% for large μ, but in between the coverage is neither 90% nor 68% but something between those values.

Solution: chapter 2, problem 5, part b:

We'll continue to assume a normal random variable with mean μ and variance 1, with observed value y. To apply the method, form the likelihood ratio:

$$\lambda(\mu; y) = \frac{L(\mu; y)}{\max_\mu L(\mu; y)}. \tag{2.8}$$

As in Eq. [2.39] of Narsky and Porter (2014a), we construct a table of observations y for which we accept a given value for μ at confidence level $1 - \alpha$:

$$A_\alpha(\mu) = \{y|\lambda(\mu; y) \geq \lambda_\alpha(\mu)\}, \tag{2.9}$$

where $\lambda_\alpha(\mu)$ is chosen such that $A_\alpha(\mu)$ contains a probability content of $1 - \alpha$ in the

sample space for y. That is, we still satisfy Eq. [2.40] of Narsky and Porter (2014a):

$$P\left[\lambda(\mu;Y) \geq \lambda_\alpha(\mu);\mu\right] \geq 1-\alpha. \tag{2.10}$$

We are dealing with a continuous distribution, so we can require equality here. However, in the FC [Feldman and Cousins (1998)] algorithm the likelihood ratio is restricted to "physical" $\mu \geq 0$ and sets $A_\alpha(\mu)$ are constructed only for such values of μ. Thus, if $y \leq 0$, the maximum likelihood estimator for μ is $\widehat{\mu} = 0$. Hence,

$$\lambda(\mu;y) = \begin{cases} \exp\left[-(y-\mu)^2/2\right], & y \geq 0; \\ \exp\left[\mu(y-\mu/2)\right], & y < 0. \end{cases} \tag{2.11}$$

For any $\mu \geq 0$, we compute critical value $\lambda_\alpha(\mu)$ according to:

$$1-\alpha = P\left[\lambda(\mu;Y) \geq \lambda_\alpha(\mu);\mu\right] = \int_{y_1}^{y_2} \mathsf{N}(y;\mu,1)dy, \tag{2.12}$$

The limits of integration, functions of μ and α, are to be determined; they define the set $A_\alpha(\mu)$. We could approach the bookkeeping in various equivalent ways, but the following is straightforward:

$$\begin{aligned} P\left[\lambda(\mu;Y) \geq \lambda_\alpha(\mu);\mu\right] = & P\left[\lambda(\mu;Y) \geq \lambda_\alpha(\mu), Y > 0;\mu\right] + \\ & P\left[\lambda(\mu;Y) \geq \lambda_\alpha(\mu), Y < 0;\mu\right] \end{aligned} \tag{2.13}$$

Let $l(\mu;y) \equiv \log\lambda(\mu;y)$ and $l_\alpha(\mu) \equiv \log\lambda_\alpha(\mu)$. Hence,

$$l(\mu;y) = \begin{cases} -(y-\mu)^2/2, & y \geq 0; \\ \mu(y-\mu/2), & y < 0. \end{cases} \tag{2.14}$$

We then have:

$$\begin{aligned} P\left[\lambda(\mu;Y) \geq \lambda_\alpha(\mu), Y > 0;\mu\right] &= P\left[l(\mu;Y) \geq l_\alpha(\mu), Y > 0;\mu\right] \\ &= P\left[(Y-\mu)^2/2 \leq -l_\alpha(\mu), Y > 0;\mu\right] \\ &= \int_{\max(0,\ell_>)}^{u_>} \mathsf{N}(y;\mu,1)dy, \end{aligned} \tag{2.15}$$

where $u_> = \sqrt{-2l_\alpha(\mu)} + \mu$ and $\ell_> = -\sqrt{-2l_\alpha(\mu)} + \mu$. Similarly, for the $Y < 0$ case:

$$\begin{aligned} P\left[\lambda(\mu;Y) \geq \lambda_\alpha(\mu), Y < 0;\mu\right] &= P\left[\mu(Y-\mu/2) \geq l_\alpha(\mu), Y < 0;\mu\right] \\ &= \int_{\min(0,\ell_<)}^{0} N(y;\mu,1)dy, \end{aligned} \tag{2.16}$$

where $l_< = \mu/2 + l_\alpha(\mu)/\mu$.

We may summarize the limits y_1 and y_2: For y_2 we have

$$y_2 = u_> = \sqrt{-2l_\alpha(\mu)} + \mu. \tag{2.17}$$

In the case of y_1, it is a little more complicated, because the desired probability content may require dipping into the $y < 0$ region. If not, we have simply

$$y_1 = l_> = -\sqrt{-2l_\alpha(\mu)} + \mu, \quad y_1 \geq 0; \tag{2.18}$$

if so, we have instead

$$y_1 = l_< = \mu/2 + l_\alpha(\mu)/\mu, \quad y_1 < 0. \tag{2.19}$$

The resulting sets $A_\alpha(\mu)$ are shown in Fig. 2.3.

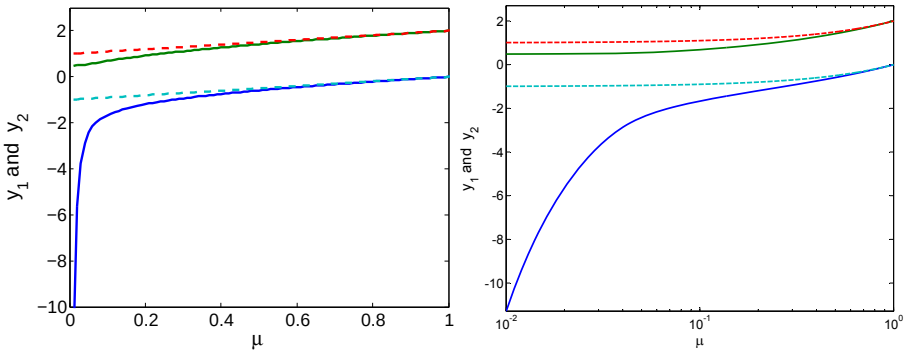

Figure 2.3 The dependence of y_1 (blue) and y_2 (green) on μ, for 68% CL. The dashed curves show the corresponding values for conventional two-sided interval estimation. While y_1 and y_2 are only defined for positive μ, note that the conventional lines extend to negative μ. Left: linear axes; Right: logarithmic axis for μ.

Given an observation y, we use our sets $A_\alpha(\mu)$ to construct a $1-\alpha$ confidence interval for μ according to the algorithm in Section 2.2 of Narsky and Porter (2014a): Look for all sets $A_\alpha(\mu)$ that contain y. The union of the μ values labeling all of these sets is a $1-\alpha$ confidence interval for μ. This corresponds to drawing a horizontal line at the observed y on Fig. 2.3 (for 68% CL) and including all μ values between the intersections with the green and blue solid lines. For small enough y, there is no intersection with the green line, in which case the lower bound on the interval is at $\mu = 0$. Effectively we have an upper limit in this case. We show the 68% confidence intervals as a function of observation y in Fig. 2.4. This is really just a regraphing from Fig. 2.3; now a vertical line is drawn at the observed y, and the intersections with the curves give the confidence intervals in μ. We have explicitly extended the μ axis into the negative region so that it may be seen that nothing special happens for the conventional intervals, while the FC intervals never go below zero.

It should be remarked that the problem statement asks for the 68% FC intervals as a function of $\widehat{\mu}$. If we choose $\widehat{\mu} = y$, we have the result in Fig. 2.4. However, if we restrict to the physical region, $\widehat{\mu} = \max(y, 0)$, then $\widehat{\mu}$ is not sufficient to uniquely specify the FC interval whenever $y < 0$, hence we plot y instead of $\widehat{\mu}$.

We may also compute Bayes intervals at the same confidence level, assuming a uniform prior. These intervals, specified as (u, l) are computed by solving:

$$1 - \alpha = \int_l^u \mathsf{N}(\mu; y, 1)\theta(\mu)d\mu \Big/ \int_0^\infty \mathsf{N}(\mu; y, 1)d\mu, \tag{2.20}$$

where $\theta(\mu)$ is the unit step function. Denote the normalizing denominator in Eq. 2.20 by A. We may compute l and u by ordering on the likelihood as follows: Let δ be given by

$$\int_0^\delta \mathsf{N}(\mu; 0, 1)d\mu = A\frac{1-\alpha}{2}. \tag{2.21}$$

If $y - \delta > 0$ then $l = y - \delta$ and $u = y + \delta$. Otherwise, $l = 0$ and u is the soltution to

$$\int_0^u \mathsf{N}(\mu; y, 1)d\mu = A(1 - \alpha). \tag{2.22}$$

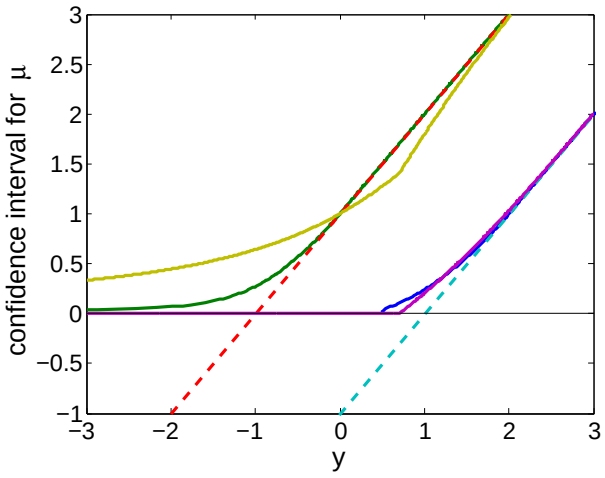

Figure 2.4 The confidence interval for μ as a function of observation y, for $\alpha = 0.68$ (see text). The green and blue curves show the FC intervals. The dashed curves show the confidence inteval for the conventional two-sided interval estimation. The yellow-green and magenta curves show the Bayes interval with a uniform prior.

The MATLAB code used to compute the FC and other intervals is provided in Narsky and Porter (2014c).

y	FC	conventional	Bayes
-2	(0,0.07)	(-3,-1)	(0,0.45)
-1	(0,0.27)	(-2,0)	(0,0.64)
0	(0,1)	(-1,1)	(0,1)
1	(0.24, 2)	(0,2)	(0.2,1.8)

Table 2.1 Comparison of FC and conventional intervals for some small values of observation y. The last column shows the Bayes intervals with a uniform prior. Units are standard deviations of the sampling normal. All are 68% CL.

Discussion:

First, let us be explicit that we restrict present discussion to the case of Gaussian sampling. The methodology can be used for other distributions; however we think it is crucial to understand the "simple" case of the Gaussian before embarking on other complications. Even with this restriction, the discussion is of practical relevance. With the central limit theorem we know that the normal is a good approximation as long as the sample size is large enough. The expected "signal" need not be large (it could be zero, or negative, for that matter) for this to obtain, as long as the overall sample size (that is, including "background") is large. In practice, for 68% confidence sets, "large" may sometimes be fairly small, as illustrated for example by the case study in Section 2.3.1 of Narsky and Porter (2014a).

Next, we contrast the FC and conventional (two-sided) intervals, first without the flip-flop. By construction, both methods provide intervals with correct frequencies. We must look to other properties to determine which might be preferred. For large y, the intervals are essentially the same. Table 2.1 compares the intervals in a region of small y, where the intervals differ.

The conventional intervals are always the same size (2 standard deviations), and reflect the resolution of our measurement independently of the result. In contrast, the FC intervals become smaller as the observation decreases. The information about the resolution is obscured. It is still there, but not evident, and a casual observer might think that the resolution (or the sensitivity) is better than it really is. An interval of size 0.27 looks pretty good, but the standard deviation is really about four times larger. In addition, averaging of results is more straightforward with the conventional intervals. It can still be done with the FC intervals, but effectively the inverse mapping must be performed to extract the standard deviation from the quoted interval. Thus, in the absence of the flip-flop motivation behind FC, the conventional intervals are preferable; it is no accident that they are widely used.

Another example may also be helpful, taking this discussion to a more extreme situation. Suppose we were interested in constructing FC intervals with some given CL less than 50%. In this case, there will be a set of observations y such that the quoted interval is null. For example, consider $1-\alpha = 0.4$. Then values $y < -1.2816$ do not appear in $A_{0.6}(\mu)$ for any μ. Hence, whenever $y < -1.2816$ the confidence interval contains no values of μ. In the case with $\mu = 0$, this means that 10% of observations yield null confidence intervals at 40% CL. Here, we actually do lose

the information that would be needed to go back and permit averaging except as an inequality. This does not mean our intervals are "wrong"; they still have correct frequentist coverage (including the null cases). But their interpretation is not simple.

However, the flip-flop is a real concern, and this is what the FC intervals are designed to mitigate. On the other hand, there is a more elegant solution to this problem: Just quote the conventional two-sided interval independently of the observed value. This has the advantages mentioned above, and is what we recommend. Some people actually do this, but why doesn't everyone? The answer lies in a mindset remarked long ago by no less than Neyman himself [Neyman (1941)]:

> "In spite of the complete simplicity of the above definition, certain persons have difficulty in following it. These difficulties seem to be due to what Karl Pearson (1938) used to call routine of thought. In the present case the routine was established by a century and a half of continuous work with Bayes's theorem..."

This remark comes immediately after writing down the definition of a confidence interval in the frequency sense, and is really about the confusion between thinking like a frequentist and thinking like a Bayesian. It is important to remember that neither way of thinking is wrong; they just have different intents. Frequency statistics is descriptive; it does not attempt to represent "truth", even probabilistically. Bayesian statistics is interpretive; it does attempt to portray truth, in the sense of degree-of-belief, which unavoidably must include prior knowledge. A nice discussion of the distinction may be found in James and Roos (1991).

Why is this relevant here? It is relevant because the flip-flop is indicative of a lack of clarity in what the analyst is trying to convey. On the one hand, s/he wants to present the result of the measurement, preferably concisely and without subjective bias. The conventional two-sided frequency interval does an admirable job of this. On the other hand, physicists, like anyone else, are interested as well in the interpretation, what we infer about reality. This is the domain of Bayesian statistics and is conditioned on prior experience. The physicist who sees a result that is not "significantly" greater than zero is tempted to abandon the objective description in order to quote a result that is not obviously "wrong" in the sense of interpretation. This may be at the subconscious level, but is a symptom of Neyman's remark.

It is discouraging that exactly the same mind-set problem persists with use of the FC. Very often we encounter people who, upon finding a result that is not significantly greater than zero, decide that they are going to quote an FC upper limit. If the result had been significant, they would have quoted a conventional two-sided 68% confidence interval. With a non-significant result, they quote a 90% confidence FC upper limit. The flip-flop is still with us!

It is better to stick with a uniform descriptive approach, and if desired present a distinct interpretive discussion in addition. While both the FC and conventional two-sided intervals provide uniform approaches, the conventional intervals have the advantages noted earlier. The FC interval was introduced because of lack of consistency in applying the conventional intervals. We suggest that, instead of changing the method, the conventional method should be applied consistently. Again, we stress

that we are only discussing normal sampling here, other distributions should also be investigated before generalizing this recommendation.

It might be thought that the FC intervals could be used as interpretive intervals, since they avoid making "wrong" statements in the sense of interpretation. In Table 2.1 we include a column with 68% Bayesian intervals (see also Fig. 2.4), assuming a flat prior distribution. The FC intervals are quite different. If a uniform prior is a reasonable description of prior odds, then the FC intervals do not provide generally comparable posteriors. This shouldn't be surprising; like the conventional intervals, the FC intervals were never intended to be used as posteriors.

A further comment about the Bayesian prior is in order. In this problem we were told to use a flat prior. In fact, if we really tried to put in our prior degree-of-belief, we most likely would use a prior that trends to decrease with increasing μ. Probably we have some prior knowledge such that if μ were very large, we would have noticed it somewhere else (e.g., maybe μ is the rate for some process). The reader may wish to try other priors. The above discussion does not really change, however.

Bibliography

Feldman, G.J. and Cousins, R.D. (1998) A unified approach to the classical statistical analysis of small signals. *Phys. Rev. D*, **57**, 3873–3889.

James, F. and Roos, M. (1991) Statistical notes on the problem of experimental observations near an unphysical region. *Phys.Rev.D*, **44** (1), 299–301.

Narsky, I. and Porter, F.C. (2014a) *Statistical Analysis Techniques in Particle Physics*, Wiley-VCH.

Narsky, I. and Porter, F.C. (2014b) Statistical analysis techniques in particle physics, online supplements. URL `http://www.hep.caltech.edu/~NarskyPorter/Solutions/NPsolution2.5a.m`.

Narsky, I. and Porter, F.C. (2014c) Statistical analysis techniques in particle physics, online supplements. URL `http://www.hep.caltech.edu/~NarskyPorter/Solutions/NPsolution2.5b.m`.

Neyman, J. (1941) Fiducial argument and the theory of confidence intervals. *Biometrika*, **32** (2), 128–150.