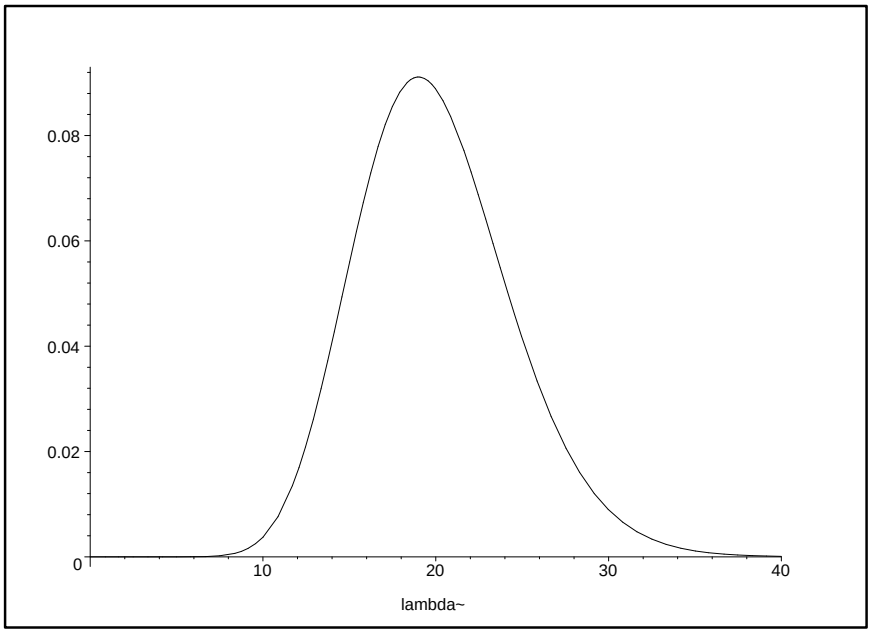
Diffuse Priors (continued)

```
> plot( p( lambda, 20, 1 ), lambda = 0 .. 40, color = black );
```



$\Gamma(20, 1)$ does indeed look **not far from Gaussian**, and at first it may appear that it is indeed **relatively flat** in the region where the likelihood is appreciable ($\lambda \in (1.0, 3.3)$), but we'll see below that it's actually **rather more informative** than we intend.

Recalling that the **mean** and **SD** of a $\Gamma(\alpha, \beta)$ random quantity are $\frac{\alpha}{\beta}$ and $\sqrt{\frac{\alpha}{\beta^2}}$, respectively, and that when used as a prior with the Poisson likelihood the $\Gamma(\alpha, \beta)$ distribution acts like a dataset with **prior sample size** β, you can construct the following table:

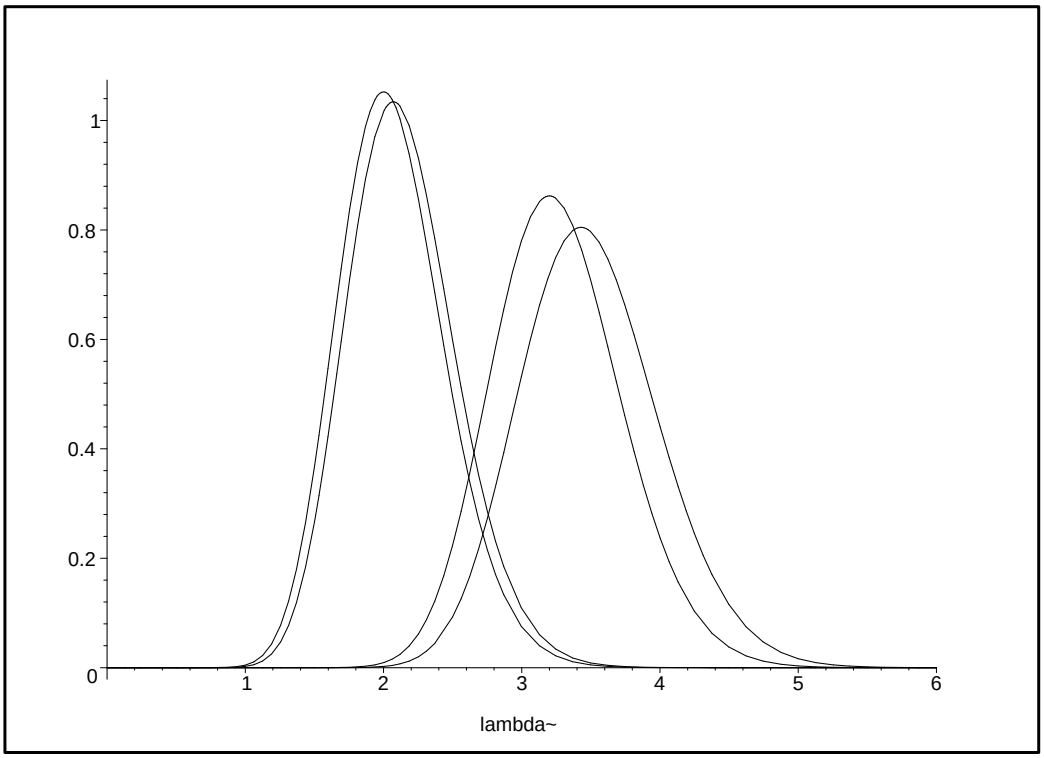
Prior				Posterior			
α	$\beta =$ Sample Size	Mean	SD	α^*	β^*	Mean	SD
0.001	0.001	1	31.6	29.001	14.001	2.071	0.385
1	0.001	1000	1000	30	14.001	2.143	0.391
20	1	20	4.47	49	15	3.267	0.467
20	0.001	20000	4472	49	14.001	3.500	0.500
$U(0, C)$ for $C > 4$		$\frac{C}{2}$	$\frac{C}{\sqrt{12}}$	30	14	2.143	0.391

Diffuse Priors (continued)

The $\Gamma(1, \epsilon)$ prior leads to an analysis that's **essentially equivalent** to the **integrated likelihood (fiducial)** approach back on p. 72, and the $U(0, C)$ prior for $C > 4$ (say) produces similar results: $U(0, C)$ yields the $\Gamma(s + 1, n)$ posterior **truncated** to the right of C (and this truncation has **no effect** if you choose C big enough).

You might say that the $U(0, C)$ distribution has a **prior sample size of 0** in this analysis, and its prior mean $\frac{C}{2}$ and SD $\frac{C}{\sqrt{12}}$ (both of which can be made arbitrarily large by letting C grow without bound) are **irrelevant** (an example of how intuition can change when you depart from the class of **conjugate** priors).

```
> plot( { p( lambda, 29.001, 14.001 ), p( lambda, 30, 14.001 ),
    p( lambda, 49, 15 ), p( lambda, 49, 14.001 ) }, lambda = 0 .. 6,
    color = black );
```



The **moral** is that with only $n = 14$ observations, some care is needed (e.g., through **pre-posterior** analysis) to achieve a prior that **doesn't affect the posterior very much**, if that's your goal.

2.8 Continuous Outcomes

For **continuous outcomes** there's an analogue of de Finetti's Theorem that's **equally central** to Bayesian model-building (e.g., Bernardo and Smith, 1994):

de Finetti's Theorem for Continuous Outcomes.
If $Y_1, Y_2, \ldots$ is an infinitely exchangeable sequence of **real-valued** random quantities with probability measure p, there exists a probability measure Q over $\mathcal{D}$, the space of all distribution functions on R, such that the joint distribution function of $Y_1, \ldots, Y_n$ has the form

$$p(y_1, \ldots, y_n) = \int_{\mathcal{D}} \prod_{i=1}^{n} F(y_i)\, dQ(F), \tag{77}$$

where $Q(F) \stackrel{P}{=} \lim_{n\to\infty} p(F_n)$ and F_n is the **empirical distribution function** based on $Y_1, \ldots, Y_n$.

In other words, exchangeability of real-valued observables is **equivalent** to the hierarchical model

$$\begin{array}{rcll} F & \sim & p(F) & \text{(prior)} \\ (Y_1, \ldots, Y_n | F) & \stackrel{\text{IID}}{\sim} & F & \text{(likelihood)} \end{array} \tag{78}$$

for some **prior distribution** p on the **set** $\mathcal{D}$ **of all possible distribution functions**.

This prior makes the continuous form of de Finetti's Theorem **considerably harder to apply**: to take the elicitation task seriously is to try to specify a probability distribution on a **function space** (F is in effect an **infinite-dimensional** parameter).

(**NB** This task is not unique to Bayesians—it's equivalent to asking **"Where does the likelihood come from?"** in frequentist analyses of observational data.)

Continuous Outcomes (continued)

What people often do in practice is to appeal to considerations that narrow down the field, such as an *a priori* judgment that the Y_i ought to be **symmetrically** distributed about a measure of center μ, and then try to use a fairly **rich parametric family** satisfying (e.g.) the symmetry restriction as a substitute for all of $\mathcal{D}$.

Strictly speaking you're not supposed to look at the Y_i while specifying your prior on $\mathcal{D}$—this can lead to a failure to fully assess and propagate **model uncertainty**—but not doing so can permit the data to surprise you in ways that would make you want to go back and revise your prior (an example of **Cromwell's Rule** in action).

As mentioned earlier, I'll suggest two potential ways out of this dilemma, based on **out-of-sample predictive validation** (the model-checking in the LOS data above was an example of this) and **Bayesian nonparametrics**.

Case Study: *Measurement of physical constants*. What used to be called the National Bureau of Standards (NBS) in Washington, DC, conducts extremely high precision measurement of physical constants, such as the actual weight of so-called **check-weights** that are supposed to serve as reference standards (like the official kg).

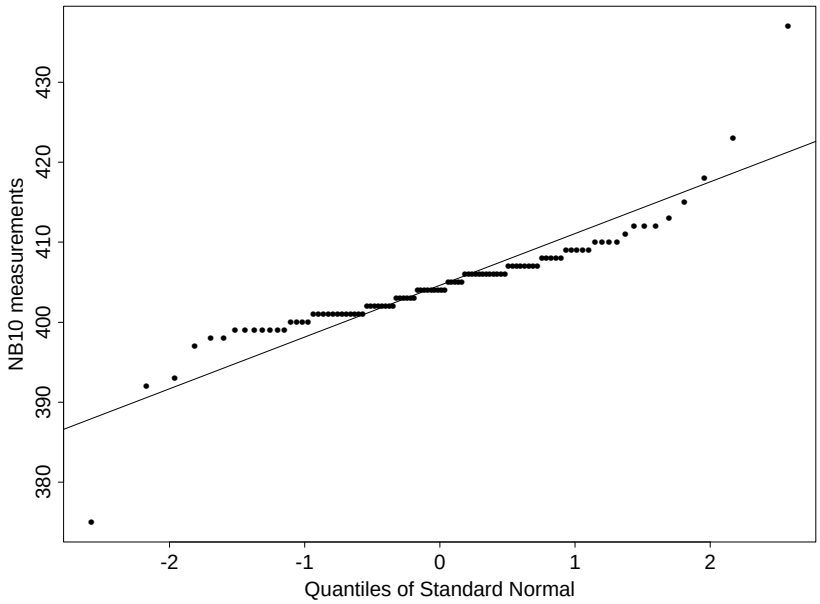
In 1962–63, for example, $n = 100$ weighings (listed below) of a block of metal called **NB10**, which was supposed to weigh exactly 10g, were made under conditions **as close to IID as possible** (Freedman et al., 1998).

Value	375	392	393	397	398	399	400	401
Frequency	1	1	1	1	2	7	4	12
Value	402	403	404	405	406	407	408	409
Frequency	8	6	9	5	12	8	5	5
Value	410	411	412	413	415	418	423	437
Frequency	4	1	3	1	1	1	1	1

NB10 Modeling

Q: (a) How much does NB10 **really weigh**? (b) How certain are you given the data that the true weight of NB10 is **less than** (say) 405.25? And (c) How accurately can you **predict** the 101st measurement?

The graph below is a **normal qqplot** of the 100 measurements $y = (y_1, \ldots, y_n)$, which have a mean of $\bar{y} = 404.6$ (the units are **micrograms below 10g**) and an SD of $s = 6.5$.



Evidently it's plausible in answering these questions to assume **symmetry** of the "underlying distribution" F in de Finetti's Theorem.

One standard choice, for instance, is the **Gaussian:**

$$\begin{aligned} (\mu, \sigma^2) &\sim p(\mu, \sigma^2) \\ (Y_i | \mu, \sigma^2) &\overset{\text{IID}}{\sim} N\left(\mu, \sigma^2\right). \end{aligned} \tag{79}$$

Here $N\left(\mu, \sigma^2\right)$ is the familiar **normal density**

$$p(y_i | \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{y_i - \mu}{\sigma}\right)^2\right]. \tag{80}$$

Gaussian Modeling

Even though you can see from the previous graph that (79) is **not a good model** for the NB10 data, I'm going to fit it to the data for practice in working with the normal distribution from a Bayesian point of view (later we'll **improve** upon the Gaussian).

(79) is more **complicated** than the models in the AMI and LOS case studies because the parameter θ here is a **vector**: $\theta = (\mu, \sigma^2)$.

To warm up for this new complexity let's first consider a **cut-down version of the model** in which we pretend that σ is known to be $\sigma_0 = 6.5$ (the sample SD).

This **simpler model** is then

$$\left\{ \begin{array}{ccc} \mu & \sim & p(\mu) \\ (Y_i|\mu) & \overset{\text{IID}}{\sim} & N\left(\mu, \sigma_0^2\right) \end{array} \right\}. \tag{81}$$

The **likelihood function** in this model is

$$\begin{aligned} l(\mu|y) &= \prod_{i=1}^{n} \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma_0^2}(y_i-\mu)^2\right] \\ &= c\,\exp\left[-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(y_i-\mu)^2\right] \\ &= c\,\exp\left[-\frac{1}{2\sigma_0^2}\left(\sum_{i=1}^{n} y_i^2 - 2\,\mu\sum_{i=1}^{n} y_i + n\mu^2\right)\right] \\ &= c\,\exp\left[-\frac{1}{2\left(\frac{\sigma_0^2}{n}\right)}(\mu-\bar{y})^2\right]. \end{aligned} \tag{82}$$

Thus the likelihood function, when thought of as a **density** for μ, is a **normal distribution** with mean $\bar{y}$ and SD $\frac{\sigma_0}{\sqrt{n}}$.

Gaussian Modeling (continued)

Notice that this SD is the same as the frequentist **standard error** for $\bar{Y}$ based on an IID sample of size n from the $N\left(\mu, \sigma_0^2\right)$ distribution.

(82) also shows that the sample mean $\bar{y}$ is a **sufficient statistic** for μ in model (81).

In finding the conjugate prior for μ it would be nice if the **product of two normal distributions is another normal distribution**, because that would demonstrate that the conjugate prior is normal.

Suppose therefore, to see where it leads, that the **prior for** μ is (say) $p(\mu) = N\left(\mu_0, \sigma_\mu^2\right)$.

Then **Bayes' Theorem** would give

$$\begin{aligned} p(\mu|y) &= c\, p(\mu)\, l(\mu|y) \qquad (83) \\ &= c \exp\left[-\frac{1}{2\sigma_\mu^2}(\mu-\mu_0)^2\right] \exp\left[-\frac{n}{2\sigma_0^2}(\mu-\bar{y})^2\right] \\ &= c \exp\left\{-\frac{1}{2}\left[\frac{(\mu-\mu_0)^2}{\sigma_\mu^2} + \frac{n(\mu-\bar{y})^2}{\sigma_0^2}\right]\right\}, \end{aligned}$$

and we want this to **be of the form**

$$\begin{aligned} p(\mu|y) &= c \exp\left\{-\frac{1}{2}\left[A(\mu-B)^2 + C\right]\right\} \\ &= c \exp\left\{-\frac{1}{2}\left[A\mu^2 - 2AB\mu + (AB^2 + C)\right]\right\} \qquad (84) \end{aligned}$$

for some B, C, and $A > 0$.

Maple can help **see if this works**:

```
> collect( ( mu - mu0 )^2 / sigmamu^2 +
    n * ( mu - ybar )^2 / sigma0^2, mu );
```

```
                                                                2          2
/   1          n   \   2   /     mu0        n ybar \         mu0        n ybar
|-------- + -------| mu  + |-2 -------- - 2 -------| mu + -------- + -------
|       2         2|       |          2           2|              2          2
\sigmamu    sigma0 /       \   sigmamu      sigma0 /      sigmamu    sigma0
```

Gaussian Modeling

Matching coefficients for A and B (we don't really care about C) gives

$$A = \frac{1}{\sigma_\mu^2} + \frac{n}{\sigma_0^2} \quad \text{and} \quad B = \frac{\frac{\mu_0}{\sigma_\mu^2} + \frac{n\bar{y}}{\sigma_0^2}}{\frac{1}{\sigma_\mu^2} + \frac{n}{\sigma_0^2}}. \tag{85}$$

Since $A > 0$ this demonstrates two things: (1) the **conjugate prior** for μ in model (81) is **normal**, and (2) the **conjugate updating rule** (when σ_0 is assumed known) is

$$\left\{ \begin{array}{c} \mu \sim N\left(\mu_0, \sigma_\mu^2\right) \\ (Y_i|\mu) \overset{\text{IID}}{\sim} N\left(\mu, \sigma_0^2\right), \\ i = 1, \ldots, n \end{array} \right\} \rightarrow (\mu|y) = (\mu|\bar{y}) = N\left(\mu_*, \sigma_*^2\right), \tag{86}$$

where the **posterior mean and variance** are given by

$$\mu_* = B = \frac{\left(\frac{1}{\sigma_\mu^2}\right)\mu_0 + \left(\frac{n}{\sigma_0^2}\right)\bar{y}}{\frac{1}{\sigma_\mu^2} + \frac{n}{\sigma_0^2}} \quad \text{and} \quad \sigma_*^2 = A^{-1} = \frac{1}{\frac{1}{\sigma_\mu^2} + \frac{n}{\sigma_0^2}}. \tag{87}$$

It becomes useful in understanding the meaning of these expressions to define the **precision** of a distribution, which is just the **reciprocal** of its variance: whereas the variance and SD scales measure **uncertainty**, the precision scale quantifies **information** about an unknown.

With this convention (87) has a series of **intuitive interpretations**, as follows:

- The **prior**, considered as an **information source**, is Gaussian with mean μ_0, variance σ_μ^2, and **precision** $\frac{1}{\sigma_\mu^2}$, and when viewed as a data set consists of n_0 (to be determined below) observations;

- The **likelihood**, considered as an **information source**, is Gaussian with mean $\bar{y}$, variance $\frac{\sigma_0^2}{n}$, and **precision** $\frac{n}{\sigma_0^2}$, and when viewed as a data set consists of n observations;

Gaussian Modeling (continued)

• The **posterior**, considered as an **information source**, is Gaussian, and the posterior mean is a **weighted average** of the prior mean and data mean, with weights given by the **prior** and **data precisions**;

• The **posterior precision** (the reciprocal of the posterior variance) is just the **sum** of the prior and data precisions (this is why people invented the idea of precision—on this scale **knowledge** about μ in model (81) is **additive**); and

• **Rewriting** μ_* as

$$\mu_* = \frac{\left(\frac{1}{\sigma_\mu^2}\right)\mu_0 + \left(\frac{n}{\sigma_0^2}\right)\bar{y}}{\frac{1}{\sigma_\mu^2} + \frac{n}{\sigma_0^2}} = \frac{\left(\frac{\sigma_0^2}{\sigma_\mu^2}\right)\mu_0 + n\bar{y}}{\frac{\sigma_0^2}{\sigma_\mu^2} + n}, \tag{88}$$

you can see that the **prior sample size** is

$$n_0 = \frac{\sigma_0^2}{\sigma_\mu^2} = \frac{1}{\left(\frac{\sigma_\mu}{\sigma_0}\right)^2}, \tag{89}$$

which makes sense: the **bigger** σ_μ is in relation to σ_0, the **less prior information** is being incorporated in the conjugate updating (86).

Bayesian inference with multivariate θ. Returning now to (79) with σ^2 unknown, (as mentioned above) this model has a $(p = 2)$-dimensional **parameter vector** $\theta = (\mu, \sigma^2)$.

When $p > 1$ you can still use Bayes' Theorem directly to obtain the **joint posterior distribution**,

$$\begin{aligned} p(\theta|y) &= p(\mu, \sigma^2|y) = c\, p(\theta)\, l(\theta|y) \\ &= c\, p(\mu, \sigma^2)\, l(\mu, \sigma^2|y), \end{aligned} \tag{90}$$

Multivariate Unknown θ

where $y = (y_1, \ldots, y_n)$, although making this calculation directly requires a p-dimensional **integration** to evaluate the normalizing constant c; for example, in this case

$$\begin{aligned} c &= [p(y)]^{-1} = \left(\iint p(\mu, \sigma^2, y)\, d\mu\, d\sigma^2 \right)^{-1} \\ &= \left(\iint p(\mu, \sigma^2)\, l(\mu, \sigma^2 | y)\, d\mu\, d\sigma^2 \right)^{-1} . \end{aligned} \tag{91}$$

Usually, however, you'll be more interested in the **marginal posterior distributions**, in this case $p(\mu|y)$ and $p(\sigma^2|y)$.

Obtaining these requires p integrations, each of dimension $(p-1)$, a process that people refer to as **marginalization** or **integrating out the nuisance parameters**—for example,

$$p(\mu|y) = \int_0^\infty p(\mu, \sigma^2|y)\, d\sigma^2 \ . \tag{92}$$

Predictive distributions also involve a p-dimensional integration: for example, with $y = (y_1, \ldots, y_n)$,

$$\begin{aligned} p(y_{n+1}|y) &= \iint p(y_{n+1}, \mu, \sigma^2|y)\, d\mu\, d\sigma^2 \\ &= \iint p(y_{n+1}|\mu, \sigma^2)\, p(\mu, \sigma^2|y)\, d\mu\, d\sigma^2. \end{aligned} \tag{93}$$

And, finally, if you're interested in a **function of the parameters**, you have some more hard integrations ahead of you.

For instance, suppose you wanted the posterior distribution for the **coefficient of variation** $\lambda = g_1(\mu, \sigma^2) = \frac{\sqrt{\sigma^2}}{\mu}$ in model (79).

Multivariate Unknown θ

Then one fairly direct way to get this posterior (e.g., Bernardo and Smith, 1994) is to (a) introduce a **second function** of the parameters, say $\eta = g_2(\mu, \sigma^2)$, such that the mapping $f = (g_1, g_2)$ from (μ, σ^2) to (λ, η) is **invertible**; (b) compute the joint posterior for (λ, η) through the usual **change-of-variables formula**

$$p(\lambda, \eta|y) = p_{\mu,\sigma^2}\left[f^{-1}(\lambda, \eta)|y\right] \, |J_{f^{-1}}(\lambda, \eta)| \,, \tag{94}$$

where $p_{\mu,\sigma^2}(\cdot, \cdot|y)$ is the joint posterior for μ and σ^2 and $|J_{f^{-1}}|$ is the **determinant** of the **Jacobian** of the inverse transformation; and (c) **marginalize** in λ by integrating out η in $p(\lambda, \eta|y)$, in a manner analogous to (92).

Here, for instance, $\eta = g_2(\mu, \sigma^2) = \mu$ would create an invertible f, with **inverse** defined by $(\mu = \eta, \sigma^2 = \lambda^2\eta^2)$; the **Jacobian determinant** comes out $2\lambda\eta^2$ and (94) becomes $p(\lambda, \eta|y) = 2\lambda\eta^2 \, p_{\mu,\sigma^2}(\eta, \lambda^2\eta^2|y)$.

This process involves **two integrations**, one (of dimension p) to get the normalizing constant that defines (94) and one (of dimension $(p-1)$) to get rid of η.

You can see that when p is a lot bigger than 2 all these integrals may create **severe computational problems**—this has been the **big stumbling block** for applied Bayesian work for a long time.

More than 200 years ago **Laplace** (1774)—perhaps the second applied Bayesian in history (after Bayes himself)—developed, as one avenue of solution to this problem, what people now call **Laplace approximations** to high-dimensional integrals of the type arising in Bayesian calculations (see, e.g., Tierney and Kadane, 1986).

Starting in the next case study after this one, we'll use another, computationally intensive, **simulation-based** approach: **Markov chain Monte Carlo** (MCMC).

Gaussian Modeling

Back to model (79). The conjugate prior for $\theta = (\mu, \sigma^2)$ in this model (e.g., Gelman et al., 2003) turns out to be most simply described **hierarchically**:

$$\begin{aligned} \sigma^2 &\sim \text{SI-}\chi^2(\nu_0, \sigma_0^2) \\ (\mu|\sigma^2) &\sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right). \end{aligned} \tag{95}$$

Here saying that $\sigma^2 \sim \text{SI-}\chi^2(\nu_0, \sigma_0^2)$, where SI stands for **scaled inverse**, amounts to saying that the precision $\tau = \frac{1}{\sigma^2}$ follows a **scaled** χ^2 distribution with parameters ν_0 and σ_0^2.

The scaling is chosen so that σ_0^2 can be interpreted as a **prior estimate** of σ^2, with ν_0 the **prior sample size** of this estimate (i.e., **think of a prior data set with** ν_0 **observations and sample SD** σ_0).

Since χ^2 is a special case of the Gamma distribution, SI-χ^2 must be a special case of the **inverse Gamma** family—its **density** (see Gelman et al., 2003, Appendix A) is

$$\begin{aligned} \sigma^2 &\sim \text{SI-}\chi^2(\nu_0, \sigma_0^2) \leftrightarrow \\ p(\sigma^2) &= \frac{\left(\frac{1}{2}\nu_0\right)^{\frac{1}{2}\nu_0}}{\Gamma\left(\frac{1}{2}\nu_0\right)} \left(\sigma_0^2\right)^{\frac{1}{2}\nu_0} \left(\sigma^2\right)^{-\left(1+\frac{1}{2}\nu_0\right)} \exp\left(\frac{-\nu_0\,\sigma_0^2}{2\sigma^2}\right). \end{aligned} \tag{96}$$

As may be verified with `Maple`, this distribution has **mean** (provided that $\nu_0 > 2$) and **variance** (provided that $\nu_0 > 4$) given by

$$E\left(\sigma^2\right) = \frac{\nu_0}{\nu_0 - 2}\sigma_0^2 \quad \text{and} \quad V\left(\sigma^2\right) = \frac{2\nu_0^2}{(\nu_0 - 2)^2(\nu_0 - 4)}\sigma_0^4. \tag{97}$$

Gaussian Modeling (continued)

The parameters μ_0 and κ_0 in the second level of the prior model (95), $(\mu|\sigma^2) \sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right)$, have **simple parallel interpretations** to those of σ_0^2 and ν_0: μ_0 is the **prior estimate** of μ, and κ_0 is the **prior effective sample size** of this estimate.

The **likelihood function** in model (79), with **both** μ and σ^2 **unknown**, is

$$\begin{aligned} l(\mu, \sigma^2|y) &= c \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \mu)^2\right] \\ &= c \left(\sigma^2\right)^{-\frac{1}{2}n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2\right] \qquad (98) \\ &= c \left(\sigma^2\right)^{-\frac{1}{2}n} \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} y_i^2 - 2\,\mu \sum_{i=1}^{n} y_i + n\mu^2\right)\right]. \end{aligned}$$

The **expression in brackets** in the last line of (98) is

$$\begin{aligned} [\ \cdot\ ] &= -\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n} y_i^2 + n(\mu - \bar{y})^2 - n\bar{y}^2\right] \qquad (99) \\ &= -\frac{1}{2\sigma^2}\left[n(\mu - \bar{y})^2 + (n-1)s^2\right], \end{aligned}$$

where $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$ is the **sample variance**. Thus

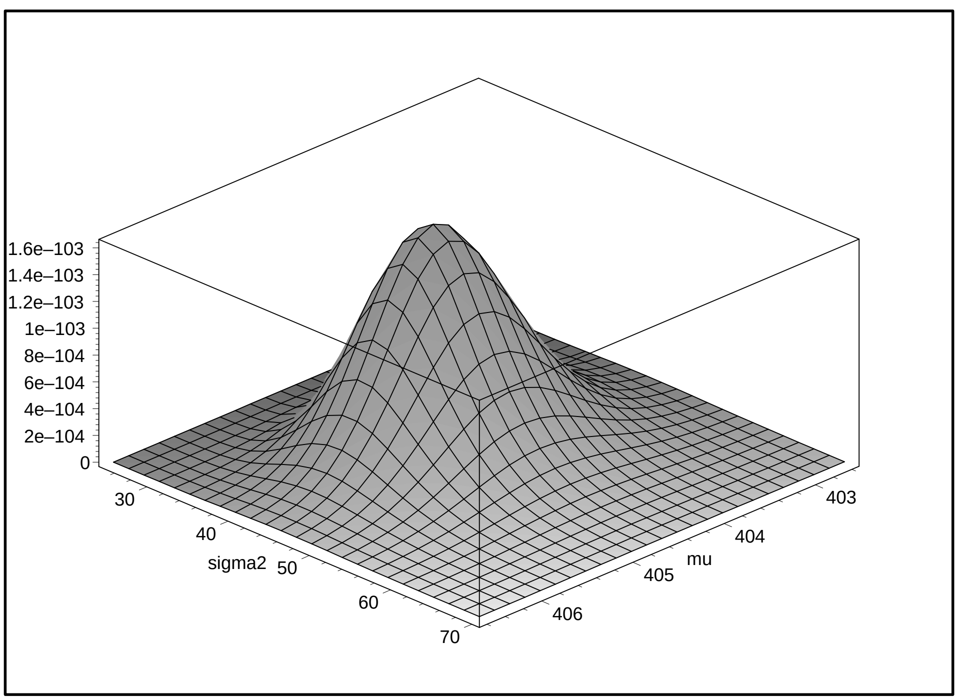
$$l(\mu, \sigma^2|y) = c \left(\sigma^2\right)^{-\frac{1}{2}n} \exp\left\{-\frac{1}{2\sigma^2}\left[n(\mu - \bar{y})^2 + (n-1)s^2\right]\right\},$$

and it's clear that the **vector** $(\bar{y}, s^2)$ is **sufficient** for $\theta = (\mu, \sigma^2)$ in this model, i.e., $l(\mu, \sigma^2|y) = l(\mu, \sigma^2|\bar{y}, s^2)$.

Gaussian Analysis

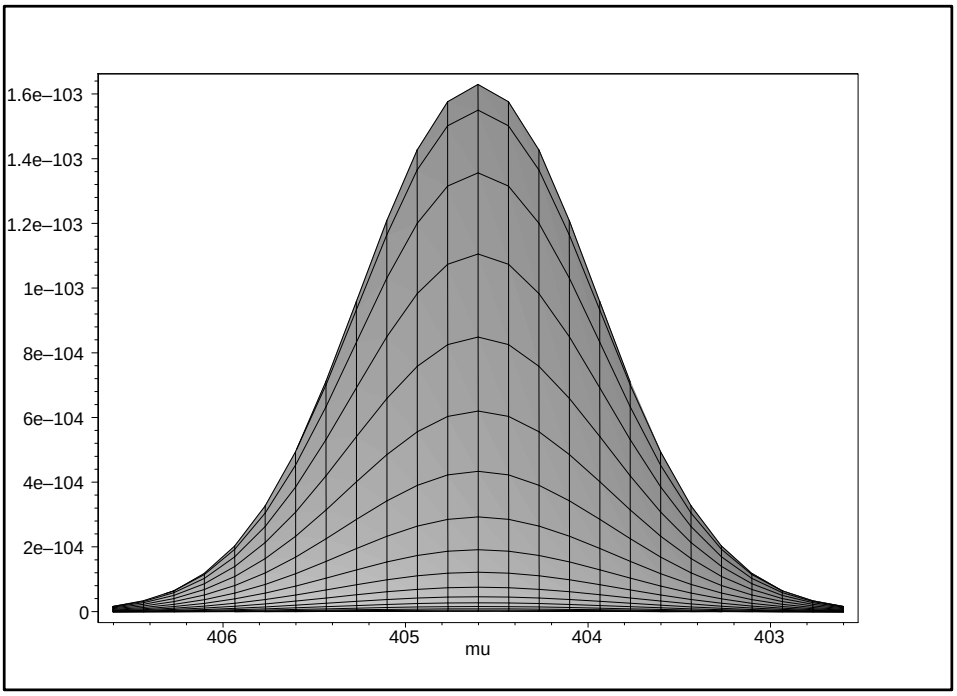
`Maple` can be used to make **3D** and **contour plots** of this likelihood function with the NB10 data:

```
> l := ( mu, sigma2, ybar, s2, n ) -> sigma2^( - n / 2 ) *
    exp( - ( n * ( mu - ybar )^2 + ( n - 1 ) * s2 ) / ( 2 * sigma2 ) );

l := (mu, sigma2, ybar, s2, n) ->

                                                   2
          (- 1/2 n)              n (mu - ybar)   + (n - 1) s2
    sigma2          exp(- 1/2 ----------------------------)
                                         sigma2

> plotsetup( x11 );

> plot3d( l( mu, sigma2, 404.6, 42.25, 100 ), mu = 402.6 .. 406.6,
    sigma2 = 25 .. 70 );
```

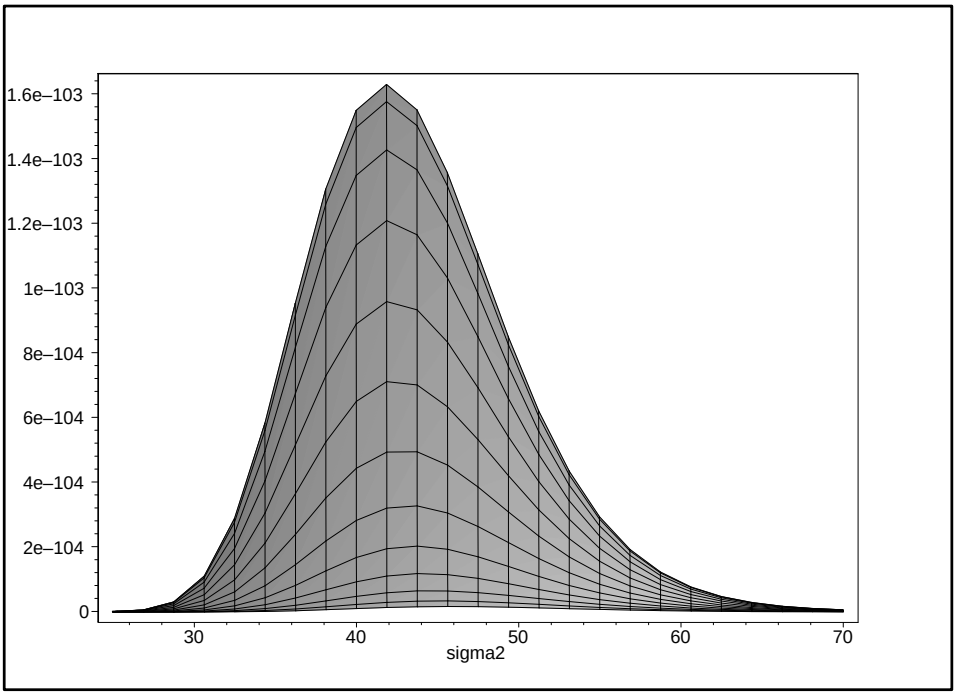


You can use the mouse to **rotate** 3D plots and get **other useful views** of them:

Gaussian Analysis



The **projection** or **shadow plot** of μ looks a lot like a **normal** (or maybe a t) distribution.



And the shadow plot of σ^2 looks a lot like a **Gamma** (or maybe an **inverse Gamma**) distribution.