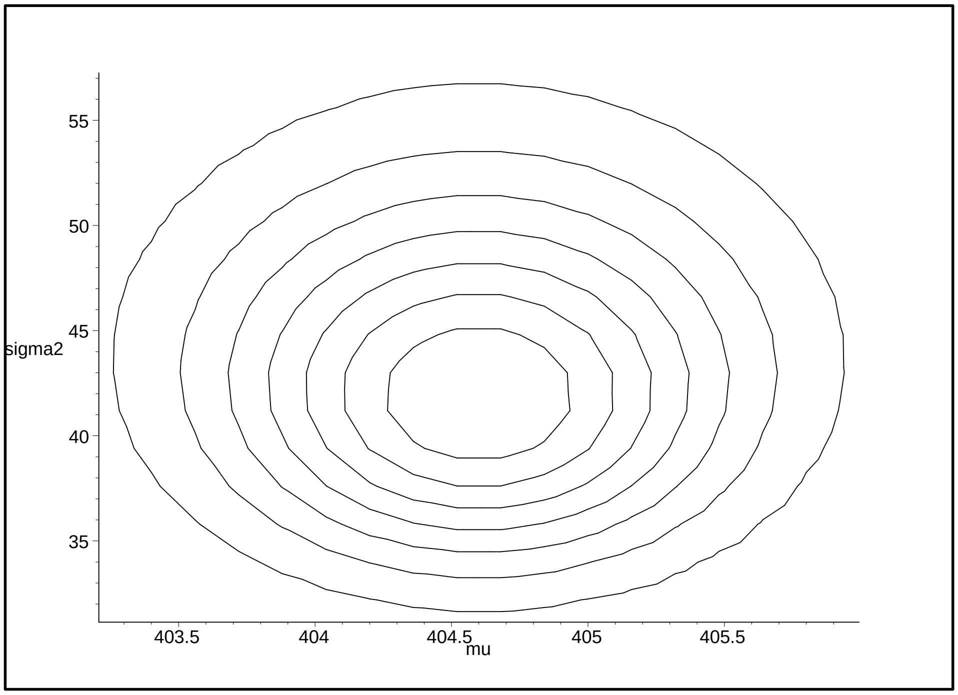
Gaussian Analysis

```
> plots[ contourplot ]( 10^100 * l( mu, sigma2, 404.6, 42.25, 100 ),
    mu = 402.6 .. 406.6, sigma2 = 25 .. 70, color = black );
```



The **contour plot** shows that μ and σ^2 are **uncorrelated** in the likelihood distribution, and the **skewness** of the marginal distribution of σ^2 is also evident.

Posterior analysis. Having adopted the **conjugate prior** (95), what I'd like next is simple expressions for the **marginal posterior distributions** $p(\mu|y)$ and $p(\sigma^2|y)$ and for **predictive distributions** like $p(y_{n+1}|y)$.

Fortunately, in model (79) all of the **integrations** (such as (92) and (93)) may be done **analytically** (see, e.g., Bernardo and Smith 1994), yielding the following results:

$$\begin{aligned} (\sigma^2|y,\mathcal{G}) &\sim \text{SI-}\chi^2(\nu_n,\sigma_n^2), \\ (\mu|y,\mathcal{G}) &\sim t_{\nu_n}\left(\mu_n,\frac{\sigma_n^2}{\kappa_n}\right), \quad \text{and} \\ (y_{n+1}|y,\mathcal{G}) &\sim t_{\nu_n}\left(\mu_n,\frac{\kappa_n+1}{\kappa_n}\sigma_n^2\right). \end{aligned} \tag{100}$$

NB10 Gaussian Analysis

In the above **expressions**

$$\begin{aligned} \nu_n &= \nu_0 + n, \\ \sigma_n^2 &= \frac{1}{\nu_n}\left[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0+n}(\bar{y}-\mu_0)^2\right], \quad (101) \\ \mu_n &= \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y}, \quad \text{and} \\ \kappa_n &= \kappa_0 + n, \end{aligned}$$

$\bar{y}$ and s^2 are the usual **sample mean** and **variance** of y, and $\mathcal{G}$ denotes the assumption of the **Gaussian model**.

Here $t_\nu(\mu, \sigma^2)$ is a **scaled** version of the usual t_ν distribution, i.e., $W \sim t_\nu(\mu, \sigma^2) \iff \frac{W-\mu}{\sigma} \sim t_\nu$.

The scaled t distribution (see, e.g., Gelman et al., 2003, Appendix A) has **density**

$$\eta \sim t_\nu(\mu, \sigma^2) \leftrightarrow p(\eta) = \frac{\Gamma\left[\frac{1}{2}(\nu+1)\right]}{\Gamma\left(\frac{1}{2}\nu\right)\sqrt{\nu\pi\sigma^2}}\left[1 + \frac{1}{\nu\sigma^2}(\eta-\mu)^2\right]^{-\frac{1}{2}(\nu+1)}. \quad (102)$$

This distribution has **mean** μ (as long as $\nu > 1$) and **variance** $\frac{\nu}{\nu-2}\sigma^2$ (as long as $\nu > 2$).

Notice that, as with all previous conjugate examples, the posterior mean is again a **weighted average** of the prior mean and data mean, with weights determined by the **prior sample size** and the **data sample size**:

$$\mu_n = \frac{\kappa_0}{\kappa_0+n}\mu_0 + \frac{n}{\kappa_0+n}\bar{y}. \quad (103)$$

NB10 Gaussian Analysis (continued)

NB10 Gaussian Analysis. *Question (a):* I don't know anything about what NB10 is supposed to weigh (down to the nearest microgram) or about the accuracy of the NBS's measurement process, so I want to use a **diffuse prior** for μ and σ^2.

Considering the meaning of the **hyperparameters**, to provide little prior information I want to choose both ν_0 and κ_0 **close to 0**.

Making them exactly 0 would produce an **improper** prior distribution (which doesn't integrate to 1), but choosing positive values as close to 0 as you like yields a **proper and highly diffuse prior**.

You can see from (100, 101) that the result is then

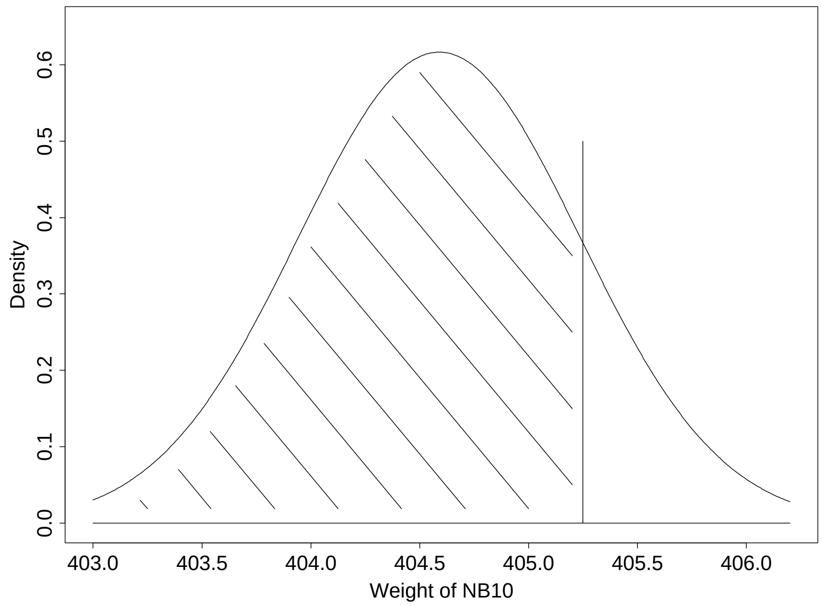
$$(\mu | y, \mathcal{G}) \sim t_n \left[\bar{y}, \frac{(n-1)s^2}{n^2} \right] \doteq N\left(\bar{y}, \frac{s^2}{n} \right), \qquad (104)$$

i.e., with diffuse prior information (as with the Bernoulli model in the AMI case study) the 95% central Bayesian interval **virtually coincides** with the usual frequentist 95% confidence interval
$\bar{y} \pm t_{n-1}^{.975} \frac{s}{\sqrt{n}} = 404.6 \pm (1.98)(0.647) = (403.3, 405.9)$.

Thus both {frequentists who assume $\mathcal{G}$} and {Bayesians who assume $\mathcal{G}$ with a diffuse prior} conclude that **NB10 weighs about** 404.6μg **below 10g, give or take about** 0.65μg.

Question (b). If interest focuses on whether NB10 weighs **less than some value** like 405.25, when reasoning in a Bayesian way you can answer this question directly: the posterior distribution for μ is shown below, and $P_B(\mu < 405.25 | y, \mathcal{G}, \text{diffuse prior}) \doteq .85$, i.e., your **betting odds** in favor of the proposition that $\mu < 405.25$ are about 5.5 to 1.

NB10 Gaussian Analysis (continued)



When reasoning in a frequentist way $P_F(\mu < 405.25)$ is **undefined**; about the best you can do is to test $H_0\colon \mu < 405.25$, for which the p-value would (approximately) be $p = P_{F,\mu=405.25}(\bar{y} > 405.59) = 1 - .85 = .15$, i.e., **insufficient evidence to reject** H_0 at the usual significance levels (note the **connection** between the p-value and the posterior probability, which arises in this example because the null hypothesis is **one-sided**).

NB The significance test tries to answer a **different question**: in Bayesian language it looks at $P(\bar{y}|\mu)$ instead of $P(\mu|\bar{y})$.

Many people find the latter quantity **more interpretable**.

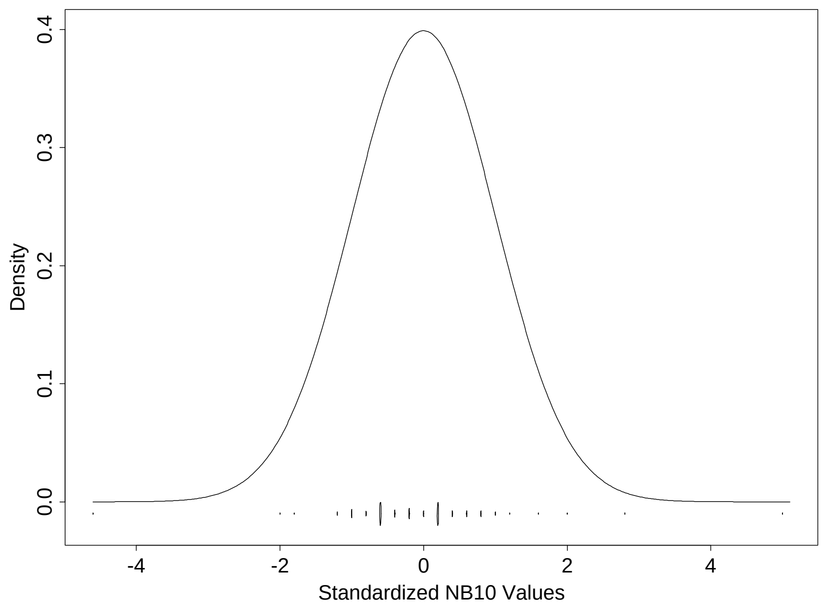
Question (c). We saw earlier that **in this model**

$$(y_{n+1}|y, \mathcal{G}) \sim t_{\nu_n}\left[\mu_n, \frac{\kappa_n + 1}{\kappa_n}\sigma_n^2\right], \qquad (105)$$

and for n large and ν_0 and κ_0 close to 0 this is $(y_{n+1}|y, \mathcal{G}) \mathrel{\dot\sim} N(\bar{y}, s^2)$, i.e., a **95% posterior predictive interval** for y_{n+1} is $(392, 418)$.

Model Expansion

A **standardized version** of this predictive distribution is plotted below, with the standardized NB10 data values **superimposed**.



It's evident from this plot (and also from the normal qqplot given earlier) that the Gaussian model provides a **poor fit** for these data—the three most extreme points in the data set in standard units are $-4.6, 2.8$, and 5.0.

With the **symmetric heavy tails** indicated in these plots, in fact, the empirical CDF looks quite a bit like that of a t distribution with a rather small number of **degrees of freedom**.

This suggests revising the previous model by **expanding** it: **embedding** the Gaussian in the t family and adding a parameter k for **tail-weight**.

Unfortunately there's no standard **closed-form conjugate** choice for the prior on k.

A more **flexible** approach to computing is evidently needed—this is where **Markov chain Monte Carlo** methods (our next main topic) come in.

Postscript on the Exponential Family

It will be helpful in what follows to close part 2 of the notes with two more examples of the use of the **exponential family**.

(1) An example of a **non-regular** exponential family: suppose (as in the case study in homework 3 problem 2) that a reasonable model for the data is to take the observed values $(y_i|\theta)$ to be conditionally IID from the **uniform** distribution $U(0,\theta)$ on the interval $(0,\theta)$ for unknown θ:

$$p(y_1|\theta) = \left\{ \begin{array}{cc} \frac{1}{\theta} & \text{for } 0 < y_1 < \theta \\ 0 & \text{otherwise} \end{array} \right\} = \frac{1}{\theta} I(0,\theta), \quad (106)$$

where $I(A) = 1$ if A is true and 0 otherwise.

θ in this model is called a **range-restriction** parameter; such parameters are fundamentally different from **location** and **scale** parameters (like the mean μ and variance σ^2 in the $N(\mu,\sigma^2)$ model, respectively) or **shape** parameters (like the degrees of freedom ν in the t_ν model).

The Exponential Family (continued)

(106) is an **example of (51)** with $c = 1, f_1(y) = 1, g_1(\theta) = \frac{1}{\theta}, h_1(y) = 0$, and $\phi_1(\theta) =$ anything you want (e.g., 1), but only when the set $\mathcal{Y} = (0, \theta)$ is taken to depend on θ.

(**Truncated** distributions with **unknown truncation point** also lead to non-regular exponential families.)

As you'll see in homework 3, inference in non-regular exponential families is **similar** in some respects to the story when the exponential family is regular, but there are some **important differences** too.

(2) For an example with $p > 1$, take $\theta = (\mu, \sigma^2)$ with the **Gaussian likelihood**:

$$\begin{aligned} l(\theta|y) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \mu)^2\right] \quad (107) \\ &= c\left(\sigma^2\right)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} y_i^2 \right.\right. \\ &\qquad \left.\left. -2\mu \sum_{i=1}^{n} y_i + n\mu^2\right)\right]. \end{aligned}$$

The Exponential Family (continued)

This is **of the form (53)** with $k = 2,\ f(y) = 1, g(\theta) = \left(\sigma^2\right)^{-\frac{n}{2}} \exp\left(-\frac{n\mu^2}{2\sigma^2}\right), \phi_1(\theta) = -\frac{1}{2\sigma^2}, \phi_2(\theta) = \frac{\mu}{\sigma^2}, h_1(y_i) = y_i^2$, and $h_2(y_i) = y_i$, which shows that $[h_1(y) = \sum_{i=1}^n y_i^2, h_2(y) = \sum_{i=1}^n y_i]$ or equivalently $(\bar{y}, s^2)$ is **sufficient** for θ.

Some **unpleasant algebra** then demonstrates that an application of the **conjugate prior** theorem (54) in the exponential family leads to (95) as the conjugate prior for the Gaussian likelihood when both μ and σ^2 are unknown.

2.10 References

Bernardo JM, Smith AFM (1994). *Bayesian Theory*. New York: Wiley.

Craig PS, Goldstein M, Seheult AH, Smith JA (1997). Constructing partial prior specifications for models of complex physical systems. *The Statistician*, **46**, forthcoming.

Draper D (1995). Inference and hierarchical modeling in the social sciences (with discussion). *Journal of Educational and Behavioral Statistics*, **20**, 115–147, 233–239.

Draper D, Hodges JS, Mallows CL, Pregibon D (1993). Exchangeability and data analysis (with discussion). *Journal of the Royal Statistical Society, Series A*, **156**, 9–37.

de Finetti B (1930). Funzione caratteristica di un fenomeno aleatorio. *Mem. Acad. Naz. Lincei*, **4**, 86–133.

de Finetti B (1964). Foresight: its logical laws, its subjective sources. In *Studies in Subjective Probability*, HE Kyburg, Jr., and HE Smokler, eds., New York: Wiley (1980), 93–158.

de Finetti B (1974/5). *Theory of Probability*, **1–2**. New York: Wiley.

Fisher RA (1922). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London A*, **222**, 309–368.

Fisher RA (1956). *Statistical Methods and Scientific Inference*. London: Oliver and Boyd.

Freedman D, Pisani R, Purves R (1998). *Statistics*, third edition. New York: Norton.

Gelman A, Carlin JB, Stern HS, Rubin DB (2003). *Bayesian Data Analysis*, second edition. London: Chapman & Hall.

Hacking I (1975). *The Emergence of Probability*. Cambridge: Cambridge University Press.

Johnson NL, Kotz S (1970). *Distributions in statistics: Continuous univariate distributions*, **1**. New York: Wiley.

Kadane JB, Dickey JM, Winkler RL, Smith WS, Peters SC (1980). Interactive elicitation of opinion for a normal linear model. *Journal of the American Statistical Association*, **75**, 845–854.

References (continued)

Kadane JB, Wolfson LJ (1997). Experiences in elicitation. *The Statistician*, **46**, forthcoming.

Kahn K, Rubenstein L, Draper D, Kosecoff J, Rogers W, Keeler E, Brook R (1990). The effects of the DRG-based Prospective Payment System on quality of care for hospitalized Medicare patients: An introduction to the series. *Journal of the American Medical Association*, **264**, 1953–1955 (with editorial comment, 1995–1997).

Laplace PS (1774). Mémoire sur la probabilité des causes par les évenements. *Mémoires de l'Academie des Sciences de Paris*, **6**, 621–656. English translation in 1986 as "Memoir on the probability of the causes of events," with an introduction by SM Stigler, *Statistical Science*, **1**, 359–378.

O'Hagan A (1997). Eliciting expert beliefs in substantial practical applications. *The Statistician*, **46**, forthcoming.

Samaniego FJ, Reneau DM (1994). Toward a reconciliation of the Bayesian and frequentist approaches to point estimation. *Journal of the American Statistical Association*, **89**, 947–957.

Tierney L, Kadane JB (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**, 82–86.