# Bayesian Modeling, Inference, Prediction and Decision-Making

## 2c: Continuous Outcomes; Gaussian Modeling

**David Draper**

Department of Applied Mathematics and Statistics
University of California, Santa Cruz

draper@ams.ucsc.edu
www.ams.ucsc.edu/∼draper

*eBay/Google*

10 Fridays, 11 Jan–22 Mar 2013 (except 25 Jan)

**Short course web page:**
www.ams.ucsc.edu/∼draper/eBay-Google-2013.html

# 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

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

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

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

$$\boxed{\begin{array}{rclr} F & \sim & p(F) & \text{(prior)} \\ (Y_1, \ldots, Y_n | F) & \overset{\text{IID}}{\sim} & F & \text{(likelihood)} \quad (2) \end{array}}$$

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 measure 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 $\mu$, 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, in this short course 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; also see topic 5) and **Bayesian nonparametrics/semi-parametrics** (which we will examine in topic 5).

**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).

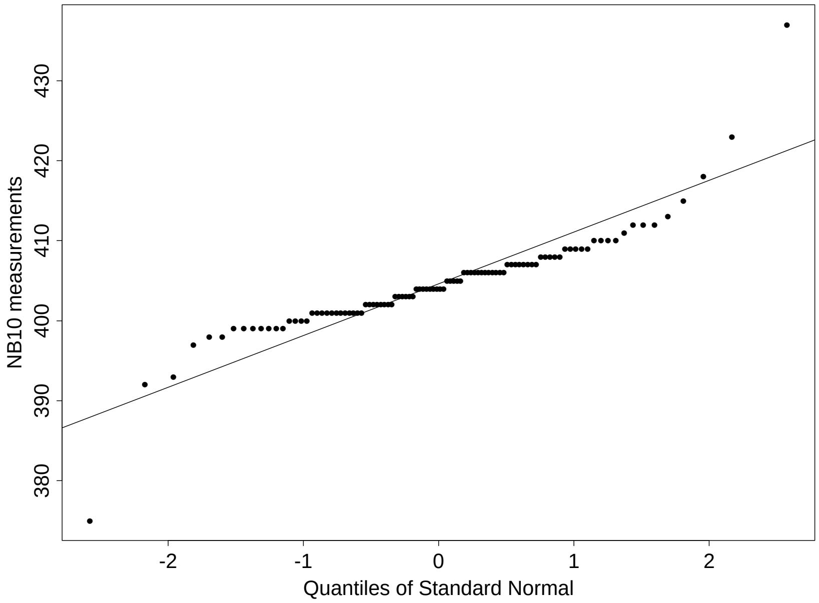
In 1962–63, for example, $n = 100$ weighings (listed below) of a block of metal called **NB10**, which was supposed to weigh exactly 10g, were made under conditions **as close to IID as possible** (Freedman et al., 1998).

| 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 $\boxed{\textbf{Gaussian:}}$

$$\begin{aligned} (\mu, \sigma^2) & \sim p(\mu, \sigma^2) \\ (Y_i | \mu, \sigma^2) & \overset{\text{IID}}{\sim} N\left(\mu, \sigma^2\right). \end{aligned} \qquad (3)$$

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

$$p(y_i | \mu, \sigma^2) = \frac{1}{\sigma \sqrt{2\pi}} \exp\left[ -\frac{1}{2} \left( \frac{y_i - \mu}{\sigma} \right)^2 \right]. \qquad (4)$$

# Gaussian Modeling

Even though you can see from the previous graph that (3) 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).

(3) is more **complicated** than the models in the AMI and LOS case studies because the parameter $\theta$ 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 $\sigma$ is known to be $\sigma_0 = 6.5$ (the sample SD).

This **simpler model** is then

$$\left\{ \begin{array}{ccc} \mu & \sim & p(\mu) \\ (Y_i|\mu) & \stackrel{\text{IID}}{\sim} & N\left(\mu, \sigma_0^2\right) \end{array} \right\}. \qquad (5)$$

The **likelihood function** in this model is

$$\begin{aligned} l(\mu|y) &= \prod_{i=1}^{n} \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma_0^2}(y_i - \mu)^2\right] \\ &= c \exp\left[-\frac{1}{2\sigma_0^2}\sum_{i=1}^{n}(y_i - \mu)^2\right] \qquad (6) \\ &= c \exp\left[-\frac{1}{2\sigma_0^2}\left(\sum_{i=1}^{n} y_i^2 - 2\,\mu\sum_{i=1}^{n} y_i + n\mu^2\right)\right] \\ &= c \exp\left[-\frac{1}{2\left(\frac{\sigma_0^2}{n}\right)}(\mu - \bar{y})^2\right]. \end{aligned}$$

Thus the likelihood function, when thought of as a **density** for $\mu$, 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.

(6) also shows that the sample mean $\bar{y}$ is a **sufficient statistic** for $\mu$ in model (5).

In finding the conjugate prior for $\mu$, 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** $\mu$ is (say) $p(\mu) = N\left(\mu_0, \sigma_\mu^2\right)$.

Then **Bayes's Theorem** would give

$$\begin{aligned} p(\mu|y) &= c\, p(\mu)\, l(\mu|y) \qquad (7) \\ &= c \exp\left[-\frac{1}{2\sigma_\mu^2}(\mu-\mu_0)^2\right] \exp\left[-\frac{n}{2\sigma_0^2}(\mu-\bar{y})^2\right] \\ &= c \exp\left\{-\frac{1}{2}\left[\frac{(\mu-\mu_0)^2}{\sigma_\mu^2} + \frac{n(\mu-\bar{y})^2}{\sigma_0^2}\right]\right\}, \end{aligned}$$

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

$$\begin{aligned} p(\mu|y) &= c \exp\left\{-\frac{1}{2}\left[A(\mu-B)^2 + C\right]\right\} \\ &= c \exp\left\{-\frac{1}{2}\left[A\mu^2 - 2AB\mu + (AB^2 + C)\right]\right\} \qquad (8) \end{aligned}$$

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

$$A = \frac{1}{\sigma_\mu^2} + \frac{n}{\sigma_0^2} \quad \text{and} \quad B = \frac{\frac{\mu_0}{\sigma_\mu^2} + \frac{n \bar{y}}{\sigma_0^2}}{\frac{1}{\sigma_\mu^2} + \frac{n}{\sigma_0^2}}. \tag{9}$$

Since $A > 0$ this demonstrates two things: (1) the **conjugate prior** for $\mu$ in model (5) is **normal**, and (2) the **conjugate updating rule** (when $\sigma_0$ is assumed known) is

$$\left\{ \begin{array}{c} \mu \sim N\left(\mu_0, \sigma_\mu^2\right) \\ (Y_i|\mu) \stackrel{\text{IID}}{\sim} N\left(\mu, \sigma_0^2\right), \\ i = 1, \ldots, n \end{array} \right\} \rightarrow (\mu|y) = (\mu|\bar{y}) = N\left(\mu_*, \sigma_*^2\right), \tag{10}$$

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

$$\mu_* = B = \frac{\left(\frac{1}{\sigma_\mu^2}\right) \mu_0 + \left(\frac{n}{\sigma_0^2}\right) \bar{y}}{\frac{1}{\sigma_\mu^2} + \frac{n}{\sigma_0^2}} \quad \text{and} \quad \sigma_*^2 = A^{-1} = \frac{1}{\frac{1}{\sigma_\mu^2} + \frac{n}{\sigma_0^2}}. \tag{11}$$

It becomes useful in understanding the meaning of these expressions to define the $\boxed{\textbf{precision}}$ 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, (10) and (11) have a series of nice **intuitive interpretations**, as follows:

- The **prior**, considered as an **information source**, is Gaussian with mean $\mu_0$, variance $\sigma_\mu^2$, 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 Bayesians invented the idea of precision—on this scale **information** about $\mu$ in model (5) is **additive**); and

- **Rewriting** $\mu_*$ as

$$\mu_* = \frac{\left(\frac{1}{\sigma_\mu^2}\right)\mu_0 + \left(\frac{n}{\sigma_0^2}\right)\bar{y}}{\frac{1}{\sigma_\mu^2} + \frac{n}{\sigma_0^2}} = \frac{\left(\frac{\sigma_0^2}{\sigma_\mu^2}\right)\mu_0 + n\bar{y}}{\frac{\sigma_0^2}{\sigma_\mu^2} + n}, \qquad (12)$$

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

$$n_0 = \frac{\sigma_0^2}{\sigma_\mu^2} = \frac{1}{\left(\frac{\sigma_\mu}{\sigma_0}\right)^2}, \qquad (13)$$

which makes sense: the **bigger** $\sigma_\mu$ is in relation to $\sigma_0$, the **less prior information** is being incorporated in the conjugate updating (10).

**Bayesian inference with multivariate $\theta$.** Returning now to model (3) with $\sigma^2$ unknown, (as mentioned above) this model has a ($k = 2$)-dimensional **parameter vector** $\theta = (\mu, \sigma^2)$.

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

$$\begin{aligned} p(\theta|y) &= p(\mu, \sigma^2|y) = c\,p(\theta)\,l(\theta|y) \\ &= c\,p(\mu, \sigma^2)\,l(\mu, \sigma^2|y), \end{aligned} \qquad (14)$$

# Multivariate Unknown $\theta$

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

$$\begin{aligned} c &= [p(y)]^{-1} = \left( \iint p(\mu, \sigma^2, y) \, d\mu \, d\sigma^2 \right)^{-1} \\ &= \left( \iint p(\mu, \sigma^2) \, l(\mu, \sigma^2 | y) \, d\mu \, d\sigma^2 \right)^{-1} . \qquad (15) \end{aligned}$$

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

$$p(\mu|y) = \int_0^\infty p(\mu, \sigma^2 | y) \, d\sigma^2 \ . \qquad (16)$$

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

$$\begin{aligned} p(y_{n+1}|y) &= \iint p(y_{n+1}, \mu, \sigma^2 | y) \, d\mu \, d\sigma^2 \qquad (17) \\ &= \iint p(y_{n+1}|\mu, \sigma^2) \, p(\mu, \sigma^2 | y) \, d\mu \, d\sigma^2 . \end{aligned}$$

And, finally, if you're interested in a **function of the parameters**, you also 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 (3).

# Multivariate Unknown $\theta$

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 $(\mu, \sigma^2)$ to $(\lambda, \eta)$ is **invertible**; (b) compute the joint posterior for $(\lambda, \eta)$ through the usual **change-of-variables formula**

$$p(\lambda, \eta | y) = p_{\mu, \sigma^2}\big[ f^{-1}(\lambda, \eta) | y \big] \, |J_{f^{-1}}(\lambda, \eta)| \, , \qquad (18)$$

where $p_{\mu, \sigma^2}(\cdot, \cdot | y)$ is the joint posterior for $\mu$ and $\sigma^2$ and $|J_{f^{-1}}|$ is the **determinant** of the **Jacobian** of the inverse transformation; and (c) **marginalize** in $\lambda$ by integrating out $\eta$ in $p(\lambda, \eta | y)$, in a manner analogous to (16).

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 (18) 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 $k$) to get the normalizing constant that defines (18) and one (of dimension $(k-1)$) to get rid of $\eta$.

You can see that when $k$ 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 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

$\boxed{\textbf{Back to model (3).}}$ The conjugate prior for $\theta = \left(\mu, \sigma^2\right)$ in this model (e.g., Gelman et al., 2003) turns out to be most simply described **hierarchically**:

$$\begin{array}{rcl} \sigma^2 & \sim & \text{SI-}\chi^2(\nu_0, \sigma_0^2) \\ (\mu|\sigma^2) & \sim & N\left(\mu_0, \dfrac{\sigma^2}{\kappa_0}\right). \end{array} \qquad (19)$$

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** $\chi^2$ distribution with parameters $\nu_0$ and $\sigma_0^2$.

The scaling is chosen so that $\sigma_0^2$ can be interpreted as a **prior estimate** of $\sigma^2$, with $\nu_0$ the **prior sample size** of this estimate (i.e., **think of a prior data set with** $\nu_0$ **observations and sample SD** $\sigma_0$).

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

$$\begin{array}{rcl} \sigma^2 & \sim & \text{SI-}\chi^2(\nu_0, \sigma_0^2) \leftrightarrow \\ p(\sigma^2) & = & \dfrac{\left(\frac{1}{2}\nu_0\right)^{\frac{1}{2}\nu_0}}{\Gamma\left(\frac{1}{2}\nu_0\right)} \left(\sigma_0^2\right)^{\frac{1}{2}\nu_0} \left(\sigma^2\right)^{-\left(1+\frac{1}{2}\nu_0\right)} \exp\left(\dfrac{-\nu_0\,\sigma_0^2}{2\sigma^2}\right). \end{array} \qquad (20)$$

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

$$E\left(\sigma^2\right) = \frac{\nu_0}{\nu_0 - 2}\sigma_0^2 \quad \text{and} \quad V\left(\sigma^2\right) = \frac{2\nu_0^2}{(\nu_0-2)^2(\nu_0-4)}\sigma_0^4. \quad (21)$$

# Gaussian Modeling (continued)

The parameters $\mu_0$ and $\kappa_0$ in the second level of the prior model (19), $(\mu|\sigma^2) \sim N\left(\mu_0, \frac{\sigma^2}{\kappa_0}\right)$, have **simple parallel interpretations** to those of $\sigma_0^2$ and $\nu_0$: $\mu_0$ is the **prior estimate** of $\mu$, and $\kappa_0$ is the **prior effective sample size** of this estimate.

The **likelihood function** in model (3), with **both** $\mu$ and $\sigma^2$ **unknown**, is

$$\begin{aligned} l(\mu, \sigma^2|y) &= c \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(y_i-\mu)^2\right] \\ &= c \left(\sigma^2\right)^{-\frac{1}{2}n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right] \qquad (22) \\ &= c \left(\sigma^2\right)^{-\frac{1}{2}n} \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} y_i^2 - 2\,\mu\sum_{i=1}^{n} y_i + n\mu^2\right)\right]. \end{aligned}$$

The **expression in brackets** in the last line of (22) is

$$\begin{aligned} [\ \cdot\ ] &= -\frac{1}{2\sigma^2}\left[\sum_{i=1}^{n} y_i^2 + n(\mu-\bar{y})^2 - n\bar{y}^2\right] \qquad (23) \\ &= -\frac{1}{2\sigma^2}\left[n(\mu-\bar{y})^2 + (n-1)s^2\right], \end{aligned}$$

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

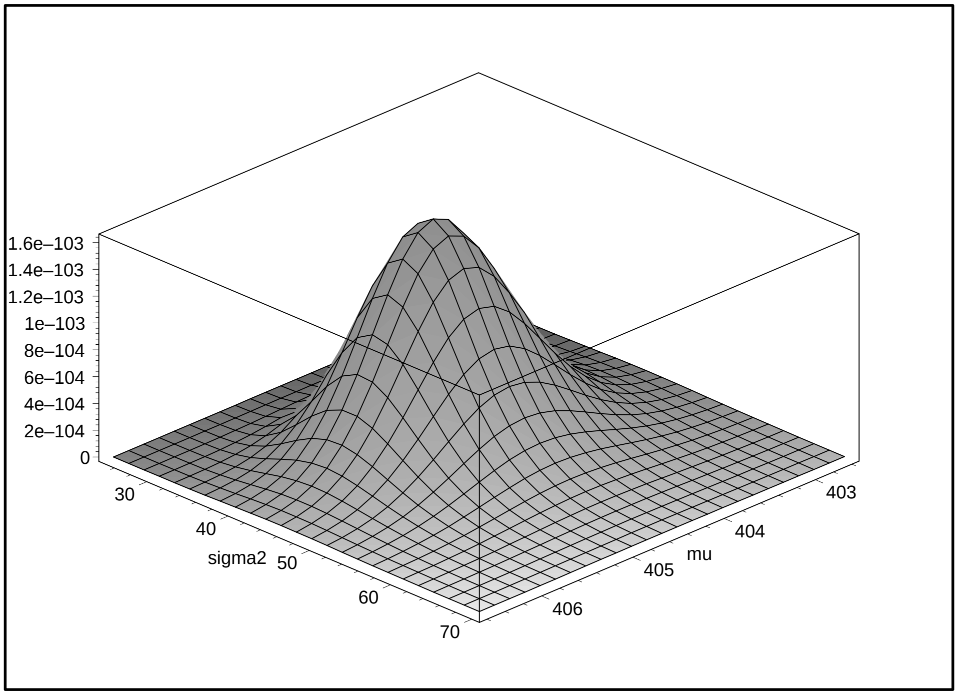
$$l(\mu, \sigma^2|y) = c \left(\sigma^2\right)^{-\frac{1}{2}n} \exp\left\{-\frac{1}{2\sigma^2}\left[n(\mu-\bar{y})^2 + (n-1)s^2\right]\right\},$$

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

# Gaussian Analysis

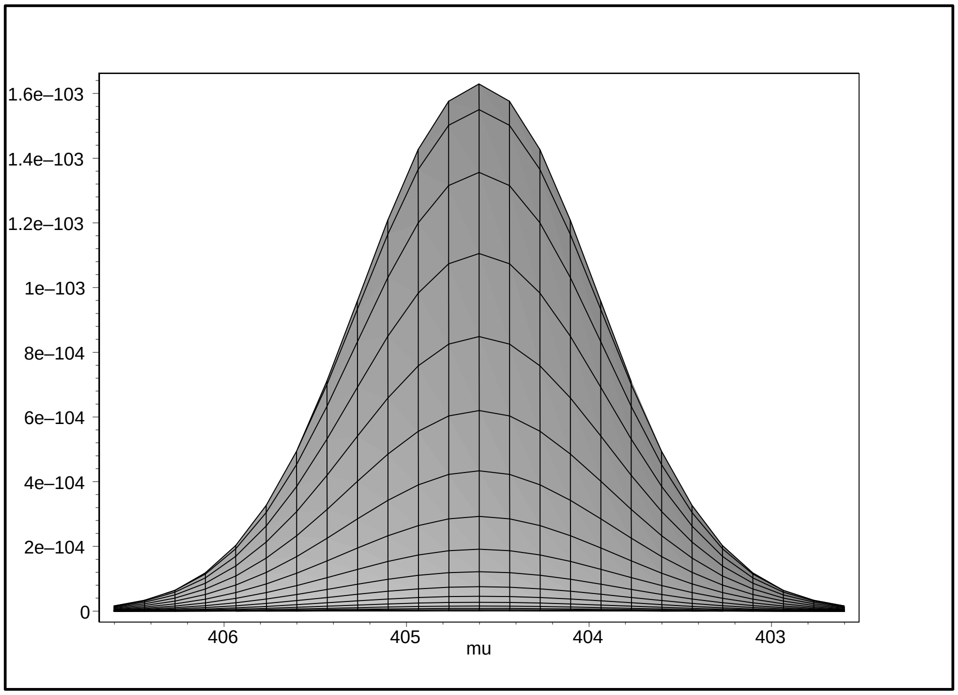
`Maple` can be used to make **3D** and **contour plots** of this likelihood function with the NB10 data:

```
> 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 );
```

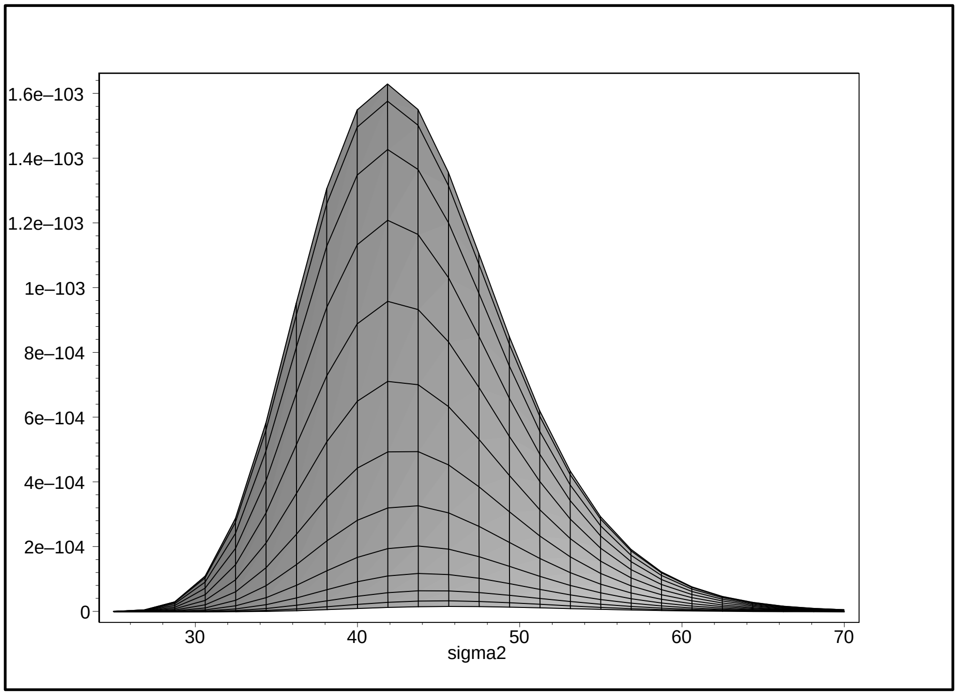


You can use the mouse to **rotate** 3D plots and get **other useful views** of them:

# Gaussian Analysis



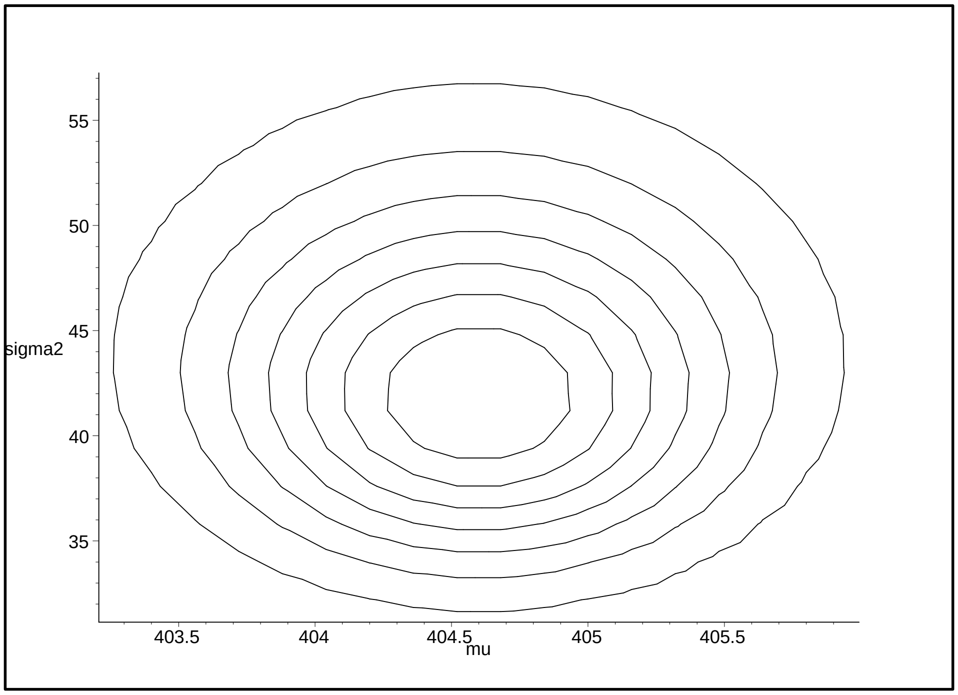
The **projection** or **shadow plot** of $\mu$ looks a lot like a **normal** (or maybe a $t$) distribution.



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

# 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 $\mu$ and $\sigma^2$ are **uncorrelated** in the likelihood distribution, and the **skewness** of the marginal distribution of $\sigma^2$ is also evident.

**Posterior analysis.** Having adopted the **conjugate prior** (19), 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 (3) all of the **integrations** (such as (16) and (17)) may be done **analytically** (see, e.g., Bernardo and Smith 1994), yielding the following results:

$$\begin{aligned} (\sigma^2|y,\mathcal{G}) &\sim \text{SI-}\chi^2(\nu_n,\sigma_n^2), \\ (\mu|y,\mathcal{G}) &\sim t_{\nu_n}\left(\mu_n,\frac{\sigma_n^2}{\kappa_n}\right), \quad \text{and} \\ (y_{n+1}|y,\mathcal{G}) &\sim t_{\nu_n}\left(\mu_n,\frac{\kappa_n+1}{\kappa_n}\sigma_n^2\right). \end{aligned} \tag{24}$$

# NB10 Gaussian Analysis

In the above **expressions**

$$\begin{aligned}
\nu_n &= \nu_0 + n, \\
\sigma_n^2 &= \frac{1}{\nu_n}\left[\nu_0\sigma_0^2 + (n-1)s^2 + \frac{\kappa_0 n}{\kappa_0 + n}(\bar{y} - \mu_0)^2\right], \\
\mu_n &= \frac{\kappa_0}{\kappa_0 + n}\,\mu_0 + \frac{n}{\kappa_0 + n}\,\bar{y}, \quad \text{and} \\
\kappa_n &= \kappa_0 + n,
\end{aligned} \qquad (25)$$

$\bar{y}$ 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 $t_\nu$ 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**

$$\eta \sim t_\nu(\mu, \sigma^2) \leftrightarrow p(\eta) = \frac{\Gamma\left[\frac{1}{2}(\nu+1)\right]}{\Gamma\left(\frac{1}{2}\nu\right)\sqrt{\nu\pi\sigma^2}}\left[1 + \frac{1}{\nu\sigma^2}(\eta - \mu)^2\right]^{-\frac{1}{2}(\nu+1)}. \qquad (26)$$

This distribution has **mean** $\mu$ (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**:

$$\mu_n = \frac{\kappa_0}{\kappa_0 + n}\,\mu_0 + \frac{n}{\kappa_0 + n}\,\bar{y}. \qquad (27)$$

# NB10 Gaussian Analysis (continued)

$\boxed{\textbf{NB10 Gaussian Analysis.}}$ *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 $\mu$ and $\sigma^2$.

Considering the meaning of the **hyperparameters**, to provide little prior information I want to choose both $\nu_0$ and $\kappa_0$ **close to 0**.

Making them exactly 0 would produce an $\boxed{\textbf{improper}}$ 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 (24, 25) that the result is then

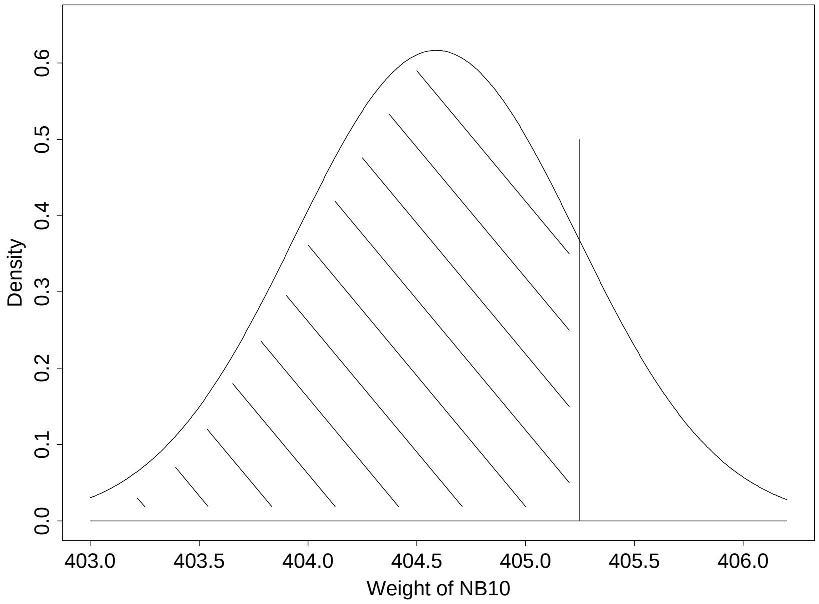
$$(\mu|y, \mathcal{G}) \sim t_n \left[ \bar{y}, \frac{(n-1)s^2}{n^2} \right] \doteq N\left( \bar{y}, \frac{s^2}{n} \right), \qquad (28)$$

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** 404.6$\mu$g **below 10g, give or take about** 0.65$\mu$g.

<u>*Question (b)*.</u> 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 $\mu$ 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**.

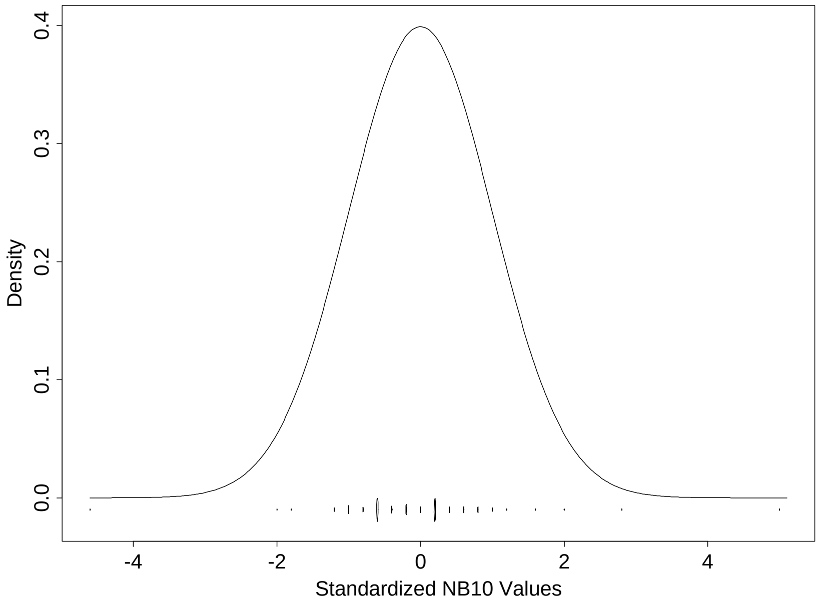
*Question (c).* We saw earlier that **in this model**

$$(y_{n+1}|y, \mathcal{G}) \sim t_{\nu_n}\left[\mu_n, \frac{\kappa_n + 1}{\kappa_n}\sigma_n^2\right], \tag{29}$$

and for $n$ large and $\nu_0$ and $\kappa_0$ close to 0 this is $(y_{n+1}|y, \mathcal{G}) \dot{\sim} N(\bar{y}, s^2)$, 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.

# The Exponential Family

It was noticed a long time ago that many of the standard sampling distributions that you're likely to want to use in constructing likelihood functions have the **same general form**, which is referred to as the **exponential family**:

**Definition** (e.g., Bernardo and Smith, 1994): Given data $y_1$ (a sample of size 1) and a parameter vector $\theta = (\theta_1, \ldots, \theta_k)$, the (marginal) sampling distribution $p(y_1|\theta)$ belongs to the $k$-**dimensional exponential family** if it can be expressed in the form

$$p(y_1|\theta) = c\, f_1(y_1)\, g_1(\theta)\, \text{exp}\left[ \sum_{j=1}^{k} \phi_j(\theta)\, h_j(y_1) \right] \qquad (30)$$

for $y_1 \in \mathcal{Y}$ and 0 otherwise; if $\mathcal{Y}$ does not depend on $\theta$ the family is called **regular**.

$(\phi_1(\theta), \ldots, \phi_k(\theta))$ in (30) is referred to as the **natural parameterization** of the exponential family.

In this case the **joint distribution** $p(y|\theta)$ of a **sample** $y = (y_1, \ldots, y_n)$ of size $n$ which is conditionally IID from (30) (which also defines, as usual, the **likelihood function** $l(\theta|y)$) will be

$$\begin{aligned} p(y|\theta) &= l(\theta|y) = \prod_{i=1}^{n} p(y_i|\theta) \qquad (31) \\ &= c \left[ \prod_{i=1}^{n} f_1(y_i) \right] [g_1(\theta)]^n \, \text{exp} \left[ \sum_{j=1}^{k} \phi_j(\theta) \sum_{i=1}^{n} h_j(y_i) \right] . \end{aligned}$$

# The Exponential Family (continued)

This leads to **another way** to define the exponential family: in (30) take $f(y) = \prod_{i=1}^n f_1(y_i)$ and $g(\theta) = [g_1(\theta)]^n$ to yield

**Definition**: Given data $y = (y_1, \ldots, y_n)$ (a conditionally IID sample of size $n$) and a parameter vector $\theta = (\theta_1, \ldots, \theta_k)$, the (joint) sampling distribution $p(y|\theta)$ belongs to the $k$-**dimensional exponential family** if it can be expressed in the form

$$p(y|\theta) = c\, f(y)\, g(\theta) \text{ exp} \left[ \sum_{j=1}^k \phi_j(\theta) \sum_{i=1}^n h_j(y_i) \right] . \qquad (32)$$

Either way you can see that $\{\sum_{i=1}^n h_1(y_i), \ldots, \sum_{i=1}^n h_k(y_i)\}$ is a set of **sufficient** statistics for $\theta$ under this sampling model, because the likelihood $l(\theta|y)$ depends on $y$ only through the values of $\{h_1, \ldots, h_k\}$.

I bring up the exponential family in part because, if the likelihood $l(\theta|y)$ is of the form (32), then in searching for a **conjugate** prior $p(\theta)$ — that is, a prior of the same functional form as the likelihood — you can see directly what will work:

$$p(\theta) = c\, g(\theta)^{\tau_0} \text{ exp} \left[ \sum_{j=1}^k \phi_j(\theta)\, \tau_j \right] , \qquad (33)$$

for some $\tau = (\tau_0, \ldots, \tau_k)$.

# The Exponential Family (continued)

With this choice the **posterior** for $\theta$ will be

$$p(\theta|y) = c\, g(\theta)^{1+\tau_0} \exp\left\{ \sum_{j=1}^{k} \phi_j(\theta) \left[ \tau_j + \sum_{i=1}^{n} h_j(y) \right] \right\}, \tag{34}$$

which is indeed of the **same form** (in $\theta$) as (33).

As a first example, with $s = \sum_{i=1}^n y_i$, the **Bernoulli/binomial** likelihood (equation (6) in part 2) can be written

$$\begin{aligned} l(\theta|y) &= \theta^s (1-\theta)^{n-s} \\ &= (1-\theta)^n \left( \frac{\theta}{1-\theta} \right)^s \qquad (35) \\ &= (1-\theta)^n \exp\left[ s \log\left( \frac{\theta}{1-\theta} \right) \right], \end{aligned}$$

which shows (a) that this sampling distribution is a member of the **exponential family** with $k = 1$, $g(\theta) = (1-\theta)^n$, the natural parameterization $\phi_1(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$ (**NB** the basis of **logistic regression**), and $h_1(y_i) = y_i$, and (b) that $\sum_{i=1}^n h_1(y_i) = s$ is sufficient for $\theta$.

Then (33) says that the **conjugate prior** for the Bernoulli/binomial likelihood is

$$\begin{aligned} p(\theta) &= c\,(1-\theta)^{n\tau_0} \exp\left[ \tau_1 \log\left( \frac{\theta}{1-\theta} \right) \right] \\ &= c\,\theta^{\alpha-1}(1-\theta)^{\beta-1} = \text{Beta}(\alpha,\beta) \qquad (36) \end{aligned}$$

for some $\alpha$ and $\beta$, as we've already seen is **true**.

# The Exponential Family (continued)

As an example of a **non-regular** exponential family, suppose 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 $\theta$:

$$p(y_1|\theta) = \left\{ \begin{array}{cc} \frac{1}{\theta} & \text{for } 0 < y_1 < \theta \\ 0 & \text{otherwise} \end{array} \right\} = \frac{1}{\theta} I(0, \theta), \quad (37)$$

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

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

(37) is an **example of (30)** 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 $\theta$.

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

It turns out that 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 (e.g., with a conditionally IID sample of size $n$ from (37), $V(\theta|y) = O\big(n^{-2}\big)$ (!) instead of the more familiar $O\big(n^{-1}\big)$).

# The Exponential Family (continued)

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

$$\begin{aligned} l(\theta|y) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \mu)^2\right] \quad (38) \\ &= \sigma^{-n}(2\pi)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} y_i^2 \right.\right. \\ & \qquad \left.\left. -2\mu \sum_{i=1}^{n} y_i + n\mu^2\right)\right]. \end{aligned}$$

This is **of the form (32)** with $k = 2$, $c = (2\pi)^{-\frac{n}{2}}$, $f(y) = 1, g(\theta) = \sigma^{-n} \exp\left(-\frac{n\mu^2}{2\sigma^2}\right), \phi_1(\theta) = -\frac{1}{2\sigma^2}, \phi_2(\theta) = \frac{\mu}{\sigma^2}, h_1(y_i) = y_i^2$, 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 $\theta$.

Some **unpleasant algebra** then demonstrates that an application of (33) leads to (19) as the **conjugate prior** for the Gaussian likelihood when both $\mu$ and $\sigma^2$ are unknown.

# In Dispraise of Hypothesis Testing

$\boxed{\textbf{Setup:}}$ **Controlled experiment** of **new** versus **old treatment**, with $n$ (human) subjects **randomized**, $\frac{n}{2}$ to old, $\frac{n}{2}$ to new, $n$ (fairly) large.

$\theta =$ the **mean difference** (new − old), on the **most important outcome** of interest (scaled, without loss of generality, so that large values are better than small), in the **population** $\mathcal{P}$ of subjects judged **exchangeable** with those in the trial.

(This is like imagining that the $n$ trial subjects were **randomly sampled** from $\mathcal{P}$ [of course this is typically **not** how subjects are actually enlisted in the trial] and then **randomized to new or old**, which gives $\theta$ a **causal** interpretation as the **mean improvement per person caused** by receiving the new treatment instead of the old.)

As we've noted earlier, **two frequentist schools of inference** about $\theta$ developed in the twentieth century:

- The **Fisherian** approach, which has two parts:

(a) **Point** and **interval estimates** of $\theta$ based on the **likelihood function**; and

(b) Summarization of the **evidence against** a **null hypothesis** like $H_0$: $\theta = 0$ via $P$-**values** (the chance, if the null is true, of getting data **as extreme as**, or **more extreme than**, what you got).

# Hypothesis Testing (continued)

- The **Neyman-Pearson** approach, which also has two parts:

(c) **Testing** $H_0$: $\theta = 0$ against $H_1$: $\theta \neq 0$ by developing rules (as a function of $n$) that **reject** $H_0$ with a pre-specified **Type I error probability** $\alpha$ (the chance of **incorrectly rejecting** $H_0$), and then (having first specified $\alpha$) choosing $n$ so that the **Type II error probability** $\beta$ (the chance of **incorrectly failing to reject** $H_0$) is no more than some **pre-specified threshold** when $\theta$ actually is some **pre-specified positive value** $\theta_1$ (this is equivalent to choosing $n$ so that the **power** $(1 - \beta)$ of the test is not less than a pre-specified threshold when $\theta = \theta_1$); and

(d) Constructing a **confidence interval** for $\theta$ with some pre-specified **confidence level** $100(1 - \gamma)\%$.

In practice a **combined frequentist approach** has somehow evolved in which randomized trials are often **designed** from the **Neyman-Pearson** point of view (c) but then **summarized** with **Fisherian** $P$**-values** (b) as **measures of evidence against** $H_0$.

From a **Bayesian** point of view this approach **perversely emphasizes the worst** of both the Fisherian and Neyman-Pearson schools, by failing to focus on the **most scientifically relevant summary** of any given trial: an (interval) **estimate** of $\theta$ **on the scale of the most important outcome variable** (recall de Finetti's Bayesian emphasis on **predicting** data values **on the scales on which they're measured**).

# Hypothesis Testing (continued)

A good **rule of thumb**: don't wander off onto the **probability scale** (as $P$-values do) when you can stay on the **data scale** (as interval estimates do), because it's harder to think about whether **probabilities** are important scientifically ("Is $P = 0.03$ small enough?") than it is to think about whether **changes on the main outcome scale of interest** are real-world relevant ("Would it positively affect eBay's **bottom line** if the **change to the web experience** we're now studying increased the **percentage of visits to the eBay web page that end in a sale** from **10%** to **12%**?").

$\boxed{\textbf{Standard example:}}$ I've run my experiment and the $P$-value comes out **0.02**, which is **"small enough to publish"**; but can I tell from this whether the difference I've found is **real-world meaningful**?

In a **two-tailed** test of $H_0$: $\theta = 0$ against $H_1$: $\theta \neq 0$ I can work backwards from $P = 0.02$ to figure out that the value of the standard **test statistic**

$$z = \frac{\overline{\text{new}} - \overline{\text{old}}}{\widehat{SE}\left(\overline{\text{new}} - \overline{\text{old}}\right)} \tag{39}$$

that gave rise to $P = 0.02$ was ±2.3 (taking $n$ to be **large**), but

(1) I can't even tell from the $P$-value whether the new treatment was **better or worse than the old**,

(2) the thing I really want to know to judge the **practical significance** of this finding is the **numerator** of (39),

(3) the thing I really want to know to judge the **statistical significance** of this finding is the **denominator** of (39), and

(4) the $P$-value has **thrown away crucial information** by (in effect) specifying only the **ratio** of (2) and (3) rather than their **separate**, and **separately important**, values.

# Hypothesis Testing (continued)

If I have to work out the **numerator** and **denominator** of (39) **separately** to pin down both the **practical** and **statistical** significance of my result, both of which are **key scientific summaries**, then **what's the point** of calculating the $P$-value at all?

Why not **dispense with it altogether** and go directly to the (e.g., 95%) interval estimate

$$\left(\overline{\text{new}} - \overline{\text{old}}\right) \pm 2\,\widehat{SE}\left(\overline{\text{new}} - \overline{\text{old}}\right)? \qquad (40)$$

(This is a large-$n$ **approximation** to the **Bayesian solution to the inference problem** when **prior information** is **diffuse**.)

For me the above argument **demolishes the use of $P$-values in inference** (although in part 5 I'll make better use of them in **diagnostic checking** of a **statistical model**, which is another task altogether).

The **Fisherian point** and **interval estimates** (a) and the **Neyman-Pearson confidence intervals** (d) are much more in keeping with the scientifically compelling idea of **staying on the data scale**, but they have the following two **drawbacks** in relation to the Bayesian approach:

- They **fail to incorporate relevant prior information** about $\theta$ when it's available, and

- They **don't necessarily work very well** (i.e., they don't necessarily live up to their **advertised frequentist properties**) when the **likelihood function** is **heavily skewed** and/or when $n$ is **small**.

# 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, Adhikari A (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.