2006-12-05

A Bayesian perspective on estimating mean, variance, and standard-deviation from data

Travis E. Oliphant

A Bayesian perspective on estimating mean, variance, and standard-deviation from data

Travis E. Oliphant

December 5, 2006

Abstract

After reviewing some classical estimators for mean, variance, and standard-deviation and showing that un-biased estimates are not usually desirable, a Bayesian perspective is employed to determine what is known about mean, variance, and standard deviation given only that a data set in-fact has a common mean and variance. Maximum-entropy is used to argue that the likelihood function in this situation should be the same as if the data were independent and identically distributed Gaussian. A non-informative prior is derived for the mean and variance and Bayes rule is used to compute the posterior Probability Density Function (PDF) of (μ, σ) as well as $\left(\mu, \sigma^2\right)$ in terms of the sufficient statistics $\bar{x} = \frac{1}{n}\sum_i x_i$ and $C = \frac{1}{n}\sum_i \left(x_i - \bar{x}\right)^2$. From the joint distribution marginals are determined. It is shown that $\left(\frac{\mu-\bar{x}}{\sqrt{C}}\right)\sqrt{n-1}$ is distributed as Student-t with $n-1$ degrees of freedom, $\sigma\sqrt{\frac{2}{nC}}$ is distributed as generalized-gamma with $c=-2$ and $a=\frac{n-1}{2}$, and $\sigma^2\frac{2}{nC}$ is distributed as inverted-gamma with $a=\frac{n-1}{2}$. It is suggested to report the mean of these distributions as the estimate (or the peak if n is too small for the mean to be defined) and a confidence interval surrounding the median.

1 Introduction

A standard concept encountered by anyone exposed to data is the idea of computing a mean, a variance, and a standard deviation from the data. This paper will explore various approaches to computing estimates of mean, standard-deviation, and variance from samples and will conclude by recommending a Bayesian approach to inference about these values from data.

Typically it is assumed that the data are realizations of a collection of independent, identically distributed (*i.i.d.*) random variables. This random vector is denoted $\mathbf{X} = [X_1, X_2, \ldots, X_n]$, and the joint Probability Density Function (PDF) of $\mathbf{X}$ is

$$f_{\mathbf{X}}\left(\mathbf{X}\right) = \prod_{i=1}^{n} f_X\left(x_i\right).$$

1.1 Traditional mean estimate

Commonly, the mean of $\mathbf{X}$ is estimated as the sample average:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

One can then show in a rather satisfying fashion that

$$\begin{aligned} E\left[\hat{\mu}\right] &= E\left[X\right] \\ \mathrm{Var}\left[\hat{\mu}\right] &= \frac{1}{n}\mathrm{Var}\left[X\right]. \end{aligned}$$

These statements are typically used to justify this choice of estimator for the mean as unbiased and consistent. Sometimes this estimator is further justified by noticing that it is also the Maximum Likelihood (ML) estimate for the mean assuming the noise comes from an exponential family (*e.g.* Gaussian).

1.2 Traditional variance estimate

The ML estimate for variance assuming Gaussian noise is

$$\hat{\sigma}^2_{\mathrm{ML}} = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \hat{\mu}\right)^2.$$

It is sometimes suggested to use instead

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \hat{\mu}\right)^2$$

to ensure that

$$E\left[\hat{\sigma}^2\right] = \hat{\sigma}^2.$$

We are supposed to believe that this is preferable to an estimator that instead minimizes some other metric such as the mean-squared error (which includes both bias and variance).

A good discussion of these concepts will also mention that if the X_i are all normal random variables, then $\hat{\mu}$ and $\hat{\sigma}^2$ are independent, $\hat{\mu}$ is normal, and $(n-1)\,\hat{\sigma}^2/\sigma^2$ is chi-squared with $n-1$ degrees of freedom. Confidence intervals can then be determined from these facts in a straightforward way.

1.3 Standard-deviation estimates

Typically, standard-deviation estimates are obtained using $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$. Typically, little is then said about the uncertainty of this estimate. Often, the square-root of the un-biased variance is taken with little justification other than convenience and despite the fact that $\hat{\sigma}$ is generally not an un-biased estimate of σ even when $\hat{\sigma^2}$is.

1.4 Outline of the paper

In this paper, the mean-square error of modified classical estimators for the variance and standard-deviation will be compared. The point of this comparison will be to elucidate which normalization factor gives the smallest error (under the hypothesis of normally-distributed data). While instructive, this comparison does not end the discussion as it does not address the question of whether or not the normalization constant should be the only issue in dispute.

As a result, the problem will be addressed from a Bayesian perspective. Under this perspective, I begin with the assumption that the data has a common mean and variance and use maximum entropy (with a flat prior) to assert that the likelihood function is normal. Using a flat prior for μ and a Jeffrey's prior [2] for σ (and σ^2), the posterior probability of (μ, σ) and $\left(\mu, \sigma^2\right)$ is derived. From this joint posterior, the posterior probability for μ, σ, and σ^2 can be given which leads to simple rules for an estimate and confidence interval calculations.

2 Comparing various estimators

Assuming the X_i come from a standard normal population with mean μ and variance σ^2, three estimators for σ and σ^2 will be compared in terms of the mean-squared error and bias: 1) the unbiased estimator: $\hat{\sigma}_{\mathrm{UB}}$ and $\hat{\sigma}^2_{\mathrm{UB}}$, 2) the maximum-likelihood estimator: $\hat{\sigma}_{\mathrm{ML}}$ and $\hat{\sigma}^2_{\mathrm{ML}}$, and 3) the Minimum Mean-Squared Estimator (MMSE) (among those of a certain class): $\hat{\sigma}_{\mathrm{MMSE}}$ and $\hat{\sigma}^2_{\mathrm{MMSE}}$.

All three estimators of both quantities are of the form

$$\begin{aligned}\hat{\sigma}^2 &= a\sum_{i=1}^{n}\left(X_i - \hat{\mu}\right)^2 \\ \hat{\sigma} &= \sqrt{a\sum_{i=1}^{n}\left(X_i - \hat{\mu}\right)^2}.\end{aligned}$$

For both classes of estimators, the bias, $E\left[\hat{\theta}\right]$, and the mean-square error, $E\left[\left(\hat{\theta}-\theta\right)^2\right] = E\left[\hat{\theta}^2\right] - 2\theta E\left[\hat{\theta}\right] + \theta^2$, will be calculated assuming X_i comes from a normal distribution with mean μ and variance σ^2. The identity

$$\mathrm{MSE}\left[\hat{\theta}\right] \equiv E\left[\left(\hat{\theta}-\theta\right)^2\right] = \mathrm{Var}\left[\hat{\theta}\right] + \left(E\left[\hat{\theta}-\theta\right]\right)^2$$

will be useful in what follows.

2.1 Estimators of variance

For all three estimators of variance it is known that under the hypothesis of normally distributed data, $\hat{\sigma}^2/a\sigma^2$ is χ^2_{n-1} and therefore has mean $n-1$ and variance $2(n-1)$. Consequently,

$$\begin{aligned}
E\left[\hat{\sigma}^2\right] &= a\sigma^2(n-1) \\
E\left[\hat{\sigma}^4\right] &= a^2\sigma^4\left[(n-1)^2 + 2(n-1)\right] \\
&= a^2\sigma^4\left(n^2-1\right) \\
E\left[\left(\hat{\sigma}^2-\sigma^2\right)^2\right] &= E\left[\hat{\sigma}^4\right] - 2\sigma^2 E\left[\hat{\sigma}^2\right] + \sigma^4 \\
&= \sigma^4\left[a^2\left(n^2-1\right) - 2a(n-1) + 1\right].
\end{aligned}$$

It can be shown that the maximum-likelihood estimator for σ^2 requires $a_{\mathrm{ML}} = \frac{1}{n}$. The unbiased estimator for σ^2 is obviously $a_{\mathrm{UB}} = \frac{1}{n-1}$. The minimum mean-square error estimator is found by differentiating Eq. (**??**) and setting the result equal to zero. This procedure results in $a_{\mathrm{MMSE}} = \frac{1}{n+1}$.

The three estimators and their performance are summarized in the following table:

	$\hat{\sigma}^2$	$E\left[\hat{\sigma}^2\right]$	$\mathrm{MSE}\left[\hat{\sigma}^2\right]$
UB	$\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\hat{\mu})^2$	σ^2	$\frac{2\sigma^4}{n-1}$
ML	$\frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{\mu})^2$	$\frac{n-1}{n}\sigma^2$	$\frac{(2n-1)\sigma^4}{n^2}$
MMSE	$\frac{1}{n+1}\sum_{i=1}^{n}(X_i-\hat{\mu})^2$	$\frac{n-1}{n+1}\sigma^2$	$\frac{2\sigma^4}{n+1}$

It is not difficult to show that for $n > 1$

$$\frac{2}{n+1} < \frac{2n-1}{n^2} < \frac{2}{n-1},$$

and therefore in a mean-square sense, the MMSE and ML estimators are both better than the unbiased estimator. This example serves to show a general property that improved estimators are usually possible in a mean-square sense by using biased estimators.

Figure 1 shows $\sqrt{\mathrm{MSE}\left[\hat{\sigma}^2\right]/\sigma^4}$ and $E\left[\hat{\sigma}^2\right]/\sigma^2$ for the three estimators when $n > 1$.

3 Estimators for σ

Estimators for σ are not often discussed, but are often used and should, therefore, receive better treatment. For normally distributed data, the maximum likelihood estimator for σ is

$$\hat{\sigma}_{\mathrm{ML}} = \sqrt{\hat{\sigma}^2_{\mathrm{ML}}} = \sqrt{\frac{1}{n}\sum_i (X_i-\hat{\mu})^2},$$

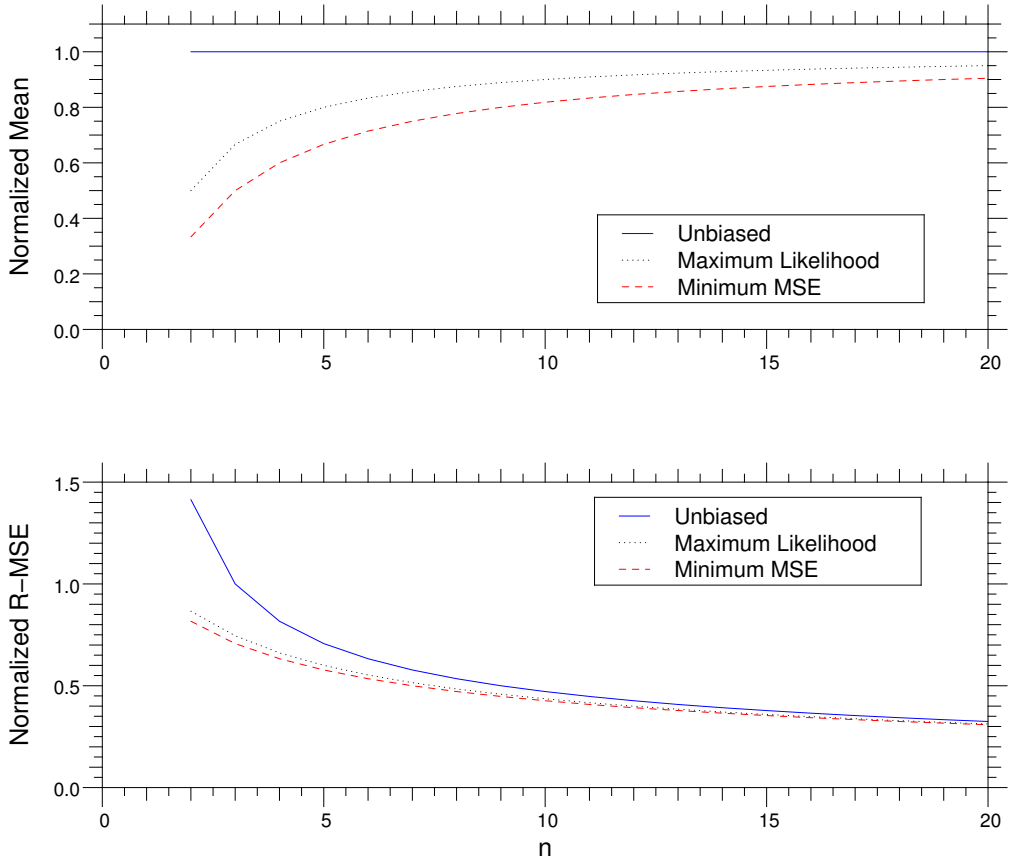

Figure 1: Normalized mean, $E\left[\hat{\sigma}^2\right]/\sigma^2$ and normalized root-mean-square error (R-MSE), $\sqrt{\mathrm{MSE}\left[\hat{\sigma}^2\right]/\sigma^4}$, of several estimators of σ^2.

and thus $a_{\mathrm{ML}} = \frac{1}{n}$. The mean and mean-square error for all three estimators can be computed by noticing that $\sqrt{\hat{\sigma}^2/\sigma^2 a}$ is χ_{n-1} (a chi random variable with $n-1$ degrees of freedom). Because

$$\begin{aligned} E\left[\chi_{n-1}\right] &= \frac{\sqrt{2}\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} \\ \mathrm{Var}\left[\chi_{n-1}\right] &= n-1-2\left[\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\right]^2 \end{aligned}$$

we can conclude that

$$\begin{aligned} E\left[\hat{\sigma}\right] &= \sigma\frac{\sqrt{2a}\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)} = t_n\sigma\sqrt{2a} \\ E\left[\left(\hat{\sigma}-\sigma\right)^2\right] &= \mathrm{Var}\left[\hat{\sigma}\right] + \left(E\left[\hat{\sigma}\right]-\sigma\right)^2 \\ &= a\sigma^2\left(n-1\right) - 2a\sigma^2\left[\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\right]^2 + \sigma^2\left(\frac{\sqrt{2a}\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}-1\right)^2 \\ &= \sigma^2\left[a\left(n-1\right) - 2\sqrt{2a}t_n + 1\right] \end{aligned}$$

where

$$t_n = \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}.$$

From these expressions, the unbiased estimator will result if $a_{\mathrm{UB}} = 1/2t_n^2$ while the minimum mean-square estimator can be found by differentiating with respect to a the expression for mean-square error and solving for a. The result is $a_{\mathrm{MMSE}} = 2t_n^2/\left(n-1\right)^2$.

The following table summarizes the estimators and their performance.

	$\hat{\sigma}$	$E\left[\hat{\sigma}\right]$	$\mathrm{MSE}\left[\hat{\sigma}\right]$
UB	$\frac{1}{t_n}\sqrt{\frac{1}{2}\sum_{i=1}^{n}\left(X_i-\hat{\mu}\right)^2}$	σ	$\sigma^2\left[\frac{n-1}{2t_n^2}-1\right]$
ML	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\hat{\mu}\right)^2}$	$t_n\sqrt{\frac{2}{n}}\sigma$	$2\sigma^2\left[1-t_n\sqrt{\frac{2}{n}}-\frac{1}{2n}\right]$
MMSE	$\frac{t_n}{n-1}\sqrt{2\sum_{i=1}^{n}\left(X_i-\hat{\mu}\right)^2}$	$\frac{2t_n^2}{n-1}\sigma^2$	$\sigma^2\left[1-\frac{2t_n^2}{n-1}\right]$

It can be shown (or observed from the plot below) that

$$1-\frac{2t_n^2}{n-1} < 2-2t_n\sqrt{\frac{2}{n}}-\frac{1}{n} < \frac{n-1}{2t_n^2}-1.$$

Therefore, comparing the estimators on the basis of mean-squared error results in the MMSE and the ML estimator outperforming the unbiased estimator.

Figure 2 shows plots of $E\left[\hat{\sigma}\right]/\sigma$ and $\sqrt{\mathrm{MSE}\left[\hat{\sigma}\right]/\sigma^2}$ to give some idea of the small-sample performance of these different estimators on normal data.

4 Bayesian Perspective

Given data $\{x_1, x_2, x_3, \ldots, x_n\}$, the task is to find the mean μ, variance $\sigma^2 = v$ and standard-deviation σ of these data. As stated the problem doesn't have a solution. More information is needed in order to work

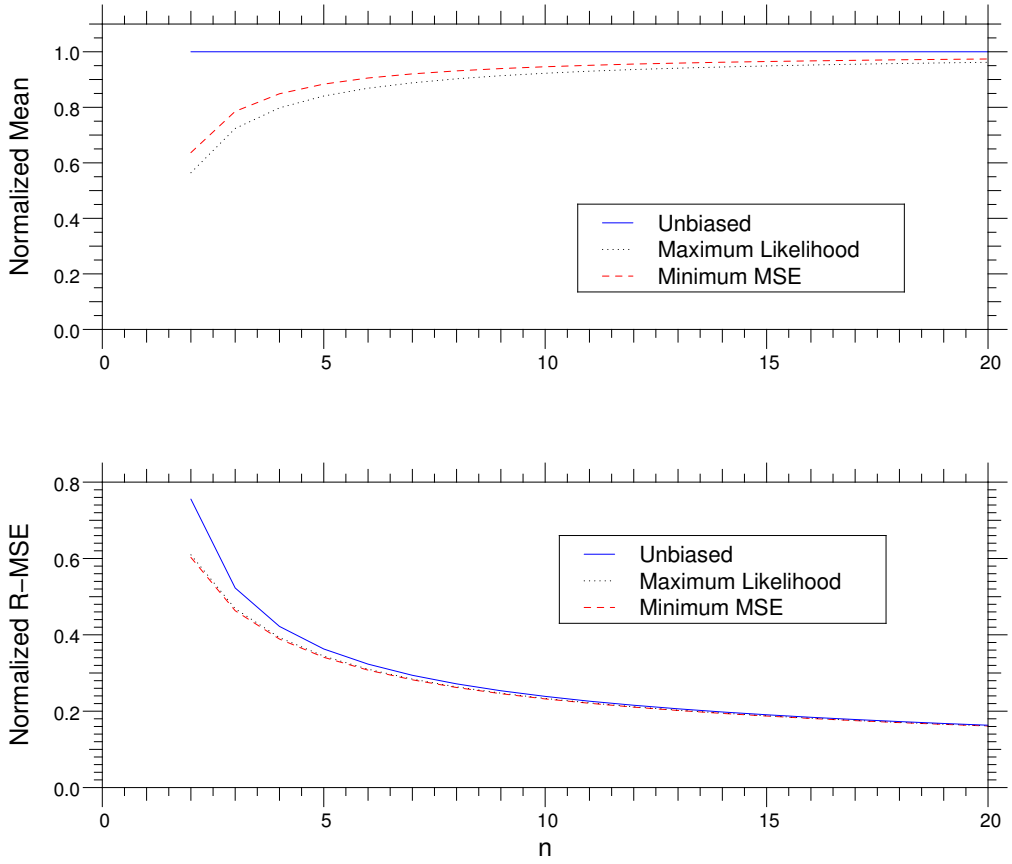

Figure 2: Normalized mean, $E\left[\hat{\sigma}\right]/\sigma$ and normalized root-mean-square error (R-MSE), $\sqrt{\mathrm{MSE}\left[\hat{\sigma}\right]/\sigma^2}$, of several estimators of σ.

towards an answer. First, assume that data has a common mean and a common variance. The principle of maximum entropy can then be applied under these constraints (using a flat "ignorance" prior) to choose the distribution

$$f\left(\mathbf{X}|\mu,\sigma\right)=\frac{1}{\left(2\pi\right)^{n/2}\sigma^{n}}\exp\left[-\frac{1}{2\sigma^{2}}\sum_{i}\left(x_{i}-\mu\right)^{2}\right].$$

which adds the least amount of information to the problem other than the assumption of a common μ and σ^2. Notice that we can use maximum entropy (with a flat "ignorance" distribution so that entropy is $-\int f\left(x\right)\log f\left(x\right)dx$) to justify the common assumption of normal *i.i.d.* data. Using Bayes rule we find that

$$\begin{aligned}f\left(\mu,\sigma|\mathbf{X}\right) &= \frac{f\left(\mathbf{X}|\mu,\sigma\right)f\left(\mu,\sigma\right)}{f\left(\mathbf{X}\right)}\\ &= D_{n}f\left(\mathbf{X}|\mu,\sigma\right)f\left(\mu,\sigma\right)\end{aligned}$$

where D_n is a normalizing constant. This distribution tells us all the information that is available about μ and σ given the data $\mathbf{X}$. We can use this joint-PDF to estimate μ and/or σ and to report confidence in the estimates.

4.1 Choosing the prior $f\left(\mu,\sigma\right)$

Central to solving this problem is choosing the prior knowledge for μ and σ. Because we can normalize the random variables using

$$\frac{Z-\mu}{\sigma}$$

to obtain zero-mean, unit variance random variables, μ is a location parameter, and σ is a scale parameter. Following Jaynes's reasoning [1], we choose the prior which expresses complete ignorance except for the fact that μ is a location parameter and σ is a scale parameter. In other words, we consider a new problem with data $\mathbf{x}'$ which is shifted and scaled version of the old data. The prior in both of these case should be the same function. However the prior has adjusted according to well-established rules. This defines an expression that the prior should satisfy:

$$f\left(\mu,\sigma\right)=af\left(\mu+b,a\sigma\right)$$

where $a>0$ and b is an arbitrary real number. The prior that satisfies this transformation equation is the so-called "Jeffrey's" prior.

$$f\left(\mu,\sigma\right)=\frac{\text{const}}{\sigma}.$$

This prior is improper in the sense that it is not normalizable by itself. However, when used to find the posterior a total normalization constant can be found.

Specifically,

$$\begin{aligned}f\left(\mu,\sigma|\mathbf{X}\right) &= \frac{D_{n}}{\sigma^{n+1}}\exp\left[-\frac{1}{2\sigma^{2}}\sum_{i}\left(x_{i}-\mu\right)^{2}\right]\\ &= \frac{D_{n}}{\sigma^{n+1}}\exp\left[-\frac{1}{2\sigma^{2}}\left(\sum_{i}x_{i}^{2}-2\mu\sum_{i}x_{i}+n\mu^{2}\right)\right]\\ &= \frac{D_{n}}{\sigma^{n+1}}\exp\left[-\frac{\left(\mu-\overline{x}\right)^{2}+C}{2\sigma^{2}/n}\right]\end{aligned}$$

where

$$\begin{aligned}\overline{x} &= \frac{1}{n}\sum_{i}x_{i}\\ C &= \overline{x^{2}}-\overline{x}^{2}=\frac{1}{n}\sum_{i}x_{i}^{2}-\left(\frac{1}{n}\sum_{i}x_{i}\right)^{2}=\frac{1}{n}\sum_{i}\left(x_{i}-\overline{x}\right)^{2}\end{aligned}$$

$$\begin{aligned} D_n &= \left[\int_0^\infty \frac{1}{\sigma^{n+1}} \int_{-\infty}^\infty \exp\left(-\frac{\alpha^2+C}{2\sigma^2/n}\right) d\alpha d\sigma\right]^{-1} \\ &= \sqrt{\frac{n^n C^{n-1}}{\pi 2^{n-2}}} \frac{1}{\Gamma\left(\frac{n-1}{2}\right)}. \end{aligned}$$

This joint posterior PDF tells the whole story about μ and σ if only samples constrained to have the same μ and σ are given. Using this joint PDF we can compute any desired probability. Notice that $n > 1$ or else $D_n \to 0$ which is expressing the fact that with $n = 1$ there is no information about σ whatever.

Later, will be needed the joint posterior PDF of μ and $v = \sigma^2$ which is

$$f(\mu, v|\mathbf{X}) = G_n f(\mathbf{X}|\mu, v) f(\mu, v)$$

where

$$f(\mu, v) = \frac{\text{const}}{\nu}$$

so that we are just as uniformed about v as about σ. Then,

$$\begin{aligned} G_n^{-1} &= \int_0^\infty \int_{-\infty}^\infty v^{-\left(\frac{n+2}{2}\right)} \exp\left[-\frac{(\mu-\overline{x})^2+C}{2v/n}\right] d\mu dv \\ G_n &= \sqrt{\frac{n^n C^{n-1}}{\pi 2^n}} \frac{1}{\Gamma\left(\frac{n-1}{2}\right)} = \frac{1}{2} D_n. \end{aligned}$$

4.2 Marginal distributions

The joint distributions provide all of the information available about the parameters of interest using the data and the assumptions. Notice that these distributions only depend on the data (specifically, the statistics $\bar{\mathbf{x}}$ and C), and can be used easily to compute confidence intervals.

We can integrate out one of the variables and get just the marginal density function of μ or σ separately.

$$\begin{aligned} f(\mu|\mathbf{X}) &= \int_0^\infty \frac{D_n}{\sigma^{n+1}} \exp\left[-\frac{(\mu-\overline{x})^2+C}{2\sigma^2/n}\right] d\sigma \\ &= \frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{\pi C}} \left[1+\frac{(\mu-\overline{x})^2}{C}\right]^{-n/2} \end{aligned}$$

so that $\left(\frac{\mu-\overline{x}}{\sqrt{C}}\right)\sqrt{n-1}$ is Student-t distributed with $n-1$ degrees of freedom. We naturally need $n > 1$ for this distribution to provide information. When $n = 1$ we have an improper distribution for μ proportional to $\frac{1}{|x-\mu|}$. For other cases we can deduce:

$$\begin{aligned} E[\mu|\mathbf{X}] &= \overline{x} \\ \text{Var}[\mu|\mathbf{X}] &= \frac{C}{n-3} \quad n > 3 \\ \arg\max_\mu f(\mu|\mathbf{X}) &= \overline{x}. \end{aligned}$$

The marginal distribution of σ is

$$\begin{aligned} f(\sigma|\mathbf{X}) &= \int_{-\infty}^\infty \frac{D_n}{\sigma^{n+1}} \exp\left[-\frac{\alpha^2+C}{2\sigma^2/n}\right] d\alpha \\ &= D_n \frac{\sqrt{2\pi}}{\sigma^n\sqrt{n}} \exp\left[-\frac{nC}{2\sigma^2}\right] \quad \sigma > 0 \\ &= \sqrt{\frac{n^{n-1}C^{n-1}}{2^{n-1}}} \frac{2\exp\left[-\frac{nC}{2\sigma^2}\right]}{\Gamma\left(\frac{n-1}{2}\right)\sigma^n} \quad \sigma > 0. \end{aligned}$$

Thus, $\frac{\sigma\sqrt{2}}{\sqrt{nC}}$ is generalized gamma distributed with shape parameters $c = -2$ and $a = \frac{n-1}{2}$. If $n = 1$, the distribution reduces to an improper distribution proportional to $1/\sigma$ (i.e. we have no additional information about σ other than what we started with. For other values of n we can find:

$$\begin{aligned} E\left[\sigma|\mathbf{X}\right] &= \sqrt{\frac{n}{2}}\frac{\Gamma\left(\frac{n-2}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\sqrt{C} \quad n > 2, \\ \mathrm{Var}\left[\sigma|\mathbf{X}\right] &= \frac{n}{2}\left[\frac{2}{n-3} - \frac{\Gamma^2\left(\frac{n}{2}-1\right)}{\Gamma^2\left(\frac{n-1}{2}\right)}\right]C \quad n > 3. \\ \arg\max_{\sigma} f\left(\sigma|\mathbf{X}\right) &= \sqrt{C}. \end{aligned}$$

This distribution does not have a well-defined mean unless $n > 2$ and it does not have a well-defined variance unless $n > 3$.

Finally, the marginal distribution of $v = \sigma^2$ is

$$\begin{aligned} f\left(v|\mathbf{X}\right) &= \int_{-\infty}^{\infty} G_n v^{-\frac{n+2}{2}} \exp\left[-\frac{\alpha^2 + C}{2\nu/n}\right] d\alpha \\ &= \frac{\left(\frac{nC}{2}\right)^{\frac{n-1}{2}}}{\Gamma\left(\frac{n-1}{2}\right) v^{(n+1)/2}} \exp\left[-\frac{nC}{2v}\right] \quad v > 0 \end{aligned}$$

When $n = 1$, this also reduces to an improper distribution proportional to $1/v$. For other values of n, $\frac{\nu 2}{nC}$ is an inverted gamma distribution with $a = \frac{n-1}{2}$. Useful parameters of this distribution are

$$\begin{aligned} E\left[\sigma^2|\mathbf{X}\right] &= \frac{n}{n-3}C \quad n > 3 \\ \mathrm{Var}\left[\sigma^2|\mathbf{X}\right] &= \frac{2n^2C^2}{(n-3)^2(n-5)} \quad n > 5 \\ \arg\max_{\sigma^2} f\left(\sigma^2|\mathbf{X}\right) &= \frac{n}{n+1}C. \end{aligned}$$

Notice that this distribution does not have a well-defined mean unless $n > 3$ and does not have a well-defined variance unless $n > 5$.

To illustrate, the posterior probabilities for various numbers of samples, Figures 3, 4, 5 show normalized plots of the Student-t, generalized Gamma, and inverted gamma distributions for n =3, 10, and 50, corresponding to the mean, standard-deviation, and variance of the data sample.

4.3 Gaussian approximations

The marginal posterior distributions for μ, σ, and σ^2 all approach Normal distributions as $n \to \infty$. In particular, the posterior distribution for μ approaches a normal distribution with mean $\bar{x}$ and variance $\frac{C}{n}$. The posterior distribution for σ approaches a normal distribution with mean $\sqrt{C}$ and variance $\frac{C}{2n}$. Finally, the posterior distribution for σ^2 approaches a normal distribution with mean C and variance $\frac{2C^2}{n}$.

4.4 Joint MAP estimators

Joint Maximum A-Posterior (MAP) estimators are sometimes useful Because $\mu|\mathbf{X}$ and $\sigma|\mathbf{X}$ (and similarly $\mu|\mathbf{X}$ and $\sigma^2|\mathbf{X}$) are not independent, the joint MAP estimator can produce different results than the marginal MAP estimators. These estimators minimize the jointly-uniform loss function. To find the joint estimator we solve

$$\begin{aligned} \hat{\mu}, \hat{\sigma} &= \arg\max_{\mu,\sigma} f\left(\mu, \sigma|\mathbf{X}\right) \\ &= \arg\min_{\mu,\sigma} \left[-\log f\left(\mu, \sigma|\mathbf{X}\right)\right] \\ &= \arg\min_{\mu,\sigma} \left[(n+1)\log\sigma + \frac{(\mu - \overline{x})^2 + C}{2\sigma^2/n}\right] \end{aligned}$$

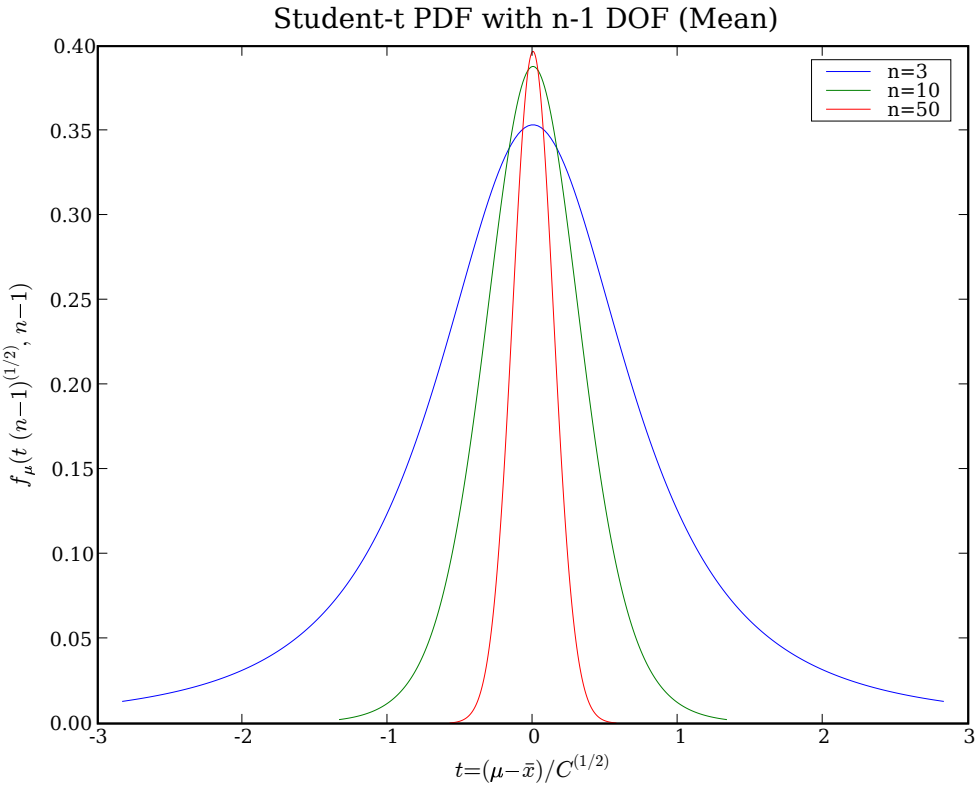

Figure 3: Graph of the posterior PDF for the mean for several values of n. The function $f_\mu(t,\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)\left[1+\frac{x^2}{\nu}\right]^{\frac{\nu+1}{2}}}$ is the PDF of the Student-t distribution with ν Degrees Of Freedom (DOF).

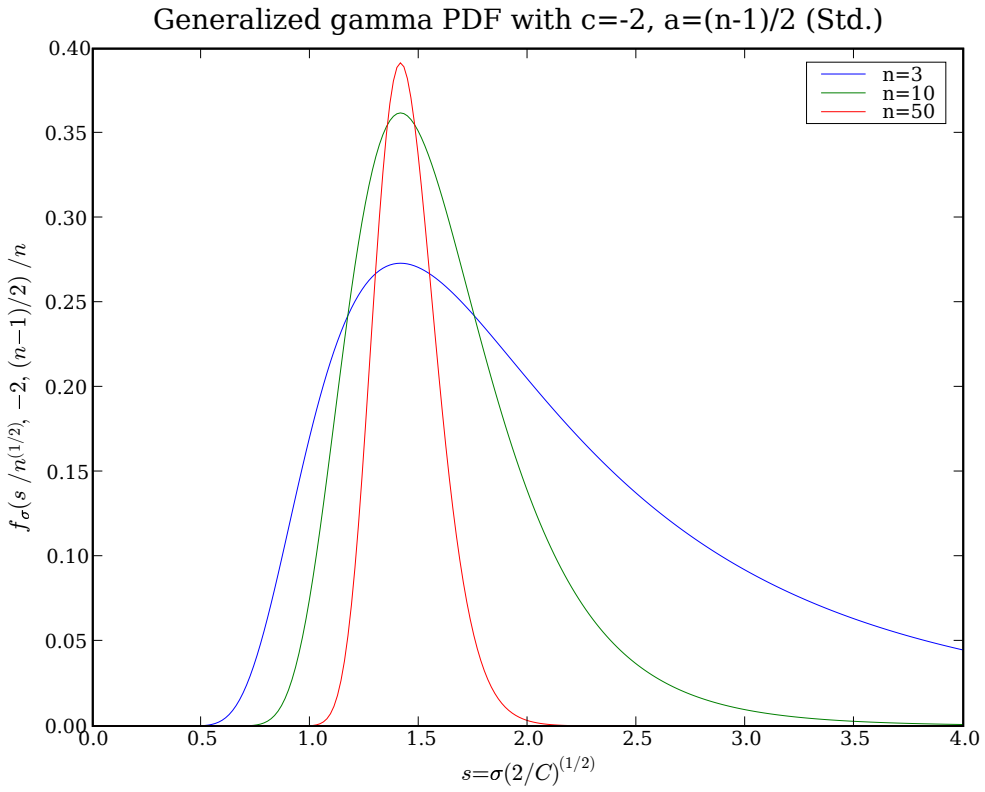

Figure 4: Graph of the posterior PDF for the standard-deviation for several values of n. The function $f_\sigma\left(s,c,a\right)=\frac{|c|x^{ca-1}}{\Gamma(a)}\exp\left(-x^c\right)\quad x>0$ is the PDF of the generalized gamma distribution with shape parameters c and a.

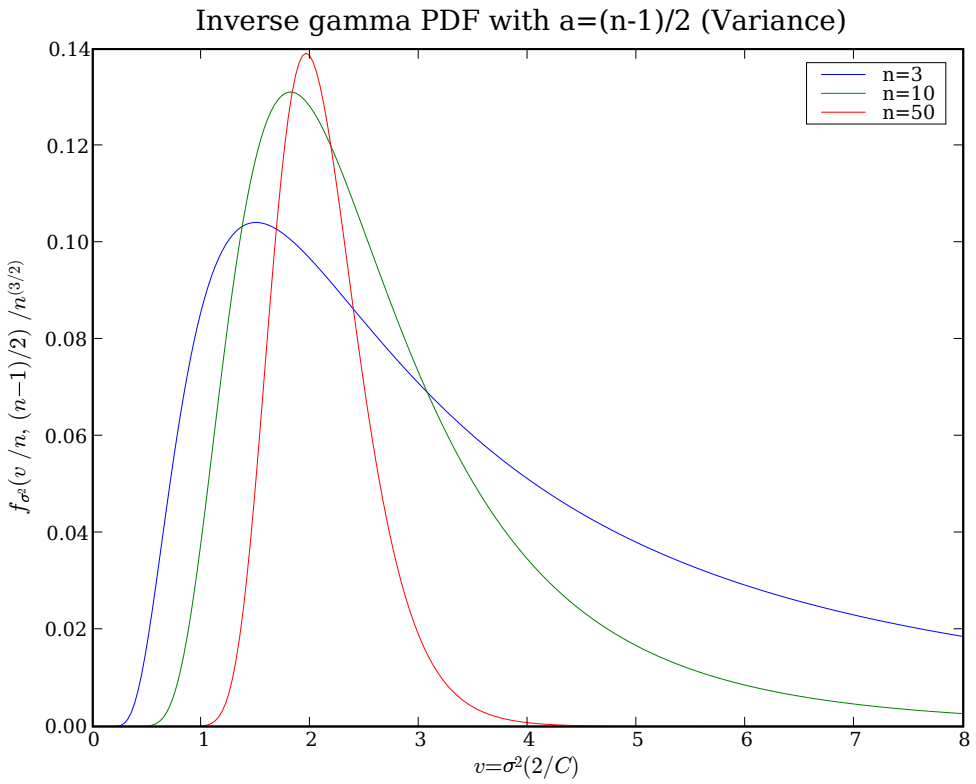

Figure 5: Graph of the posterior PDF for the standard-deviation for several values of n. The function $f_\sigma\,(s,a) = \frac{1}{\Gamma(a)} x^{-a-1} \exp\left(-\frac{1}{x}\right) \quad x > 0$ is the PDF of the inverse gamma distribution with shape parameter a.

and thus,

$$\begin{aligned} 0 &= \frac{(\hat{\mu} - \overline{x})}{\hat{\sigma}^2/n} \\ 0 &= \frac{n+1}{\hat{\sigma}} - \frac{(\hat{\mu} - \overline{x})^2 + C}{\hat{\sigma}^3/n}. \end{aligned}$$

Solving these simultaneously gives

$$\begin{aligned} \hat{\mu}_{JMAP} &= \overline{x}, \\ \hat{\sigma}_{JMAP} &= \sqrt{\frac{n}{n+1}C}. \end{aligned}$$

The joint estimator for μ and v can also be found in the same way.

$$\hat{\mu}, \hat{v} = \arg\min_{\mu,v} \left[\frac{n+2}{2} \log v + \frac{(\mu - \overline{x})^2 + C}{2v/n} \right].$$

Differentiating results in

$$\begin{aligned} 0 &= \frac{\hat{\mu} - \overline{x}}{\hat{v}/n} \\ 0 &= \frac{n+2}{2\nu} - \frac{(\hat{\mu} - \overline{x})^2 + C}{2v^2/n}. \end{aligned}$$

Solving these simultaneously gives

$$\begin{aligned} \hat{\mu}_{\text{JMAP}} &= \overline{x} \\ \hat{\sigma}^2_{\text{JMAP}} &= \frac{n}{n+2}C. \end{aligned}$$

While the estimator for the mean is un-interesting, a wide variety of normalization constants to C show up in this analysis. Using these estimates, requires a particular devotion to maximizing $f(\mu, \sigma|\mathbf{V})$ instead of other estimation approaches. The most useful approach to understanding estimates of μ, σ, and σ^2 is to determine confidence intervals which is the subject of the next section.

4.5 Confidence intervals

One of the advantages of the Bayesian perspective is that it automatically provides a method to obtain practical confidence intervals for the estimates. With the probability density function given, confidence intervals can be constructed by finding an area straddling the mean (or the peak, or the median) with equal areas on either side. Given the nature of the confidence interval as an area there is some aesthetic value in choosing the median as the middle value to surround.

4.5.1 General case

Suppose there is a parameter with probability density function $f(\theta)$ and cumulative distribution function $F(\theta)$. How is a confidence interval, $[a, b]$, constructed about the mean, peak, and/or median. The interval should be such that the probability of θ lying within the range is $\alpha \times 100$ percent, where α is a given parameter. In other words, the area under $f(\theta)$ over the confidence interval should be α. Suppose $\hat{\theta}$ is the position about which it is desired an equal-area confidence interval. Then, the two end points of the interval can be calculated from

$$\begin{aligned} P\left\{a \le \theta \le \hat{\theta}\right\} &= \frac{\alpha}{2} \\ P\left\{\hat{\theta} \le \theta \le b\right\} &= \frac{\alpha}{2}. \end{aligned}$$

These state that

$$\begin{aligned} F\left(\hat{\theta}\right) - F\left(a\right) &= \frac{\alpha}{2} \\ F\left(b\right) - F\left(\hat{\theta}\right) &= \frac{\alpha}{2} \end{aligned}$$

so that

$$\begin{aligned} a &= F^{-1}\left[F\left(\hat{\theta}\right) - \frac{\alpha}{2}\right] \\ b &= F^{-1}\left[F\left(\hat{\theta}\right) + \frac{\alpha}{2}\right]. \end{aligned}$$

For the confidence interval about the median, $F\left(\hat{\theta}\right) = \frac{1}{2}$ so that for that important case.

$$\begin{aligned} a &= F^{-1}\left[\frac{1-\alpha}{2}\right] \\ b &= F^{-1}\left[\frac{1+\alpha}{2}\right] \end{aligned}$$

4.5.2 Mean

For the case of the mean, we have seen that the distribution of $\frac{\mu|\mathbf{X}-\overline{x}}{\sqrt{C/(n-1)}}$ is Student-t with $n-1$ degrees of freedom. As a result,

$$F_{\mu}^{-1}\left(q\right) = \overline{x} + \sqrt{\frac{C}{n-1}} F_t^{-1}\left(q; n-1\right)$$

where $F_t^{-1}\left(q;\nu\right)$ is the inverse cumulative distribution function (cdf) of the Student-t distribution with ν degrees of freedom. In addition, the mean, the median, and the peak are all the same value. Note also that because the Student-t distribution is symmetric:

$$F_t^{-1}\left(\frac{1}{2} - q; \nu\right) = -F_t^{-1}\left(\frac{1}{2} + q; \nu\right)$$

so that

$$\begin{aligned} a &= \overline{x} - \sqrt{\frac{C}{n-1}} F_t^{-1}\left(\frac{1+\alpha}{2}; n-1\right) \\ b &= \overline{x} + \sqrt{\frac{C}{n-1}} F_t^{-1}\left(\frac{1+\alpha}{2}; n-1\right). \end{aligned}$$

4.5.3 Standard deviation

For the case of the standard deviation, we have seen that the distribution of $(\sigma|\mathbf{X})\sqrt{\frac{2}{nC}}$ is generalized gamma with $c = -2$ and $a = \frac{n-1}{2}$. Therefore

$$F_{\sigma}^{-1}\left(q\right) = \sqrt{\frac{nC}{2}} F_1^{-1}\left(q; \frac{n-1}{2}\right)$$

where $F_1^{-1}\left(q;a\right)$ is the inverse cumulative distribution function (cdf) of the generalized gamma distribution with parameters $c = -2$ and a:

$$F_1^{-1}\left(q;a\right) = \left\{\Gamma^{-1}\left[a, \Gamma\left(a\right) q\right]\right\}^{-1/2}$$

where $\Gamma\left(a, \Gamma^{-1}\left(a, y\right)\right) = y$. As a result:

$$\begin{aligned} a &= \sqrt{\frac{nC}{2}} F_1^{-1}\left(F_1\left(\hat{\sigma}\sqrt{\frac{2}{nC}}; \frac{n-1}{2}\right) - \frac{\alpha}{2}; \frac{n-1}{2}\right) \\ b &= \sqrt{\frac{nC}{2}} F_1^{-1}\left(F_1\left(\hat{\sigma}\sqrt{\frac{2}{nC}}; \frac{n-1}{2}\right) + \frac{\alpha}{2}; \frac{n-1}{2}\right) \end{aligned}$$

where $F_1(x;a) = 1-\Gamma\left(a, x^{-2}\right)/\Gamma(a)$ is the cumulative distribution function (cdf) of the generalized gamma with parameter a and $c=-2$. When using the median as the center point, these expressions simplify to:

$$\begin{aligned} a &= \sqrt{\frac{nC}{2}} F_1^{-1}\left(\frac{1-\alpha}{2}; \frac{n-1}{2}\right), \\ b &= \sqrt{\frac{nC}{2}} F_1^{-1}\left(\frac{1+\alpha}{2}; \frac{n-1}{2}\right). \end{aligned}$$

If the peak of the distribution is used as the center point, then

$$\begin{aligned} a &= \sqrt{\frac{nC}{2}} F_1^{-1}\left(F_1\left(\sqrt{\frac{2}{n}}; \frac{n-1}{2}\right) - \frac{\alpha}{2}; \frac{n-1}{2}\right) \\ b &= \sqrt{\frac{nC}{2}} F_1^{-1}\left(F_1\left(\sqrt{\frac{2}{n}}; \frac{n-1}{2}\right) + \frac{\alpha}{2}; \frac{n-1}{2}\right) \end{aligned}$$

4.5.4 Variance

We have seen that the distribution of $\left(\sigma^2|\mathbf{X}\right)\frac{2}{nC}$ is inverted gamma with $a=\frac{n-1}{2}$. Therefore,

$$F_{\sigma^2}^{-1}(q) = \frac{nC}{2} F_I^{-1}\left(q; \frac{n-1}{2}\right)$$

where $F_I^{-1}(q;a)$ is the inverse cdf of the inverted gamma distribution with parameter a:

$$F_I^{-1}(q;a) = \frac{1}{\Gamma^{-1}\left[a, \Gamma(a)\, q\right]}$$

where $\Gamma\left(a, \Gamma^{-1}(a,y)\right) = y$. Therefore,

$$\begin{aligned} a &= \frac{nC}{2} F_I^{-1}\left(F_I\left(\hat{\sigma}^2 \frac{2}{nC}; \frac{n-1}{2}\right) - \frac{\alpha}{2}; \frac{n-1}{2}\right) \\ b &= \frac{nC}{2} F_I^{-1}\left(F_I\left(\hat{\sigma}^2 \frac{2}{nC}; \frac{n-1}{2}\right) + \frac{\alpha}{2}; \frac{n-1}{2}\right) \end{aligned}$$

where $F_I(x;a) = 1-\Gamma\left(a, x^{-1}\right)/\Gamma(a)$ is the cdf of the inverted gamma with parameter a. Again, if the median is used as the center point, then these expression simplify to

$$\begin{aligned} a &= \frac{nC}{2} F_I^{-1}\left(\frac{1-\alpha}{2}; \frac{n-1}{2}\right), \\ b &= \frac{nC}{2} F_I^{-1}\left(\frac{1+\alpha}{2}; \frac{n-1}{2}\right). \end{aligned}$$

If the peak of the marginal distribution is used as the center point, the equations are

$$\begin{aligned} a &= \frac{nC}{2} F_I^{-1}\left(F_I\left(\frac{2}{n+1}; \frac{n-1}{2}\right) - \frac{\alpha}{2}; \frac{n-1}{2}\right) \\ b &= \frac{nC}{2} F_I^{-1}\left(F_I\left(\frac{2}{n+1}; \frac{n-1}{2}\right) + \frac{\alpha}{2}; \frac{n-1}{2}\right) \end{aligned}$$

Notice that the confidence interval for the variance is the square of the confidence interval for the standard deviation when the median of the distribution is used in both cases. Also, care must be taken for large α and small n that none of the arguments to the inverse cdf are negative. Such a situation, indicates that symmetry around the peak is impossible. Therefore, either the median should be used as the middle point, or the area should be taken from 0 to an upper bound, b.

5 Discussion

Having learned that the minimum mean-square estimator for θ from data $\mathbf{X}$ is $E[\theta|\mathbf{X}]$, one might be surprised by the fact that the expected value of the posterior marginal distribution in this case does not result in the same estimator as the minimum mean-square estimator over all classes of estimators of the form aC even though it has the same form. For example, the classic estimator for $\hat{\sigma}^2$ that gives minimum MSE is $\frac{n}{n+1}C$ but the Bayesian minimum mean-square estimator is $\frac{n}{n-3}C$. Why are these different? The difference comes in the subtle distinction between the two estimators. The former finds the value of a that minimizes the integral

$$\int \left(aC - \sigma^2\right)^2 f\left(\hat{\sigma}^2\right) d\hat{\sigma}^2$$

while the latter finds the function of $\mathbf{X}$ (which happens to be $aC(\mathbf{X})$) that minimizes the integral

$$\int\int \left(g(\mathbf{X}) - \sigma^2\right)^2 f\left(\mathbf{X}, \sigma^2\right) d\sigma^2 d\mathbf{X}.$$

This final integral includes an averaging integral against the non-informative prior as well as an integral over the data. These are two entirely different optimization problems and should not be expected to provide the same result.

It is important to understand the full probability distribution of μ, σ, and σ^2 especially when the number of data-samples is small. For example, the variance of the posterior probability distribution for σ^2 is not even defined if $n \leq 5$. As a result, it can be impossible to just report the mean and variance. A confidence interval (or high-density region) around the median of the distribution is always possible.

6 Summary and Conclusions

In this paper, a study of several estimators for the mean, variance, and standard-deviation of data was presented. In particular, it was shown that the unbiased estimator for variance so commonly used is not typically a good choice (especially for small n) because using $n+1$ as a divisor rather than $n-1$ shrinks the mean-square error of the estimator.

In addition, a fully Bayesian perspective on the problem of estimating a common mean and variance from samples was presented using maximum entropy and non-informative priors. The results provide the posterior conditional PDF of the mean, standard-deviation, and variance from which estimates and confidence intervals can be calculated.

The results also emphasize the point that calculating the Bayesian Mean-Square Error (MSE) is not necessarily the same as other non-Bayesian definitions of MSE because it involves another averaging integral over the prior information on the quantity to be estimated. The mean of the conditional PDF minimizes Bayesian MSE.

Table 1 summarizes the results for understanding μ, σ, and σ^2 from data assumed to have a common mean and variance. The table requires only the sufficient statistics

$$\begin{aligned} \bar{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i \\ C &= \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 \end{aligned}$$

It is also very useful to note that $\left(\frac{\mu-\overline{x}}{\sqrt{C}}\right)\sqrt{n-1}$ is Student-t distributed with $n-1$ degrees of freedom, $\sigma\frac{\sqrt{2}}{\sqrt{nC}}$ is generalized gamma distributed with shape parameters $c=-2$ and $a=\frac{n-1}{2}$, and $\sigma^2\frac{2}{nC}$ is inverted gamma with shape parameter $a=\frac{n-1}{2}$. These final facts can be used to compute estimates and confidence intervals from standardized tables and distributions.

Table 1: Summary of posterior probability distributions for μ, σ, and σ^2.

$\cdot$	PDF: $f(\cdot\|\mathbf{X})$	Mode	$E[\cdot]$	$\mathrm{Var}[\cdot]$
μ	$\frac{\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)\sqrt{\pi C}}\left[1+\frac{(\mu-\overline{x})^2}{C}\right]^{-n/2}$	$\bar{x}$	$\bar{x}$	$\frac{C}{n-3} \quad n>3$
σ	$\frac{2(nC/2)^{(n-1)/2}}{\Gamma\left(\frac{n-1}{2}\right)\sigma^n}\exp\left[-\frac{nC}{2\sigma^2}\right] \quad \sigma>0$	$\sqrt{C}$	$\sqrt{\frac{n}{2}}\frac{\Gamma\left(\frac{n-2}{2}\right)}{\Gamma\left(\frac{n-1}{2}\right)}\sqrt{C} \quad n>2$	$\frac{n}{2}\left[\frac{2}{n-3}-\frac{\Gamma^2\left(\frac{n}{2}-1\right)}{\Gamma^2\left(\frac{n-1}{2}\right)}\right]C \quad n>3$
σ^2	$\frac{(nC/2)^{(n-1)/2}}{\Gamma\left(\frac{n-1}{2}\right)(\sigma^2)^{(n+1)/2}}\exp\left[-\frac{nC}{2\sigma^2}\right] \quad \sigma>0$	$\frac{n}{n+1}C$	$\frac{n}{n-3}C \quad n>3$	$\frac{2n^2C^2}{(n-3)^2(n-5)} \quad n>5$

7 References

[1] Jaynes, E. T., *Probability Theory: The Logic of Science,* Cambridge University Press, 2003.

[2] Jeffreys, Sir Harold, *Theory of Probability, 3rd edition,* Oxford University Press, 1961.